

Expected Learning Outcomes / Outcomes Assessment Evaluation
Department of Sociology
Spring 2018

This report summarizes the Department of Sociology's expected learning outcomes (ELO) assessment for Spring 2018.

BACKGROUND

In February 2015, the Department of Sociology adopted five ELOs and developed an assessment plan to gauge how well students are meeting these objectives. The five ELOs for the BA and BS degrees in Sociology are as follows:

- ELO 1: Understand what sociology is, as a social science discipline;
- ELO 2: Utilize sociological theories to guide research and improve understanding of social phenomena and human behavior;
- ELO 3: Learn to use a variety of research methods as a means of understanding the social world and human interaction;
- ELO 4: Apply sociological and social-science perspectives to the understanding of real-world problems or topics (e.g., issues of diversity, health, globalization, crime & law, sustainability); and
- ELO 5: Communicate effectively about sociological issues, making well-organized arguments supported by relevant evidence.

In Spring 2017, the Undergraduate Committee assessed two of these five learning outcomes: ELO 1 and ELO 3. For each of these outcomes, the Committee had two separate reviewers evaluate one set of assignments. Although the Spring 2017 report concluded that sociology classes are meeting the learning outcomes, it also noted that reviewers often disagreed in their evaluations. It further suggested that future reports should incorporate more reviewers and include more assignments.

CURRENT EVALUATION

In Spring 2018, the Undergraduate Committee undertook to evaluate the Department of Sociology's remaining ELOs: 2, 4, and 5. Based on the recommendations of the Spring 2017 report, the Committee solicited two sets of student artifacts—obtained from different classes—for each outcome, and used three raters instead of two to evaluate these artifacts. Thus, for each ELO, three reviewers independently rated the same 12 student artifacts (six for each of two courses).¹

In December 2017, the Director of Undergraduate Studies (DUGS) asked instructors from Spring and Fall 2017 courses to submit updated ELO/Outcome Assessment matrices indicating which of the five ELOs (if any) their courses addressed and how they assessed these ELOs. The DUGS used these matrices to

¹ There is one exception: two raters for ELO 4 offered artifacts from their own courses to be evaluated. Their ratings of the artifacts from these courses are removed, leaving two sets of scores (even though their initial evaluation of "high-," "intermediate-," and "low-quality" work did not always align with their subsequent ratings of the same work).

identify a sample of suitable courses for the Spring evaluation. Six courses were identified, two for each ELO under review. For each ELO, one elective course and one course required of all majors were selected. Only one lower-division course was included (the vast majority of courses in Sociology are upper division).

ELO 2	Upper-division required course	Upper-division elective
ELO 4	Lower-division required course	Upper-division elective
ELO 5	Upper-division required course	Upper-division elective

The DUGS contacted the instructors for these courses, asking them to submit six student artifacts from assignments they used to evaluate the stated ELO. Instructors submitted two artifacts they deemed high quality, two intermediate quality, and two low quality. The artifacts consisted of written assignments, final papers, and essays. After anonymizing the artifacts, they were distributed to Committee members for their evaluations.

Evaluators were provided with a scoring sheet for rating the artifacts (see attached). These scoring sheets asked evaluators first to consider how well the assignment assesses the specified ELO. Then, for each of the six student artifacts, evaluators rated mastery of the ELO using the following four-point scale:

- 0 = Poor (There is no evidence that the ELO was addressed);
- 1 = Emerging/Low (Initial but substandard effort to address the ELO);
- 2 = Competent/Mid (ELO was achieved with reasonable proficiency); and
- 3 = Exemplary/High (Artifact demonstrates mastery of the ELO).

Evaluators were also asked to provide brief qualitative feedback to support their ratings.

ANALYSIS

The first set of results consider levels of interrater agreement for each ELO under review. Table 1 reports Krippendorff's alpha reliability coefficients for reviewers' ratings. Three sets of coefficients are reported. The first column presents reliability coefficients among raters only. The second column includes the instructors' independent assessments of their students' work. Again, instructors were asked to provide two artifacts in each of three categories—"high," "intermediate," and "low"—based on how well they thought students performed with respect to the specified ELO. For the analysis, these categories were assigned scores of 3, 2, and 1, respectively. Because raters were able to assign an additional score of 0 (indicating that a given artifact did not address the ELO at all), reviewers' and instructors' ratings are not always directly comparable. To render them comparable, the last column uses reviewers' ratings after recoding all 0's (poor) to 1's (emerging/low), where applicable.

According to Krippendorff, an α of .80 or greater represents an "acceptable" level of agreement. He further suggests that an α of .67 is the "lowest conceivable limit" for which tentative conclusions can be drawn.² On the basis of these thresholds, artifacts corresponding with ELO 4 show the highest level of

² Klaus Krippendorff, "Reliability in Content Analysis: Some Common Misconceptions and Recommendations," *Human Communication Research*, vol. 30, no. 3 (2004), p. 429.

reliability in the analysis. Across all available measures, the coefficients fall just short of the recommended threshold of .80. Conversely, reliability levels for ELOs 2 and 5 fall well below accepted thresholds for reliability.

Table 1. Krippendorff's alpha reliability coefficients

	Raters only	Raters plus instructor	Raters plus instructor [†]
ELO 2	.481	.470	.439
ELO 4	.774	.770	n/a
ELO 5	.361	.493	.504

[†] For this analysis, ratings of 0 (poor) were recoded as 1 (emerging/low) to render them comparable with instructors' assessment of high, intermediate, and low artifacts. No coefficient is estimated for ELO 4 because none of the artifacts received a rating of 0.

To corroborate these findings, a second reliability analysis was performed using kappa interrater agreement scores. These scores vary from 0 (the amount of agreement is what would be expected to be observed by chance) and 1 (perfect agreement). Intermediate values can be interpreted as follows:³

< .00	Poor
.00 – .20	Slight
.21 – .40	Fair
.41 – .60	Moderate
.61 – .80	Substantial
.81 – 1.0	Almost perfect

As before, ELO 4 shows the highest degree of interrater reliability, with scores falling squarely in the moderate range. Both scores for the ELO 4 assessment also exceed conventional thresholds for statistical significance. Scores for the remaining ELOs were much lower by comparison, reaching only slight to fair levels of agreement.

Table 2. Combined kappa interrater agreement scores

	Raters only	Raters plus instructor	Raters plus instructor [†]
ELO 2	.120*	.246***	.249**
ELO 4	.497**	.498***	n/a
ELO 5	.130	.243***	.327***

* $p < .05$, ** $p < .01$, *** $p < .001$ (two tailed).

[†] For this analysis, ratings of 0 (poor) were recoded as 1 (emerging/low) to render them comparable with instructors' assessment of high, intermediate, and low artifacts. No coefficient is estimated for ELO 4 because none of the artifacts received a rating of 0.

³ J.R. Landis and G.G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33 (1977), p. 165.

One benefit of kappa scores is that they can be disaggregated to determine where there is more or less agreement. Table 3 reports these disaggregated kappa scores. Consider, for example, ELO 2. Raters (and instructors) showed substantial agreement on artifacts they rated poor or low, but they did not agree as to what constitutes an intermediate or high degree of mastery. Likewise, raters for ELO 4 showed very high levels of agreement regarding low-level work, but they differed in their assessment of mid- and high-quality artifacts. Agreement was weakest with respect to ELO 5, where raters showed only a fair amount of agreement for artifacts deemed “low” but disagreed quite extensively across all other ratings.

Table 3. Disaggregated kappa interrater agreement scores

	Rating	Raters only	Raters plus instructor†
ELO 2	Poor	.768***	n/a
	Low	.518**	.484***
	Mid	-.178	.063
	High	.000	.188
ELO 4	Poor	n/a	n/a
	Low	.822**	.880***
	Mid	.250	.250
	High	.395	.345*
ELO 5	Poor	-.091	n/a
	Low	.308*	.484***
	Mid	-.037	.121
	High	.182	.348**

* $p < .05$, ** $p < .01$, *** $p < .001$ (two tailed).

† For this analysis, ratings of 0 (poor) were recoded as 1 (emerging/low) to render them comparable with instructors’ assessment of high, intermediate, and low artifacts.

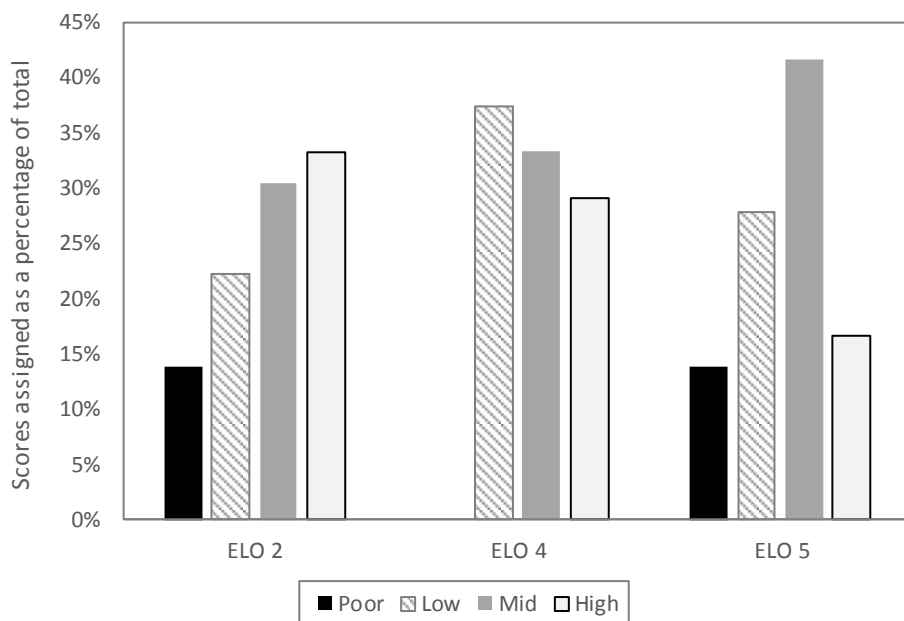
Do instructors and raters evaluate student mastery of ELOs similarly? Table 4 addresses this question by presenting rank-order correlations between instructors’ and raters’ scores. This metric confirms that agreement tends to be strongest for ELO 4, followed by ELO 5. Instructors and raters in these categories tended to assess student work in a similar fashion. Agreement is lowest, again, for ELO 2, where two raters in particular evaluated student artifacts differently from the instructors who submitted them.

Table 4. Rank-order correlations between instructors’ and raters’ scores

ELO	Instructor & Rater 1	Instructor & Rater 2	Instructor & Rater 3	Average correlation
2	.32	.45	.68	.48
4	.75	.89	.71	.78
5	.83	.67	.60	.70

Interrater reliability analyses tell us the extent to which raters and instructors agree in their evaluations of student work, but they do not provide information about the substance of these evaluations. Based on the samples of work provided, how well are students performing with respect to each ELO? Figure 1 plots the percentage of ratings falling into each category—poor, low, mid, and high—for each learning outcome. Although raters showed little agreement in their evaluations of artifacts for ELO 2, the modal rating was nevertheless “high,” followed closely by intermediate scores. That is, most submitted artifacts for this learning outcome were deemed either proficient or exemplary. In contrast, although raters showed much higher agreement in the scores they assigned to ELO 4 artifacts, they also agreed that the majority of this work evinced substandard levels of achievement. Artifacts for ELO 5 were most likely to be scored proficient, followed by substandard.

Figure 1. Composition of ratings across ELOs



SUMMARY AND RECOMMENDATIONS

These analyses suggest an acceptable level of interrater agreement for ELO 4, regarding the application of sociological and social-science perspectives to the understanding of real-world problems or topics. However, raters tended to agree that students were not always successful in achieving this learning outcome. Agreement was much weaker with respect to ELOs 2 and 5—pertaining, respectively, to the utilization of sociological theories to guide research and effective communication about sociological issues.

Going forward, the faculty may need to discuss how these ELOs are interpreted and assessed, with the goal of generating more agreement across instructors, courses, and raters. There seems to be at least tacit agreement regarding what constitutes low-quality work, but it is apparently more difficult to identify work that is proficient or exemplary.

As part of this discussion, it may be fruitful to revisit the ELOs with an eye toward distinguishing them more forcefully from one another. This seems especially relevant for ELOs 2 and 4. What does it mean to “utilize sociological theories” and “apply sociological perspectives”? How are theories and perspectives

different? In qualitative feedback, reviewers noted that assignments designed to assess ELO 2 were perhaps not always well-suited to the task, precisely because they asked students to consider sociological perspectives and concepts rather than theories per se.

It may also be useful to standardize the department's assessment procedures and methodologies, which would permit longitudinal analyses in the future. It is essential to compile and analyze comparable data over time to establish trends and track progress (or the lack thereof). In light of this suggestion, the Committee recommends reanalyzing ELOs 1 and 3 next spring using the framework established in this report.

Evaluations should continue to use at least two samples of student artifacts for each ELO under review, and to assign at least two (and preferably three) raters to each set of artifacts.

The Committee might also consider abandoning the practice of soliciting "stratified samples" of student artifacts from instructors. This strategy primes reviewers to expect patterned variation in the artifacts they evaluate, and they may attempt to rate those artifacts accordingly. Raters know *a priori* that they will be given two examples each of high-, intermediate-, and low-quality work for review, and this knowledge may bias their own independent assessments of the artifacts. If anything, this approach may overstate levels of interrater agreement. It might be useful to inject more variation into the samples, if only to keep reviewers "guessing." For example, without being told, reviewers might be given three examples of substandard work, two examples of exemplary work, and only one example of work deemed intermediate in quality. These ratios could then be varied across samples.

It is also advisable to develop explicit rubrics for evaluating student mastery of ELOs, in order to improve interrater reliability but also to guide instructors as they design student assessments.

Finally, we offer a few points to consider based on our review of the ELO/Outcomes Assessment matrices submitted by instructors. These matrices sometimes indicated that class discussions were used to evaluate mastery of ELOs. Classroom observations may therefore need to be incorporated into the yearly evaluation. Moreover, some matrices indicated that multiple-choice exams were used to assess ELO 5, pertaining to communication. It is difficult to see, *prima facie*, how a multiple-choice format is suited to the evaluation of this type of learning objective. Some discussion as to what kinds of assessments best correspond with different ELOs might be useful.

Your name: _____

Date of assessment: _____

ELO 2: Utilize sociological theories to guide research and improve understanding of social phenomena and human behavior

Artifact #1

How well does the assignment assess the specified ELO?

Rate each of the six student “artifacts” (examples of student work) below.

Rate the student’s mastery of the ELO.	<u>0 = Poor</u> There is no evidence that the ELO was addressed	<u>1 = Emerging/Low</u> Initial but substandard effort to address the ELO	<u>2 = Competent/Mid</u> ELO was achieved with reasonable proficiency	<u>3 = Exemplary/High</u> Artifact demonstrates mastery of the ELO
--	---	---	---	--

Artifact 1-A	Circle a rating and provide feedback to justify it	0	1	2	3
Artifact 1-B	Circle a rating and provide feedback to justify it	0	1	2	3
Artifact 1-C	Circle a rating and provide feedback to justify it	0	1	2	3
Artifact 1-D	Circle a rating and provide feedback to justify it	0	1	2	3
Artifact 1-E	Circle a rating and provide feedback to justify it	0	1	2	3
Artifact 1-F	Circle a rating and provide feedback to justify it	0	1	2	3

Additional comments regarding your assessment	
---	--