General Education Assessment Review

Foreign Language

Hirosuke Honda
Institutional Research Analyst

Center for Planning and Institutional Effectiveness
9/15/2012
Introduction

In 1998, the State University of New York’s Board of Trustees passed a resolution that was the basis for forming the Provost’s Advisory Task Force on the assessment of student learning outcomes. This group established the framework for the SUNY Assessment Initiative, which mandated assessment of all academic majors and general education programs across the State University system, with the focus on student learning outcomes. As a result, task forces were convened at colleges across the state with the responsibility of developing an assessment plan detailing procedures and guidelines as well as a timeline for implementation. A group consisting of SUNY faculty, students, and representatives from System Administration began reviewing campus assessment plans during the 2001-02 academic year, with the first year of campus-based general education assessment beginning the following year.

Empire State College’s original GEAR assessment plan outlined a schedule in which one assessment was conducted in 2003-04, three in 2004-05, and four assessments in 2005-06. That plan was later revised and changed to reflect current practices, which requires the college to conduct four General Education Assessment Reviews per year, on a three-year cycle.

Rating Instrument

Prior to the inaugural assessment in each general education area, Empire State College convened task forces consisting of faculty teaching courses in these respective disciplines. These faculty members were asked to design rubrics based on SUNY learning outcomes that could be used to rate samples of student work against specified outcomes. These rubrics consist of explicitly stated criteria describing student work that “Does Not Meet,” “Approaches,” “Meets,” or “Exceeds” student learning objectives. In instances where raters perceive that a learning objective is not relevant to a given sample of student work, a fifth category of not applicable or “NA,” is utilized. For reporting purposes, each score is coded numerically from 1 to 4, where 1=“Does Not Meet,” 2=“Approaches,” 3=“Meets,” and 4=“Exceeds.” These rubrics were approved by the Undergraduate Student Policies Committee as well as administrators within the Office of Academic Affairs prior to their use.

The Foreign Language rubric consists of two student learning outcomes:

1. L2 (second language) comprehension and written ability
2. Cultural competence in the L2

Student Work Sample

The 2012 GEAR in Foreign Language was conducted in May 2012. The prior assessment was conducted in May 2009. A total of 100 work samples were collected from CDL archives for both assessments. Each student work sample was rated by two faculty raters, producing a total of 200 ratings for both assessments.

Procedure

Prior to the formal assessment activity, a team of four faculty members were sent a copy of the rubric and one sample of student work. They were asked to review and rate the sample against the assessment rubric and then record their results. Once the faculty rating group in the subject area was convened via Elluminate, faculty members were asked to share their ratings on each of the rubric’s learning objectives during a facilitated norming session. If there were discrepancies in ratings, faculty members discussed these differences with the goal of establishing consensus or at least narrowing differences among raters. During the three-week assessment session, teams of faculty raters reviewed and rated student work
samples against the assessment rubric in the same way they did during the norming session. Samples were rated twice so that inter-rater reliability could be assessed.

At the conclusion of the three-week assessment period, C-PIE staff reviewed faculty ratings and identified major discrepancies between raters that needed to be resolved. This information was provided to the team of raters who were responsible for convening a consensus discussion session with the other member of the team with whom they had a rating discrepancy on a learning objective for a single piece of student work.

Reliability
The two types of reliability (consistency) calculated for GEAR are internal consistency and inter-rater reliability.

Internal consistency refers to how the individual learning objectives which make up the rubric hold together statistically. A rubric has internal consistency to the extent that the individual items or objectives on the rubric are rated similarly for a single piece of student work. The reliability statistic used to determine internal consistency is Cronbach’s alpha. An alpha level of 0.70 or higher is considered acceptable for a new scale. The Cronbach’s alpha level was above the threshold of acceptability in 2009 (0.77) and approaching this threshold in 2012 (0.67), indicating that the Empire State College Foreign Language GEAR Rubric possesses acceptable internal consistency.

Inter-rater reliability refers to agreement on ratings using the rubric by two faculty raters. For the rubric to be reliable, multiple judges should generally agree in their evaluation of a writing sample. Inter-rater reliability is estimated using intra-class correlations, which have properties similar to Cronbach’s alpha. Again, an alpha level of 0.70 or higher is considered to be acceptable. The reliability estimates appearing in Table 1 depict intra-class correlations after the consensus discussion. An examination of the post-consensus discussion intra-class correlations reveals that coefficients fall within acceptable ranges for each learning objective on Foreign Language.

Table 1: Intra-class Correlations from 2009 and 2012 GEAR in Foreign Language

<table>
<thead>
<tr>
<th>Learning Outcomes</th>
<th>2009 (Spanish Only)</th>
<th>2012 (Spanish Only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 (second language) comprehension and written ability</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>Cultural competence in the L2</td>
<td>0.91</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Results
In the 2012 assessment, a large majority of ratings were “Meets” and “Exceeds” on “Second language comprehension and written ability” (70.4%) and “Cultural competence in the L2” (96.8%). The mean scores of the two learning objectives indicated that student work samples fell somewhere between the “Meets” and “Exceeds” categories of the rubric regarding faculty expectations for a rising junior.

Complete results for 2009 and 2012 assessments appear below in Table 2. Both assessments were based solely on student work samples from Spanish language and culture courses. The comparison of mean scores between 2009 and 2012 indicated that the mean score for the “Second language comprehension and written ability” objective remained the same, while the mean score on the “Cultural competence” objective increased slightly.
However, it should be noted that the differences in the mean scores can be affected by various factors including the student work samples and the members of the rating groups across the two GEAR assessments.

### Table 2: GEAR Foreign Language Comparison Between 2009 and 2012 Assessments

<table>
<thead>
<tr>
<th>Learning Outcome</th>
<th>Valid N</th>
<th>Does not meet</th>
<th>Approaches</th>
<th>Meets</th>
<th>Exceeds</th>
<th>NA</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 (second language) comprehension and written ability 2009*</td>
<td>200</td>
<td>5 2.5</td>
<td>35 17.5</td>
<td>128</td>
<td>64.0</td>
<td>32</td>
<td>16.0</td>
<td>0</td>
</tr>
<tr>
<td>L2 (second language) comprehension and written ability 2012 *</td>
<td>147</td>
<td>5 3.4</td>
<td>37 25.2</td>
<td>70</td>
<td>47.6</td>
<td>35</td>
<td>23.8</td>
<td>53</td>
</tr>
<tr>
<td>Cultural competence in the L2 2009**</td>
<td>42</td>
<td>0 0.0</td>
<td>10 23.8</td>
<td>22</td>
<td>52.4</td>
<td>10</td>
<td>23.8</td>
<td>158</td>
</tr>
<tr>
<td>Cultural competence in the L2 2012**</td>
<td>96</td>
<td>0 0.0</td>
<td>4 4.2</td>
<td>55</td>
<td>57.3</td>
<td>37</td>
<td>38.5</td>
<td>104</td>
</tr>
</tbody>
</table>

Note: Percentages are based on the number of valid cases. Mean ratings based on the following scale: 1="does not meet," 2="approaches," 3="meets," 4="exceeds." Ratings of “NA” were excluded from mean calculations.

\* \( \chi^2 (3, \text{N}=347) = 9.30, p < .05 \)

\* \( \chi^2 (3, \text{N}=138) = 13.10, p < .01 \)

### Summary and Discussion

The rubric used in both the 2009 and 2012 General Education Assessment Reviews in Foreign Language has two distinct student learning outcomes: 1) L2 (second language) comprehension and written ability; and 2) Cultural competence in the L2.

These outcomes were written by faculty prior to the 2009 assessment and were determined to be in alignment with SUNY learning outcomes in the area of Foreign Language.

A test to determine the rubric’s internal consistency after the consensus sessions produced a Cronbach’s alpha level of 0.77 in 2009, and 0.67 in 2012. These calculations were either above or approaching 0.70, which is generally recognized as the threshold for acceptability.

The inter-rater reliability for both of the rubric’s objectives for the 2012 GEAR in Foreign Language was satisfactory. These results indicated that faculty raters were interpreting the rubric and scoring work samples in a similar manner on both of the rubric’s learning objectives.

In 2012, the mean scores on the rubric’s two outcomes were 2.92 and 3.34, indicating that student work samples fell somewhere between the “Approaches” and “Exceeds” categories against raters’ expectations of a rising junior. The results from 2012 indicated an improvement from 2009 on the second learning objective, Cultural competence. Although this result was statistically significant, various other factors including student work samples and the composition of the rating groups across the two assessments could account for these differences.

Despite efforts to collect student writing samples that were of manageable length to review (seven to ten pages), there were many studies that required papers less than or in excess of the desired page...
length. Some studies required multiple mini-papers and discussion board postings as well. The lack of uniformity in assignments rated within and across assessments should be recognized.

During the assessment sessions, raters strongly expressed that student work samples from regional centers need to be included in future Foreign Language GEARs. Furthermore, at the closing the loop session, faculty discussed a need to ensure the qualifications of foreign language instructors across the college.

One approach that could strengthen the assessment process at Empire State College would be to implement an electronic portfolio system that could serve as a repository for student work. Students could designate work samples from each study to be displayed in the portfolio. This would allow work samples to be collected from undergraduate students enrolled at regional centers across the state, rather than primarily from students enrolled at the Center for Distance Learning. This sample would be much more reflective of the undergraduate population and would better serve the college in terms of generalizability of assessment findings. An electronic portfolio also would allow staff from C-PIE to select student work samples that are more relevant to each learning outcome or at least most of the rubric’s learning outcomes.
Appendix A: Empire State College Foreign Language Rubric

Please fill in the circle that best reflects your assessment of each of the following expectations for student’s work. Refer to the rubric for the appropriate evaluative criteria for each student learning outcome.

<table>
<thead>
<tr>
<th>Expectations</th>
<th>Does Not Meet</th>
<th>Approaches</th>
<th>Meets</th>
<th>Exceeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 (second language) comprehension and written ability</td>
<td>Recognizes just a few words in the L2. Does not answer questions correctly or completely, indicating a misunderstanding or incomprehension of the L2. Makes several written errors including grammatical, spelling, and word choice. Sentences are often incoherent.</td>
<td>Recognizes many words in the L2 but may answer questions incompletely or not fully correctly. Writes basic sentences grammatically with some errors in orthography and spelling.</td>
<td>Understands most words in their context and responds appropriately to questions and other assignments. Writes grammatical sentences with minimal errors in spelling and orthography.</td>
<td>Understands the L2 in any given context and responds well accordingly. Writes well developed, grammatically correct sentences and/or paragraphs. Word choice and sentence structure is excellent as is spelling and orthography.</td>
</tr>
<tr>
<td>Cultural competence in the L2</td>
<td>Knowledge about cultural issues for the L2 is vague or stereotypical.</td>
<td>Knowledge of cultural issues for the L2 exists on a surface level.</td>
<td>Showcases a working knowledge of multiple cultural aspects that take place in lands of the L2, including but not limited to the history, art, peoples, values and current events of L2 countries or communities.</td>
<td>Significantly knowledgeable about issues that evoke cultural significance in the L2, including but not limited to the history, art, peoples, values and current events of L2 countries or communities.</td>
</tr>
</tbody>
</table>

Comments: O O O O O