# Assessing Essential Learning In Ethnic Studies Courses

James A. Wollack, PhD
Research Director
Director, Testing & Evaluation Services
Associate Professor, Educational Psychology

Sissel Schroeder, PhD
Professor, Anthropology

Elaine Klein
Chair, University General Education Committee
Assistant Dean, College of Letters & Sciences

Katie Lindstrom, PhD
General Education Research Assistant

Sarbani Chakraborty
General Education Research Assistant

April 7, 2014

# Assessing Essential Learning In Ethnic Studies Courses

## Introduction

The UW-Madison College of Letters & Sciences implemented a requirement in 1989 that all students enrolled in the College would complete a minimum of three credits devoted to either

(1) *The study of the experience of discrimination by some ethnic, racial, or religious group so affected in American society; or*

(2) *The thorough examination of aspects of the culture and historical experience of an ethnic, racial, or religious group that remains on the margin in the United States; or*

(3) *The study of discrimination, cultural differences, and ethnicity in other settings in ways which help in the understanding of cultural and ethnic problems in the United States.*

This requirement, referred to as the Ethnic Studies requirement (ESR), was adopted by the entire campus for all incoming freshmen and transfer students in 1994.

In the years that have followed, the ESR has been reviewed by the Board of Regents and by various campus-level committees on several occasions, for purposes of evaluating the requirement and improving its implementation. In 2003, the ESR was narrowed to courses that focused specifically on "ethnic/racial minorities that have been marginalized or discriminated against in the U.S." A couple years later, a number of changes aimed at increasing the status of ESR on campus and improving the meaningfulness of the requirement were recommended. One of the important changes that followed these recommendations was that the oversight of the

administration of ESR was formally assigned to the University General Education Committee, with an Ethnic Studies Subcommittee being assigned the task of determining which courses students could take to satisfy the requirement.

Since the ESR was first implemented in 1994, no systematic, objective attempt has been made to evaluate the extent to which ESR courses are achieving their stated purposes. As a first step in this process, the Ethnic Studies Subcommittee convened a meeting of ESR instructors in March, 2010. The primary purposes of this meeting were to identify the common goals present in existing ESR courses and to define and adopt a set of measurable learning outcomes appropriate for all courses that satisfy the ESR, but distinct from the course-specific learning outcomes that exist for each individual class.

More than 50 individuals attended and participated in the meeting, including 12 group facilitators and 36 faculty and staff involved in courses and departments that are connected with the ESR. The meeting resulted in a set of four learning outcomes that, when taken collectively, would produce students with cultural competence who are able to effectively participate in a multicultural society. These four learning objectives are as follows:

1. Awareness of history's impact on the present
2. Ability to recognize and question assumptions
3. A consciousness of self, other, and difference
4. Effective participation in a multicultural society

The significance of these essential learning outcomes (ELOs) cannot be understated. Although ESR courses are tremendously diverse with respect to their specific foci and educational objectives, as might be expected of any large collection of classes spanning dozens of academic departments, these learning outcomes describe the shared educational goals that unite these courses into a single, cohesive program. Not only do the ELOs provide a statement about the expected and defining characteristics of ESR courses, but they also create a foundation for future evaluation of student learning across the entire Ethnic Studies curriculum.

In 2011, the General Education Committee, in conjunction with the ESR subcommittee and the ESR faculty, decided to conduct an empirical evaluation of courses satisfying the ESR. The specific purpose of the study was to evaluate the extent to which students completing ESR courses satisfy the ESR ELOs. It was made clear from the outset that this study was to focus on programmatic learning gains; it was not to focus on the extent to which students mastered the ELOs in specific courses, nor was it to evaluate different ESR instructors.

## Study Design

The ESR Subcommittee felt it was important that the evaluation of the ESR be based on actual samples of student work (what we will call artifacts) and be evaluated solely with respect to the extent to which that work provided evidence that the ELOs were adequately mastered. Therefore, it was determined that course grades or grades on the individual artifacts were inappropriate for purposes of this study, since those likely reflect many course-specific ELOs and other criteria that are unrelated to the ESR ELOs. Therefore, it was decided that the ESR faculty should develop a scoring rubric providing the rules for assigning scores, based on the evidence that each ELO was mastered. It was further decided that the student artifacts, which

were assigned for purposes of measuring course-specific ELOs rather than the ESR ELOs, should be evaluated with respect to the ELO rubric by independent raters who are unaffiliated with the particular classes in which the work samples were assigned, but are intimately familiar with the principles underlying ESR, and with the ESR ELOs.

*Development of ESR Rubrics*

ESR faculty met several times during the 2012 spring semester to discuss the ELOs and create the scoring rubrics. During the first meeting, two important decisions were made. First, it was decided that the fourth ELO—Effective participation in a multicultural society—did not lend itself to being evaluated with samples of student work from current ESR courses. This particular objective is a distal outcome that can only be evaluated some years down the road, quite likely after the participants are no longer students at UW-Madison. For this reason, it was decided that the fourth ELO would not be evaluated as part of this study. The second decision was that the remaining three ELOs were all sufficiently distinct that it was unlikely that any one student work sample would be able to adequately provide students with an opportunity to demonstrate mastery of all three. Consequently, the ESR faculty felt it best to develop a separate rubric for each of the three studied ELOs, and to allow different artifacts to be selected depending on which ELO was being assessed.

The ELO rubrics were modeled after and adapted from the American Association of Colleges and Universities' VALUE rubrics in "Intercultural Knowledge and Competence," "Critical Thinking," and "Civic Engagement." A single rubric framework was used for all ELOs. Mastery of ELOs was evaluated on a 4-point scale. For all three ELOs, scores of 4 were awarded for artifacts that exhibited "**sophisticated and substantial** cognitive, affective, and

behavioral skills and characteristics that support effective and appropriate understanding and interaction in a variety of cultural contexts." Scores of 3 were awarded for artifacts exhibiting "**developing and consistent** cognitive, affective, and behavioral skills and characteristics that support effective and appropriate understanding and interaction in a variety of cultural contexts." Ratings of 2 were reserved for artifacts with "**emerging and inconsistent** cognitive, affective, and behavioral skills and characteristics that support effective and appropriate understanding and interaction in a variety of cultural contexts." Finally, a rating of 1 indicated that the artifact demonstrated "**minimal and surface-level** cognitive, affective, and behavioral skills and characteristics that support effective and appropriate understanding and interaction in a variety of cultural contexts." Each ELO was evaluated separately and included specific, unique verbiage to explicate the meaning of each scale point. The agreed-upon Ethnic Studies Rubric is provided as Appendix A.

*Participants*

*Classrooms.* All faculty teaching a course that satisfied the ESR during the Spring, 2012 semester were send a letter by the research team in which the study was described and they were asked if they would be willing to cooperate. Cooperation on the part of the instructors involved three components. First, the research team asked the instructor to set aside 15 minutes at their convenience early in the semester during which one of the Principal Investigators would come to discuss the study with their class and distribute consent forms. During this time, it was explained to students that to be eligible to participate, they were required to be at least 18 years old and to be enrolled in the current course on a graded basis. Also, students were informed that this study was limited to those who had not previously satisfied the ESR through another course.

In addition, instructors were asked to identify the single artifact in their class that best allowed students an opportunity to demonstrate their mastery of each ELO. Instructors were told that a single artifact could be identified for two or even all three ELOs, but it was also possible that a different artifact could be identified for each learning outcome. The research team would then randomly select the ELO to be assessed for each class selected to participate. This approach was implemented to avoid having instructors choose which ELO they wanted assessed for their classes. Although one ELO might be emphasized more in a class or might be easier to assess than the others, ethnic studies courses are supposed to measure *all* the ELOs. If it can be demonstrated that a randomly selected ELO is well satisfied, we may generalize that *any* randomly selected ELO would have been well satisfied.

Finally, instructors were asked to share electronic copy of the artifacts for all consenting students at the end of the semester. In the interest of limiting the risk to the students, instructors were not notified which students had provided consent until after the deadline to submit final grades.

A total of 21 instructors provided the research team with access to their classes, and approximately two-thirds of students in those classes signed consent forms (though some of those students failed to meet the eligibility requirements). Due to limits to both budget and time, not all classes were able to participate. Instead, 15 classes were randomly selected from among the classes with at least 12 eligible students. These classes were further randomly assigned to one of the three ELO groups, subject to the constraint that each ELO group would include exactly five classes. The ELO group to which a class was assigned determined which artifact would be solicited from each instructor. At the end of the semester, instructors of these courses were notified to which ELO group their class had been assigned, and were asked to provide to

the research team the single best measure of the targeted ELO for each consenting student. From each class, artifacts for 10 consenting students were randomly selected for this study.

*Rubric Training Materials*. Even with a very carefully crafted scoring rubric, it is common (and generally regarded as best practice) to accompany that rubric with a set of artifacts that have been evaluated with the rubric by the rubric developers (or other primary stakeholders), for purposes of illustrating how the rubric should be applied. Over the month leading up to the rating event, a subcommittee of the ESR Committee met to read and score artifacts, for purposes of refining the instructions to raters and creating a set of training artifacts for each ELO. Because we received consent from many more students than were able to participate in the study, artifacts from consenting students not selected to participate were used for training purposes. Therefore, none of the training artifacts were included in the actual rating project.

For each ELO, the group was asked to find at least one example of an artifact that was half-way between adjacent score points (e.g., a good measure of a 2.5). By measuring the mid-points between allowable scores, the research team hoped to create a benchmark that raters could easily compare to live artifacts. Artifacts that were judged to be better than the benchmark would receive the higher score and those that were judged to be of lesser quality would receive the lower score.

*Rating Event.* The rating of all student artifacts was done over a period of two full days, on August 15-16, 2012. Raters were all graduate students who had previously taught an ESR course. A total of 13 raters were selected to participate. In some cases, the raters were teaching assistants during the Fall, 2012 for one of the courses participating in the study, but that was neither a requirement nor an exclusionary condition.

Raters were randomly assigned to their target ELO such that four raters were assigned to review the artifacts that were selected to measure ELO-1, four were assigned to review the artifacts selected to measure ELO-2, and five were assigned to review the artifacts selected to measure ELO-3. All raters in each group were asked to read and rate the same 50 artifacts, though the order in which each received them was randomized to control for order effects.

Prior to any live rating, all raters were required to undergo training. The entire morning of the first day was dedicated to learning the rubric and learning to apply it in a way consistent with the expectations of the ESR faculty. Raters received a packet of training artifacts, as well as other carefully selected artifacts, to read, rate, and discuss with others in their group. In addition, each group was assigned a table leader, a representative from the ESR Subcommittee who had participated in the crafting of the rubric and the selection of the training artifacts. The purpose of these discussions was to attempt to understand the rationale behind the scores assigned to the anchor artifacts by the ESR Subcommittee, and to come to a common understanding about how various aspects of the artifacts should be considered in assigning ratings. Some time was also spent discussing the types of artifacts they might encounter, though no information was available about the specifics of each particular assignment.

After the training period was over, the raters were asked to begin live rating. Raters were each given a stack of artifacts and asked to review and evaluate them in the order in which they were presented. Raters were instructed to focus primarily on issuing a good rating for the target ELO (i.e., the one their artifacts were selected to measure), but they were also asked to provide a rating for each of the other ELOs. Raters were asked to provide a score for all three ELOs for

each artifact; however the scoring sheet also allowed raters to indicate N/A if the artifact did not provide the opportunity to evaluate mastery of the ELO.

## Methods

*Descriptive Methods*. For each of the three sets of artifacts (i.e., those primarily intended to measure ELO-1, ELO-2, or ELO-3), we computed mean ELO ratings for the target and ancillary ELOs across all raters. In addition, we computed adjusted means that treated as missing the ratings for any categories scored as N/A by at least two raters. Because it is unclear whether ratings are meaningful when raters indicated N/A, for all remaining analysis, the adjusted means were used as the dependent variables of interest. Prior to the study, the ESR faculty agreed that it was their hope that students would, on average, achieve a level of 3 or higher on the ESR rubric for each of the three ELOs.

In addition, we examined the distribution of ratings across the four score points, both for the target ELO and the two ancillary ELOs. Examining the distributions allow for a more criterion-referenced interpretation of ratings, and allows us to see what percentage of students were satisfying the ELOs to different degrees.

*Inferential Methods*. Within each ELO, paired *t*-tests were conducted to explore whether mean ratings were higher for the targeted ELO than for the ancillary ELOs. Recall that the particular artifacts in each group were selected because the instructor felt they allowed students to demonstrate their mastery of the target ELO well. Therefore, it is reasonable to expect that mean ratings would be higher for the target ELO than for the others. Hence, the two comparisons between the target ELO and an ancillary ELO were directional, one-tailed tests.

However, because there was no *a priori* reason to expect differences between the two ancillary ELOs, those differences were tested using a two-tailed, non-directional test. A separate three pairwise comparisons were conducted for each group of artifacts. For all comparisons, the α-level was set at .05.

To help understand the sources of variability in mean rating score within each group of artifacts, a two-way split plot analysis of variance (ANOVA) was conducted. For each analysis, we examined the impact of ELO and Class on the ratings. ELO rating was treated as a within-subjects variable, because each artifact received three different scores—one for each ELO. Class, on the other hand, was a between-subjects variable, because the 50 artifacts within each ELO group were drawn from five different classes. In addition to providing a statistical test of whether mean ratings varied across either main effect or as a result of the interaction of the two, an effect size measure, $\eta^2$, was estimated for each effect[1]. According to Cohen (1988), $\eta^2$ values between .01 and .05 are considered small, between .06 and .13 are considered medium, and those at or above .14 are considered large.

*Rater Consistency*. Pearson correlations between ratings for all pairs of raters evaluating the same sets of artifacts were computed to evaluate the extent to which raters interpreted and applied the rubrics similarly. For each group, median, minimum, and maximum correlations were computed to facilitate making comparisons across groups (i.e., to see whether rubrics for some ELOs were easier to apply consistently than others).

[1] $\eta^2$ is computed as the ratio between the sums of squares for the effect of interest and the total sums of squares.

Results

Table 1 shows the mean and adjusted mean ratings for each of the three ELOs. Results are shown separately for each target ELO. That is, the first two columns show the mean and adjusted mean scores for the set of artifacts that were intended to measure ELO-1. The middle two columns show data on those artifacts intended to measure ELO-2. And the last two columns provide data relating only to the artifacts for which ELO-3 was the target.

Table 1

Mean and Adjusted Mean ELO ratings

| | Target ELO | | | | | |
|---|---|---|---|---|---|---|
| | ELO-1 | | ELO-2 | | ELO-3 | |
| Assessed ELO | Average Rating | Excluding N/A | Average Rating | Excluding N/A | Average Rating | Excluding N/A |
| ELO-1 | 2.355 (N = 50) | 2.355 (N = 50) | 2.290 (N = 50) | 2.364 (N = 46) | 2.124 (N = 50) | 2.395 (N = 37) |
| ELO-2 | 2.080 (N = 50) | 2.082 (N = 49) | 2.395 (N = 50) | 2.413 (N = 49) | 2.188 (N = 50) | 2.188 (N = 50) |
| ELO-3 | 1.440 (N = 50) | 2.25 (N = 12) | 2.145 (N = 50) | 2.202 (N = 42) | 2.120 (N = 50) | 2.263 (N = 35) |

In looking at these data, several interesting patterns emerge. First, mean ratings are uniformly higher after excluding scores for artifacts that were determined to be inappropriate measures of that ELO. This result is not surprising, as it is reasonable to expect that ratings would generally be low for artifacts that were classified as N/A. By systematically omitting those spuriously low scores, the average rating (across artifacts that could reasonable be assessed with respect to the ELO) was increased. As a rule, the magnitude of the difference between mean and adjusted

mean scores was heavily influenced by the number of scores that were dropped. The more scores dropped, the greater the difference between the two measures.

The second observation is that mean ratings were fairly similar across ELOs. This was especially true for ELO-1 and ELO-3. In the case of ELO-1, adjusted mean ratings were between 2.355 and 2.395 (a difference of just .04), regardless of the target ELO. Similarly, for ELO-3, the adjusted mean ratings varied from 2.202 to 2.263, a difference of just .061. In the case of ELO-2, there was little difference between ratings when either ELO-1 or ELO-3 was the target ELO; however, when ELO-2 was the target ELO, mean scores were noticeably higher.

A final observation is that raters felt that the artifacts were often inappropriate for assessing ELO-3. This can be seen by the markedly lower sample sizes associated with adjusted means for ELO-3. Curiously, this pattern was even evident when ELO-3 was the target. It is, of course, perfectly reasonable to think that a particular ELO might be sufficiently distinct from the others that assignments intended to evaluate ELO-1 or ELO-2 would fail to provide an opportunity to assess ELO-3. However, the data here suggest that artifacts that were specifically selected because they were the best available measures for ELO-3 were still poorly suited for that task.

Table 2 shows the distribution of ELO ratings for each of the three target ELOs. Note that the column percentages all add to 1.0 because all raters were required to issue a rating for each artifact on each of the three ELOs. It is clear from Table 2 that the percentage of artifacts earning a rating of 3 or higher was fairly low. In fact, for no artifact was the percentage of ratings of scores of 3 and 4 greater than 50%. Not surprisingly, it is also the case that very few artifacts received the highest ratings of 4. Overall, it appears as though fewer 4's were issued for ELO-1 than for the other ELOs.

Table 2

Distributions of Ratings Across Target ELOs

| | Target ELO | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ELO-1 | | | ELO-2 | | | ELO-3 | | |
| Rating | ELO-1 | ELO-2 | ELO-3 | ELO-1 | ELO-2 | ELO-3 | ELO-1 | ELO-2 | ELO-3 |
| 1 | 0.110 | 0.184 | 0.208 | 0.223 | 0.219 | 0.333 | 0.308 | 0.252 | 0.291 |
| 2 | 0.505 | 0.571 | 0.396 | 0.348 | 0.311 | 0.262 | 0.195 | 0.408 | 0.303 |
| 3 | 0.305 | 0.224 | 0.333 | 0.272 | 0.306 | 0.274 | 0.292 | 0.240 | 0.257 |
| 4 | 0.080 | 0.020 | 0.063 | 0.158 | 0.163 | 0.131 | 0.205 | 0.100 | 0.149 |

However, it is not the case that the low percentage of 4s reflects a bias against using the extremes of the score scale. Raters were much more willing to issue scores of 1 than they were to issue 4s. In fact, in several cases, the proportion of 1s not only exceeded the proportion of 4s, but the proportion of 3s, as well. Although ELO-1 was the hardest ELO on which to earn a 4, it was also the hardest on which to earn a score of 1.

Results from the paired-*t* tests are shown in Table 3, separately for each target ELO. The last row of the table shows the pattern of statistically significant findings. The pattern shows some support for the hypothesis that ratings for the target ELO would be higher than ratings for ancillary ELOs. For each group of artifacts, there was one instance where the mean ratings for

the target ELO was higher than that for an ancillary ELO, but there was also one instance where there was no significant difference between the target ELO mean and that of another ELO mean. In all cases, the two ancillary ELOs produced mean ratings that were not statistically different from each other.

Table 3

Paired Comparisons Among ELO Scores

| | Target ELO | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ELO-1 | | | ELO-2 | | | ELO-3 | | |
| Comp | 1 *v.* 2 | 1 *v.* 3 | 2 *v.* 3 | 1 *v.* 2 | 1 *v.* 3 | 2 *v.* 3 | 1 *v.* 2 | 1 *v.* 3 | 2 *v.* 3 |
| Mean Diff | 0.27 | −0.21 | −0.10 | −0.08 | 0.07 | 0.23 | 0.10 | −0.25 | −0.03 |
| St. Dev | 0.45 | 0.71 | 0.42 | 0.63 | 0.74 | 0.29 | 0.58 | 0.70 | 0.42 |
| df | 48 | 11 | 11 | 44 | 38 | 41 | 36 | 23 | 34 |
| t | 4.14 | −1.02 | −0.86 | −0.83 | 0.59 | 5.21 | 1.03 | −1.74 | −0.40 |
| p-value | $p < .05$ | NS | NS | NS | NS | $p < .05$ | NS | $p < .05$ | NS |

Results from the 2-way split-plot ANOVA are shown in Table 4. Asterisks in the "p" column denote that the effect was statistically significant at the $\alpha = .05$ level. When ELO-1 was the target ELO, ratings with respect to ELO-3 were not included in the analysis because the number of artifacts that were appropriate for ELO-3 was prohibitively small. This explains why the

numerator degrees of freedom for the ELO main effect is 1, instead of 2, as it is for the other two target ELOs.

As can be seen, both main effects and the interaction effect were statistically significant when the target ELO was 1 or 2; when ELO-3 was the target, only the interaction was significant. Significant main effects for ELO in groups 1 and 2 indicates that the average ratings were different depending on which ELO was being evaluated. Similarly, the class main effect suggests that, regardless of the ELO being evaluated, the average ratings were higher in some classes than in others. It is not clear from this analysis whether this reflects differences in instruction, student quality, or the nature of the artifacts (some may be better than others at allowing students to demonstrate their level of mastery). In addition, the significant ELO × class interactions mean that the pattern of average means across ELOs was different for different classes. Again, this is not a surprising result. It is easy to imagine that artifacts from different classes would differ with respect to their abilities to provide meaningful information about the ancillary ELOs, if not also about the target ELO.

Although most of the effects studied in Table 4 were statistically significant, it is also important to consider the magnitudes of the effects. The effects for ELO were fairly small-to-moderate, suggesting that it was not a major contributor to the variation in ratings. The effects for class, on the other hand, were much larger and were between three and ten times bigger than the effects for ELO. The magnitude of these effects is sufficiently large that it would be difficult to explain them entirely as a function of instructor effects or student effects. It seems likely that the difference is also attributable to the artifacts themselves, with some artifacts not providing

students equal opportunity to demonstrate their mastery of the ELOs. Effects for the interaction were generally moderate, but were very large when ELO-3 was the target ELO.

Table 4

2-Way Split Plot ANOVA

| | Target ELO | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ELO-1[#] | | | ELO-2 | | | ELO-3 | | |
| | $F_{df}$ | p | $\eta^2$ | $F_{df}$ | p | $\eta^2$ | $F_{df}$ | p | $\eta^2$ |
| Class | 5.07 (4, 44) | * | 0.23 | 3.43 (4, 34) | * | 0.22 | 1.72 (4, 19) | NS | 0.13 |
| ELO | 29.57 (1, 44) | * | 0.07 | 4.20 (2, 68) | * | 0.02 | 2.74 (2, 38) | NS | 0.02 |
| ELO × Class | 8.574 (4, 44) | * | 0.08 | 5.83 (8, 68) | * | 0.09 | 8.91 (8, 38) | * | 0.31 |

#Only ratings for ELO-1 and ELO-2 were analyzed because of small N for ELO-3

With any type of a study involving application of rubrics or judges' ratings, it is necessary to verify that the rubric was applied consistently across raters and that the different individuals involved in ratings were evaluating the same artifacts in ways that were reasonably similar. Table 5 presents the median, minimum and maximum inter-rater correlations between artifact ratings for all pairs of raters within a group. Because it is the expectation that experts will generally evaluate the same work similarly, it was our expectation that the correlations would be positive and moderately high. However, that expectation was not consistently met. It appears as though raters had a very difficult time agreeing on how artifacts should be rated. In all cases, the median correlations were less than 0.5, and the maximum correlations (between the two raters

Table 5

Inter-rater Correlations

| | Target ELO | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ELO-1 | | | ELO-2 | | | ELO-3 | | |
| | ELO-1 | ELO-2 | ELO-3 | ELO-1 | ELO-2 | ELO-3 | ELO-1 | ELO-2 | ELO-3 |
| N | 50 | 49 | 12 | 46 | 49 | 42 | 37 | 50 | 35 |
| Median Correlation | 0.39 | 0.19 | 0.39 | 0.29 | 0.42 | 0.48 | 0.13 | 0.06 | 0.31 |
| Max Correlation | 0.49 | 0.39 | 0.76 | 0.43 | 0.58 | 0.62 | 0.41 | 0.29 | 0.47 |
| Min Correlation | 0.03 | 0.03 | 0.17 | 0.14 | 0.27 | 0.34 | −0.20 | −0.14 | −0.11 |

who were most similar) never exceeded 0.76. Median correlations were particularly low when ELO-3 was the target ELO. Similarly, the minimum inter-rater correlations were all very low. For artifacts intended to measure ELO-1, the minimum inter-rater correlations were as low as 0.03, indicating no linear relationship between raters' scores. For artifacts intended to measure ELO-3, inter-rater correlations for all three ELOs were negative. Negative correlations mean that artifacts that were rated highly by one judge tended to be rated poorly by the other, and vice versa. The extreme lack of agreement between judges rating common artifacts is cause for concern, as it suggests that judges were either not attending to the same artifact details in the same manner or that they were, in effect, each using their own internal rubric.

Small amounts of rater inconsistency are expected. In fact, it is for this very reason that studies such as this often insist on having multiple raters per artifact. However, in this case, the extreme lack of consistency raises some questions about the appropriateness of averaging scores across raters, since it is not clear that the items being averaged are all measures of the same construct.

## Discussion

The results of this study have raised many questions: Are less than half the UW students really mastering the ESR ELOs to the desired degree? Is the ESR rubric sufficiently well refined? Was the research design used in this study appropriate for evaluating ELOs in courses satisfying the ESR? Is it simply too challenging to parse evidence of ESR ELOs from assignments designed to principally measure other outcomes? Were the artifacts too varied and decontextualized to allow raters to determine whether responses provided evidence of the ELOs? Did instructors understand the basis on which to provide artifacts to the research team? Are ESR courses collecting evidence of programmatic ELO mastery, or are assessments overwhelmingly focused on course-specific ELOs?

The UW-Madison ESR has not previously been evaluated. This study provided a first glimpse at student learning in ESR courses. At first blush, the results from this study would seem to be disappointing. The magnitude of ELO mastery demonstrated was fairly low, even in the target ELO. Target ELO mean ratings were below the scale midpoint for all three ELOs and greater than 20% of the students achieved ratings of 1.0 on ELO-2 and ELO-3 when those were the target ELOs. If the findings of this study are true and the average level of ELO mastery is generally fairly low, it would be interesting to conduct a pre-post study or a comparison study with a non-ESR control group to assess whether mastery of the ELOs increase as a result of the

course. If it could be demonstrated that ESR courses raise students' scores, but that those scores are still lower-than-desired, it might suggest that more than one ESR course is needed to achieve the desired mastery levels.

In addition, ELO ratings varied considerably from class-to-class, suggesting a lack of uniformity across classes with respect to the match between the assignment and the rubric. Furthermore, 30 percent of the artifacts selected by instructors as the best measures in those classes of ELO-3 were flagged as inappropriate for evaluation with the ELO-3 rubric. There are many plausible explanations for these findings. Perhaps the rating team was not provided with enough information about the specific assignments to allow them to evaluate the ELOs. It might have been very helpful to have provided the rating team with syllabi for the courses or the descriptions of the assignments that were given to students. It might also have been helpful to have the instructors or TAs for the courses attend the rater training and spend a few minutes talking about the assignments and the types of evidence the raters should be looking for and expecting to find. Or perhaps the task is sufficiently complex that it is unreasonable to expect graduate students to appreciate all the nuances in students' arguments or to differentiate misunderstandings about content from misunderstandings about ESR ELOs. It may be that the job of assigning scores to ESR artifacts would be best done by a team of ESR faculty.

An alternative explanation is that, in some classes, none of the assignments were particularly well-suited for measuring the program-level ELOs. We must recall that the assignments selected for use in this study were all designed as regular classroom projects. Although the selected projects were, presumably, the ones that mapped most closely onto the ESR ELOs, none were written exclusively for that purpose. It is quite likely that the artifacts were more closely aligned

with course-specific ELOs than with ESR-specific ELOs. Given that the students' task was not specifically to speak to the ELOs, finding evidence of mastery of the programmatic ELOs in an assignment written for another purpose altogether may have presented a needle-in-a-haystack problem for reviewers. Note that this is not to necessarily suggest that the ESR ELOs are being assessed inadequately (or not at all), but it is entirely reasonable that evidence of ELO mastery may be distributed across multiple assignments, such that no one single assignment gives a very complete picture of a student's level. It is also a possibility that the artifacts that were, in theory, most closely aligned with the ESR ELOs were administered too early in the semester, meaning that artifact scores reflect a work-in-progress, rather than the end-state for those students.

If it is true that typical classroom assignments, even at their best, do not provide adequate evidence of ELO mastery, it is worthwhile to ask whether there is anything that can be done to alter that reality and potentially improve ESR assessment. One possibility may be to request that ESR instructors develop an assignment, to be administered towards the end of the semester, the purpose of which is to directly assess the intersection between the course's objectives and the ESR ELOs. Such an artifact would provide students the opportunity to demonstrate their mastery of the ELOs, while remaining highly relevant to the course. However, contextualizing the artifacts in course-specific content would still mean that raters (a) might require content knowledge across an unreasonably broad spectrum, and (b) would be tasked with evaluating artifacts that varied considerably across courses. Perhaps it is worth considering if even greater standardization in assignments is even possible. Given the breadth of topics in courses that satisfy the ESR, it is not clear that it is feasible to develop a standardized assignment which would be reasonable to administer in all classes satisfying the ESR. But if such an assignment

were possible, it could go a long ways towards eliminating some of the challenges we faced in this study.

A third possibility is that the problem is not with the raters or the artifacts, but with the rubric itself. While much care was given to articulating the types of characteristics expected of artifacts at each of the four score points within each ELO, it is possible that the nature of this task did not lend itself well to holistic scoring. The general education assessment project most similar to this one, a recent assessment of the Communications-A requirement (Wollack, Young, Klein, & Westphal-Johnson, 2011), also asked raters to evaluate artifacts of student learning, but the writing rubrics developed for that project required raters to record a score on several different dimensions, in addition to a holistic score. Perhaps the ESR rubrics would have benefited from having been further refined and explicating the core elements comprising each ELO.

Perhaps, also, tasking raters to evaluate each artifact with respect to all three ELOs led to a lack of focus and uncertainty over how to differentiate them. Raters were clearly told to focus on their target ELO. However, this also meant that raters were assigning scores to other ELOs with very little training on those rubrics. The rationale behind the decision to ask raters to assign scores for all three ELOs was sound—it is commonly held that raters are naturally holistic scorers and have difficulty focusing on a single, isolated component without being influenced by other components. Therefore, by allowing them the opportunity to express satisfaction or dissatisfaction with aspects of the artifact that are irrelevant to the task at hand, we allow them to better provide a score that is focused on the relevant aspects. However, in this case, given that so many of the artifacts were inappropriate for scoring with respect to the ancillary ELOs, this may not have been the best decision. The groups of raters all spent considerable time discussing

aspects of the ancillary ELOs, time that likely could have been better spent focusing on the ELO of interest. By focusing only on the target ELO, the rating event could have more easily been divided into separate events, allowing each ELO group to meet separately, making it easier to invite faculty to talk about details of their assignments and enabling all raters and trainers to have engaged in a single, common discussion about the rubrics.

As mentioned previously, this study represented the first time that the ESR has been evaluated. The process involved leading up to and throughout this study was fantastically collaborative and educational and in spite of the limitations of this study, represented a monumental step forward for the ethnic studies program. It will be important to build on the momentum of this project and use the information we learned from this study to help improve future assessments.

References

Cohen, J. (1988) Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Wollack, J. A., Young, M., Klein, E., & Westphal-Johnson, N. (2011). *An Assessment of Writing Outcomes in the First Semester of College at the University of Wisconsin-Madison: A Pilot Study*. Madison, WI: University of Wisconsin.

# Appendix A: ETHNIC STUDIES RUBRIC

## Essential Learning in the Ethnic Studies Requirement

Adapted from Association of American Colleges and Universities "Intercultural Knowledge and Competence," "Critical Thinking" and "Civic Engagement" VALUE Rubrics

| SCORE → | 4 | 3 | 2 | 1 |
|---|---|---|---|---|
| **ESR Outcomes** ↓ | *Artifact exhibits **sophisticated and substantial** cognitive, affective, and behavioral skills and characteristics that support effective and appropriate understanding and interaction in a variety of cultural contexts.* | *Artifact exhibits **developing and consistent** cognitive, affective, and behavioral skills and characteristics that support effective and appropriate understanding and interaction in a variety of cultural contexts* | *Artifact exhibits **emerging and inconsistent** cognitive, affective, and behavioral skills and characteristics that support effective and appropriate understanding and interaction in a variety of cultural contexts* | *Artifact exhibits **minimal and surface-level** cognitive, affective, and behavioral skills and characteristics that support effective and appropriate understanding and interaction in a variety of cultural contexts.* |
| **ELO 1: Awareness of History's Impact on the Present** | Demonstrates sophisticated understanding of the complexity of elements important to members of another culture in relation to its history, politics, economy or beliefs and practices. | Demonstrates adequate understanding of the complexity of elements important to members of another culture in relation to its history, politics, economy or beliefs and practices. | Demonstrates partial understanding of the complexity of elements important to members of another culture in relation to its history, politics, economy or beliefs and practices. | Demonstrates surface understanding of the complexity of elements important to members of another culture in relation to its history, politics, economy or beliefs and practices. |
| **ELO 2: Ability to recognize and questions assumptions related to culture** | Thoroughly analyzes own and others' assumptions regarding culture and carefully evaluates the relevance of contexts when presenting a position. | Demonstrates identification of own and others' assumptions regarding culture and several relevant contexts when presenting a position. | Questions some assumptions regarding culture. May be more aware of others' assumptions than one's own (or vice versa). | Shows a minimal awareness of present assumptions regarding culture. Begins to identify some superficial contexts when presenting a position. |
| **ELO 3: Consciousness of "Self" and "Other" *(cultural self-awareness)*** | Articulates insights into own cultural rules and biases (e.g. aware of how her/his experiences have shaped these rules, and how to recognize and respond to cultural biases, resulting in a shift in self-perception. | Recognizes multiple perspectives about own cultural rules and biases (e.g. not looking for sameness, comfortable with the complexities that multiple perspectives offer). | Identifies own cultural rules and biases (albeit with a strong preference for those rules shared with own cultural group and seeks the same in others). | Show minimal awareness of own cultural rules and biases (even those shared with own cultural groups, e.g. uncomfortable with identifying possible cultural difference with others). |

*NOTE: A N/A should also be indicated for any ELO for which the artifact provides no opportunity to evaluate skills and characteristics that support effective and appropriate understanding and interaction in a variety of cultural contexts.*