"What Do I Do with the Data Now?": Analyzing Assessment Information for Accountability and Improvement

Suzanne L. Pieper
Northern Arizona University

Keston H. Fulcher
Christopher Newport University

Donna L. Sundre and T. Dary Erwin
James Madison University

Abstract

Most colleges and universities have implemented an assessment program of some kind in an effort to respond to calls for accountability from stakeholders as well as to continuously improve student learning on their campuses. While institutions administer assessment instruments to students and receive reports, many campuses do not reap the maximum benefits from their assessment efforts. Oftentimes, this is because the data have not been analyzed in a way that answers questions that are important to the institution or other stakeholders. This paper describes four useful analytical strategies that focus on the following key educational research questions: (a) Differences: Do students learn or develop more if they participate in a course or program compared to other students who did not participate?; (b) Relationships: What is the relationship between student assessment outcomes and relevant program indicators (e.g., course grades, peer ratings)?; (c) Change: Do students change over time?; and (d) Competency: Do students meet our expectations? Each of these strategies is described, followed by a discussion of the advantages and disadvantages of each method. These strategies can be effectively adapted to the needs of most institutions. Examples from the general education assessment program at James Madison University are provided.

Introduction

In response to calls for accountability as well as the desire to improve student learning and development on college campuses, many institutions implement assessment programs of some kind. Furthermore, institutions that endeavour to demonstrate the quality of their programs, as well as continuously improve them, focus on assessment of student learning outcomes. In other words, they attempt to measure what their schools contribute to students' knowledge, skills, and attitudes. While assessment of student learning poses many challenges, perhaps the most significant challenge is analyzing and drawing meaningful conclusions from assessment data.

Let's examine an all-too-familiar assessment scenario played out on college campuses across our nation and beyond. In the scenario, learning objectives are stated, an instrument selected, and data collected, but the data remain grossly under-analyzed and therefore, under-utilized. The analyses "used" for assessment consist of a summary report provided by a test scoring service or perhaps the instrument vendor. These reports generally provide descriptive statistics summarizing student performance, such as the average score. In addition, individual student scores are provided, which may be used to give feedback to students – a potentially good strategy for enhancing student motivation for testing. However, a listing of student scores is of no assistance for program assessment purposes, and for ethical and legal reasons, it cannot be reported. Descriptive statistics on the student group may be of interest when compared to normative data. It is important to keep in mind, however, that no truly representative norms exist upon which the assessment performances of our students can be compared (Baglin, 1981). In other words, normative data are based on samples from schools that agree to use the tests, not from a random selection of students in higher education. Descriptive statistics of this kind may find utility when considered longitudinally at a given institution; however, other important opportunities to learn from the data were lost.

The issue at hand is that the data were not used in a way that answered the questions "To what extent were the stated objectives achieved?" and "What components of the curriculum contributed to achievement of these objectives?" Typically, such assessment reports gather dust on a shelf, are not read, and do not contribute to meaningful discussions about our programs. It would not be uncommon or surprising on campuses where this occurs for assessment to be legitimately referred to as a "waste of time and money."

The scenario above illustrates our frequent inability to provide compelling evidence of program quality as well as our failure to effectively use campus assessment for continuous improvement. At the same time, the scenario underscores the importance of asking good questions about program effectiveness via establishing clear learning objectives and then addressing these questions with complementary analytical strategies. More broadly, the scenario also demonstrates the importance of creating critical linkages between program goals, actions, instrumentation, data analysis, and interpretation of results (Erwin, 1991). The process of creating these assessment linkages is often called "alignment" by experts in the assessment field (Allen, 2004; Maki, 2004).

The purpose of this paper is to describe some effective analytical strategies that are designed to respond to some of the most important research questions we might wish to pose about program quality and impact. These analytical methods have been tested and successfully used for outcomes assessment at James Madison University (JMU) and a growing number of other institutions. We anticipate that these strategies may be useful for other institutions. It should be noted that no single analytical method will provide sufficient information about the quality of our programs; however, all of the methods taken together will more fully illuminate the meaning of student test performances and the value of our educational programs. In addition, if the answers to our research questions conform to expectations, they provide greater validation of our assessment methods and designs.

Four basic analytical strategies have been developed. While the use of all four strategies is highly recommended, it may take time for assessment practitioners to fully implement them because they require a robust institutional assessment infrastructure. The important first step is to ask the research questions of interest and then gather the necessary data to respond. The four analytical strategies focus on the following key educational research questions:

1. Differences: Do students learn or develop more if they participate in a course or program compared to other students who did not participate?

2. Relationships: What is the relationship between student assessment outcomes and relevant program indicators (i.e., course grades, peer ratings)?

3. Change: Do students change over time?

4. Competency: Do students meet our expectations?

Each of these strategies will be described and examples provided along with the advantages and disadvantages of each method. Note that while we encourage (and personally engage in) the use of appropriate statistical analyses to examine significance and effect size, in this paper we treat the analytical strategies from a more general and conceptual level. What we are trying to do is to demonstrate how these strategies can be used to stimulate conversations among teachers and assessment practitioners about student learning.

Differences

The first analytical strategy involves outlining expected differences in student performance that should result if our program is effective. Our research question might ask, "Do students learn or develop more if they have participated in a course or program compared to students who did not participate?" There are many ways to develop such questions. Essentially, we are asking about the impact of an educational treatment. We expect that greater exposure to the educational program should result in enhanced performance on our assessment measure. For example, when assessing the impact of a general education program in science, we might frame our question around the expectation that as students complete more relevant science courses, they will perform better on the assessment than students who did not complete coursework. This strategy could also be used with students participating in a co-curricular leadership program. Our expectation here might be that if our program is effective, students who participate in the leadership program on campus would be expected to show stronger assessment performances when compared with other students who did not participate in the program. There are many naturally occurring groups

that can be identified to frame highly meaningful contrasts. Table 1 illustrates an example from JMU of this analytic strategy. In this example, differences in scientific reasoning assessment performances are compared in relation to the number of relevant courses completed. Although the expectation that assessment scores should increase with additional course completion was met, JMU faculty noted that these increases were small. A lively discussion ensued about student learning and performance standards.

Table 1
Differences in Student Scientific Reasoning Test Scores
by Number of Science-Related Courses Taken

Science-Related Courses	N	Total Test Score
None	16	52.2
One	131	55.4
Two	201	57.4
Three	251	58.6
Four	145	60.7
Five or more	41	61.4

Note. Total Test Score *SD*=12.9

The advantage of this strategy is that it is intuitively straightforward and answers a general question, generally. If the curriculum in a certain program impacts student learning, then students who take more courses should demonstrate more student learning via a higher assessment score. Like the other methods that follow, results from this method encourage faculty thought and conversation about student learning. Instead of being an abstract or philosophical exercise, faculty dialog has now become grounded in empirical data.

A disadvantage of this strategy is the difficulty in collecting data that reflect various strata of student course experiences. For example, because of the science requirement at JMU, very few sophomores who were assessed fit into the no-science-courses-taken category. Another difficulty to consider is that the number of courses students complete may be confounded with other variables, most notably ability and interest. For instance, it is entirely possible that students with higher ability may opt to take more courses in science. In such an event, the meaning of higher course exposure with higher assessment performances becomes obscured, hampering the ability to make inferences about program quality. This confounding problem can be addressed statistically by using an ability measure such as SAT or ACT scores as a covariate in the analysis. A third issue is that the results lack specificity regarding courses. Because courses are aggregated together, it is impossible to determine to what degree individual courses contributed to student learning. Fortunately, the next strategy addresses this issue.

Relationships

The second analytical strategy seeks to answer questions such as, "What is the relationship between student assessment outcome measures and course grades?" The logic here is that if a course is included as part of a program requirement, we should expect to see a positive correlation between course outcomes as measured by grades and performances on our assessment instrument. Correlation coefficients range from -1.00 to +1.00. Correlations near 0 indicate no relationship, while correlations closer to +1.00 indicate a strong, positive relationship between assessment outcomes and course grades. It should be noted that correlations between course grades and assessment scores are not expected to be perfect. In this context, correlations of +.30 and +.40 seem strong. As Phillips (2000) points out, assessment scores and grades in courses measure, at least to some extent, different aspects of a student's educational experience. Assessment covers achievement of skills; grades may cover many other factors in addition to achievement, such as participation, attendance, attitude, timeliness, and effort. Further, many general education programs require completion of more than one course to fulfill an area requirement, suggesting that a single course may not address all relevant program objectives. However, we would not expect to see negative relationships between course grades and assessment performances, which would mean that students who score better on the assessment tend to receive lower grades in particular classes. Table 2 provides an example from JMU of this analytical strategy.

Table 2

Correlations of Scientific Reasoning Test Scores with University Science Course Grades Over a Three-Year Period

Course	Year 1		Year 2		Year 3	
	r	*N*	*r*	*N*	*r*	*N*
Physics, Chemistry & the Human Experience	.28	352	.24	370	.20	252
Environment: Earth	.13	130	.29	107	.20	69
Discovering Life	.45	91	.28	76	.37	57
Scientific Perspectives	.15	128	.09	164	.15	109

The correlations presented in Table 2 generated considerable conversation among JMU faculty regarding the association between grades earned in courses considered relevant to the material tested and assessment scores. Although no single course can be expected to cover all of the objectives targeted on the test, faculty did expect that each course should contribute to student learning of the goals and objectives. Clearly, some course grades were more strongly related to assessment scores than others. Correlations were calculated over three separate assessment administrations over a three-year period; thus, the stability of correlations over time were also a part of the discussion.

The primary advantage of this strategy is that, similar to the first strategy, it is fairly easy to understand conceptually. Second, in terms of program improvement, it yields diagnostic information. From this strategy, we can pinpoint which classes are contributing to student learning in a particular educational area and which are not. It also may provide evidence that the assessment method and relevant course grades are measuring the same constructs (i.e., convergent validity).

The major disadvantage of this strategy is that, like other correlational studies, inferences about causation should be made with caution. In addition, this strategy requires adequate sample sizes to produce stable correlation coefficients. Unfortunately, many general education programs include a plethora of courses purported to contribute to our assessment outcomes in a specific area, which makes it very difficult to collect sufficient data to calculate stable correlations based on individual courses. Note that when this is the case, strategy one can be employed by counting the number of course exposures a student has completed with the expectation that more course completions should result in higher assessment performances. An additional concern is that a third variable, such as general ability, might obscure the meaning of the relationship between assessment performances and course grades. Again, as with strategy one, this problem can be statistically controlled with a partial correlation procedure that removes the effect of general ability, as measured by SAT or ACT, from the correlation. Last, because course grades are considered unreliable, their use as criterion variables is questionable (Erwin & Sebrell, 2003).

Change

The third analytical strategy, "Do students change over time?" has been used by a variety of programs and services across many campuses. Also called the "value-added" or longitudinal approach, the expectation is that, as a result of a course or program, students will show marked improvement from pretest to posttest. For most faculty members, this strategy provides the most direct route to understanding the efficacy of their programs. Table 3 shows an example from JMU of this analytical strategy.

Table 3

Pre- and Post-Scores of Scientific Reasoning Test

	N	*SD*	Score
Freshmen (Pre)	148	10.2	56.8
Sophomore/Juniors (Post)	148	11.9	62.7

Note. The Freshmen and Sophomore/Juniors groups reflect the same cohort of students at two points in time.

While the faculty at JMU were very happy to see that the difference between performances were statistically significant, they were disappointed by the magnitude of the overall change. They clearly would have preferred to see greater change than they observed. These findings led to discussions of several

important topics about JMU's assessment design including the sensitivity of the instrument, student motivation to perform well in a low-stakes assessment condition, the timing of tests in relation to coursework, and the nature of general education itself. All of these topics were important in providing appropriate interpretations of assessment results, and they also led to improvements in data collection and review of the instrument.

The major advantage of this powerful strategy is that we can look at program effectiveness more directly because there is a baseline with which to compare. A statistical advantage exists as well. Because the same students are being assessed twice, extraneous variables and error are more carefully controlled.

The major disadvantage of this strategy is that when students are studied longitudinally, some positive changes may occur as a result of maturation, not necessarily as a result of any contribution of the coursework or program. Using a control group as part of the design can provide some statistical control for changes resulting due to maturation or other factors; however, such control groups are difficult to find. Additionally, bias may be introduced when students "drop out," "stop out," or transfer from the campus. These are not random events; therefore, it is likely that the students remaining at the end of a program might be systematically stronger than those choosing to depart or delay completion. Moreover, two testing times are required for this longitudinal design, which requires stability in the data collection process and highly reliable measurement. As Erwin (1991) points out, any measurement errors in pretest or posttest measures are compounded in change scores, further justifying the need for reliable assessment tools.

<div style="text-align:center">Expectations</div>

The fourth analytic strategy seeks to answer the research question, "Do our students meet our expectations?" This analytical question is also exceptionally important, because establishment of standards indicates quality (Shepard, 1980). All stakeholders in higher education-- faculty, students, parents, taxpayers, employers, and policy makers-- are interested in whether students have met established and credible standards. Table 4 provides a JMU example of this analytical strategy.

At JMU, sophomore registration is held until students have passed all technology proficiency requirements, attaching high stakes consequences to the standards. The approach taken at JMU has been to assure that all students will achieve these expectations by providing additional tutorials and assistance to those who need it.

Table 4
Percent and Number of Students Meeting Standard on Information Literacy Computer-Based Test

	%	# of students
Met the standard	98	3044

Note. Figures reflect number of freshmen passing all three components of the information technology standards before a specified date.

The major advantage of this analytical strategy is that it demonstrates to all interested stakeholders that students have been measured with a common instrument and held to a common standard. Those inside the institution are assured that students have attained designated knowledge and skills before progressing. Those outside of the institution value the certification of skills as more meaningful than course grades or even assessment scores. However, high stakes tests may introduce new concerns, particularly liability issues. An institution must be prepared to defend its entire standard setting process in the face of possible legal challenges. See Phillips (2000) for a full discussion of the legal issues pertaining to high stakes tests and the precautions an institution should take.

It should be noted that it is not necessary to implement high stakes testing to introduce faculty expectations for student performance. When faculty establish their expectations for student performances on a given test they can do so within a particular context, such as a low stakes testing condition after student coursework is completed. The key issue is providing a framework for appropriate interpretation of assessment results. We have noticed that faculty pay much closer attention to assessment results when they have played a role in establishing performance expectations. These performance expectations must be established prior to review of the results, not after. Moreover, these performance expectations must be

meaningful and defensible; for more information on establishing expectations, also known as standard setting, see Shepard (1980).

Although most of the above examples are related to general education assessment, these four strategies could be effectively applied to any program assessment—curricular or co-curricular—of student learning and development. Whatever the assessment context, the relationship between analytical strategies and establishment of program goals and objectives cannot be overemphasized. Their compatibility is essential for an effective assessment program. As Erwin (1991) points out, when establishing program objectives, questions will naturally arise about the quality of the program. These questions, Erwin notes, lead faculty and staff to seek out evidence that will answer their questions. This is the time, before information is collected, to think about how the assessment information collected will be examined. The research questions that faculty and staff pose at the beginning of the assessment initiative should guide how the data will later be analyzed. Palomba and Banta (1999) concur fully and suggest that anticipating the way data will be analyzed, "helps assessment planners identify the types of information needed, appropriate methods and sources to obtain this information, and the number of cases to be examined" (p. 313). In other words, explicitly stating your research questions can ensure that data collection and the subsequent analytical methods are linked and viable.

Conclusion

These strategies are, of course, just a few of the many potential strategies an institution might choose to analyze outcomes assessment information. Again, it is important to design the analytical strategies to answer specific questions of faculty and staff on a particular campus. Every institution will necessarily pose different questions. It is also important to note that data analysis is a recursive process that begins with questions in the early designing of outcome objectives. As Erwin (1991) noted, after the data is analyzed still more questions are generated: Have the early questions changed? Do other questions need to be added? Are students learning according to faculty expectations?

In sum, data analysis is the critical connection between what comes before-- establishing objectives for outcome assessments, selecting assessment methods or designing assessment methods to suit institutional needs, and collecting and maintaining information--and what comes after-- reporting and using assessment information. Assessment information cannot be used to either demonstrate accountability or improve learning and development if it is not analyzed or if it does not answer the right questions. It is more important now than ever for colleges and universities to take a closer look at this weakest assessment link.

References

Allen, M.J. (2004). *Assessing academic programs in higher education.* Bolton, MA: Anker Publishing.

Baglin, R.F. (1981). Does 'nationally' normed really mean nationally? *Journal of Educational Measurement, 18*, 97-107.

Erwin, T. D. (1991). *Assessing student learning and development: A guide to the principles goals, and methods of determining college outcomes.* San Francisco: Jossey-Bass.

Erwin, T. D., & Sebrell, K. W. (2003). Assessment of critical thinking: ETS's tasks in critical thinking [Electronic version]. *The Journal of General Education, 52*(1), 50-70.

Maki, P.L. (2004). *Assessing for learning: Building a sustainable commitment across the institution* Sterling, VA, Stylus Publishing.

Palomba, C., & Banta, T. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education.* San Francisco: Jossey-Bass.

Phillips, S. E. (2000). GI Forum v. Texas Education Agency: Psychometric evidence, *Applied Measurement in Education*, *13*(4), 343-385.

Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement*, *4*, 447-467.