UCDAVIS
UNDERGRADUATE EDUCATION

Academic

English 10 Writing Assessment – Results and Analysis

OVERVIEW

This study is part of a multi-year effort undertaken by the Department of English to develop sustainable outcomes assessment practices to inform curricular, budget, and strategic planning efforts, as well as program self-study processes.

The goal of this portion of the project is to provide information about the validity and reliability of the pilot rubric, which was designed to articulate criteria of value that are fully portable across literary-historical period and genres. Specifically, the rubric is intended to advance the larger goal of determining whether expectations and outcomes are clearly defined and consistently deployed in final essays in the series known as the “10 Series,” which comprises ENL 10A, ENL 10B, and ENL 10C.

METHOD

Participants

Seven end-of-term research papers were randomly selected from each of two sections of ENL 10A, ENL 10B, and ENL 10C offered in Winter quarter 2015 for a total n = 42 papers. Identifying information from the papers was removed.

Materials

The Curriculum Committee of the English department developed a rubric to assess general qualities of writing. The rubric contains five items: Prose, Argument, Thesis, Context and Research. Each item is measured on a 4-point scale with a performance level of 1 being the lowest, or least developed, and 4 being the highest, or most developed. It is important to note that the iteration of the rubric used in this study did not include descriptions at each performance level.

Data Collection

Seven faculty members from the English department volunteered to read and rate the papers using the rubric developed by the English Department Curriculum Committee. The faculty members were assigned the papers from two sections of ENL 10 with two faculty readers assigned to each paper, and no readers paired together for more than eight papers. The seven faculty readers participated in a “norming” session where the rubric criteria were discussed and applied to six papers randomly chosen from across the six sections and representing grade levels of A, B and C (two from each level). These papers were not included in the sample rated by the readers. Once normed, the faculty readers were given links to the papers they were to rate and they recorded their ratings in a Google spreadsheet.

Data Analysis

Evaluating Interrater Reliability

Interrater reliability is an important estimate when independent raters are used to objectively measure traits that are inherently subjective. Generally, interrater reliability quantifies the extent to which multiple raters agree when rating or ranking subjects on established criteria. Reliability is an important criteria of instrument validity, as an instrument cannot be considered valid without an acceptable level of reliability. The type of interrater reliability used for a particular study should be chosen with suitability of the assessment purpose in mind.

There are three types of interrater reliability: consensus estimates, consistency estimates, and measurement estimates (Stemler, 2004). Consensus estimates (one of the estimates used in this study) are based on the idea that two independent judges should be able to come to exact agreement on the application of various levels of a scoring rubric. If exact agreement can be reached, it can be concluded that the judges share a common interpretation of the construct(s) being measured. Cohen's kappa is a consensus estimate and modifies a simple percent agreement statistic by correcting for the amount of agreement that could be expected by chance. Cohen's weighted kappa (used in this study) is also based on a percent agreement calculation, but differentially penalizes discrepant scores based on the magnitude of the discrepancy. For instance, a subject rated 1 by one rater and 2 by another would not decrease the interrater reliability statistic as much as a subject who had been rated 1 by one rater and 4 by another.

The other estimate of reliability used in this study is Cronbach's alpha, a consistency estimate. Consistency estimates are based on the assumption that raters do not have to agree on a common meaning of the rating scale as long as each rater is consistent in applying the scale based on his/her definition. For example, one rater may consistently give a rating of 1 where another consistently gives a rating of 3 and although the raters may not agree on how to apply the rating scale, the difference in their application is predictable.

Evaluating Student Performance

In order to evaluate student performance, ratings were tabulated at two different performance levels: the first considered a rating of 2 or better to be acceptable, the second considered a rating of 3 or better to be acceptable, regardless of perfect agreement between raters. For instance, if both raters scored the paper at or above *x* (whatever the acceptable performance level, 2 or 3), the rubric item was considered met, if both raters scored below *x*, the item was considered unmet. If one rater scored below *x*, and one scored *x* or above, the item was considered discrepant.

Evaluating Instrument Validity

The third question, regarding instrument validity, will be analyzed in future analyses by comparing the ratings given by the faculty raters using the rubric to the paper grades officially given by the instructors of each ENL 10 section.

RESULTS

Interrater Reliability

The interrater reliability estimates indicate low reliability between all rater pairs on most of the criteria. The consensus estimates in Table 1 shows that there was little agreement on application of the rubric to student work. Widely accepted guidelines for interpreting kappa values (Landis & Koch, 1977) are as follows: values between 0.0 to 0.2 indicate slight agreement, values between 0.21 to 0.4 indicate fair agreement, values between 0.41 to 0.6 indicate moderate agreement, values between 0.61 to 0.80 indicate substantial agreement, and values between 0.81 to 1.0 indicate almost perfect, or perfect agreement. Two pairs of raters had predominantly negative values of kappa indicating that they agreed less than would be expected by chance. The other four pairs each had at least one kappa value in the moderate agreement range, though not all of the values were statistically significant. Lack of statistical significance indicates that the range of possible values for a pair of raters could be negative or positive and is therefore inconclusive. Since the sample sizes for each group were very small, a low Cohen's weighted kappa value would indicate large discrepancies between ratings (see Appendix A for percent agreement and magnitude of discrepancy).

Table 1. Cohen's weighted kappa for each rater pair, by criterion

Rater Pair	Prose	Argument	Thesis	Context	Research
1	0.727*	0.037	0.458	0.333	0.545
2	0.292	0.722*	0.077	0.408	0.705*
3	0.090	0.495	0.624*	0.157	N/A
4	0.778	0.424	0.297	0.174	0.600
5	-0.429	-0.667	0.160	-0.293	-0.565
6	-0.641	0.556	-0.078	-0.148	0.098

**significant at the .05 level*

The consistency estimates also indicate low reliability among raters. Generally, a Cronbach's alpha value of 0.7 is considered satisfactory, and as shown in Table 2, each pair had at most 2 criteria for which they reached a satisfactory level, with one pair having none. Again, two of the rater pairs produced negative values of alpha indicating very weak correlations.

Table 2. Cronbach's alpha for each rater pair, by criterion

Rater Pair	Prose	Argument	Thesis	Context	Research
1	0.842	0.071	0.629	0.500	0.706
2	0.452	0.839	0.143	0.580	0.827
3	0.164	0.662	0.769	0.271	N/A[1]
4	0.875	0.596	0.458	0.296	0.750
5	-1.5	-4	0.276	-8.28	-2.6
6	-3.574	0.714	-0.17	-0.348	0.179

Student Performance

The results regarding student performance should be read with the understanding that they are not generalizable due to the lack of interrater reliability.

For future studies, and program review, the English department will need to decide which performance level is appropriate for each course in which the rubric is used. Generally, scores of 2 or below on a 4-point rubric indicate that student performance is approaching expectations. The tables below illustrate the percentages of pairs of scores, disaggregated by criterion, in which both raters gave a rating of 2 or higher (Table 3), and 3 or higher (Table 4) where perfect agreement between raters was not necessary. For both performance levels, Argument had the highest percentage of students meeting the standard and Research had the lowest percentage of students meeting the standard. Additionally, the number of discrepancies greatly increases from performance level 2 to performance level 3 which supports the evidence of low reliability between raters.

Disaggregation of the criteria can indicate general areas of strengths and weaknesses in students' writing, however, it is also useful to see how many students meet the desired performance level across multiple criteria (Table 5).

[1] In this group one rater marked N/A for every paper in the Research item, which is considered missing data and is therefore excluded from analysis.

Table 3: Percentage of pairs of scores rated 2 or higher

Criterion	% score ≥ 2	% score 1	% Discrepancy[2]
Prose	74%	2%	24%
Argument	88%	5%	7%
Thesis	81%	7%	12%
Context	83%	2%	17%
Research	57%	5%	33%

Table 4: Percentage of pairs of scores rated 3 or higher

Criterion	% score ≥ 3	% score ≤ 2	% Discrepancy
Prose	33%	26%	40%
Argument	45%	17%	38%
Thesis	33%	26%	40%
Context	29%	24%	45%
Research	26%	24%	45%

Table 5: Percentage of students who met *x* number of criteria at performance levels 2 and 3 and percentage of discrepant pairs at each level

Number of criteria Met (*x*)	**1**	**2**	**3**	**4**	**5**
Percent of students with pairs of scores ≥ 2 on *x* or more of the criteria (*n* = 40).	98%	95%	85%	73%	35%
Percent of students with discrepant pairs on *x* or more of the criteria (*n* = 40, performance level = 2).	65%	20%	10%	0%	0%
Percent of students with pairs of scores ≥ 3 on *x* or more of the criteria (*n* = 40).	78%	45%	25%	18%	5%
Percent of students with discrepant pairs on *x* or more of the criteria (*n* = 40, performance level = 3).	85%	65%	38%	15%	10%

[2] Discrepancy = the number of times pairs of ratings were split between Met and Not Met

Next Steps:

Interrater reliability can be influenced by myriad factors. Bresciani (2009) cites relevant research that lists many of these factors, among them inadequate detail of rubric, inadequate training of raters, and differing levels of understanding of the scoring criteria. Interrater reliability can be improved by discussing discrepant scores and appropriately amending the rubric, as well as ensuring that performance levels are clearly and specifically differentiated. Since the English department has already re-designed the rubric to include specific definitions of criteria at each performance level, the next step would be to assess reliability and validity in order to be able to report accurate student performance results.

There are multiple approaches the English department could take to this endeavor, depending on the number of available readers and time available to read. Following the assumption that time is preciously limited, the design of the next study should take that into consideration as well as increasing the sample size for pairs (or preferably groups) of readers.

APPENDIX A – ADDITIONAL DATA

Table A1. Percent Agreement Between Rater Pairs

	Rater Pairs					
Cumulative agreement between rater pairs up to +/- 1 point	1 (n=35)	2 (n=35)	3 (n=35)	4 (n=35)	5 (n=30)	6 (n=40)
Percent of perfect agreement between raters	37%	49%	23%	34%	37%	23%
Percent of agreement between raters +/- 1 point	89%	89%	74%	86%	67%	75%
Percent of discrepant pairs						
Percent of rating pairs discrepant by 2 points	11%	9%	3%	14%	23%	23%
Percent of rating pairs discrepant by 3 points	0%	3%	23%	0%	10%	3%

Table A2. Percent Agreement Between Rater Pairs by Rubric Criteria

	Rubric Criteria				
Cumulative agreement between rater pairs up to +/- 1 point	Prose	Argument	Thesis	Context	Research
Percent of perfect agreement between raters	26%	33%	40%	38%	29%
Percent of agreement between raters +/- 1 point	83%	93%	86%	71%	67%
Percent of discrepant pairs					
Percent of rating pairs discrepant by 2 points	12%	5%	14%	19%	19%
Percent of rating pairs discrepant by 3 points	5%	2%	0%	10%	14%

Notes: This gives some indication as to the criteria that may be most confusing to raters. Context and Research are particularly discrepant, Argument and Thesis, much less so.

REFERENCES

Bresciani, Marilee J., Oakleaf, Megan, Kolkhorst, Fred, Nebeker, Camille, Barlow, Jessica, Duncan, Kristin, and Hickmott, Jessica (2009). Examining Design and Inter-Rater Reliability of a Rubric Measuring Research Quality across Multiple Disciplines. *Practical Assessment, Research & Evaluation*, 14(12). Available online: http://pareonline.net/getvn.asp?v=14&n=12.

Landis, J. R., & Koch, G. G. (1977). The measure of observer agreement for categorical data. Biometrics, 33.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating inter-rater reliability. *Practical Assessment, Research, and Evaluation, 9*(4). Available online: http://pareonline.net/getvn.asp?v=9&n=4.