Chapter XXIV
Method Development for Assessing a Diversity Goal

Joan Hawthorne
University of North Dakota, USA

Tatyana Dumova
University of North Dakota, USA

April Bradley
University of North Dakota, USA

Daphne Pedersen
University of North Dakota, USA

ABSTRACT

In this chapter the authors describe a method developed to assess the outcome of a “cultural familiarity” general education goal. Challenges in defining, measuring, and providing summary information on variables of interest are discussed. We review the process of developing our own “oral examination” assessment method, explain our rationale for using this particular method, and suggest that locally-developed methods – this one and others – may have particular benefits that make them especially useful for program review and revision. In addition, we provide insight about how this specific method could be adapted to provide meaningful data for other goals that are similarly difficult to assess in a higher education environment.
INTRODUCTION

Four years ago, the University of North Dakota (UND) went through the once-every-ten-years process of institutional re-accreditation by a regional accrediting agency, in our case, the Higher Learning Commission of the North Central Association. In recent years, the accreditation process has come to routinely include close attention to assessment of student learning. Assessing learning within degree programs is a challenge with which many institutions still struggle. But assessing general education outcomes is a particular challenge (Aloi, Gardner, & Lusher, 2003). The accreditation team’s report to the university in the wake of our site visit emphasized the need to improve our work in assessment of learning, citing assessment within the general education program as a particular weakness.

In this chapter, we describe one project that was developed in response to the clear need, pointed out by accreditors but also recognized by internal constituencies, to overhaul and improve our assessment within general education. Many of the earliest and most productive efforts to assess general education outcomes have focused around goals for which learning outcomes can be relatively easily documented, most notably in the areas of critical thinking, written communication, and oral presentation (Banta, 2007, Blattner & Frazier, 2004; Eder, 2004; Greene, 2003; Lusher, 2003; Morreale, Rubin, & Jones, 1998). But what about other goals like cultural understanding or ethical development? How do we document learning in these areas?

The project outlined in this chapter was a locally-developed solution to the need for a direct assessment of student learning around UND’s “cultural familiarity” goal, one of a handful of general education goals that are common nationwide, but for which student learning is difficult to document. What methods can generate data about outcomes for a goal that’s abstract and somewhat ambiguous? What methods will result in information that faculty will find interesting, informative, persuasive, and useful?

These questions are explored in this chapter. We describe our process of developing a measure that proved useful for examining student learning around a diversity goal, and we identify lessons learned from that process. Furthermore, we describe the method developed and provide our thoughts about how that method could be adapted to provide meaningful data for other goals that are similarly difficult to assess. Finally, we argue that “home-grown” qualitative methods of assessment, while perhaps out of step with the national drive toward comparability of measures, can often be more meaningful and more useful for curriculum revision than commercially available standardized assessments.

NAMING A DIVERSITY GOAL

Most university systems have a goal pertaining to diversity and student learning. Yet diversity itself is an evolving and contested term that invites debate. Shifting cultural, racial, and national boundaries encourage us to continually rethink how we frame diversity (AAC&U, 2006); thus it comes as no surprise that, while campuses almost uniformly have diversity goals, they also have quite divergent language describing these goals. Diversity goals may be very general or quite specific. For example, while some students may be encouraged to “respond thoughtfully to diversity,” (West Chester University, 2008) others may be charged with studying “the interrelationships of individuals, racial groups, and cultural groups to understand and appreciate issues of diversity, equality, and structured inequality in the U.S., its institutions, and its cultures” (San Jose State University, 2005). In this context, diversity is understood to “include the experiences and/or contributions of those varying in (including but
not limited to): accent, age, ancestry, citizenship status, color, creed, disability, ethnicity, gender, marital status, medical condition, national origin, race, religion or lack thereof, sex, sexual orientation, transgender, and veteran’s status.”

Diversity is often viewed as occurring at both the domestic and international levels, such that some students have a two-pronged requirement to develop an awareness of both American and global diversity. The University of Wyoming, for example, requires students to take courses in “American Diversities” and “Global Awareness,” with the latter focusing on either “a single culture, or… a regional cluster of cultures” (AAC&U, 2008).

Frequently, diversity and culture are used as interchangeable terms framing these general education plans. Noting the combination of imprecise language and controversy around desired outcome, Levine (2001) points out that few campuses have fully sorted out the pieces or defined their goals for diversity or cultural awareness. The generality of a number of diversity goals contributes to both varied interpretations of what these goals actually mean and a broad range of approaches to fulfilling them. As a result, learning experiences may be disconnected at worst and varied at best, and assessment plans are difficult to create and implement. Lacking a uniform and cohesive definition of diversity, it is clearly difficult for any institution to develop a systematic plan for assessing students’ achievement of diversity goals. Thus, it should be no surprise that universities have been slower to implement direct assessments for diversity goals than for other general education goals like communication or critical thinking.

**ASSESSING A DIVERSITY GOAL**

Still, valuing diversity has long been viewed as an integral outcome of general education along with the development of ethical conduct, leadership skills, and effective citizenship (UC Commission on General Education, 2007). During the past three decades, educational practitioners, administrators, and scholars have examined different means of integrating diversity into higher education curricula, with a variety of integration levels ranging from individual courses to campus-wide initiatives. Some institutions, for example, the University of Minnesota, have gone as far as establishing a position of Chief Diversity Officer to coordinate campus diversity efforts (Barceló, 2007). Assessing student achievement has become an important part of these initiatives.

For example, given the mission of the University of Michigan-Flint to “promote and defend academic freedom, diversity, equality and justice” (University of Michigan-Flint, 2003a, pp. 3-4), the broad intent of “increasing an awareness and appreciation of diversity” is addressed through three general education goals (University of Michigan-Flint, 2003b):

- Enhancing one’s understanding of one’s position within and relationship to the global community.
- Enhancing understanding of diversity as a national and global phenomenon with a particular focus on the American experience.
- Students will develop an appreciation for cultural, racial, and gender differences, similarities and contributions at both national and global levels. (p. 1)

These goals, in turn, are operationalized through specific student learning outcomes. As stated in the University of Michigan-Flint guidelines for outcomes assessment, students graduating from the university should be able to state the meaning of diversity, identify multiple dimensions of human diversity at societal and global levels, and recognize and articulate the benefits of and barriers to diversity (p. 2). Assessment of diver-
Method Development for Assessing a Diversity Goal

The “cultural diversity” requirement outcomes at the University of Illinois at Chicago are assessed through frequency analyses of cultural diversity courses, Study Abroad Program analyses, Student Affair and alumni surveys, as well as examination of the demographics of the student body and ethnic composition of cultural diversity courses (University of Illinois at Chicago, 1999). Indiana University Purdue University Indianapolis developed six Principles of Undergraduate Learning, with “understanding society and culture” as one of them (Hamilton, 2003). Evidence about student performance is gathered electronically through an ePort web-based system where each student is asked to upload three to five documents in response to faculty-developed “scenarios” and write a reflective essay synthesizing his or her learning experiences. Performance assessment is complemented with surveys of freshman students and graduating seniors. Student competencies are rated on a 3-point scale: exceeds expectations, meets expectations, or does not meet expectations.

Overall, despite the ubiquity of diversity goals on academic campuses, there is clearly little consensus about the meaning of “diversity” as a general education goal and even less consensus about effective means of assessment. Techniques range from direct and indirect qualitative analyses of various student work products to indirect quantitative analyses of survey data and frequency counts for diversity exposure. No tool has thus far gained widespread acceptance for purposes of cross-institutional comparability, and, perhaps more significantly, no single method seems to have “caught on” across higher education as an effective means of collecting information that might be useful for program improvement.

Development of an Assessment Model

Our work on assessing our own diversity goal, named as “cultural familiarity” at our institution (students will “develop some familiarity with cultures other than their own”), began without a great deal of thought given to the complexities of definitions (University of North Dakota, 2005). Faculty recognized cultural familiarity as an important learning outcome, perhaps especially given the nature of the institution. Located in a northern and rural state, the university attracts a student population that is sometimes view as “mainstream,” “culturally homogenous,” and heavily Caucasian, although, in fact, the University also offers an extensive array of programs which draw significant numbers of Native American students. In such an environment, there is a risk that the superficial appearance of uniformity could reduce the perceived importance of developing an understanding of and respect for difference. But a diversity or cultures goal, however labeled and defined, is important at any institution today, both for students individually (leading to enrichment of their personal lives as well as enhancement of their professional and civic futures) and for the benefit of society generally.

Previous Assessments of Diversity Outcomes at University of North Dakota

At the time of our most recent accreditation site visit, one significant indirect assessment of learning within general education had already been initiated. This was a qualitative longitudinal study of students using a semi-structured interview protocol. The study followed about 100 students throughout their time at the university, and specifically focused on learning outcomes for the general education program (The General Education Study Team, 2004). That study was...
nearing conclusion by the time of the accreditation site visit, and a final report was prepared shortly thereafter. The findings from the longitudinal study, although rooted in indirect data, were of great interest campus-wide.

Faculty who served as interviewers for the longitudinal study noted one finding that was particularly disturbing; that finding was regarding the cultural familiarity goal. Students generally saw cultural familiarity as less important than the other goals, frequently citing it as the “least important” goal (The General Education Study Team, 2004, p. 4). Furthermore, students perceived it as less emphasized within their UND experience and, therefore, less valued by the faculty and the institution. Faculty involved with the study, on the other hand, believed that achievement of that cultural familiarity goal was highly important to students’ personal, civic, and professional lives. Valuing the goal as they did, members of the interview team were dismayed to learn that many students viewed it as comparatively unimportant.

Data from another indirect assessment, this one with a quantitative design, lent support to concerns that students might not be achieving intended learning around diversity. A goals inventory of general education courses had been conducted, and, based on the results of that inventory, a map was created that showed the general education program goals that were addressed in each of the 300 courses that could be used to meet program requirements. Using that map, a transcript analysis was conducted. Academic transcripts for 50 randomly selected graduating seniors were pulled by the registrar’s office, and all general education courses were marked. By checking marked courses on each transcript against the map of program goals, it was possible to see how many “hits” each of the 50 students had on each general education goal. The results of that analysis (included in Appendix A) showed that the cultural familiarity goal, for an average student, was addressed far less frequently than any other goal (an average of just more than six hits on cultural familiarity, with all other goals receiving hits about twice as frequently).

Furthermore, a few students completed almost no courses that attempted to address this goal (in the case of a single student, only one relevant course occurred on the transcript) while every other goal was addressed in a minimum of five courses for each of the 50 students. Although the transcript analysis could not provide any information about student learning, it did indicate that there was a dearth of exposure to courses in which the goal was purposefully addressed.

Based on results of the longitudinal study and the transcript analysis, as well as the ensuing discussions on campus, a decision was made to seek out some means of directly assessing student learning around the diversity goal. We knew that students placed less value on the diversity goal than on other goals and we knew that students often had less exposure to courses addressing this goal, neither of which was a good sign. But we needed to know more about their actual learning.

**Review of Existing Methodological Options**

Such an assessment effort was particularly well-timed in that a task force recently had been appointed to conduct an extensive review of the general education program. Task force members were authorized to recommend changes in program goals, revisions in the language with which goals were named or described, changes in the curriculum itself, and plans for future assessment methodologies for program goals. Given the existence of that task force, findings from a direct assessment could potentially have a long-term impact on general education at the university.

The question, of course, was how to conduct such an assessment. Those involved in general education and its assessment were aware of a small number of methods, both quantitative and qualitative, that had been used elsewhere to collect data about students’ learning around diversity or
cultures goals. Some institutions use portfolios (Banta, 2003), usually electronic (see, for example, Hamilton, 2003; Rogers & Chow, 2003; Schechter, Testa, & Eder, 2003), to assess various general education outcomes. The portfolio process could be implemented by requiring students to save materials demonstrating their learning around each goal, although another possibility was to ask faculty to save student work products on behalf of their students. For a portfolio assessment of general education, materials saved would need to demonstrate student learning around diversity, as well as learning around critical thinking, written communication, and any other program goals (Schwartz, 2006). Students might also be asked to include reflections that described their own strengths and weaknesses around each goal, formative experiences in their own learning, and an analysis of the submitted documentation (e.g., Spicuzza, 2003).

A portfolio method can be extremely useful if in place – and if there is a means of ensuring that the portfolios are created consistently, with care and thoughtfulness, and a sample of portfolios is then systematically reviewed to collect information about student learning (Eder, 2004). However, there was no existing portfolio requirement for most UND students, and no means of ensuring that students would prepare such portfolios thoughtfully (if they could be persuaded to create them at all). It would take some years to amass a suitable collection of materials if plans were developed and implemented immediately. Given the complexities of developing an appropriate portfolio process and the likely timeline for implementation, this method seemed ill-suited to our immediate need.

Capstone courses provide another convenient means of collecting student work products to assess learning outcomes (e.g., Rhodes & Agre-Kippenhan, 2007). However, our institution did not have a capstone course requirement in place, and the few capstones that existed (often in fields like engineering) rarely included an emphasis on culture or diversity. It was clear that capstones, like portfolios, would not meet our need.

Some institutions rely on indirect assessments for difficult goals, collecting student perception data via surveys. Data from commonly used tools like the National Survey of Student Engagement (NSSE) were already collected at UND, and additional home-grown surveys were also administered (e.g., the Sophomore Satisfaction Survey). Some of these existing tools include questions which could aid in the examination of student attitudes around diversity; other tools could be modified to provide relevant data by adding specially designed questions. However, such tools would provide far less detailed information about general education outcomes than was already available from our longitudinal study. Useful though surveys can be (e.g., for comparability across institutions or across years within a single institution), collecting additional indirect assessment data was unlikely to provide significant new information that would stimulate additional discussion and aid future decision-making.

Furthermore, our interest was in understanding more about actual student learning, which requires direct assessment. Indirect methods can only provide information about the students’ perceptions of learning. One quandary specific to a “cultural familiarity” goal is that individuals may not be consciously aware of when their knowledge and familiarity with diverse groups is being developed or utilized. Therefore, there was a strong potential for methods based on student perceptions to miss actual learning related to the goal. To obtain the information we were looking for we would need a direct assessment method.

A few colleges and universities use commercially available tests like the Intercultural Development Inventory™ (IDI). Such tests have the advantage of coming with studies of the instrument’s validity and reliability (Jones, 2002), which may be important to potential users. However, these measures are only useful if what’s validly and reliably measured is also well-aligned
with the intended learning outcomes described within a particular university’s goals (Erwin, 2000). Further, commercial instruments of any sort come with their own disadvantages: they are often expensive, cumbersome to administer at the outcomes stage, complicated in terms of obtaining participation, and of questionable value for loop-closing if faculty themselves are not completely persuaded that the instrument is a meaningful measure of student learning, well-aligned with institutional goals.

**Methodological Criteria**

With a limited budget, a short time line, and plenty of rich data from indirect assessments already available to us, we turned our attention to the possibility of developing our own method of direct assessment. A first decision was whether quantitative or qualitative data would be most appropriate for our purposes.

The goal of qualitative research is to understand a phenomenon, to systematically describe it within context, and to make replicable interpretations based on this description (Kazdin, 2003). It is often used when a topic area has little previous research to aid in understanding the phenomenon, or when quantitative approaches (e.g., group comparisons) are not feasible (e.g., in the case of rare phenomena, lack of appropriate quantitative measures, resource constraints, etc.). Qualitative methodology is also appropriate when the purpose is the investigation of the human experience within a context (e.g., one’s internal thoughts and feelings, and how these reactions may have an impact on overt and observable phenomena). This information is not likely to be measured (or controlled for) in a quantitative paradigm. More information about a phenomenon within context is yielded from qualitative analysis.

A wide variety of methods can yield qualitative data, including recording narratives from individuals and observing the phenomenon of interest within context. One common methodology is for the investigator to participate in data collection in a systematic but often unstructured way (e.g., via interviews or focus groups) in order to bring forth more relevant information than could be captured through quantitative methods such as surveys. For example, an interviewer will often ask follow-up questions to help elicit additional information that may be helpful in understanding the complexities of the phenomenon under study (Kazdin, 2003; Schilling, 2006).

Quantitative research, on the other hand, seeks to answer very specific questions through the use of operational definitions, control of “extraneous variables,” systematic and replicable methods of data collection, and, often times, the systematic manipulation of the phenomenon of interest. Interpretations are made via statistical analyses. Quantitative and qualitative research can be distinguished in a variety of ways. Quantitative methods often use much larger samples (although newer single-case design methodology is an exception), comparison groups are used, and data are (usually) in numerical form and obtained from measures that operationalize the phenomenon of interest. While there are benefits to the use of quantitative methods, qualitative methods have their own advantages (e.g., the descriptive nature of the data). Furthermore, qualitative methods may be better suited in cases of limited resources and sample availability (Kazdin, 2003).

These methodological distinctions were important considerations; however, we also recognized that any suitable method needed to meet specific criteria regarding time, cost, and feasibility for assessment purposes, each of which merits additional explanation:

- Data needed to be collectable near the time of graduation;
- Students needed to willingly participate with minimal incentive, which meant that the process needed to be relatively brief and “painless” from a student perspective;
Method Development for Assessing a Diversity Goal

- Data from about 50 students would be desirable, which meant that it would be extremely helpful if data could be collected within groups to reduce the amount of time spent on data collection;
- Despite the desirability of a group method, data needed to be individually scorable.

First and most troublesome, we needed a means to collect data near the time of students’ graduation. We were already collecting data from direct assessments administered within general education courses themselves. But they told us little about the state of student learning by the time of graduation, a distinction that we had learned was important through our longitudinal study. Students had noted that much of their learning around general education goals occurred near the time of graduation, sometimes within the major and sometimes through activities outside of the formal curriculum. Those experiences occurred when students were more focused and more mature. They had a “real world” element that caused students to take them more seriously. So it was important to find out how students were doing on this goal as they neared the end of their college experience – that is, near graduation.

The need for data from near-graduates made collection more challenging. Senior level courses are commonly taken in the major. There were no senior level courses that were widely taken by students from a range of majors through which data could be collected or within which a test might be embedded. Students, therefore, would need to be willing to participate in a non-course-based assessment. Although it would be possible to provide some small stipend or token of appreciation for participation, there was nothing we could offer that would entice busy seniors to spend more than an hour or so of time on an outcomes assessment for the general education program. Furthermore, we needed to ensure that students put forth genuine effort and engagement even within that relatively short amount of time, which meant, we concluded, that the process needed to be intrinsically interesting to participants. This consideration led us to reject a paper-and-pencil assessment, which we believed would feel uncomfortably like a “test” to students.

Our decision about sample size was driven by previous experience assessing general education learning outcomes. During the past year, a team of faculty at our institution had collected papers from senior courses in various disciplines and, using rubrics, conducted a session in which the papers were scored on critical thinking and written communication. As that session was being planned, a consultant in research design pointed out that the scoring session would certainly generate direct assessment data in a numeric form. But the data would not be quantitative in the typical sense, as it would not meet the standards required for statistical generalizability. There was no reasonable means, for example, of collecting a strictly randomized pool of papers from all seniors across campus. Therefore, we had been advised to consider our outcomes assessment as essentially qualitative, generating a “snapshot” which would be very useful in terms of allowing us to see the kinds of performances our institution’s students were producing by their senior year. This snapshot approach would yield sufficient information to inform curricular decision-making, which is the primary purpose for assessing learning.

The scoring session would also allow us to see the aspects of critical thinking, for example, on which our students were often strong and aspects where student performance was frequently weak. So for assessment purposes, the scoring would be meaningful in helping us understand the learning of our graduates, despite the lack of statistical generalizability. The findings would be instructive in that they would allow us to see the range of achievement, typical areas of strength, and typical areas of weakness among almost-graduates. Given those aims, we had been advised to examine the work of about 50 senior students in our scoring session. That number, although
insufficient for statistical analysis, would likely be enough (assuming we obtained papers from a broad range of seniors) to provide a good look at student learning.

The 50-student mark had proven useful and quite adequate for the study of written communication and critical thinking. In view of that, we felt that it would be appropriate to seek a similarly sized pool for the cultures assessment. The need to achieve that level of participation became an additional criterion as we considered possible methods of data collection. Since we had rejected a paper-and-pencil test as implausible given the need for voluntary and serious student participation, it would probably not be possible to have a single “administration” of an assessment. However, it would be tremendously time-consuming to conduct and score 50 separate administrations. It therefore would be important to be able to collect data from students in groups, if possible, while retaining the ability to score the learning demonstrated by individual student participants.

Our four criteria were important constraints. But there were other considerations as well, and those additional considerations led us to favor a method which would yield qualitative data. We were interested in understanding the individualized application of cultural familiarity in the context of a group, and in a systematic way. This is best understood through a qualitative analysis of the phenomenon within context.

One of the goals of this assessment process was to provide extensive information to administrators and faculty for use within future curriculum development. Different faculty from different disciplines may define culture and diversity differently and, therefore, use the information in a variety of ways when considering general education requirements and course construction. Given this, the extensive detail provided through a qualitative method could prove especially important.

Finally, we needed the ability to elicit relevant information that might not be revealed in a pencil-and-paper activity conducted largely under the student’s own discretion. Oral methods allowed for a facilitator to be involved. Questionnaires, tests, essays, and other similar tasks (whether quantitative or qualitative) would not provide the additional depth that might be generated through such a methodology with the assistance of such an individual.

The Focus Group “Oral Examination”

Based on these considerations, we focused on the idea of a group “oral examination,” which could be designed to feel more like a focus group discussion than a test. The notion of groups developed as a means of limiting the amount of faculty time: putting students in groups of five, for example, would mean significantly less time from both facilitator and scorers (in that scoring of five students could occur at a time), making the process considerably more manageable.

In addition, team members recognized that cultural familiarity (like most diversity goals) is socially enacted. Putting students in a social context for data collection could actually improve the quality of the data over the more sterile approach of interviewing or “testing” students individually. Each student’s performance could be influenced by the presence of peers; however, the application of cultural familiarity skills and knowledge in any real world context has a similar potential for influence by others. For our purpose, which was to improve our understanding of students’ achievement of the general education goal rather than to grade student performance in a class, the “disadvantage” of contextual influences would actually be an advantage, aiding us in our understanding of student learning and behavior.

The group context offered the additional advantage of being motivational for participants. Participating in a group discussion would be interesting and potentially engaging, as long as questions were carefully constructed. Appropri-
Method Development for Assessing a Diversity Goal

ate questions would elicit enough information to allow scoring of individual students, but would not be experienced as hostile or judgmental. If the right balance could be struck, students would experience the process not as a “test” but a “discussion.”

The task was daunting but also intriguing, and we agreed that it could be done. We learned that one department on campus owned a state-of-the-art laboratory for research on interpersonal behavior. That room (with its built-in video recording equipment) could be borrowed, meaning that each group session could be conducted by a single facilitator, with scorers reviewing DVDs of the session at a later time.

With logistics of video recording resolved, we turned our attention to development of questions and a scoring rubric. We began by developing questions that would be asked during semi-structured group interviews. Questions (see Appendix A) were designed to provide students with opportunities to demonstrate the following:

1. Their knowledge of culture and diversity,
2. Their awareness of themselves as persons within cultures,
3. Their ability to make comparisons related to diversity, and
4. Their ability to apply knowledge and awareness of cultural diversity.

“Scenario” questions appeared at first to be a compelling approach for assessing application skills, since such questions would allow students to demonstrate their learning as they analyzed and deconstructed various situations. As we considered wording, however, it became clear that scenarios can easily be perceived as “loaded,” with some answers that are clearly right (or socially acceptable) and others that are equally clearly wrong. No student would willingly give a “wrong” or socially unacceptable answer, especially in front of peers, and we concluded that these questions would be unlikely to serve our purposes. We settled on a list of ten questions, with follow-up prompts, that would enable students to demonstrate their cultural familiarity conceptually and attitudinally, but also through application.

After finalizing the questions, we began to develop a rubric to score each student’s responses. A rubric was important for several reasons. First, it was important to be systematic in our assessment of student learning. Second, we needed a mechanism for categorizing and describing the information about students’ familiarity with culture in a way that would allow us to present and share the information with colleagues across the university. Finally, rubrics help scorers be consistent in their interpretation of the data, allowing replicability across groups and scorers.

As discussion about rubrics began, we quickly realized that it would have been wiser to let the questions flow from the rubric rather than tailoring the rubric to already-developed questions. However, the process of developing the method and the specific questions had led to considerable conversation about the meaning of the goal, and returning to that meaning led us to rough criteria that were later reworded and revised in the light of a theoretical model that seemed well-aligned with both the intent of the goal and the questions themselves (see the rubric in Appendix B). The model was developed by Campinha-Bacote (1999), and it was originally designed to address cultural competence in the context of health care. This model integrated five components: (a) cultural awareness, (b) cultural knowledge, (c) cultural skill, (d) cultural encounters, and (e) cultural desire. According to this model, cultural awareness encompassed self-awareness and self-examination, as well as in-depth examination of one’s own cultural background. Cultural knowledge was viewed as the process of creating a solid educational foundation about diverse ethnic and cultural groups and their history, arts, science, languages, and worldviews. The ability to identify culturally-based differences and to recognize cultural beliefs, values, and behaviors composed...
Method Development for Assessing a Diversity Goal

The model viewed cultural encounters as the process involving direct interactions with individuals from other cultures and diverse backgrounds, while cultural desire referred to the motivation to become a culturally competent and skillful individual (Campinha-Bacote, 1999, 2002). This model seemed appropriate for our purposes, and the final version of the rubric was rooted in Campinha-Bacote’s five categories.

Implementation

With the method determined, questions developed, and rubric prepared, we were ready to collect data. The registrar’s office was able to provide us with a randomly generated list of graduating seniors, and a team member recruited student participants from that list via personal e-mails to individual students. The VPAA/Provost had agreed to provide funding for the project, which made it possible to offer $25 stipends to student participants, each of whom would commit to a single one-hour “group interview.” Students were told that the focus of the interview was on general education outcomes, but they were not told in advance that all questions would be focused on the cultural familiarity goal; there was a fear that students would self-select based on their attitudes about the goal if they were provided with that information in advance, potentially undermining the value of the data. Student volunteers signed up for times that were compatible with their personal schedules, resulting in groups that ranged in size from 3 to 6 students.

One team member, experienced with both focus groups and individual interview methodologies, facilitated all ten groups, a strategy that minimized the potential for score variations related to facilitator style or personality. All scoring was deferred until the summer, at which time a group scoring session was held, using DVDs from a single group of five student participants. The purpose of the whole-group scoring session was to provide some standardization of scoring by enabling comparison of scores and discussion of factors that influence scoring. The wisdom of this approach was proven when initial scores ranged widely with some scorers favoring a “generous” approach that gave students the benefit of the doubt and others advocating for stricter evidence of learning. After much discussion of the appropriate anchors to use in rating students’ verbal responses, we settled on a protocol of assuming that the mid-range score (3 on a Likert scale of 1-5) represented the average level of achievement that we would expect to be demonstrated by a graduate of UND. If any scorer believed a student’s performance on a particular criterion was stronger or weaker than that default score, a case needed to be made. Based on the strength of the case (as determined in group discussion), a final score would be agreed on by consensus.

Having achieved strong agreement about scoring processes and meaning, group members were divided into teams of three. Each three-person team scored three additional groups of students (a total of about 15 students scored by each scoring team, in addition to the five students who were scored by the entire team). Team members gathered in the fall to review scoring data, discuss sample student comments that justified various scores, draw conclusions, and make plans for sharing the findings with faculty across campus. Mean scores for each criterion were calculated, along with high and low scores for each and counts of students scoring at various levels.

Given our choices about methodology, we avoided extrapolating from the scores of these 50 students to the likely performance of UND graduates as a whole. However, it was immediately apparent from a scan of scores that some students graduate from the university with scores that demonstrate a very low level of achievement on all five criteria, while others excelled across the board. The presence of very low end scores on several different criteria for even a handful of students was enough to demonstrate that some
students are graduating with unsatisfactory levels of achievement on this key goal, a finding that supported what had been heard from students in the longitudinal study and what we feared based on the transcript analysis. Furthermore, the individual student scores, reviewed in their totality, constituted a snapshot of comparative strengths and weaknesses on the various criteria. We noted, for example, that most students who participated in the study were weaker on conceptual knowledge around cultural diversity than on other elements of the goal.

Scorers were able to describe examples of student answers which demonstrated impressive levels of achievement and other examples of responses which demonstrated considerably less satisfactory achievement. During the first campus-wide session for discussion of findings, those examples came from the memories of scorers. Transcripts of the interview sessions (for the first year of data collection, as well as for a subsequent year) have since been made and direct quotes can now be cited as necessary.

Even without the statistical analysis normally considered so valuable, this study has proven to be immensely useful. As anticipated, the findings of this study became important to the work of those engaged in revision of the general education program as a whole. In the previous program of general education, students were required to complete a single “world cultures” class. Within the new program, students will be required to take two classes that will be redesigned to focus on a considerably reworked “social-cultural diversity” goal. Members of the general education task force noted findings demonstrating that conceptual understanding was often an area of weakness. As a result, they agreed that one diversity course could be experiential (e.g., study abroad) but one would need to be classroom-based in order to ensure needed emphasis on theory and concepts. Finally, we recognized that lack of clarity about the goal’s meaning was interfering with both teaching and learning around the goal. New language that clarifies the meaning of the new goal draws extensively from the language of the rubric that was developed for scoring. In addition, findings from the assessment process and the rubric itself continue to be shared with faculty and to influence the development of courses created to address the new goal.

CONCLUSION

By going through this process, we learned a number of lessons regarding ways to improve the value of our data. For example, we agreed to collect demographic information during the second iteration of the method. We obtained permission to collect transcripts, which would enable us to find out whether the number of designated “cultures” courses on a student’s transcript appeared to correspond with the student’s overall score. Collecting such data could provide one indicator for the effectiveness of courses designed to address this goal.

More importantly, we learned that our locally-developed method, which was thoughtful but also improvisational, was useful and is potentially applicable to assessing other goals as well. Most institutions have goals like “ethical decision-making” or “civic engagement and participation” (in addition to diversity goals) that are difficult to measure. And yet it is important to document learning around such goals. Assessing learning around these goals is a way to understand more about program outcomes, a means of taking seriously the outcomes we say we value, a mechanism for provoking campus discussion of and attention to goals which are important although not concrete—and a necessity for accreditation, given the mood of the age.

Many of these important goals are not addressed through commercial assessments. Even when commercial assessments exist, there may be reasons (e.g., expense, difficult in obtaining student participation, difficulty in motivating
students to put forth their best work, misalignment with the institution’s own goal definitions, etc.) why these methods are not appropriate. Yet it is important to assess in ways that go beyond surveys and student perception data. In universities that make extensive use of capstone courses or student portfolios, a variety of kinds of data may be readily collectable. But in other universities, home-grown assessments may be the best or only practical means of finding out how fully students are achieving intended learning.

The group interview/oral examination methodology is a highly promising method for examining outcomes around many general education goals. One set of interview questions could provoke responses that would allow students to demonstrate their creative thinking, and a different set could allow demonstration of their ability to reach ethical decisions and describe the bases of those decisions. Scoring could be done with faculty viewing through a one-way glass (perhaps using a facility developed for faculty who do research in marketing or psychology), with video or audio-recordings, or with faculty scorers present in the room during the interview/examination process itself.

Local faculty are by necessity involved as scorers with this process, which is a benefit in terms of stimulating campus interest and discussion. Using local scorers demonstrates faculty ownership of the goals and the findings. Developing rubrics that can be used for such scorings can be an extremely useful faculty development exercise, while also serving as a means of establishing campus-wide definitions for campus-wide goals.

Although there are other means of achieving these same purposes, it is probably easier to achieve them when using locally-developed tools, measures, and rubrics than when collecting data via nationally normed tools. If we believe—as assessment textbooks suggest—that the primary purpose of assessment is for the collection of information that can inform local decision-making, then this is a powerful argument in favor of adding home-grown tools to the mix of methodologies in use on any campus.

**FUTURE TRENDS**

In response to the work of the Secretary of Education’s Commission on the Future of Higher Education (known informally as the “Spellings’ Commission”), there has been a strong move toward purchased assessments that can provide comparability across institutions (U.S. Department of Education, 2006). Although recent actions by key higher education organizations suggest that this emphasis is not about to go away (NASULGC, 2008), the success of our assessment reinforces our perception that a rush to standardization may not serve higher education well.

Standardized tests offer two major benefits. The first is that they provide inter-institutional comparability, and this is a benefit that is not to be under-valued. Such cross-institution measures are not likely to have much value for the public, despite Secretary Spellings’ comments about the need for more information when helping her daughter select an institution that could cost upwards of $40,000 per year for attendance. But they can be tremendously valuable, if used properly, for the institution itself. If an institution really values student writing as a general education outcome, for example, it could be useful for faculty to see how the writing of their students compares with the writing of students at similar institutions.

The second benefit of standardized tests is due to the simple fact that the assessments are standardized. With standardization, we can be confident of psychometric reliability and validity. We may not always be sure what the findings mean for our own curriculum or in relation to our own goals. But we do know that they represent something that carries meaning beyond the judgment of a single individual or a small group of individuals.
On the other hand, standardized measures of various sorts also come with built-in disadvantages. One of the most serious is that standardized assessments are by definition not customized to the specific values or goals of individual institutions. However reliable or valid the tests, they are useless if findings and conclusions are not directly relevant to an institution’s own purposes and curriculum.

A related problem is that, even apart from concerns about alignment, faculty’s previous experiences with standardized assessments have often failed to engender trust and respect. There is sound reason for this: for example, it was typical until quite recently that standardized tests measured students’ “written communication” skills via multiple choice tests of usage and grammar. There was often a concern about whether such tests measured what was most valuable (which is not to say that usage and grammar have no value at all).

Even today, with tests like the Collegiate Learning Assessment available to offer a considerably more nuanced measure of written communication (Klein, Benjamin, Shavelson, & Bolus, 2007), faculty must place their faith in anonymous scorers far removed from an institution’s own culture and priorities. Are they valuing formulaic answers? Are they primarily rewarding correctness? Are they measuring critical thinking? Or are they really measuring “written communication” in the ways that faculty themselves might define it? Until faculty are confident in the answers to such questions, it is unlikely that they will make curricular decisions based on data from such tests. And if the data aren’t trusted by faculty, then the primary value of assessment – to collect information for use in programmatic decision-making – is lost.

The goals we pursue in general education programs are lofty and difficult to measure. General education programs often have no convenient access to students near the point of graduation, which increases the difficulty of documenting student learning. And yet we know that desired general education outcomes are complex and developmental. Documenting learning within general education courses themselves provides only a partial picture of desired outcomes (amply demonstrated when students routinely pass courses, even with satisfyingly high grades, but are unable to demonstrate adequate understanding of key concepts in capstone courses or experiences which might take place a year or more later). When the learning in question is around skills (as is often the case with general education) rather than concepts, the gap can be especially worrisome.

In such a circumstance, it makes great sense for institutions to be creative in their assessments. Several different kinds of measures should be used whenever possible, and both direct and indirect assessments should be considered. Some measures may provide information about learning within general education courses, but some should provide understanding of learning apart from those courses if the institutional intent is for students to achieve these outcomes by the time of graduation. Some measures may be standardized, to allow cross-institutional comparability, but some should be specifically designed to maximize alignment with an institution’s own goals and definitions – and to maximize faculty ownership of and involvement with assessment. Locally-developed measures are not the only approach to assessment, but they will be an important component of effective assessment at most institutions. The best assessment plans, especially for a complex program like general education, will rely on data from an array of methods, carefully selected for maximum decision-making value.

**ACKNOWLEDGMENT**

This chapter is based on an assessment project developed by a team including (in alphabetical order): Abdallah Badahdah, April Bradley, Tatyana Dumova, Joan Hawthorne, Anne Kelsch, Evelyn
REFERENCES


*Intercultural Development Inventory™ (IDI)*. Developed by M. R. Hammer & M. J. Bennett. Available from the Intercultural Communication Institute (8835 S.W. Canyon Lane, Suite 238, Portland, Oregon 97225, U.S.A. Phone: 503-297-4622, e-mail: idi@intercultural.org, Web: http://www.intercultural.org)
Method Development for Assessing a Diversity Goal


The University of California Commission on General Education. (2007, April 1). *General education in the 21st century: A report of the University of
Method Development for Assessing a Diversity Goal


KEY TERMS

Cultural Diversity: Differences between individuals and groups such as language, religious beliefs, traditions, and social behaviors, among others.

Ethnic Identity: The degree to which an individual self-identifies membership with a given ethnic/cultural group. Membership is often defined as one’s comfort within a specific group, adherence to group norms, and participation in rituals/traditions of the group.

Indirect Assessment: Any assessment process which provides perception data about the outcome in question.

Direct Assessment: Any assessment process which is based on examination of an actual performance (written or oral) which demonstrates the outcome in question.

Qualitative Research: An investigational paradigm in which information is obtained through narrative accounts and/or observations. Information obtained is summarized and/or interpreted.

Quantitative Research: An investigational paradigm involving efforts to operationally define relevant variables, to control the influence of extraneous variables, and to quantify results through the use of statistical analysis.

External Validity: The extent to which the results obtained can be generalized to other individuals and/or contexts not studied.

Semi-Structured Interview: A process of conducting oral interviews of subjects by using a brief prepared list of questions, along with follow-up probe questions, to maximize the interviewer’s opportunity to obtain detailed information in response to each question.

Rubric: A tool for systematically and consistently scoring written or oral responses according to pre-determined criteria.
**APPENDIX A: TRANSCRIPT ANALYSIS RESULTS**

<table>
<thead>
<tr>
<th>GE Goal</th>
<th>Mean hits per transcript</th>
<th>Range of hits per transcript</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communicate effectively</td>
<td>12.28</td>
<td>5-27</td>
</tr>
<tr>
<td>Think critically/creatively</td>
<td>15.82</td>
<td>6-28</td>
</tr>
<tr>
<td>Make informed choices</td>
<td>14.42</td>
<td>6-26</td>
</tr>
<tr>
<td>Understand conclusions in natural sciences, social sciences, arts and humanities</td>
<td>15.34</td>
<td>5-28</td>
</tr>
<tr>
<td>Acquire knowledge</td>
<td>12.98</td>
<td>5-26</td>
</tr>
<tr>
<td>Familiar with cultures</td>
<td>6.42</td>
<td>1-20</td>
</tr>
</tbody>
</table>

NOTE: This chart is based on an analysis of 50 student transcripts in which each course “counting” towards the general education (GE) program requirements, was marked. The mean number of hits indicates the number of courses on an average transcript that addressed a particular goal. The range of hits indicates the range from least to highest number of hits on each particular goal.

**APPENDIX B: CULTURES INTERVIEW SCRIPT**

1. What does the phrase, “cultures other than your own,” mean to you? What kinds of cultures does UND seem to have in mind when that gets listed as a gen ed goal?

2. People sometimes say that understanding other cultures helps us better understand our own. Can you think of an example of a time when you experienced that?

3. Do you think that there are cultural universals, things that we all share? If so, what might some of those things be?

4. In your own professional future, or that of another person in your same field, how important do you think this goal could be? Explain. How would you use it? Where would you see it?

5. Are there negative things that could occur as a result of a lack of familiarity with other cultures? Do you recall any times recently when you or someone you know has run into difficulty because they were NOT familiar with a culture other than their own?

6. We talk sometimes about pluralism or multi-culturalism, sometimes using examples of a melting pot or a salad bowl as analogies for what the United States is like. What’s good about being a diverse society? What problems might that create? How would you expect a country with a very homogeneous population to be different from one with a diverse population?

7. People talk a lot about globalization today. What do you think are the implications of a globalized world? What might be good about it? What concerns would you have?
8. [NOTE: Present students with handout of population data showing projected growth of various ethnic groups and minorities as a percentage of the U.S. population; data taken from U.S. Census Bureau, International Data Base, Table 094, <http://www.census.gov/ipc/www.idbprint.html>.] Think forward a few years. What are the implications of the information you see presented here? Why might it be important? What changes are likely to follow?

9. Is there a course here at UND that’s been particularly helpful in terms of developing a familiarity with cultures? Tell me about that. What made it so useful? Are there other experiences that have helped you?

10. Has anything you’ve learned in your classes or through your UND experiences changed how you think about your own culture?

APPENDIX C: SCORING RUBRIC

GENERAL EDUCATION “CULTURES” GOAL

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Lowest</th>
<th>Highest</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Culture as a concept:</strong> Student demonstrates an understanding of the concept of culture. (Questions 1, 3)</td>
<td>1 2 3 4 5</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Cultural self-awareness:</strong> Student recognizes that it’s not just other people who have cultures and that there are things about one’s own culture to understand. (Questions 2, 10)</td>
<td>1 2 3 4 5</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Cultural desire:</strong> Student demonstrates open-mindedness and receptiveness to cultures. (Questions 6, 7)</td>
<td>1 2 3 4 5</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Application of cultural knowledge:</strong> Student is able to apply cultural knowledge and UND/classroom experiences to societal and global issues. (Questions 4, 6, 7, 8, 10)</td>
<td>1 2 3 4 5</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Cultural encounter:</strong> Student can identify experiences that have involved interaction with other cultures. (Questions 5, 9)</td>
<td>1 2 3 4 5</td>
<td>N/A</td>
</tr>
</tbody>
</table>