

Speech Assessment Report – Spring 2014

Author: Joseph F. van Gaalen, Ph.D., Coordinator, Academic Assessment

1 INTRODUCTION

Common course assessments can measure student learning of course level objectives (Hall, 2010). Florida SouthWestern's Speech Department has employed a common rubric used by all faculty as a means to evaluate an agreed upon series of student level outcomes as part of the college's voluntary regulation through SACSCOC and compliance with Florida Statute 1007.271. With a goal towards increasing student oral communication assessment scores on the public speaking rubric, the faculty has focused on five Student Learning Objectives (SLOs) through the following rubric criteria: Introduction, Organization, Oral Citations, Conclusions, and Nonverbal Physical Behaviors. While scores do yield some error related to the target subject such as grade level or demographic, many can be accounted for in small sub-samples (individual classes). Moreover, those correlative measures that cannot be accounted for can be better understood through assessment (Cole et al., 2011).

In conjunction with common course assessment employ, a norming session was conducted by faculty to assess variation among scorers using common criteria. The results of the norming session and the common course assessment results are herein described to gauge progress towards the SLOs and provide support toward instructive improvement, therefore allowing assessment to drive instruction as defined by Elder and Paul (2007).

2 OUTCOMES AND RESULTS

2.1 STUDENT LEARNING OBJECTIVES (SLOs)

Using common rubric criterion as an assessment method, the FSW Speech faculty defined five areas of interest for evaluation. The rubric criterion is modeled on a 4-point scale where a score of 1 indicates "Beginning" level, 2 indicates the "Developing" level, 3 indicates the "Accomplished" level, and 4 indicates the "Exemplary", or highest level. The SLOs and their measure of success are:

SLO1 – Students will be able to prepare and deliver an effective speech introduction that gains the attention of the audience, establishes speaker credibility, relates the topic to the audience, identifies the topic and previews the main points of the speech. The faculty established measure of success for this SLO is a rating of "Developing" or higher a for 60% of the students.

SLO2 – Students will be able to effectively organize the body of a speech by identifying main points, using appropriate support material and connecting ideas with appropriate transitions. The faculty established measure of success for this SLO is a rating of "Developing" or higher (≥ 2) for 70% of the students.

SLO3 – Students will know how to avoid plagiarizing when speaking by incorporating an oral citation that includes appropriate information. The faculty established measure of success for this SLO is a rating of “Developing” or higher (≥ 2) for 70% of the students.

SLO4 – Students will be able to provide effective speech closure by delivering a conclusion that signals the end, reviews purpose and main points and that ends smoothly and memorably. The faculty established measure of success for this SLO is a rating of “Developing” or higher (≥ 2) for 60% of the students.

SLO5 – Students will be able to apply nonverbal physical behaviors appropriately to deliver an effective extemporaneous presentation. The faculty established measure of success for this SLO is a rating of “Developing” or higher (≥ 2) for 75% of the students.

2.2 ASSESSMENT ANALYSIS & SIGNIFICANCE TESTS

2.2.1 Norming Analysis

Nine speech faculty participated in a norming exercise to determine variation among scoring from the common rubric. The following results will serve two purposes going forward: (1) A normalization factor can be applied either program-wide or within specific faculty to provide a more robust statistical analysis of the results, and (2) act as instructional support by serving as a baseline for instructor cognizance of collective interpretation and application of the rubric in cases where individual faculty measure of success is significantly different than that of the department mean.

Each of the nine faculty was placed into groups of three and scored the three unique sets of 23 student artifacts. Faculty (rater) names were replaced with numbers for anonymity. Faculty can obtain their rater index number if they wish by contacting the author directly. A radar plot of mean scores for each rubric criteria of Group A is shown in Figure 1.

Raters must assign each artifact to one of the four measurement of success levels, or score a 0 if none of the conditions are met (e.g. scores of 0, 1, 2, 3, or 4 only). In Figure 1, it is clear that Raters 1 and 4 are in consensus with six of eight criteria to a reasonable degree of certainty. The raters show some disagreement, however, with Supporting Materials and Nonverbal Communication with mean score differences of 0.75 and 0.79. By comparison, the largest difference in the mean between these two raters across the six similarly scored criteria is Vocal Expression at 0.23. All raters of the group are in consensus with Choice of Words criterion with a mean score range between the three raters of 0.24. With the exception of Choice of Words, Rater 7 consistently scores well below both Raters 1 and 4 in all criteria.

Rater 7, on average, scores more than one full rubric cell level below the next comparison rater in both Organization and Conclusion at 1.02 and 1.22 points, respectively. This is nearly the case with Introduction and Vocal Expression at 0.94 and 0.81, respectively. For details on the consistency of scoring among raters for Group A, see Table 1.

From reviewing Table 1 we can see that with respect to the six criteria in consensus between Raters 1 and 4 according to mean score, five of those six show similar rubric scoring patterns. In short, Raters 1 and 4 appear to agree on artifact placement in rubric level the majority of the time with Introduction, Organization, Conclusion, Choice of Words, and Vocal Expression. The mean score similarities of Oral

Citations is merely a statistical artifact. In other words, the raters scored very differently, but the way in which they scored differently offset their differences when averaged.

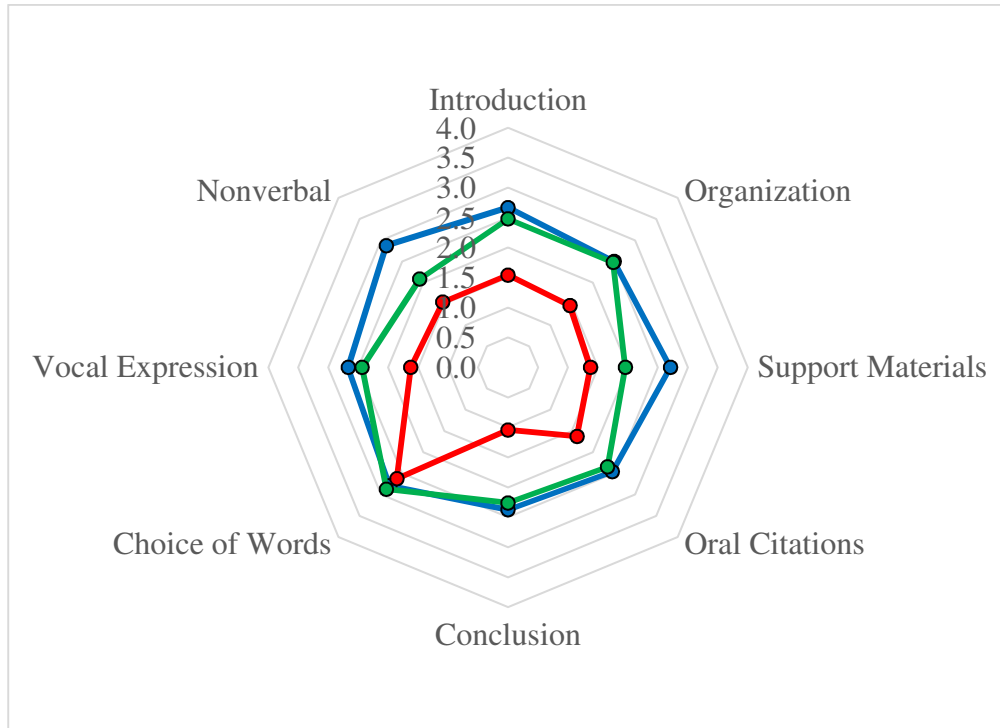


Figure 1. Radar plot of norming Group A depicting average rater score for each rubric criteria of 23 common artifacts. Blue is Rater 1, green is Rater 4, and red is Rater 7. Possible scores for each artifact are 0, 1, 2, 3, and 4. Note: Only 1 artifact scored a 0 in any criteria and it was universally scored a 0 by all three raters.

	Rubric Cell	1	2	3	4	5	6	7	8	Total
Rater 1	0	0	0	0	1	0	0	0	0	1
	1	0	0	0	5	0	0	0	0	5
	2	11	14	8	4	15	6	8	8	74
	3	10	8	15	10	9	17	16	11	96
	4	3	2	1	4	0	1	0	5	16
Rater 4	0	0	0	0	1	0	0	0	0	1
	1	1	0	5	3	2	0	1	7	19
	2	11	12	14	3	14	5	12	8	79
	3	10	11	4	3	6	16	9	7	66
	4	1	0	0	9	1	2	1	1	15
Rater 7	0	0	0	0	1	0	0	0	0	1
	1	16	15	11	7	15	0	11	10	85
	2	3	7	8	5	2	10	11	12	58
	3	5	2	2	6	2	13	2	1	33
	4	0	0	0	1	0	1	0	0	2

Table 1. Distribution of rubric cell level scores for each Group A rater. Cell 1 is Introduction, 2 is Organization, 3 is Support Materials, 4 is Oral Citations, 5 is Conclusion, 6 is Choice of Words, 7 is Vocal Expression, and 8 is Nonverbal. Green shaded cell marks the greatest total for the highest score given in each rubric criterion among the three raters, red shaded cell marks the greatest total for the lowest score given in each rubric criterion among the three raters. Note: The single 0 score for criteria 4 is not included in this tally.

The disagreement between Rater 7 with Raters 1 and 4 is more apparent in Table 1. Rater 7 consistently scores a majority of artifacts for multiple criteria in rubric cell level 1 when comparison raters scored the bulk of their artifacts for that criteria in level 2 or 3. This is the case for Introduction, Organization, Supporting Materials, Conclusion, Vocal Expression, and Nonverbal Communication. The red shaded cells in the table show that Rater 7 has the largest amount of lowest scores recorded in all criteria. Rater 1 and 4 share the largest amount of highest scores recorded in all criteria as indicated by the green shaded cells.

Figure 2 depicts a radar plot of mean scores for each rubric criteria of Group B. The raters (2, 8, and 9) show fairly good agreement in five of the eight criteria (Organization, Supporting Materials, Conclusion, Choice of Words, and Nonverbal Communication) where the largest mean score range between the three raters of 0.29. Rater 2 is consistently the highest mean score and is in disagreement with comparison raters 8 and 9 in Oral Citations. Rater 8 differs with respect to comparison raters 2 and 9 in Introduction and Vocal Expression.

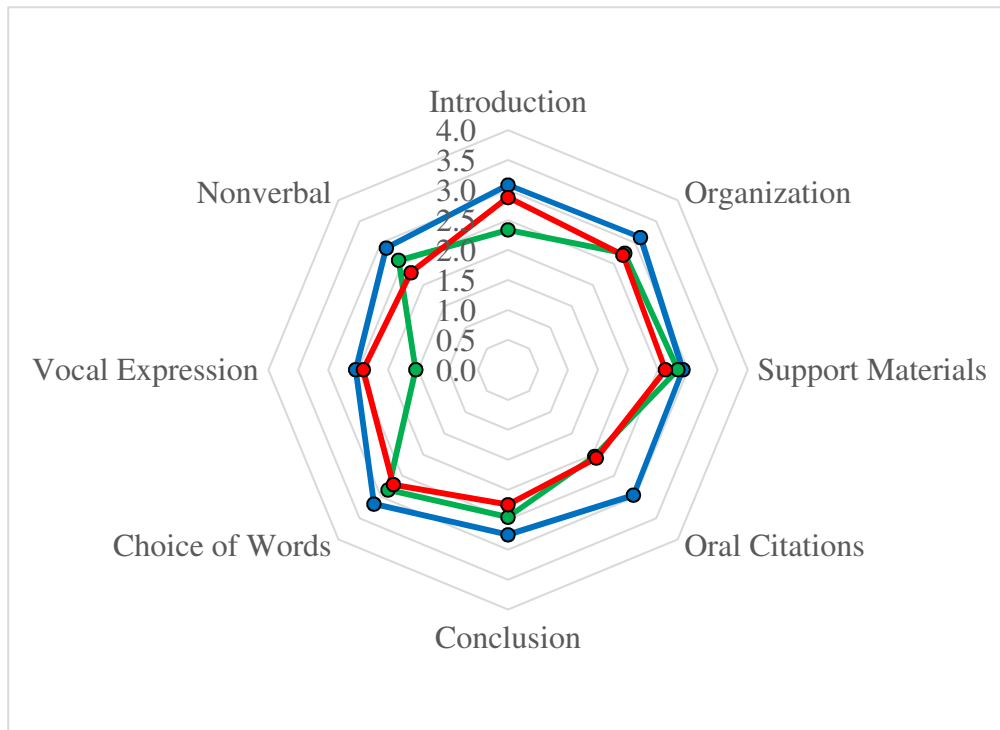


Figure 2. Radar plot of norming Group B depicting average rater score for each rubric criteria of 23 common artifacts. Blue is Rater 2, green is Rater 8, and red is Rater 9. Possible scores for each artifact are 0, 1, 2, 3, and 4.

With a review of individual scores from each rater, the cause of the disagreement of Rater 8 at Introduction and Vocal Expression becomes more apparent (Table 2). Rater 8's disagreement in each criteria stems from a tendency to score lower and hesitate to score a 4. Where Rater 2 and 9 recorded 40, and 33 level 4s, respectively, over the eight rubric criteria, Rater 8 recorded just 7.

Rater 9 exhibits a trend of not hesitating to score 0s, particularly in Oral Citations, where eight level 0s were recorded, none were recorded by Rater 2, and only two were recorded by Rater 8. The disparity between mean scores for Oral Citations is a result of this distribution as Rater 2 does not record 0s and has an inflated grade relative to comparison raters as a result.

	Rubric Cell	1	2	3	4	5	6	7	8	Total
Rater 2	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	1	1	0	0	0	2
	2	5	3	4	8	10	1	11	8	50
	3	12	15	18	6	7	18	13	11	100
	4	7	6	2	9	6	5	0	5	40
Rater 8	0	0	0	0	2	0	0	0	0	2
	1	1	0	0	7	1	0	13	0	22
	2	14	9	5	5	11	4	9	11	68
	3	9	12	18	8	12	20	2	12	93
	4	0	3	1	2	0	0	0	1	7
Rater 9	0	1	2	1	8	2	1	1	1	17
	1	1	2	2	1	4	0	1	5	16
	2	2	5	8	4	6	4	10	8	47
	3	16	7	7	3	10	19	11	6	79
	4	4	8	6	8	2	0	1	4	33

Table 2. Distribution of rubric cell level scores for each Group B rater. Cell 1 is Introduction, 2 is Organization, 3 is Support Materials, 4 is Oral Citations, 5 is Conclusion, 6 is Choice of Words, 7 is Vocal Expression, and 8 is Nonverbal. Green shaded cell marks the greatest total for the highest score given in each rubric criterion among the three raters, red shaded cell marks the greatest total for the lowest score given in each rubric criterion among the three raters.

The red shaded cells in the table indicate that Rater 9 has the largest amount of lowest scores recorded in all criteria. Rater 2, however, has the largest amount of highest scores recorded in all criteria as indicated by the green shaded cells. Based on the results of Group B norming, the interpretation of the rubric measurements for success may need to be further honed.

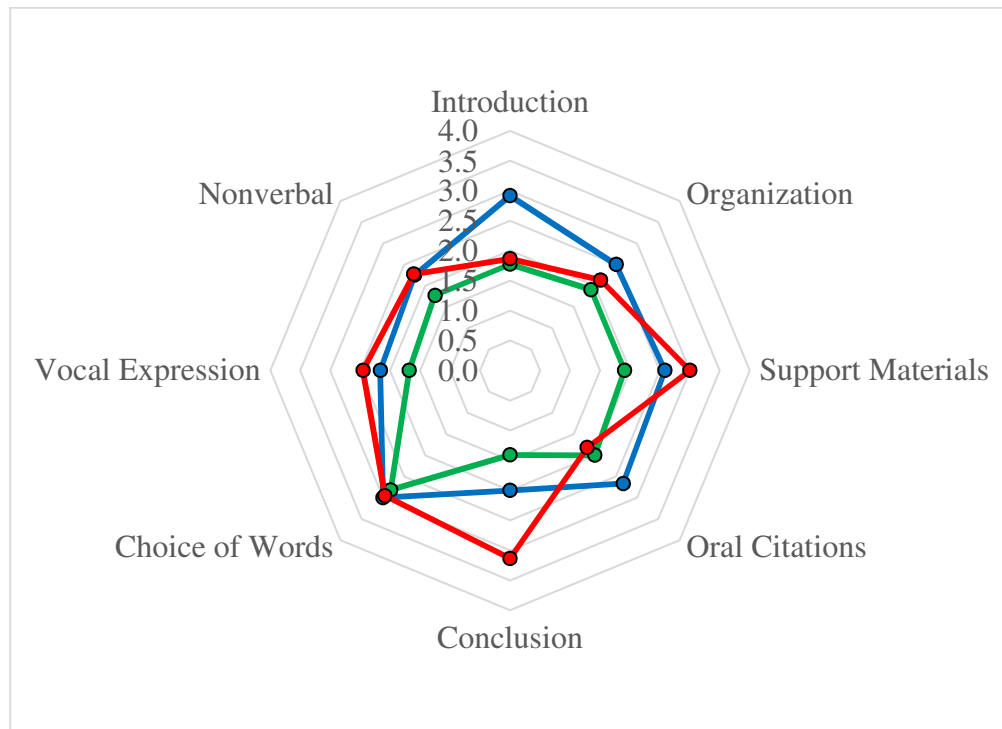


Figure 3. Radar plot of norming Group C depicting average rater score for each rubric criteria of 23 common artifacts. Blue is Rater 3, green is Rater 5, and red is Rater 6. Possible scores for each artifact are 0, 1, 2, 3, and 4.

Figure 3 depicts a radar plot of mean scores for each rubric criteria of Group C. The raters (3, 5, and 6) show strong agreement in one of the eight criteria (Choice of Words), with a mean score range of 0.18, the highest across all three groups. The group also shows fairly good agreement in two additional criteria (Organization, and Nonverbal Communication) with mean ranges of 0.59 and 0.50, respectively. Rater 3 differs from the other two raters in the Introduction criterion, while there is little agreement between any of the three raters in the Conclusion criteria.

The disagreement between Rater 3 in Introduction mean score from Raters 5 and 6 stems from a vastly differing opinion of rubric level assignments among all three raters (Table 3). Rater 3 scores 15/23 artifacts as Level 3, while Rater 5 and Rater 6 score 3/23 and 7/23, respectively. The similar mean scores of Raters 5 and 6 is a result of Rater 6 scoring 11/23 as level 1, while Rater 5 scores 7/23, thus effectively minimizing the difference caused by disagreement in level 3 assignment.

In other cases where there is little agreement between rater mean scores, the differing opinion among raters is not offset by such a circumstance and so is evident in the radar plot (Figure 3). The red shaded cells in the table show that all three raters have tallied the largest amount of lowest scores recorded in at least one criteria.

	Rubric Cell	1	2	3	4	5	6	7	8	Total
Rater 3	0	0	0	0	6	2	0	0	0	8
	1	3	5	4	0	4	0	3	2	21
	2	4	10	9	2	8	5	8	13	59
	3	15	7	9	13	7	17	10	6	84
	4	1	1	1	2	2	1	2	2	12
Rater 5	0	0	0	0	3	5	0	0	0	8
	1	7	8	5	3	4	0	8	6	41
	2	13	8	15	7	12	4	14	16	89
	3	3	6	3	10	1	19	1	1	44
	4	0	1	0	0	1	0	0	0	2
Rater 6	0	0	0	0	3	1	0	0	0	4
	1	11	3	0	6	2	0	2	6	30
	2	4	13	1	7	3	3	9	6	46
	3	7	6	21	7	4	17	10	8	80
	4	1	1	1	0	13	3	2	3	24

Table 3. Distribution of rubric cell level scores for each Group B rater. Cell 1 is Introduction, 2 is Organization, 3 is Support Materials, 4 is Oral Citations, 5 is Conclusion, 6 is Choice of Words, 7 is Vocal Expression, and 8 is Nonverbal. Green shaded cell marks the greatest total for the highest score given in each rubric criterion among the three raters, red shaded cell marks the greatest total for the lowest score given in each rubric criterion among the three raters. Note: Where tallies equal, the next cell level down is reviewed for shading.

When viewed across all groups, it appears the problem may be a universal interpretation of the rubric measurement of success levels (cell levels). For example, Rater 9 recorded 17 0-level scores while Rater 2 scored none, and Rater 8 scored two. Group A recorded just one 0-level from each rater, and Group C raters scored four, eight, and eight, respectively. Results are similar when comparing the highest rubric cell, level 4, where the raters tallies in the groups are Group A: 16-15-2, Group B: 40-33-7, and Group C: 24-12-2. A more complete comparison of overall scores among all three groups can be found in Figure 4.

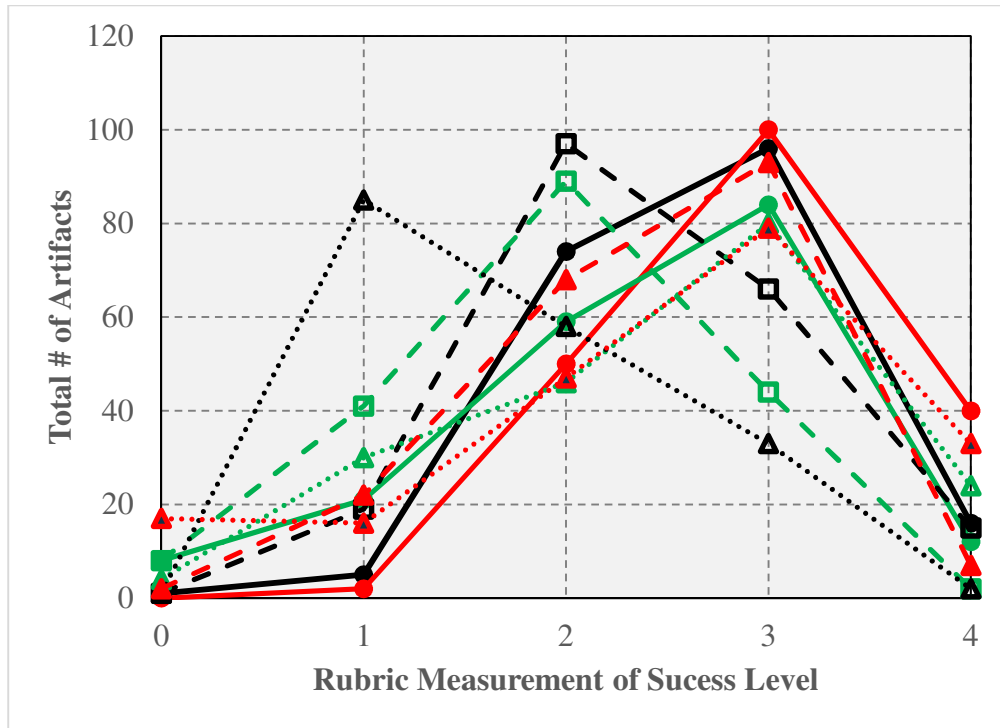


Figure 4. Histogram of total number of artifacts scored in rubric levels for each rater. Black denotes Group A (solid is Rater 1, dash is Rater 4, dot is Rater 7), red denotes Group B (solid is Rater 2, dash is Rater 8, dot is Rater 9), and green denotes Group C (Solid is Rater 3, dash is Rater 5, dot is Rater 6).

Figure 4 depicts the disparity between raters across the three norming groups. Group A (black) indicates each of the three raters has a different mode grade (location of the peak of the curve). In other words, each rater in Group A, Raters 1, 4, and 7, tend to favor scores of 3, 2, and 1, respectively. Group B (red) indicates each of the three raters has a different kurtosis (steepness of the peak), meaning while they all tend to favor scores of 3, the extent of their tendency varies. And finally, Group C (green) shows fairly good agreement between Rater 3 and 6, but Rater 5 has a different mode (favors Level 2 scores, rather than 3).

2.2.2 Descriptive Statistics & SLO Achievement

2.2.2.1 SPC1017

During the Fall 2013 semester, 71 total artifacts were recorded for SPC 1017. All 71 artifacts were non-Dual Enrollment (nonDE), all were Traditional (TD) students as opposed to Online (OnL). Also, all SPC1017 artifacts were used as part of the norming process described above. For course-level assessment, random raters were selected for use in the study and duplicates of those raters not selected were removed. Basic descriptive statistics of SPC 1017 artifacts are shown in Table 4.

The average overall score for the 71 artifacts is 21.1/32. Rubric criterion Choice of Words exhibits the highest average score of 2.93/4, while Conclusion is the lowest at 2.45/4.

N = 71	Introduction	Organization	Supporting Materials	Oral Citations	Conclusion	Choice of Words (Language and Grammar)	Vocal Expression and Paralinguage	Nonverbal	Total Score
mean	2.70	2.58	2.66	2.52	2.45	2.93	2.54	2.72	21.10
median	3	3	3	3	2	3	3	3	21
mode	3	3	3	3	2	3	3	2	20
standard deviation	0.83	0.84	0.67	1.21	0.92	0.52	0.63	0.80	4.30
Kurtosis	-0.35	-0.50	0.45	-0.26	0.20	0.85	-0.15	-0.79	0.45
Skewness	-0.30	-0.10	-0.62	-0.71	-0.13	-0.11	-0.31	0.21	-0.18
SLO goal (≥ 2)	65%	64%	67%	58%	64%	71%	68%	69%	n/a

Table 4. Basic descriptive statistics of Fall 2013 SPC1017 artifacts (71 samples). Percentage of artifacts above SLO goal (≥ 2) is listed for all rubric criteria although areas of interest outlined by faculty only include Introduction, Organization, Oral Citation, Nonverbal, and Conclusion. Shaded cells indicate SLO goal was achieved.

Table 4 also includes statistics on the achievement of faculty established SLOs. The faculty established measure of success for SLO1 was a rating of “Developing” or higher (≥ 2) in rubric criteria Introduction for 60% of the students. Fall 2013 artifacts successfully meet the goal as results indicate 65% of artifacts scored level 2 or greater. For SLO2, a rating of “Developing” or higher (≥ 2) in rubric criteria Organization for 70% of the students, Fall 2013 artifacts did not successfully meet the goal as results indicate 64% of artifacts scored level 2 or greater. For SLO3, a rating of “Developing” or higher (≥ 2) in rubric criteria Oral Citation for 70% of the students, Fall 2013 artifacts did not successfully meet the goal as results indicate 58% of artifacts scored level 2 or greater. For SLO4, a rating of “Developing” or higher (≥ 2) in rubric criteria Conclusion for 60% of the students, Fall 2013 artifacts successfully met the goal as results indicate 64% of artifacts scored level 2 or greater. For SLO5, a rating of “Developing” or higher (≥ 2) for 70% of the students in rubric criteria Nonverbal, Fall 2013 artifacts did not successfully meet the goal as results indicate 69% of artifacts scored level 2 or greater. For additional details on distribution of artifacts across rubric cell levels see Section 2.3.

2.2.2.2 SPC2023

During the Fall 2013 semester, 24 total artifacts were recorded for SPC2023. All 24 artifacts were nonDE and Trad students. Also, all SCP2023 artifacts were used as part of the norming process described above. For course-level assessment, random raters were selected for use in the study and duplicates of those raters not selected were removed.

During the Spring 2014 semester, 172 total artifacts were recorded for SPC2023. Of those, 141 were nonDE while 31 were Dual Enrollment (DE). Additionally, 158 artifacts were TD compared with 9 OnL. Basic descriptive statistics of SPC2023 for Fall 2013 and Spring 2014 artifacts are shown in Table 5. Descriptive statistics of SPC2023 binned according to nonDE/DE and TD/OnL are shown in Table 6.

The average overall score for the 24 artifacts in Fall 2013 was 25.5/32, while the average for the 172 artifacts in Spring 2014 was 27.12/32. Seven of eight rubric criteria show increases in means, although not all are significant (see Section 2.2.3 for details). In most cases, the Spring 2014 artifacts reflect an increase in standard deviation (spread of data distribution), kurtosis (steepness of data peak when graphed). Additionally, all criteria exhibited a more negatively skewed distribution in Spring 2014 data compared with Fall 2013.

Fall 2013: N = 24 Spring 2014: N = 172		Introduction	Organization	Supporting Materials	Oral Citations	Conclusion	Choice of Words (Language and Grammar)	Vocal Expression and Paralanguage	Nonverbal	Total Score
Fall 2013	mean	3.54	3.33	3.25	3.79	2.96	3.17	2.54	2.92	25.50
	median	4	3	3	4	3	3	2.5	3	26
	mode	4	4	3	4	3	3	2	3	26
	standard deviation	0.66	0.70	0.53	0.51	0.62	0.48	0.59	0.72	3.48
	Kurtosis	0.35	-0.70	-0.10	6.26	-0.11	1.06	-0.59	1.22	1.10
	Skewness	-1.16	-0.58	0.24	-2.54	0.02	0.52	0.53	-0.65	-0.79
	SLO goal (% ≥ 2)	100%	100%	100%	100%	100%	100%	100%	96%	n/a
Spring 2014	mean	3.67	3.49	3.45	3.41	3.22	3.48	3.28	3.18	27.12
	median	4	4	4	4	4	4	3	3	28
	mode	4	4	4	4	4	4	4	4	32
	standard deviation	0.52	0.69	0.75	1.00	1.02	0.60	0.83	0.95	4.12
	Kurtosis	0.56	1.01	1.15	3.00	1.82	0.63	0.08	0.52	0.35
	Skewness	-1.25	-1.21	-1.28	-1.84	-1.45	-0.84	-0.94	-1.03	-0.87
	SLO goal (% ≥ 2)	100%	99%	98%	95%	93%	99%	96%	94%	

Table 5. Basic descriptive statistics of SPC2023 artifacts for Fall 2013 (24 samples), and Spring 2014 (172 samples). Measured increases from Fall 2013 to Spring 2014 denoted with shaded cell. Percentage of artifacts above SLO goal (≥ 2) is listed for all rubric criteria although areas of interest outlined by faculty only include Introduction, Organization, Oral Citation, Nonverbal, and Conclusion.

The increased standard deviation may simply be a result of a larger sample size reflecting greater variance than captured in the small Fall 2013 sample size. An increased kurtosis is indicative of an increased tendency of artifacts to fall into the same rubric level. More negatively skewed data distribution means that artifacts are tending more towards higher scores with a tail towards lower rubric levels (Figure 5).

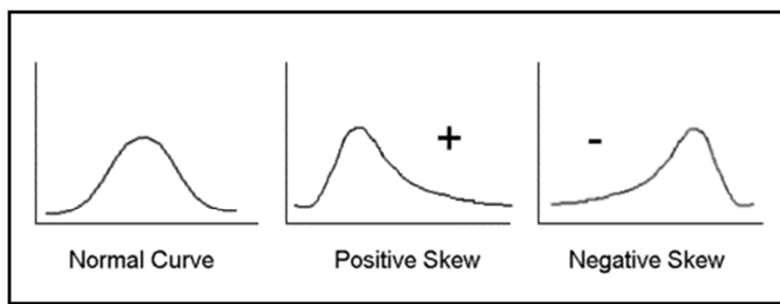


Figure 5. Example of skewness. The normal curve (left) has a skewness of 0.0. A positive value skewness (center) and negative value skewness (right) depict an ideal scenario (Starkweather, 2010).

Table 5 also includes statistics on the achievement of faculty established SLOs. The faculty established measure of success for SLO1 was a rating of “Developing” or higher (≥ 2) in rubric criteria Introduction for 60% of the students. Both Fall 2013 and Spring 2014 saw 100% of artifacts scored level 2 or greater. For SLO2, a rating of “Developing” or higher (≥ 2) in rubric criteria Organization for 70% of the students, results are 100% and 99%, for Fall 2013 and Spring 2014, respectively. For SLO3, a rating of “Developing” or higher (≥ 2) in rubric criteria Oral Citation for 70% of the students, results are 100% and 95%, for Fall 2013 and Spring 2014, respectively. For SLO4, a rating of “Developing” or higher (≥ 2) in rubric criteria Conclusion for 60% of the students, results are 100% and 93%, for Fall 2013 and Spring 2014, respectively. For SLO5, a rating of “Developing” or higher (≥ 2) for 70% of the students in rubric criteria

Nonverbal, results are 100% and 98%, for Fall 2013 and Spring 2014, respectively. All faculty established SLOs were met. For additional details on distribution of artifacts across rubric cell levels see Section 2.3.

The average overall score for DE students (26.99) was lower than that of nonDE (27.71) (see Section 2.2.3 for details on significance) (Table 6). Seven of eight rubric criteria show increases in means, although again, see 2.2.3 for significance. The average overall score for OnL students (27.00) was slightly lower than that of TD students (27.12). The high mean scores for each criteria are evenly split between the two groups.

nonDE: N = 141 DE: N = 31 TD: N = 158 OnL: N = 9		Introduction	Organization	Supporting Materials	Oral Citations	Conclusion	Choice of Words (Language and Grammar)	Vocal Expression and Paralanguage	Nonverbal	Total Score
nonDE / DE	mean	3.67 / 3.71	3.46 / 3.61	3.44 / 3.48	3.38 / 3.58	3.22 / 3.19	3.47 / 3.52	3.26 / 3.35	3.16 / 3.26	26.99 / 27.71
	median	4 / 4	4 / 4	4 / 4	4 / 4	4 / 4	4 / 4	3.5 / 3	3 / 3	28 / 29
	mode	4 / 4	4 / 4	4 / 4	4 / 4	4 / 4	4 / 4	4 / 3	4 / 4	32 / 29
	standard deviation	0.53 / 0.46	0.70 / 0.62	0.75 / 0.77	1.05 / 0.72	1.02 / 1.01	0.61 / 0.57	0.87 / 0.66	0.98 / 0.82	4.24 / 3.50
	Kurtosis	0.69 / -1.13	0.99 / 0.98	0.98 / 2.45	2.68 / 0.65	1.98 / 1.49	0.83 / -0.57	-0.02 / -0.60	0.46 / 0.39	0.30 / -0.11
	Skewness	-1.29 / -0.97	-1.18 / -1.38	-1.23 / -1.59	-1.79 / -1.45	-1.51 / -1.24	-0.88 / -0.64	-0.94 / -0.53	-1.03 / -0.92	-0.85 / -0.80
TD / OnL	mean	3.69 / 3.33	3.47 / 3.78	3.44 / 3.67	3.40 / 3.56	3.20 / 3.44	3.50 / 3.11	3.28 / 3.22	3.19 / 2.89	27.12 / 27.00
	median	4 / 3	4 / 4	4 / 4	4 / 4	4 / 3	4 / 3	4 / 3	3 / 3	28 / 28
	mode	4 / 3	4 / 4	4 / 4	4 / 4	4 / 3	4 / 3	4 / 4	4 / 3	32 / 28
	standard deviation	0.52 / 0.50	0.70 / 0.44	0.77 / 0.50	0.99 / 1.33	1.04 / 0.53	0.59 / 0.60	0.85 / 0.83	0.95 / 0.78	4.21 / 3.35
	Kurtosis	1.11 / -1.71	0.94 / 0.73	0.95 / -1.71	2.71 / 9.00	1.55 / -2.57	0.90 / 1.13	0.12 / -1.28	0.68 / -1.04	0.30 / -0.09
	Skewness	-1.46 / 0.86	-1.20 / -1.62	-1.24 / -0.86	-1.75 / -3.00	-1.41 / 0.27	-0.96 / 0.02	-0.95 / -0.50	-1.08 / 0.22	-0.88 / -0.38

Table 6. Basic descriptive statistics of SPC2023 artifacts for Spring 2014 with respect to nonDE vs. DE students (N=141, N=31, respectively) and TD vs. OnL students (N=158, N=9, respectively). Higher values for DE over nonDE and OnL over TD denoted with shaded cell.

2.2.3 Significance Testing

2.2.3.1 SPC1017

No significance testing can be conducted on SPC1017 artifacts as all were nonDE, all were Trad, and no sections were offered for Spring 2014 with which to compare.

2.2.3.2 SPC2023

Study goals demanded significance tests be conducted to determine whether the difference in the means of nonDE to DE, TD to OnL, and Fall 2013 to Spring 2014 is solely due to chance. Each rubric criterion and the overall score was tested for significance using a Welch's t-test according to standard methods (Davis, 1973; McDonald, 2009; Wilkinson, 1999). The results of significance testing for each dimension are shown in Table 7. Additional details of the distribution of the results are explored in

subsequent sections to provide further information into the variations between dataset relationships as foundation for potential future causal studies, if necessary.

nonDE / DE: df = 49 TD / OnL: df = 9 F '13 / Sp '14: df = 27		Introduction	Organization	Supporting Materials	Oral Citations	Conclusion	Choice of Words (Language and Grammar)	Vocal Expression and Paralanguage	Nonverbal	Total Score
Spring 2014 nonDE /DE	mean	-0.04	-0.15	-0.04	-0.20	0.03	-0.04	-0.09	-0.10	-0.72
	t _{crit}	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01	2.01
	t _{obs}	0.46	1.21	0.29	1.31	-0.13	0.39	0.65	0.60	1.00
	p-value	0.650	0.231	0.773	0.195	0.897	0.698	0.519	0.551	0.322
Spring 2014 TD / OnL	mean	-0.36	0.31	0.23	0.15	0.24	-0.39	-0.06	-0.30	-0.12
	t _{crit}	2.26	2.26	2.26	2.26	2.26	2.26	2.26	2.26	2.26
	t _{obs}	-2.10	1.94	1.30	0.33	1.25	-1.89	-0.22	-1.11	-0.11
	p-value	0.065	0.079	0.221	0.746	0.236	0.092	0.834	0.292	0.918
F '13 / Sp '14	mean	0.13	0.16	0.20	-0.38	0.26	0.31	0.74	0.26	1.62
	t _{crit}	2.05	2.05	2.05	2.05	2.05	2.05	2.05	2.05	2.05
	t _{obs}	0.95	1.02	1.61	-2.94	1.72	2.89	5.43	1.58	2.08
	p-value	0.351	0.318	0.116	0.005	0.092	0.007	3.54x10 ⁻⁶	0.122	0.045

Table 7. Significance test for nonDE vs. DE, TD vs. OnL, and Fall 2013 vs. Spring 2014. Positive mean scores indicate DE > nonDE, OnL > TD, and Spring 2014 > Fall 2013, respectively.

The Welch's t-test results indicate that when comparing nonDE to DE students of the Spring 2014 semester there is no significant difference in any rubric criterion or the overall score. That is to say, we cannot reject the null hypothesis that the difference in the means of the nonDE and DE artifacts are equal to 0, and we cannot rule out the possibility that the differences in scores are not solely due to chance.

The Welch's t-test results indicate that when comparing TD to OnL students of the Spring 2014 semester there is no significant difference in any rubric criterion or the overall score. That is to say, we cannot reject the null hypothesis that the difference in the means of the nonDE and DE artifacts are equal to 0, and we cannot rule out the possibility that the differences in scores are not solely due to chance.

The Welch's t-test results of the Fall 2013 to Spring 2014 students indicate that for the Oral Citations and Vocal Expression rubric criterion as well as for the overall score we must reject the null hypothesis that the difference in the means of the two semesters' artifacts equal to 0, and we can conclude this with a 95% confidence that the differences in scores are not solely due to chance. For the remaining rubric criteria we cannot reject the null hypothesis, meaning the differences in mean scores for those artifacts can be a result of chance.

The overall mean score difference, which was statistically significant, exhibited an increase of 1.62 points (on a 32-point scale). This is an increase of 5.1%. Of the two rubric criteria which exhibited a statistically significant difference in mean scores, only one showed an increase, Vocal Expression, with an increase of 0.74 points (on a 4-point scale). The other, Oral Citations, exhibited a decline of 0.38 points (on a 4-point scale).

2.3 SUPPORTING STATISTICAL ANALYSES

2.3.1 SPC 1017

Since significance tests only provide information on the rejection of a null hypothesis and not on specific details of the changes mean score groups, it is necessary that exploratory analyses be performed such that further information of value can be extracted if an evaluation of the program methods effects is to be quantitatively understood. Therefore, each rubric criteria was rigorously analyzed using multiple standard processes for support of significance testing in order to most effectively apply the results toward instructive improvement, therefore allowing assessment to drive instruction as defined by Elder and Paul (2007).

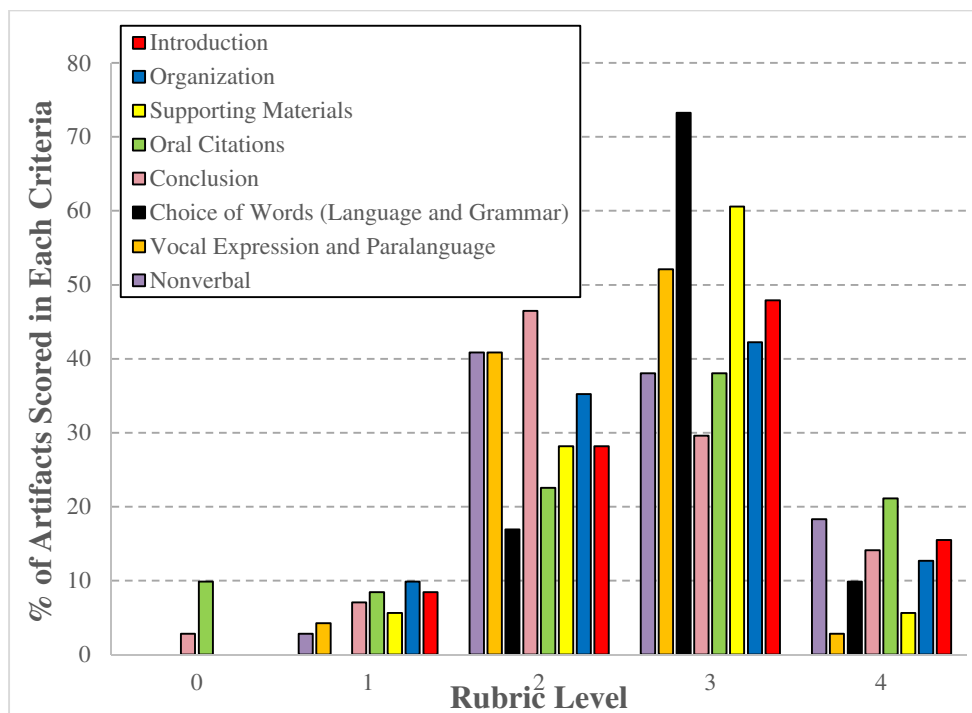


Figure 6. Histogram of Fall 2013 SPC1017 data distribution across rubric cell levels by rubric criteria.

Figure 6 depicts the distribution of scores based on rubric criteria. The Conclusion and Nonverbal criteria are the only two criteria in which rubric level 2 exhibits the largest proportion of artifacts with 46% and 41% of the total number of artifacts, respectively. In all other criteria rubric level 3 exhibits the largest proportion of artifacts. Both Conclusion and Nonverbal, however, exhibited disagreement among raters in the norming session with a range of mean scores of approximately 1.5 (on a 4-point scale) for Conclusion among Groups A and C, and 1.5 for Nonverbal among Group A. As a result, based on this data, an interpretation cannot be made with certainty as to the cause of the lower scores.

2.3.2 SPC 2023

Figure 7 depicts the distribution of scores based on rubric criteria for Fall 2013. Oral Citations exhibits the highest percentage of artifacts scored “Exemplary”, or rubric level 4, with 83%. Introduction is second highest with 63%. Fifty percent of artifacts scored rubric level 3 or better across all rubric levels. Vocal Expression is the lowest of the criteria, with just 50% greater than or equal to 3. Vocal Expression is also the only criteria in which rubric level 2 exhibits the largest proportion of artifacts with 50% of

artifacts scored a 2. Vocal Expression exhibited disagreement across all norming groups with a range of mean scores never less than 0.77. As a result, based on this data, an interpretation cannot be made with certainty as to the cause of the lower scores in this criterion.

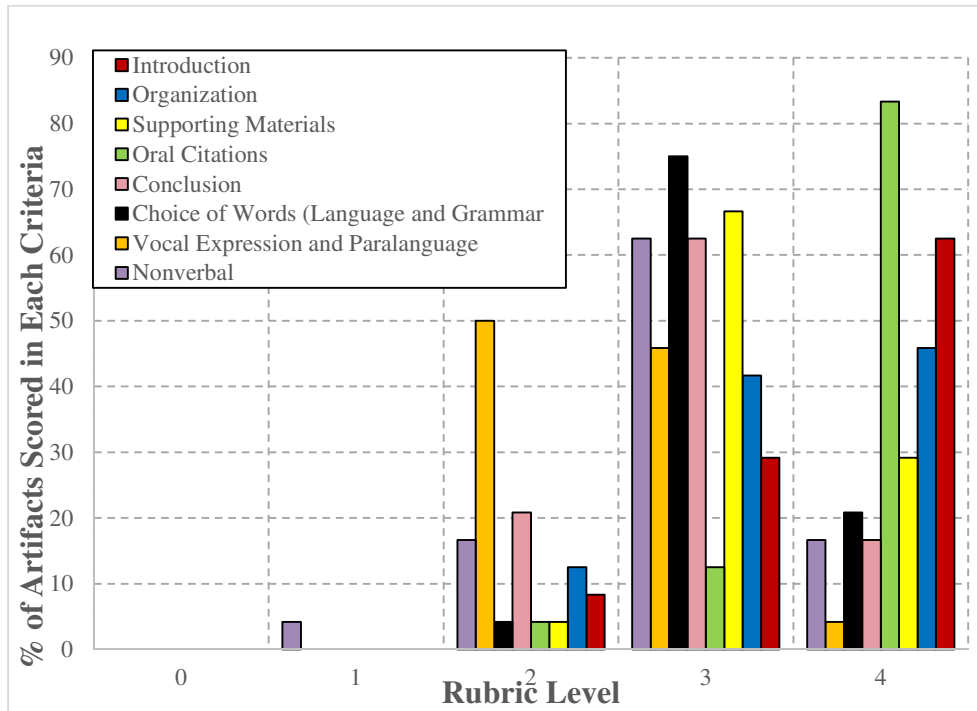


Figure 7. Histogram of Fall 2013 SPC2023 data distribution across rubric cell levels by rubric criteria.

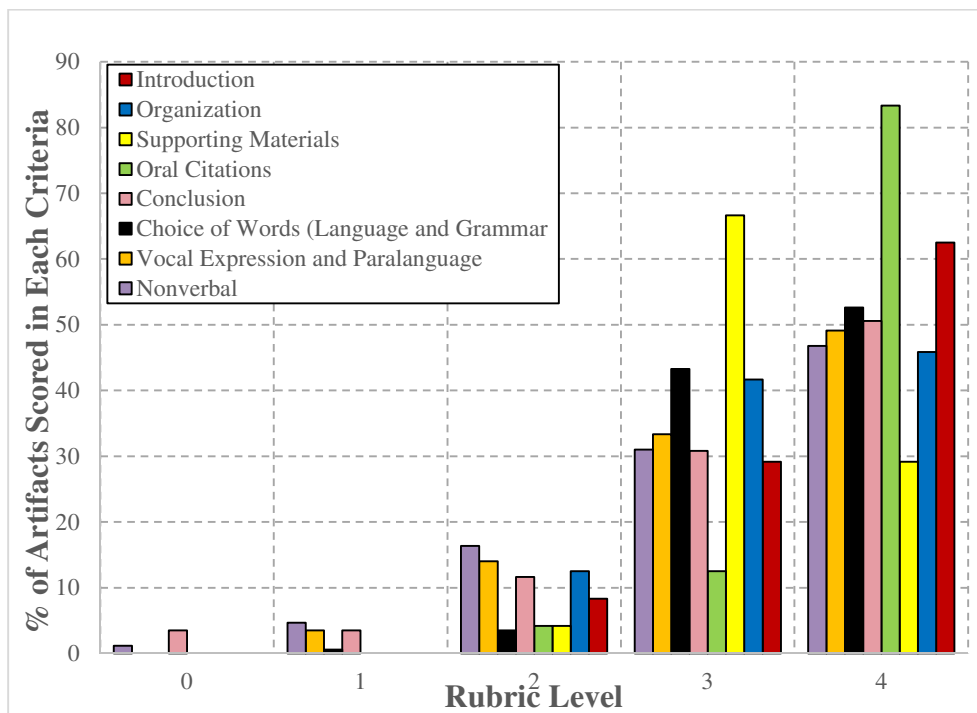


Figure 8. Histogram of Spring 2014 SPC2023 data distribution across rubric cell levels by rubric criteria.

Figure 8 depicts the distribution of scores based on rubric criteria for Spring 2014. Vocal Expression mean score improvement of 0.74 points (on a 4-point scale) is evident in the marked decrease in level 2 scores and shift towards in level 3 and 4. A similar shift from level 3 to level 4 is evident in both Choice of Words and Nonverbal. Seventy-five percent of artifacts scored rubric level 3 or better across all rubric levels.

3 CONCLUSIONS

The Florida SouthWestern's Speech Department employed a common rubric to evaluate an agreed upon series of five SLOs as the focus of the QEP-driven course level assessment plan. For SPC1017 in Fall 2013, 2 of 5 goals were met (SLO1, a rating of "Developing" or higher (≥ 2) in rubric criteria Introduction for 60% of the students; a rating of "Developing" or higher (≥ 2) in rubric criteria Conclusion for 60% of the students). For SPC2023, both Fall 2013 and Spring 2014, all five SLOs were met.

A norming session was completed to assess variation among scorers using common criteria. The results support an interpretation that there may be a problem with agreement between faculty interpretation of the rubric measurement of success levels (cell levels). For example, Rater 9 recorded 17 0-level scores while Rater 2 scored 0, and Rater 8 scored 2 on the same set of 23 artifacts. Results are similar when comparing the highest rubric cell, level 4, where three raters grading the same 23 artifacts record 40, 33, and 7 exemplary scores. These and other examples indicate that either the definition for, or use of the rubric must be further honed.

Significance testing was completed on the difference in means across the eight rubric criteria for a variety of categories: nonDE vs. DE, TD vs. OnL, and semester vs. semester. There were no DE or OnL students to respectively compare in SPC1017, nor were any courses offered in Spring 2014 with which to compare with Fall 2013 data. However, for SPC2023, there exists a significant difference in mean score from Fall 2013 to Spring 2014 in Oral Citation, Vocal Expression, and overall mean score. Of those, Oral Citation exhibited a statistically significant decrease from Fall 2013 to Spring 2014 while the others exhibited increases. There were no statistically significant changes in the mean from nonDE to DE, or from TD to OnL.

4 REFERENCES

- Cole, R., Haimson, J., Perez-Johnson, I., and May, H. 2011. Variability in Pretest-Posttest Correlation Coefficients by Student Achievement Level. NCEE Reference Report 2011-4033. Washington, DC: National Center for Education Evaluation and Regional Assistance, U.S. Department of Education.
- Davis, J.C. 1973. Statistics and Data Analysis in Geology. John Wiley & Sons, New York, New York, 564 pp.
- Elder, L, and Paul, R. 2007. Consequential Validity: Using Assessment to Drive Instruction. In: Foundation For Critical Thinking. Retrieved from <http://www.criticalthinking.org/pages/consequential-validity-using-assessment-to-drive-instruction/790>.
- Hall, J.C. 2010. Doing More with Less: Making Colleges Work Better. Springer, New York, New York, 250 pp.

- McDonald, J.H. 2009. Handbook of Biological Statistics (2nd ed.). Sparky House Publishing, Baltimore, Maryland.
- Starkweather, J. D. 2010. Introduction to Statistics for the Social Sciences. In: Research and Statistical Support. Retrieved from http://www.unt.edu/rss/class/Jon/ISSS_SC/.
- Wilkinson, L. 1999. APA Task Force on Statistical Inference. Statistical Methods in Psychology Journals: Guidelines and Explanations. American Psychologist 54 (8), 594–604.