RESEARCH & PRACTICE IN ASSESSMENT

VOLUME THREE| WINTER 2008
www.RPAjournal.com
ISSN#2161-4210

A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

Editors

Executive Editor

Keston H. Fulcher
Director of Assessment, Evaluation, and Accreditation
Christopher Newport University

Editor

Allen DuPont
Director of Assessment, Division of Undergraduate Affairs
North Carolina State University

Consulting Editors

Robin Anderson
Associate Director, Center for Assessment and Research Studies
James Madison University

Dorothy Doolittle
Associate Provost and Professor of Psychology
Christopher Newport University

John T. Willse
Assistant Professor of Educational Research Methodology
University of North Carolina at Greensboro

Table of Contents

Comments from the Editor

Keston H. Fulcher
Director of Assessment, Evaluation, and Accreditation
Christopher Newport University

This issue of Research and Practice in Assessment provides three articles that should prove beneficial to practitioners. The first article by Pieper, Fulcher, Sundre, and Erwin identifies a problem that plagues many learning outcomes assessment initiatives in higher education: the underuse of data. The authors describe a framework of analytical questions related to differences, relationships, change, and competency that can drive analyses resulting in useful information.

Considering the acceleration of globalism, the second article by Bresciani is timely. It explores cross-cultural collaboration among American and Mexican researchers. Through a qualitative methodology she uncovers several points of understanding and misunderstanding between these groups. She then postulates how these findings may be used to improve collaboration in cross-cultural endeavors in research and in the classroom.

The final article, by Thelk, examines whether a science and math test used at a four-year institution is appropriate for community colleges. She uses statistical techniques to ascertain if certain items exhibit differential item functioning (DIF). The presence of DIF could indicate bias. Indeed, many items did reveal DIF, which prompted a revision so that the instrument would be more appropriate for the community college.

“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.

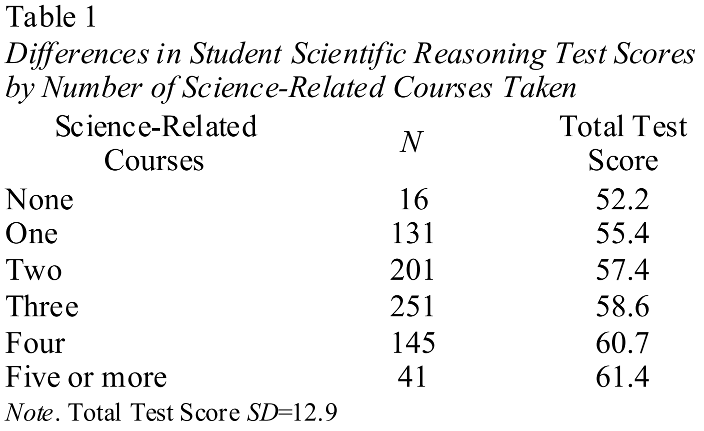

Table 1
Differences in Student Scientific Reasoning Test Scores by Number of Science-Related Courses Taken

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.

Expectations

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.

Exploring Misunderstanding in Collaborative Research Between a World Power and a Developing Country

Marilee J. Bresciani
San Diego State University

Abstract

This phenomenological study explored what principal investigators from the United States and Mexico experienced when engaging in cross-cultural collaborative research projects. Participants were asked to articulate their understanding of collaboration. While the principal investigators did not vary on how they defined quality of research; their perceptions of collaboration varied. An agreement of understanding how effective collaboration is operationalized is pertinent to the improvement of student learning.

Introduction

The United States and other countries are becoming increasingly interdependent upon one another to foster new knowledge production and economic stimulation of developing countries (Bonnema, 2006). While many institutions from countries that are considered to be economic world powers are seeking to partner with institutions from developing countries to engage in collaborative research, the expectations for such collaborative research may not be clear in all aspects of the research project and, therefore, it becomes increasingly difficult to determine whether the joint venture has been successful (Palomba & Banta, 1999; Suskie, 2004). For example, it may be clear to cross-cultural principal investigators, who sign on for a research project, what the research study will entail, however, the expectations may not be clear as to the extent or nature of the collaboration, and as to the scope of how the project's success will be identified by joint program administrators who may fund the research. As such, many misunderstandings may occur that could threaten the success of the current research project or the opportunity for continued research collaboration to occur (Montoya-Weiss, Massey, & Song, 2001).

The same concern for understanding how effective collaboration is operationalized in research can be applied to the improvement of student learning. Research has demonstrated that effective collaborations are needed in order to improve student learning (Bresciani, 2006; Maki, 2004; Palomba & Banta, 1999). If we can apply these findings to the outcomes-based assessment work that we do, particularly in the curricular and co-curricular, improvements can be made in the partnerships that are needed to inform the necessary decisions.

This study sought to understand what principal investigators and co-principal investigators experienced when engaging in cross-cultural collaborative research projects that were funded by a seed research grant project sponsored by a U.S.A. host institution. In particular, the notion of collaboration was explored from the perspectives of principal and co-principal investigators in the United States and those within several research universities in Mexico.

Methodology

A qualitative method was utilized for this study because a qualitative researcher's intent is to uncover "meaning" (Bogdan & Biklen, 1992). There are several methods by which to uncover meaning and many of them share the common goal of understanding the subject's perspective from the point of view of the subject. "Researchers in the phenomenological mode attempt to understand the meaning of events and interactions to ordinary people in particular situations." (Bogdan & Biklen, 1992). The participant's point of view thus becomes a research construct. Engaging subjective thinking, the participant's point of view becomes the reality; therefore reality comes to be understood to human beings only in the form in which it is perceived (Bogdan & Biklen, 1992).

Phenomenology was appropriate for this study as the researcher was attempting to uncover the meaning of collaboration as perceived by the participants of the collaborative grant funded research projects. Phenomenology allows the subjective view point of the study participant to be heard in the context of the participant's reality. In other words, in order to understand what collaboration is and how it would be demonstrated, the subject's perspective must be understood; her reality must be understood so that the meaning of words she uses to describe collaboration can be understood.

When Schutz (1970) developed phenomenology, he posited to depart from the experiential assumptions of the natural attitude - the "everyday interpretive stance that takes the world to be principally 'out there' separate and distinct from any act of perception or interpretation" (Holstein & Gubrium, 1994, p. 263). Because of this stance, the researcher must bracket (e.g., set aside) her orientation to the subject matter and focus on the ways in which the participants, who are living the experience, interpretively produce the collaboration they believe is real. In so doing, the participant's observations and experiences are often explained and demonstrated by the participant him or herself. "If human consciousness necessarily typifies, then language is the central medium for transmitting typifications and thereby meaning. This [epistemology] provides a methodological orientation for a phenomenology of social life concerned with the relation between language use and the objects of experience" (Holstein & Gubrium, 1994, p. 263).

Because there are often cultural misunderstandings involved in how meaning of words is defined and in how the meaning is identified (Oliva, 2000a: Oliva; 2000b), using phenomenology to extract what participants believe are the characteristics that embody collaboration based on their experiences makes sense. Phenomenology will allow the researcher to identify whether there is a context associated with certain characteristics of cross-cultural collaboration.

Social phenomenology rests on the principle that social interaction constructs as much as conveys meaning. "Schutz's social phenomenology aims for a social science that will interpret and explain human action and thought through descriptions of the reality which seems self evident to the people remaining within the natural attitude." (Holstein & Gubrium, 1994, p. 264) The goal of phenomenology is to explicate how objects and experiences, such as collaboration within research, are meaningfully constituted and communicated in the world of everyday life. Schutz's intention is to treat subjectivity as a topic for investigation in its own right, not as a methodological limitation.

In order to understand the meaning of cross-cultural collaboration, the following data were collected and analyzed using Moustakas (1994) and Polkinghorne's (1989) division of protocols into statements and then organizing the statements into clusters of meaning.

Data collection included the following:

1. Document analysis (Stake, 1995) of the U.S.A. host institution's Collaborative Research Grant Proposal criteria and the actual grant agreement (These are the formal documents that governed the review of proposals and the awarding of seed grant money for cross-cultural research projects);
2. Document analysis (Stake, 1995) of roundtable results from the October 28-30, 2004 Program Symposium on Research Outcomes held in Mexico City, where both Mexican and U.S.A. PIs and administrators were in attendance;
3. Interviews of preliminary institutional administrators from the U.S.A. host institution and the administrative liaisons in Mexico City;

Interview of principal and co-principal investigators (PIs) from the U.S.A. host research university and from the Mexico institutions

Interviews with the Mexican PIs took place during a December trip by the researcher and her graduate assistant to Mexico City. There were two single PI interviews lasting about two hours each and also a group PI meeting on the campus where the majority of research collaborations occurred, lasting approximately two and one-half hours. The three nights and three day trip also allowed for more experience and understanding of Mexican culture, understanding of work ethic, and collaboration expectations Interviews with the United States PIs took place by the researcher and her graduate assistant in two 45-minute group meetings on the U.S.A. Campus shortly after arriving back from the winter holiday break.

Sampling Procedure

It is unclear how to articulate the sampling procedure for this study as the 15 PIs and the four program administrators interviewed were selected by the U.S.A institution's Office of Latin American Programs and the selection methodology is unknown. Of further concern, is the limited time spent interviewing U.S.A. PIs and administrators, however, the researcher was not granted additional access beyond what was initially provided.

Given the nature of this sampling procedure, it may be best to describe the sampling procedure as convenience sampling (Creswell, 1998). Convenience sampling simply states that those who are available for interviewing will be interviewed. Convenience sampling is often used when it is difficult for one reason or another to access the participants. One criterion for interviewing subjects in this study is that they had to have completed the research project that was funded by the host U.S.A. institution. Another criterion for participants was that they had to speak English as all interviews were conducted in this language.

In order for the collaborative research project to be recognized by the sponsoring U.S.A. institution and thus in order for the subjects to be included in this study, the cross-cultural grant proposals must be jointly developed by principal investigators from the U.S.A. research extensive university and research consortium member institutions in Mexico. The collaborative research projects received small seed grant funding from the U.S.A. host institution. Grant funding is available in all disciplines offered by the U.S.A. host institution. The majority of disciplines requesting and receiving the seed grants represented disciplines from science, engineering, and technology.

Study Findings and Discussion

Interviews and document analysis revealed findings that can be reported in four clusters of meaning:

1. Variance in Program Goals and Outcomes for the Research Program
2. Difficulty in Alignment of the Program Delivery to Program Outcomes
3. Cultural Differences in the Meaning of Collaboration
4. Challenges in Discussing the Next Steps

Each cluster of meaning will be explored in the following paragraphs.

Variance in Program Goals and Outcomes for the Research Program

At the conclusion of the interviews and document analysis, it was apparent that there were at least four perspectives of the research program goals. They are the goals of the U.S.A. institution program directors, the Mexican Research Consortium directors, the U.S.A. PIs, and the Mexican PIs. These four entities did share understanding of some goals, yet the operationalization of those goals or, in some cases, the end results of those goals were not entirely shared.

The goals of these four groups are not fully represented in Appendix A. However, Appendix A represents agreed upon goals of the administrators since the documents analyzed to create these goals were communications between administrators. The extent to which these goals were agreed upon by Mexican and U.S.A. principal and co-principal investigators varied. Further, the extent to which these goals are interpreted for the same meaning is varied or unclear. Not unlike what researchers have shown to be the case in other cross-cultural international endeavors (Oliva, 2000a; Oliva, 2002), it can be said that there are differing rather than uniform views of what this collaborative research program was intended to achieve.

The researcher was not able to flesh out the exact variance in goals of Mexican PIs and U.S.A. PIs. In some cases, it may be that the articulation of a goal is not clear. Or, it may be that in goal implementation, varying emphasis on goals may exist. Therefore, one goal may overshadow another, or one goal may lose its value and thus not be an agreed upon goal in operation at all. For example, while goals 6 and 7 are shared goals,

6. To link research with the private sector
7. To strengthen Mexican economic development through research

These goals appeared to be emphasized more by the Mexican research consortium and less emphasized by the U.S.A. host institution administrators, U.S.A. PIs, and Mexican PIs. While U.S.A. host program directors mentioned these as program goals, they did not appear as values to U.S.A. PIs. U.S.A. and Mexican PIs primarily were focused on generating new knowledge, regardless of use by industry. However, some Mexican PIs were interested in generating research knowledge that would be applied to industry. Those Mexican PIs who had an interest in applying their research to industrial solutions came from the engineering and technology disciplines.

While it was evident that some Mexican PIs appeared to be more concerned with the applied aspect of their research than did U.S.A. PIs; the question of whether the research could be applied within the time frame to determine success for the program was of concern. In other words, would being able to identify whether research was applied to industry be possible within the time frame of a one or two year program evaluation plan to determine success of the collaborative research project?

When different goals are present within a program, it does not mean that that the program is impossible to evaluate for its success. However, it does pose challenges when attempting to identify the success of the program or more specifically, when attempting to identify where the program is successful and where it is not (Palomba & Banta, 1999; Suskie, 2004).

In Appendix A, some shared goals have been articulated. In Appendix B, are included some shared outcomes. However, as previously mentioned, emphasis on these goals and outcomes varied and thus the question of whether they were truly shared is suspect. For example, Mexican PIs felt more expectations to achieve shared outcome number 10 "a. Research findings will be presented to appropriate businesses and b. Research findings will inform at least one Mexican and U.S.A business development" than did U.S.A. PIs. While this is indeed an agreed upon program outcome, whether the collaborative research program should be evaluated on the basis of such an outcome was debated by both U.S.A and Mexican PIs.

It was inconclusive as to whether: (a) there were agreed upon goals and a lack of understanding of the operationalization of those goals, or whether (b) there was a prioritization of certain goals above others. For example, most U.S.A. PIs thought that it was a clear goal of the program for them to generate publications. Mexican PIs felt this also, but some Mexican PIs did not feel that generating publications was more important than making sure that their research was applied to industry. U.S.A. PIs did not seem to think that applied research was truly necessary. Regardless, some PIs disagreed with many of the stated outcomes in Appendix B. Both Mexican and U.S.A. PIs wanted clearer and more explicit program goals and outcomes to be delineated for them.

Align Program Delivery to Program Outcomes

In understanding the program goals and outcomes, it is very helpful to be able to tie or map the delivery of the outcome to the outcome itself (Bresciani, Zelna, & Anderson, 2004; Maki, 2004; Suskie, 2004). Doing so helps the one evaluating the program to identify naturally occurring (Ewell, 2003) means of assessing the program and more importantly, it helps the one delivering the outcomes to ensure that there is a way in which the program goal is being delivered and a way in which its success will be identified (Bresciani, 2003).

Because it was not clear how the program goals were delivered, apart from PIs completing a proposal and being funded, the research team was not able to align several program outcomes to the delivery of those outcomes. For example, it was not clear how such outcomes as "U.S.A researchers will be able to articulate the high quality of Mexican research protocols and equipment" and "United States researchers will be able to articulate values of the Mexican culture and the impact of those values on Mexican research in the sciences" (see Appendix B) are being delivered. In other words, if these are expectations of either the PIs or the joint research program directors, how do they know these expected outcomes are being taught; where are they being learned; and where are they being realized?

Cultural Differences in the Meaning of Collaboration

Even when goals were agreed upon and the priority of their importance was agreed upon, the researchers found differences in how Mexican and U.S.A. PIs defined collaboration. Confusion of what collaboration may mean and how it would be identified has been a reported phenomenon of many cross-cultural endeavors (Montoya-Weiss, et al., 2001; Simcox, Nuijens, & Lee, 2006). The criteria for collaboration found in Appendix C were formulated primarily by the Mexican PIs during this study. Mexican PIs appeared more concerned with defining and developing a collaborative relationship than did the U.S.A. PIs.

While the TAMU research team was immersed in Mexican culture for three days, it became apparent that the formation of a deep and meaningful relationship between the Mexican PIs and the research team would have enhanced the data gathered. Nonetheless, the time spent revealed that collaboration to the Mexican PIs meant more than just equal commitment to time on task; it meant spending time getting to know the people involved in the research, forming friendships of trust, and being equal partners in carrying out the research. These values have been expressed in other collaborative partnerships with Mexican scholars (Oliva, 2000a; Oliva, 2000b). To Mexican PIs, collaboration meant establishing a relationship of trust on which the research could be built. To U.S.A. PIs, it generally meant getting the work done together. In other words, U.S.A. PIs did not feel they needed to get to know the Mexican PI personally, get to know their family, or culture; they just wanted to focus on the research itself. Kyong-Jee and Bonk (2002) discovered these same phenomena in American PIs in their study of collaborative intercultural work. They noted that Americans tended to focus on the task at hand, rather than taking advantage of opportunities to build relationships with their cross-cultural colleagues.

It appeared that those who had already established relationships through international conferences considered themselves successful in collaborating on the research. Those who were trying to build collaborative relationships in order to conduct the research found the time limitations of the research project deadlines constraining on the formation of their relationships. For example, one U.S.A. PI reported, "It would have been helpful to establish a relationship prior to writing a proposal for research." A Mexican PI reported that it would have been helpful to be able to meet face to face during the writing of the proposal as well as meeting face to face in gathering the data and writing the research report.

Further, many of the Mexican PIs felt they had to spend time in the research partnership educating U.S.A. PIs that they could be equal intellectual partners. While the U.S.A. PIs shared that they did not feel any inequity in the partnership in regards to intellectual contribution or in regards to what quality research was and how it looked; they did feel that the Mexican PIs were disadvantaged with quality of research equipment. The researcher posited that it may be the U.S.A. PIs concern for shared resources and technology that may have been received by the Mexican PIs as a potential challenge to their ability to contribute to the research. To put this more bluntly, could it have been that the Mexican PIs felt that when the U.S.A. PIs questioned them about the type of technological support they would have to conduct the research that they felt their intellectual ability was being challenged?

In regards to collaborating on the research project, none of the researchers reported feeling that they did more work than their counterparts; however, there appeared to be a few challenges around the meaning of work ethic. Mexican PIs could not understand why U.S.A. PIs did not take time to better understand who they were, why the research was important to them, and how they conducted their work within their family and community. U.S.A. PIs voiced frustration in what they perceived as delayed response time from the Mexican PIs. While the U.S.A. PIs stated that the delayed response time was frustrating, they assumed it was due to the poor working conditions of their colleagues or their difficulty obtaining resources to complete their portion of the research. The Mexican PIs did not feel that they were delaying in responding. Rather, they were taking time to reflect on the work within the context of their family and culture. Thus, collaborative work ethic was not viewed in the same manner by the majority of U.S.A. PIs and Mexican PIs. Regardless of the possible interpretations of meaning that the researcher is positing, it is clear that misunderstanding of meanings was evident in many of these collaborative relationships (Oliva, 2000a; Oliva, 2000b).

Challenges in Discussing the Next Steps

Both Mexican and U.S.A. PIs considered the research project a success when their work had been accepted for publication in their discipline appropriate high quality research journal. While all PIs interviewed had successfully published their work and all reported satisfaction in the research findings and the quality of the work completed; not all were satisfied with the journals they published in, nor were they satisfied with the next steps in the study.

Some of the PIs, those who had formed relationships at professional conferences prior to the commencement of this joint research project, were planning to continue their collaborative research. The other PIs were not. Once the terms of the U.S.A. host institutional grant were completed (e.g., the successful conclusion and publication of the research that was funded by the host institution); the U.S.A. PIs were not interested in pursuing additional research as they reported very few, if any, funds available to continue to finance the joint research projects. Both Mexican and U.S.A. PIs appreciated the publications that resulted from the joint research, but they were equally frustrated by the lack of funding available from the U.S.A. federal government, the Mexican government, the U.S.A. host institution, or the Mexican research consortium to further the collaborative research. Thus, apart from the misunderstandings that prevailed, the majority of PIs wanted to continue the research since they were pleased with the quality of their results' however, additional funding (e.g., funding beyond the seed stage) was not available and therefore, PIs, particularly U.S.A. PIs did not want to pursue the collaborative work. Many of them felt they would have a better chance of getting funding if it were a U.S.A. PI led project only.

One further concern of most of the Mexican PIs was the expectation that they must apply their research to improving the Mexican economy. Many Mexican PIs were concerned with the practicality of applying their research to industry as quickly as seemed to be required by the Mexican Research Consortium, a partner in the grant agreement. The Mexican PIs just simply did not think it was possible to move from a focus of generating new knowledge to application of that knowledge. The U.S.A. PIs did not share this concern, nor did they feel the expectation to apply the research. U.S.A. PIs were simply interested in obtaining additional grant funding to continue their research.

As previously mentioned, there was interest in both Mexican and U.S.A. PIs to continue collaborative research; however, the ability to identify funding for on-going collaborative research was not apparent. Thus, the U.S.A. PIs felt that their only choice was to continue on without their Mexican partners as they perceived they had a better chance to gain grant funding without them.

Considerations for Future Research and Program Improvements

Continuing this type of research in collaborative intercultural partnerships may help those involved in such projects to identify where PIs are misunderstanding expectations and meaning of words. While phenomenology's intent is to simply understand what has occurred and therefore, findings are not generalizeable, the researcher posits a few suggestions that may be considered for administrators intending to design a program for cross-cultural research collaboration.

Variance in Program Goals and Outcomes for the Research Program

In order to address the findings and questions surrounding whether there were agreed upon goals and outcomes, it may be helpful for the program administrators, from both the Mexican research consortium and the host institution to seek consultation from PIs about the goals of the program and how those can be clearly articulated. In articulating the goals, specific outcomes could be identified as well as appropriate means to evaluate these outcomes. In doing so, program administrators may be able to identify the areas of disagreement and determine whether the disagreements in program goals will significantly hinder the expectations of the collaborative research partnership.

However, this presupposes that faculty members would be concerned with the "success of the program" in the same manner that administrators would be. It may be more typical that faculty members would be entirely focused on the success of their collaborative effort (their research project) and may not consider the nature of the program that is funding them. This may mean that more needs to be done by program administrators to communicate the importance of having the success of the program be held as a common objective by all and to all those involved in the join research projects.

It is plausible that with the refinement of an assessment plan for this program, program administrators can clarify the priority of the program goals so that PIs know whether it is more important to the success of the program for them to generate publications or generate contacts for applying their research to industry. Yet, given the aforementioned concern that faculty and administrators may not share outcomes for program success, faculty may remain primarily interested in whether their own research projects generate new knowledge and publications (e.g., the outcomes for which U.S.A. and many other countries' faculty are individually rewarded).

Once an assessment plan has been refined, and goals and outcomes clearly articulated, the communication of this assessment plan to potential PIs may help clear any misunderstandings that have occurred in the past. Once an assessment plan has been written, the goals, outcomes, and means of evaluation will help those participating to understand clearly the expected outcome of their participation and how it will be evaluated. Finally, with the clear articulation of goals and any variance in prioritization of those goals, PIs may feel less unsure about the value of their own engagement in the program, and be able to recruit additional PIs to participate in the program more easily.

Difficulty in Alignment of the Program Delivery to Program Outcomes

It may help the assessment of this program and the clarification of goals for the administrators to align the goals with the outcomes and the means to deliver those outcomes. In articulating the design to deliver the outcomes, the administrators may be able to identify additional means of evaluating the program other than those represented in Appendix A, B, and C. For example, in the host institution goal number 1, "to engage in collaborative research with Mexican scholars," the outcome tied to this goal could be, "U.S.A. researchers will report a high level of satisfaction of engagement in collaborative research with Mexican scholars." If this is an outcome of the program, how are U.S.A. and Mexican PIs taught about collaboration? What does it mean? How is it embodied? What would it look like if it occurred? How would it be identified?

Answering these types of questions could help the program administrators and PIs clarify the goals and outcomes of the program, identify whether or not they are being delivered, and provide clues as to how the outcomes can be more meaningfully evaluated. For example, if collaboration is being taught in an introductory workshop, then case studies that teach about multi-cultural collaboration can be utilized as an assessment tool as well as a teaching tool. In addition, at the end of the research experience, the PIs could rate the extent that collaboration occurred.

Cultural Differences in the Meaning of Collaboration

It may be helpful for the program to consult with both Mexican researchers and U.S.A. researchers in order to construct definitions for collaboration and quality that extend beyond the definitions that prevail in either culture. With facilitation from those who may be bi-cultural or well versed in each culture, PIs may be able to jointly define collaboration and quality in order to advance the program's goals and to continue in cross-cultural research beyond the life of the seed grant.

However, as illustrated by the earlier two points, simply defining collaboration is not enough to ensure collaboration. Articulation of collaboration as program goals and outcomes implies that PIs will have an opportunity to learn what that looks like and to practice in its identification. Further, it may be wise to have interventions available that mitigate problems that may arise from cross-cultural understanding so that research could be further enhanced.

Challenges in Discussing the Next Steps

When refining the assessment plan and implementing it, it may be helpful for program administrators to identify publications where collaborative international research is encouraged, as well as to identify additional funding resources so that PIs can determine more quickly, how suitable their proposals will be for submitting to certain publications and grants. The potentially variable needs of researchers from the two countries need to be taken into account in establishing the common outcome expectations.

To date, there is no known research grant program that fully and consistently funds collaborative international research. While there may be grants that do fund this type of research in order to build the collaboration, the lack of funding available for sustaining collaborative international research may result in seed grant money for intercultural collaborative research not being able to be continued. Grants that fund ongoing collaborative international research should be identified so that PIs can continue their research. Furthermore, several PIs shared concerns about transfer of funds. The frustration voiced by both Mexican PIs and U.S.A. PIs with the transfer of program funds illustrates that some of the inner workings of the program may need to be evaluated in order for optimal work environments for PIs to be realized. The time allotted for this research project did not allow the researchers to investigate this phenomenon further and thus it did not emerge in the cluster of meanings. However, the majority of PIs struggled with the cross-border management of the funds.

Given different business processes of the two countries and its apparent negative impact on the transfer of funds, it is important to determine whether other cross-national business and program implementation operations are keeping the program from meeting its goals for collaboration.

While this study uncovered that principal investigators in different countries perceived collaboration differently, one may wonder how scholars in varying disciplines approach effective collaboration. In an environment where collaboration is needed to improve student learning (Bresciani, 2006; Maki, 2004; Palomba & Banta, 1999), it may benefit institutions to investigate how professors perceive collaboration in order to improve student learning. In addition, institutions may be encouraged to explore how professors and co-curricular professionals view the need to work together in order to improve the whole student educational experience. It is this researcher's hope that this methodology may be replicated in a number of venues in order to uncover meaning around perceptions of collaboration that will be used to improve student learning.

References

Bogdan, R.C. & Biklen, S.K. (1992). *Qualitative research for education.* New York: Allyn and Bacon.

Bonnema, A.B. (2006). Developing institutional collaboration between Wageningen University and the Chinese Academy of Agricultural Sciences. *NJAS Wageningen Journal of Life Sciences*, *53*, 369 -386.

Bresciani, M. J. (Ed.). (2007). *Good practice case studies for assessing student learning in general education.* Bolton, MA: Anker Publishing.

Bresciani, M.J. (2003, December). Identifying projects that deliver outcomes and provide a means of assessment: A concept mapping checklist. *National Association for Student Personnel Administrators,* Inc NetResults E-Zine. http://www.naspa.org/membership/mem/nr/article.cfm?id=1291

Bresciani, M.J., Zelna, C.L., Anderson, J.A. (2004). *Techniques for assessing student learning and development: A handbook for practitioners.* Washington, D.C.: NASPA, Inc.

Creswell, J.W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Denzin, N. & Lincoln, Y. (Eds.). (1994). *Handbook for qualitative research.* Thousand Oaks, CA: Sage.

Ewell, P. T. (2003). *Specific roles of assessment within this larger vision.* Presentation given at the Assessment Institute at IUPUI. Indiana University-Purdue University- Indianapolis.

Holstein, J.A. & Gubrium, J.F. (1994). Phenomenology, ethnomethodology, and interpretive practice. *Handbook for Qualitative Research.* Thousand Oaks, CA: Sage.

Kyong-Jee, K. & Bonk, C.J. (2002, October). Cross-cultural comparisons of online collaboration. *Journal of Computer-Mediated Communication*, *8* (1). Retrieved September 25, 2006, from http://jcmc.indiana.edu/vol8/issue1/kimandbonk.html

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

Montoya-Weiss, M. M., Massey, A. P. & Song, M. (2001). Getting it together: Coordination and conflict management in global virtual teams. *Academy of Management Journal, 44*(6), 1252-1262.

Moustakes, C (1994). *Phenomenological research methods.* Thousand Oaks, CA.: Sage.

Oliva, M. (2000a). Cooperación vs. integración: Fundamentos distintos de la colaboración norteamericana en educación superior. *Interciencia: Revista de Ciencia y Tecnología de América,* *25* (2), 96-102.

Oliva, M. (2000b). Shifting landscapes/shifting langue: Qualitative research from the in-between. *Qualitative Inquiry*, *6(1)*, 33-57.

Oliva, M. (2002). Perspectives on collaboration: Deconstructing North American higher education cooperation. *Planning and Changing, 33* (1 &2), 29-52.

Palomba, C. A., & Banta, T. W. (1999). *Assessment essentials: Planning, implementing, improving.* San Francisco: Jossey-Bass.

Polkinghorne, D.E. (1994). Reaction to special section on qualitative research in counseling process and outcome. *Journal of Counseling Psychology, 41*, 510-512.

Schutz, A. (1970). *On phenomenology and social relations.* Chicago, IL: University of Chicago Press.

Simcox, A.G., Nuijens, K.L., C.C. (2006, April). School counselors and school psychologists: Collaborative partners in promoting culturally competent schools. *Professional School Counseling*, *9*(4), 272-277.

Stake, R.E. (1995). *The art of study research.* Thousand Oaks, CA: Sage Publications.

Suskie, L. (2004). *Assessing student learning: A common sense guide*. Bolton, MA: Anker Publishing Company, Inc.

Appendix A

Goals of the Collaborative Research Program

Goals:
Based on initial interviews and document analysis, the following 'represents the goals and values of the program from the USA host, from the Mexican Research Consortium, and from both organizations which are labeled as *Shared*.

Shared:

1. to provide a competitive, peer reviewed collaborative research grant program
2. to advance the inter-institutional cooperation in science, technology, and scholarly activities
3. to equally advance the research efforts of scientists and scholars from USA host institution and Mexican Institutions
4. to provide seed funding for research start-up
5. to conduct quality research so that research projects are eligible for additional funding from external funding sources
6. to link research with the private sector
7. to strengthen Mexican economic development through research
8. to engage in research that solves an industrial or governmental problem
9. to promote collaborations between the USDA and Mexican equivalent of USDA through large, multiple university projects
10. to move successful research into entrepreneurial opportunities
11. to develop the continent's capacity for improvement in technology, science, and human capital

USA Host Institution:

1. to engage in collaborative research with Mexican scholars
2. to gain experience in applying for grant funded research

Mexican Research Consortium:

1. to engage in equitable collaborative research with USA scholars
2. to develop relationships that advance Mexican institutional research
3. to educate researchers from the United States about Mexican culture
4. to establish a successful agenda of research with the USA host institution so that it can be expanded to other premiere research institutions in the USA
5. to educate USA researchers about the high quality of Mexican research protocols and equipment

Appendix B

Intended Outcomes of the Collaborative Research Program

Outcomes:

Shared:

1. a. Program administrators will provide a competitive, peer reviewed research grant program
 b. Program administrators will provide a collaborative research grant program
2. a. Funded research grants will contribute to new knowledge for both partner institutions in the areas of science and technology
 b. Researchers participating in the program will report a mutually collaborative relationship with each other
 c. Researchers participating in the program will report a mutually collaborative relationship from the partner institutions
3. a. Researchers participating in the program will report a mutually beneficial relationship with each other
 b. Researchers participating in the program will report equity in research contributions to the project
4. Researchers will receive grant funding for their proposed research
5. a. The research reports will be rated as quality according to report reviewers
 b. Researchers will apply for additional grant funding from other foundations
 c. Researchers will be awarded additional grant funding from other foundations
6. Researchers' findings will be publicly recognized by the private sector
7. Researchers' findings will aid in the solution of an industrial or governmental problem
8. a. Research findings will be presented to appropriate economic development agencies.
9. a. Appropriate research findings will be presented to the USDA and the Mexican equivalent of USDA
 b. The USDA and the Mexican equivalent of USDA will respond to the research presentations with suggestions for future developments in the research
10. a. Research findings will be presented to appropriate businesses.
 b. Research findings will inform at least one Mexican and USA business development
11. This goal will be met through shared outcomes 8, 9, and 10, and Mexican Research Consortium outcomes 2 and 4.

USA Host Institution:

1. USA researchers will report a high level of satisfaction of engagement in collaborative research with Mexican scholars
2. a. USA researchers will report a positive gain in experience in applying for grant funded research

Mexican Research Consortium:

1. Mexican researchers will report a high level of satisfaction for equitable collaborative research with USA scholars
2. a. The research results will be published in a scholarly journal
 b. The researchers will present their findings at a scholarly conference
 c. Findings from four years of collaborative research with USA Host Institution will be presented to at least five other research institutions in the hopes of securing another institutional partner
 d. Research findings will inform at least one Mexican and USA economic development
3. United States researchers will be able to articulate values of the Mexican culture and the im pact of those values on Mexican research in the sciences.
4. a. USA Host Institution administrators will report that the USA Host Institution – Mexican Research Consortium relationship has produced successful research projects
 b. Mexican Research Consortium administrators will report that the USA Host Institution - Mexican Research Consortium relationship has produced successful research projects
 c. USA Host Institution and Mexican Research Consortium administrators will co-present the findings of the summary of the collective research at scholarly conferences
 d. USA Host Institution and Mexican Research Consortium administrators will publish the

e. USA Host Institution and Mexican Research Consortium administrators will co-present the findings of the summary of the collective research to at least five other institutions annually in an n effort to promote further inter-institutional collaboratio

f. USA researchers will be able to articulate the high quality of Mexican research Appendix Criteria for Defining Collaboration in Cross-Cultural Studies

Appendix C
Criteria for Defining Collaboration in Cross-Cultural Studies

Collaborative Research

50/50 split of the work load
Good working relationship
Good friendship
High quality work produced
Understanding of each others' research goals
Trusting relationship
Time-management utilized
Positive attitude kept throughout course of research project
Task-focused
Well-prepared
Equal exchange of research
Good communication
Work funding received and utilized in a timely manner
Good work ethic demonstrated

Detecting Items That Function Differently for Two- and Four-Year College Students

Amy Thelk
James Madison University

Abstract

Differential Item Functioning (DIF) occurs when there is a greater probability of solving an item based on group membership after controlling for ability. Following administration of a 50-item scientific and quantitative reasoning exam to 286 two-year and 1174 four-year students, items were evaluated for DIF. Two-year students performed better-than-expected on 13 items and worse than expected on 10 items. Reasons for DIF are explored, along with the importance of conducting this type of study.

Introduction

As institutions commit to greater assessment activities on their campuses, the search for appropriate instrumentation ensues, especially in the measurement of student learning. Assessment professionals may opt to adopt or adapt an exam that was developed at another site to gauge student learning at their institutions. When using an exam developed at one location to assess students at a different school, the expectation is that any set of students with the same ability should perform about the same on a given test item. However, due to other factors, like on-campus culture, socioeconomic differences, and variations in exposure to material, student scores may diverge despite similar ability. Examination of differential item functioning (DIF; Hambleton, Swaminathan & Rogers, 1991) can inform consumers of tests about whether factors other than ability affect test scores.

For the community college and the 4-year institution that served as sites for this study, assessment has been incorporated into their academic schedules; students are aware of mandated testing at the time of application. Additionally, a professional partnership exists between the two schools: the four-year school serves as a transfer site for the community college, and some of the instruments developed at 4-year school are leased out to the community college.

Scientific and Quantitative Reasoning Assessment

The scientific and quantitative reasoning instrument (SR/QR) used for this study had been developed over the course of several years at the four-year institution. The items had been crafted by faculty experts in science and mathematical disciplines with the assistance of measurement experts.

At both institutions that served as data collection sites, dedicated "assessment days" were held during the spring semester; classes were cancelled for the day so that students participate in the required testing without potential schedule conflicts; the data used for this research were collected during such assessment days. For this study, both institutions administered the same version of the SR/QR.

Differential Item Functioning (DIF)

When students have the same ability level, the probability of solving a given item correctly should be the same for any student. However, sometimes factors other than ability are influential upon the score: access to information, language skills and testing conditions, for example. If different groups comprise the test-taking population, a DIF study can be designed and implemented. For this study, the data set was divided into two groups, 2-year-school students and 4-year-school students. Hambleton, et al (1991). provide a concise and useful definition of DIF: "An item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting an item right" (p. 110). Figure 1 further illustrates DIF.

DIF is instrumental in alerting test users to the possible presence of bias at the item level. The presence of DIF is a necessary component of bias, although not sufficient in itself to deduce that bias is present. If DIF is found, further investigation must take place to determine whether the differences in performance on these items are due to unfairness. A somewhat less alarming situation would be the case of an item showing DIF because students in that group have not had course exposure that would assist in solving the item successfully. In any case, a DIF analysis can provide preliminary evidence about the degree to which certain test items are biased for or against particular groups.

Method

For both institutions, testing was mandated for students and held on designated days on which classes were cancelled. Two-hundred eighty-six community college students and 1174 four-year college students participated in testing, yielding the data used for this project. The raw data were scored for each group and the two data sets were concatenated following the addition of a group-ID variable. To determine which items on the SR/QR demonstrate DIF between the community college and four-year college groups, item parameters were first estimated by item response theory (IRT). DIF was then calculated using these item parameters.

In IRT three main models are used to estimate item parameters. These models are, in order of complexity, the one-parameter logistic model (1-PL), 2-PL and 3-PL (Hambleton et al., 1991). Researchers decide which model is most appropriate for their studies by considering the sizes of their samples and evaluating model fit. The 1-PL only takes item difficulty into consideration, the 2-PL takes difficulty and discrimination into account, and the 3-PL models item difficulty, discrimination, and guessing. The first parameter is b (item difficulty), the second is a (item discrimination) and the third is c (guessing). As a general rule of thumb one should not apply a 1-PL model unless the sample has at least 200 participants. Four hundred and 1,000 are the suggested sample size minimums for the 2-PL and 3-PL models respectively. The size of our sample (1173) and the nature of our data (multiple choice items with a variety of difficulty and discrimination levels) suggested that a 3-PL model would be a logical starting point, and an analysis comparing the 1-PL, 2-PL, and 3-PL models confirmed that the 3-PL model did indeed result in the best fit.

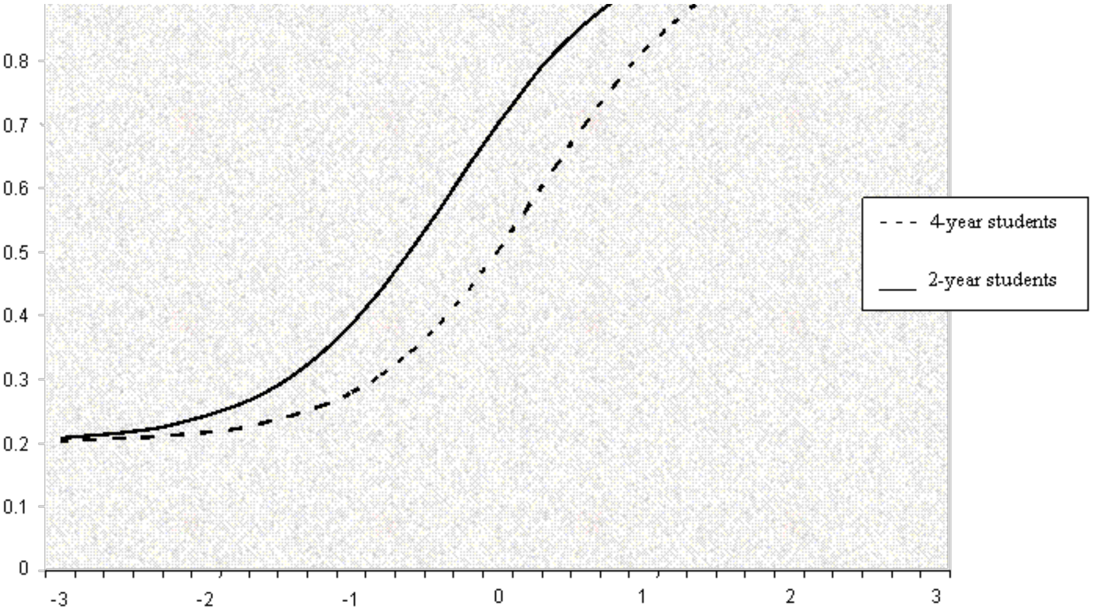

Figure 1. Example of an item showing DIF between two-year students and four-year students.

In IRT, ability (denoted by θ) is measured on a scale with 0 representing average ability and with each point above or below representing a standard deviation. For example a score of "+1" would represent ability at one standard deviation above the average and a score of "-2.5" would represent ability at two-and-a-half standard deviations below the average. The b parameter reflects at what ability level 50 percent of test takers get the item correct. When these values are aggregated over items and averaged, the result is the difficulty value for the entire test.

For this study, two methods of detecting DIF were employed. The first uses IRT to determine whether the item response characteristics look different across testing groups (Hambleton et al, 1991). Essentially, a null hypothesis is being tested to determine whether there are significant differences when groups are compared.

Using the output generated by BILOG-MG (Zimowski, Muraki, Mislevy & Bock, n.d.), the appropriate values were input into the equation (b_1 - b_2)/ σbdiff, where b_1 and b_2 are the difficulty values for groups 1 (community college students) and 2 (four-year students), respectively, and σ^bdiff is the standard error of the difference between the two b values in the numerator. The solution to this equation is distributed as a z-score (M=0, SD=1). Based on the results of the equation above for each item, any items with an absolute value z-score greater than 2.58 (corresponding to a two-tailed $p \leq .01$) were pulled out to examine for DIF, since these z-scores flagged a significant difference between the b values between groups for those item.

The second method involved the calculation of Mantel-Haenszel (M-H; Hambleton et al, 1991) statistics for each of the items that exhibited high absolute z-scores to confirm the presence of DIF. The M-H value is a common-odds ratio that represents a proportion with Group 1 in the numerator and Group 2 in the denominator. For this research, when this value was greater than 1 then the item favored Group 1, and when the value was lower than 1 it favored Group 2.

To determine effect sizes of the difference between difficulties, delta (Δ) values were evaluated. Delta values are calculated by locating the odds-ratio, or α, value on the output resulting from the M-H procedure, and substituting that value into the equation $\Delta = -2.35 \ln (\alpha)$.

Based on the Educational Testing Service scale for effect size, these Δ values are classified into A, B and C categories (Dorans, 1989). If the absolute value of Δ is less than 1, the magnitude of the effect is negligible; this is considered an "A" item. When the absolute value of Δ is between 1 and 1.5, the item is placed in the "B" category. Items that show the most DIF have an absolute Δ value greater than 1.5; these items are placed in the "C" category.

Results

Out of the 50 SR/QR items, 23 items had high absolute z-scores. Using M-H statistics, the presence of DIF was confirmed in all of these items, and the group the item favored was ascertained. Out of the 23 items that showed DIF, 13 of the items favored the two-year college, while 10 favored the four-year school. Calculation of effect sizes revealed that 22 out of 23 of the items were placed in category C connoting the highest amount of DIF. Table 1 provides a summary of these results.

A review of the item content revealed that the items biased in favor of the community college students pertained to higher order reasoning skills such as evaluating a claim or ascertaining the relationship between variables by interpreting a graph. In other words, controlling for ability, community college students did better than expected on these items. Many of these positively biased items were also part of testlets. Testlets are two or more items related to a single stimulus. Conversely, the items that two-year students missed more than expected controlling for ability (i.e., biased against the community colleges) were those related to performing routine algorithms.

Discussion

According to Anderson and Sundre (2005) examining DIF between two-year and four-year students is important because many assessments used by two-year institutions were developed for and normed on four-year students. When selecting an established instrument, colleges will want to review the fit of the items to the institution's objectives. However, exploring DIF after initial use of the instrument will assist with identifying items that have more subtle problems associated with bias. It is worth noting that just because an item favors the community college group does not necessarily mean that this group scored higher on that item. Indeed, for many items the two-year students still scored lower, but they did not score as low as expected.

Table 1
Results from both DIF Analyses (IRT and M-H) for Items Showing DIF

Item #	Group 1 *b* value	Group 2 *b* value	*b* difference	z-score	MH Odds Ratio	Group favored	Delta (Δ)	Category
31	0.524	3.500	-2.976	-9.537	0.173	1	4.120	C
10	-0.232	2.482	-2.713	-7.713	0.234	1	3.413	C
37	-0.384	2.248	-2.633	-6.169	0.264	1	3.130	C
22	0.957	2.669	-1.712	-5.388	0.390	1	2.212	C
40	0.909	8.882	-7.972	-4.773	0.120	1	4.981	C
24	-1.414	0.408	-1.822	-3.861	0.359	1	2.405	C
35	1.964	6.344	-4.380	-3.417	0.369	1	2.344	C
42	-0.949	0.575	-1.523	-3.246	0.437	1	1.944	C
26	-1.297	6.269	-7.565	-3.236	0.526	1	1.509	C
32	-0.018	1.235	-1.253	-3.156	0.467	1	1.791	C
14	-0.088	2.480	-2.569	-3.095	0.498	1	1.638	C
36	0.000	2.577	-2.577	-2.773	0.512	1	1.571	C
20	0.709	1.941	-1.232	-2.751	0.554	1	1.386	B
38	-0.781	-2.227	1.447	2.795	2.490	2	-2.144	C
21	0.585	-1.333	1.918	3.172	2.239	2	-1.895	C
30	0.688	-1.334	2.022	3.308	2.021	2	-1.654	C
39	0.763	-2.180	2.942	4.029	3.565	2	-2.987	C
43	-0.247	-2.005	1.759	4.087	3.079	2	-2.642	C
25	-0.589	-2.573	1.985	4.631	4.478	2	-3.523	C
50	11.551	-1.320	12.871	5.161	479.799	2	-14.507	C
7	8.957	-0.477	9.434	5.541	38.276	2	-8.565	C
47	3.816	-1.728	5.544	8.024	38.087	2	-8.554	C
45	0.939	-2.032	2.971	8.261	11.198	2	-5.677	C

Note. *Z*-scores ≥ |2.58| suggest DIF; **Δ** values greater than 1.5 signify high amount of DIF (category C). Categories based on Educational Testing Service classification (Zieky, 2004)

As mentioned earlier, when different groups have unequal probabilities of getting a test item correct after controlling for ability, DIF is present. Indeed, in this study many items, almost half of the total, showed DIF for and against community colleges students. They performed better than expected on 13 items and worse than expected on 10 items.

While reviewing these items, the author speculated about what factors may have contributed to DIF. Since many of the items are one part of a testlet, it is conceivable that community college students are more persistent and less likely to get bored or fatigued, and therefore do not skip items or answer carelessly as often. Persistence across groups may be worthy of future investigation.

Another factor is that these two groups represent two very different institutions, with varying curricula and objectives. So in some cases the two-year group may have actually covered certain material to a greater extent than the students at the four-year school and less of other curricular components. Relatively speaking, perhaps the community colleges spent more time on the reasoning components of science and less on applying algorithms. A counter argument is that reasoning may be acquired outside of traditional classroom learning. Given that these community college students were older and likely have had a wider array of experiences, this scenario cannot be ruled out. Such a situation would illustrate Messick's (1995) concept of construct irrelevant variance: performance on test items is due to an influence outside of the learning arena at which the instrument is aimed.

Following the administration of this test, new items were introduced to the SR/QR test form, while some of the previous items were removed due to low scoring or inappropriateness to the curriculum. Items showing DIF that were not removed were retained on a provisional basis, with the test's advisory team committing to continually analyze test data to determine the appropriateness of including these items on later versions. If these changes in the exam had not been made for the community college group, the results of this DIF study would have presented great urgency for further test review before using the exam in its original state for this population.

Summary

When performance on items is different than expected for a group, DIF is present. This article describes how DIF was indentified when comparing results of two and four-year students on the same test, and explored reasons for its presence.

As testing for the purpose of gauging student learning becomes more common, many postsecondary schools will find themselves in need of already developed instruments that are appropriate for their own testing programs. This DIF study serves as an important cautionary reminder about comparing test results of two different groups of students. Since students are exposed to a variety of instructional styles, classroom sizes, and campus cultures, it is unlikely that their performances on test items will be similar, even after controlling for any differences between the groups in overall test score. So by gathering information about differential item functioning, more appropriate comparisons can be made between or among groups.

A DIF study is a useful way to determine whether test items created for one student group yield comparable information when administered to another group. The analysis is relatively quick and only requires a data set for each diverse group, but the information that is produced is essential to the validity of the scores generated by the assessment. If the students in your school are not performing as expected as indicated by DIF, then the validity of the inferences made by the test scores, particularly comparisons among groups of students, are likely suspect.

References

American Association of Community Colleges. (2005). Fast facts: Community college fact sheet. Retrieved May 19, 2005, from: http://www.aacc.nche.edu.

Anderson, R. D., & Sundre, D. L. (2005). Assessment partnership: A model for collaboration between two-year and four-year institutions. *Assessment Update*, *17*(5), 8-16.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education*, *2*(3), 216-233.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

National Center for Education Statistics. (2003). Retrieved May 18, 2005, from http://nces.ed.gov/programs/coe/2003/section5/indicator32.asp.

Zieky, M. (2003). *A DIF Primer.* Princeton, NJ: Educational Testing Service.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (n.d.). BILOG-MG [Computer software]. St. Paul, MN: Assessment Systems Corporation.