ENC0022, 1101, 1102 Assessment Report – Fall 2014

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

1 INTRODUCTION

The English Department of Florida SouthWestern State College (FSW) outlined an initial plan for assessment in three courses: English for College Success (ENC0022), Composition I (ENC1101), and Composition II (ENC1102). In each course instructors use a common rubric with seven identified rubric dimensions in the case of ENC0022, and five dimensions for both ENC1101 and ENC1102. This assessment plan is designed to evaluate each course and inform faculty upon establishing Student Learning Objectives (SLOs) for future assessment plans.

In addition to establishing a baseline for future assessment, the common rubric assessment plan provides information on success rates of Dual Enrollment (DE) students compared with non-Dual Enrollment (nonDE) students, as well as Online (OnL) and Tradition (TD) students as highlighted in the QEP course level assessment plan. These correlative measures in conjunction with a multitude of others can be better understood through assessment and provide support toward instructive improvement (Cole et al., 2011; Elder and Paul, 2007).

For additional detail or further analysis not provided in this report, please contact Dr. Joseph F. van Gaalen, Coordinator of Academic Assessment, Academic Affairs (jfvangaalen@fsw.edu; x6965).

2 DESCRIPTIVE STATISTICS & LEARNING OBJECTIVES

2.1 ENC0022

During the Fall 2014 semester, 193 total artifacts were recorded for ENC0022. No Dual Enrollment (DE) students enrolled in the course. Additionally, all were Traditional students (TD) students with no Online (OnL) students enrolled. Of the 193 artifacts, 19 were enrolled in Mini-term semesters (A/B term), while 177 were enrolled in the full term.

ENC0022 is scored using a rubric with seven dimensions, each scored on a scale of 1 to 4 (1-Unacceptable, 2-Needs work, 3-Average, 4-Above average), with 0s if the baseline is not met. The mean overall score for the 193 artifacts is 19.3/28, or 69.0% (Table 1). The scoring of rubric dimensions is fairly evenly distributed with means for all areas between 2.5 and 2.9. Grammar and Mechanics rubric dimensions exhibit the lowest mean score (mean=2.6 and 2.5, respectively) as well as a very low percentage of artifacts were scored a 4 (just 9.8% and 6.2% compared with 20% or higher in all other dimensions) (Figure 1).

In all seven rubric dimensions, greater than 90% of artifacts were scored at level 2 or higher. In four of seven rubric dimensions (Introductory Paragraph, Support Paragraphs, Organization, and Concluding Paragraph), greater than 60% of artifacts were scored at level 3 or higher (Table 1, Figure 1). Those dimensions that did not (Grammar, Mechanics, and Research) were very near 60% at 54.4%, 52.3%, and 53.9%, respectively.

	Introductory Paragraph	Support Paragraphs	Organization	Concluding Paragraph	Grammar	Mechanics	Research
mean	2.8	2.9	2.9	2.8	2.6	2.5	2.8
standard deviation	0.86	0.78	0.78	0.83	0.75	0.70	0.96
Rubric Dimension	**%**	**%**	**%**	**%**	**%**	**%**	**%**
4	25.4	22.3	25.9	20.7	9.8	6.2	30.1
3	37.8	43.0	43.0	43.5	44.6	46.1	23.8
2	32.1	32.6	29.5	30.6	39.9	42.0	39.4
1	4.7	2.1	1.6	5.2	5.7	5.7	6.7

Table 1. Basic descriptive statistics of Fall 2014 ENC0022 artifacts. Rubric dimensions are also shown with distribution of artifacts by rubric achievement level.

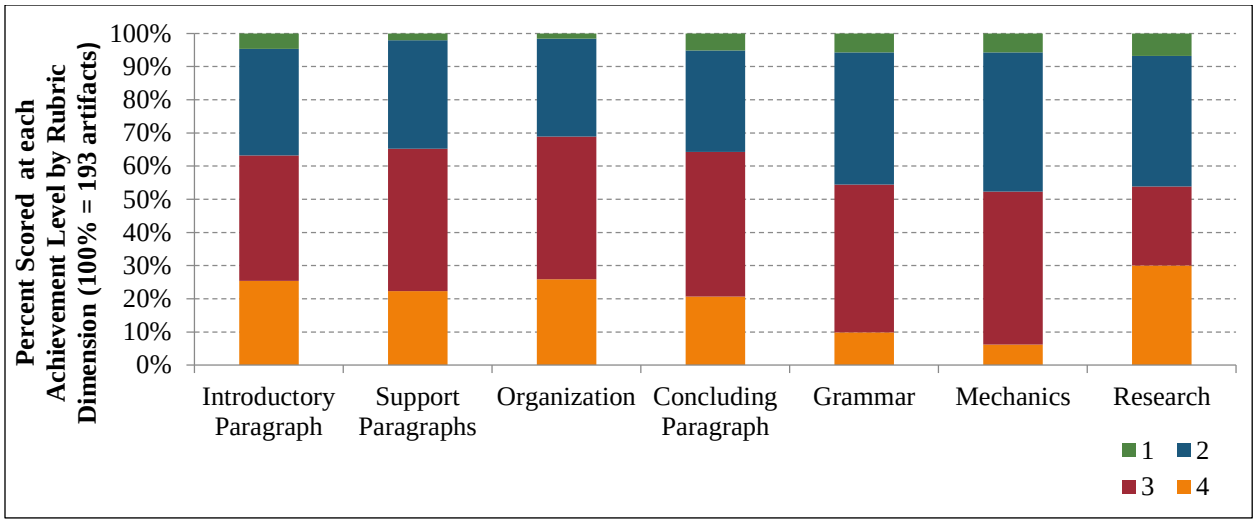

Figure 1. ENC0022 distribution of rubric scores by dimension.

2.2 ENC1101

During the Fall 2014 semester, 727 artifacts were sampled in a cluster convenience format from an enrolled ENC1101 population of 3490 (20.8% sample size). Each course section served as a cluster that was selected randomly under the confines of the convenience element which was defined to provide one section from each instructor's course load. The minimum sample size required to obtain a power of 0.70 and effect size of 0.2, conditions attainable given the target artifact size, was 620. The convenience cluster sampling exceeded this minimum condition for the target artifacts.

From the 727 samples, 236 artifacts were from DE students (32%) while the remainder, 488, were nonDE students (68%). Additionally, 56 artifacts were from OnL students (8%) while the remainder, 668, were from traditional students (92%). Finally, 116 (16%) were enrolled in mini-term (8-week) semesters, while 453 (84%) were enrolled in the full term.

ENC1101 is scored using a rubric with five dimensions, each scored on a scale of 1 to 4 (1-Does not meet standards, 2-Approaching standards, 3-Meets standards, 4-Exceeds standards), with 0s if the benchmark is not met. The mean overall score for the 727 samples is 15.1/20, or 75.7% (Table 1). The scoring of rubric dimensions is fairly evenly distributed with means for all areas between 2.9 and 3.2. The Grammar/Mechanics and Documentation rubric dimensions have the lowest scoring (means both of 2.9). The Grammar/Mechanics dimension has the lowest percentage of artifacts scored a 4, 25.1%, while the Documentation rubric dimension has the highest percentage of artifacts scored at 1, 6.7% (Figure 1).

In all five rubric dimensions, greater than 70% of artifacts were scored at achievement level (3s) or higher (Table 2, Figure 2). From these results, it is clear that the weakness in Grammar/Mechanics is apparent in the highest rubric level (4), but is negligibly different from other dimensions at lower achievement levels.

The Grammar/Mechanics dimension exhibits no significant difference in percentage of artifacts achieving 1s and 2s. The higher percentage of 3s is presumably caused by the increased number of students unable to achieve 4s. In other words, students achieve at level 4 less frequently in Grammar/Mechanics than other dimensions but achieve at level 3 or greater no different than any other dimension.

	Thesis	Evidence	Organization / Style	Grammar / Mechanics	Documentation
mean	3.2	3.0	3.1	2.9	2.9
standard deviation	0.92	0.90	0.85	0.83	1.02
Rubric Dimension	**%**	**%**	**%**	**%**	**%**
4	46.2	36.9	35.3	25.1	33.5
3	30.7	36.3	41.2	48.8	35.8
2	18.2	21.9	19.8	20.7	21.2
1	4.4	4.0	2.6	4.7	6.7

Table 2. Basic descriptive statistics of Fall 2014 ENC1101 artifacts. Rubric dimensions are also shown with distribution of artifacts by rubric achievement level.

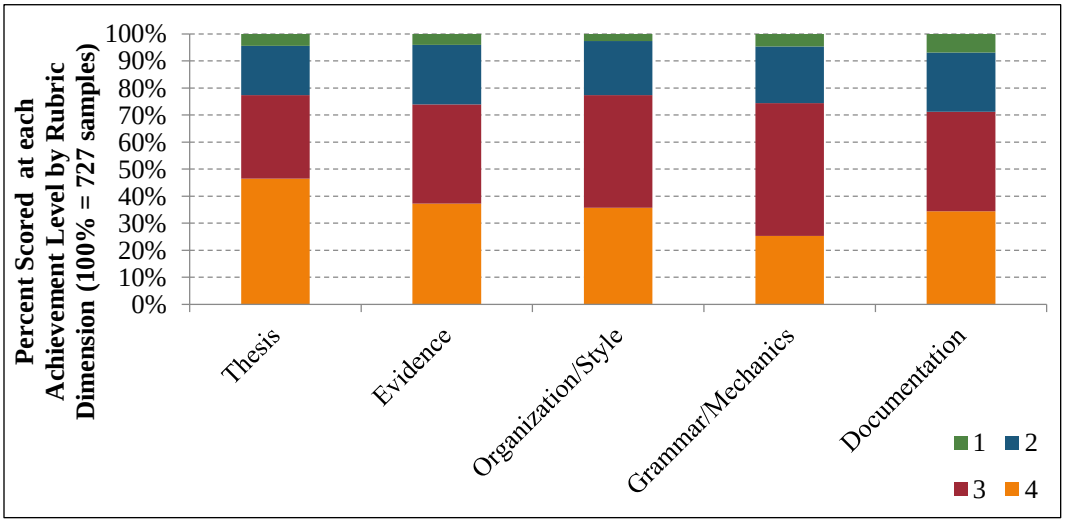

Figure 2. ENC1101 distribution of rubric scores by dimension.

2.3 ENC1102

During the Fall 2014 semester, 270 artifacts were collected in a convenience cluster format from an enrolled ENC1102 population of 1309 (20.6% sample size). Each course section served as a cluster that was selected randomly under the confines of the convenience element which was defined to provide one section from each instructor's course load. The minimum sample size required to obtain a power of 0.95 and effect size of 0.5 was 210. The convenience cluster sampling exceeded this minimum condition for the target artifacts.

From the 270 samples, 31 artifacts were from DE students (11%) while the remainder, 239, were nonDE students (89%). Additionally, 44 artifacts were from OnL students (16%) while the remainder, 226, were from traditional students (84%). Finally, 13 (5%) were enrolled in mini-term (8-week) semesters, while 257 (95%) were enrolled in the full term.

ENC1102 is scored using a rubric with five dimensions, each scored on a scale of 1 to 4 (1-Does not meet standards, 2-Approaching standards, 3-Meets standards, 4-Exceeds standards), with 0s if the benchmark is not met. The mean overall score for the 270 samples is 14.9/20, or 74.4% (Table 1). Like the results of the ENC1101 study, the scoring of rubric dimensions is fairly evenly distributed with means for all areas between 2.9 and 3.2. The Grammar/Mechanics and Documentation rubric dimensions again have the lowest scoring (means both of 2.9). The Grammar/Mechanics dimension again has the lowest percentage of artifacts scored a 4, 20.7%, while the Documentation rubric dimension again has the highest percentage of artifacts scored at 1, 8.1% (Figure 1).

In four of five rubric dimensions (Thesis, Evidence, Organization/Style, and Grammar/Mechanics), greater than 70% of artifacts were scored at level 3 or higher (Table 3, Figure 3). The dimension that did not score greater than 70% at 3 or higher was Documentation (69.3%). The Grammar/Mechanics dimension, while again exhibiting the lowest percentage of level 4 scores, exceeds both the Evidence and Documentation dimensions for level 3.

	Thesis	Evidence	Organization / Style	Grammar / Mechanics	Documentation
mean	3.2	2.9	3.1	2.9	2.9
standard deviation	0.80	0.87	0.77	0.81	0.88
Rubric Dimension	**%**	**%**	**%**	**%**	**%**
4	37.4	27.0	33.0	20.7	24.1
3	45.2	43.0	46.3	50.4	45.2
2	13.3	23.7	18.5	23.0	22.6
1	4.1	6.3	2.2	5.9	8.1

Table 3. Basic descriptive statistics of Fall 2014 ENC1102 artifacts. Rubric dimensions are also shown with distribution of artifacts by rubric achievement level.

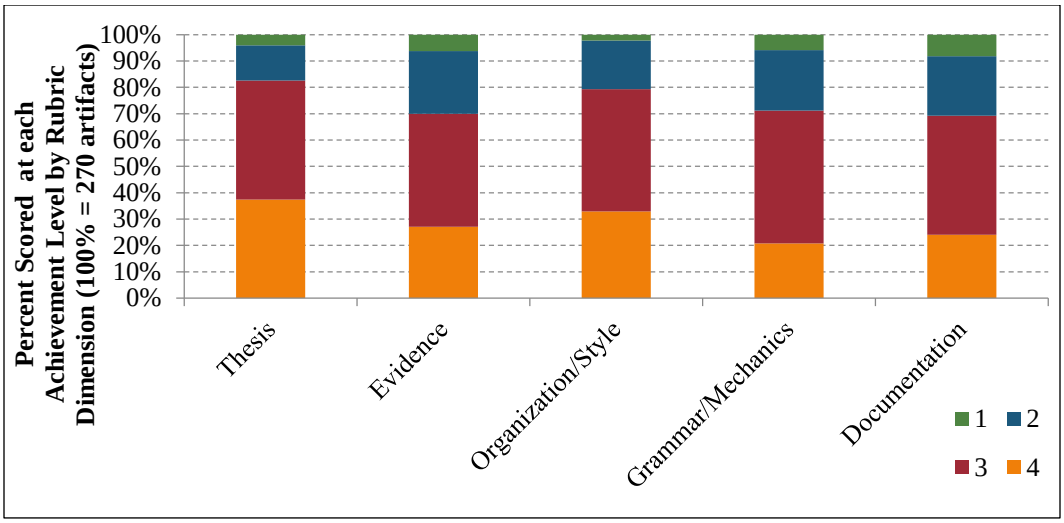

Figure 3. ENC1102 distribution of rubric scores by dimension.

3 EXPLORATORY ANALYSIS & SIGNIFICANCE TESTING

3.1 ENC0022

3.1.1 Comparison by Site, Format, or Student type

3.1.1.1 Full term to Mini-term Comparison

Since no DE students were enrolled in ENC0022, no comparison study between DE and nonDE was completed. Similarly, no online sections of the course were run during Fall 2014, so no comparison study between OnL and TD was completed. However, a small cohort of artifacts originate from a mini-term (8 week) schedule course so a comparison of full term (177 artifacts) and 8-week schedule artifacts (16) was completed.

Each rubric dimension 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 Introductory Paragraph and Research dimensions exhibit a statistically significant difference in mean scores (see Table 4); therefore, we must reject the null hypothesis that the difference in the means of the 8-week and full artifacts are equal to 0, and we can conclude with a 95% confidence that the differences in scores are not solely due to chance. For the remaining rubric dimensions and the overall score, we cannot reject the null hypothesis, meaning the differences in mean scores for those artifacts can reasonably be a result of chance.

Effect size was calculated using the Rosenthal and Rosnow (1991) for meta-analytical purposes to serve as a common thread across institutions (Lipsey and Wilson, 1993). The statistically significant results exhibit the largest effect sizes with the Introductory Paragraph and Research dimensions exhibiting a 0.45 and 0.82 effect sizes, respectively. In terms, this means that for the Introductory Paragraph there is approximately 31% of non-overlap of score distribution from full to 8-week. For the Research dimension, this value is approximately 48%.

	Overall	Introductory Paragraph	Support Paragraphs	Organization	Concluding Paragraph	Grammar	Mechanics	Research
8-week mean	17.6	2.0	2.8	3.1	2.4	2.6	2.8	1.9
full mean	19.5	2.9	2.9	2.9	2.8	2.6	2.5	2.8
Effect size	*0.21*	*0.45*	*0.03*	*-0.15*	*0.24*	*-0.03*	*-0.16*	*0.82*

Table 4. Difference in means for each rubric dimension and combined score (overall) of ENC0022 for 8-week and full. Statistically significant results indicated by shaded cell. Parameters for significance tests are: Introductory Paragraph: t(191)=3.106, p=0.0067, Research: t(191)=5.679, p=8.48x10^{-6}. Positive effect sizes indicate a higher mean score for full artifacts.

3.1.1.2 Comparison of Full-time and Part-time

During the Fall 2014 semester, 105 artifacts originate from courses taught by adjuncts while 88 artifacts originate from courses taught by full-time faculty. A comparison of the means for each rubric dimension and overall score was conducted. Each rubric dimension 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 Support Paragraphs, Research, and overall score exhibit a statistically significant difference in mean scores (see Table 5). Therefore, we must reject the null hypothesis that the differences in the means of the adjunct and full-time artifacts for these two dimensions and the overall score are equal to 0, and we can conclude with a 95% confidence that the differences in scores are not solely due to chance.

Effect size was calculated using the Rosenthal and Rosnow (1991) for meta-analytical purposes to serve as a common thread across institutions (Lipsey and Wilson, 1993). The statistically significant results exhibit what Cohen (1988) would consider medium to large effect sizes ranging from 0.28 to 1.69. In other words, non-overlap from adjunct artifacts to full-time artifacts range from approximately 27% in the case of the overall score to 73% in the case of the Research dimension.

	Overall	Introductory Paragraph	Support Paragraphs	Organization	Concluding Paragraph	Grammar	Mechanics	Research
Adjunct	17.0	3.0	3.0*	3.1	3.1	2.8	2.7	3.3
Full-time	14.7	2.7	2.6*	2.7	2.5	2.3	2.3	2.1
Effect size	*-0.28*	*-0.39*	*-0.50*	*-0.61*	*-0.83*	*-0.63*	*-0.54*	*-1.69*

*Table 5. Difference in means for each rubric dimension and combined score (overall) of ENC0022 for Adjunct and Full-time faculty. Statistically significant results indicated by shaded cell. Parameters for significance tests are: Support Paragraphs: t(191)=3.471, p=0.037, Research: t(191)=11.676, p=1.49x10^{-44}, and Overall: t(191)=2.151, p= p=2.62x10^{34} Positive effect sizes indicate a higher mean score for Full-time faculty artifacts. *Denote marginal significance as defined by Johnson (2013).*

3.1.2 Data distribution

Results from section 2.1 briefly described the distribution in scores among rubric dimensions. The Grammar, Mechanics, and Research dimensions each exhibit somewhat different scoring distributions both with each other and with the other four dimensions in which 60% of artifact scores were 3 or greater. For a clearer representation of these variations, a histogram of the results is presented in Figure 4 below.

From Figure 4 it is clear that both the Grammar and Mechanics dimensions exhibit a substantially lower number of artifacts scored at 4 while the Research dimension exhibits a substantially higher number of artifacts scored at 4 when compared with the four dimensions in which 60% of artifact scores were 3 or greater. In the Grammar and Mechanics dimensions, it is clear that the body of results is simply shifted

lower as there are more level 2 scores for those two dimensions than any other. The Research dimension, however, exhibits a bimodality in which a larger number of students score 2s and 4s but fewer 3s.

A histogram of the upper quartile of possible scores (highest 25% of possible scores) was created in order to provide more information into the varied distributions of the three dimensions mentioned above. The upper quartile distribution, overall scores of 22 through 28, is shown in Figure 5. Note that this is not the upper quartile of scored artifacts, but the upper quartile of possible scores.

The distribution of both the Grammar and Mechanics dimensions for the upper quartile of artifacts behave similarly to each other when compared with other dimensions. Both exhibit lower 4s than other dimensions and higher 3s than other dimensions. Further, both dimensions have no artifacts scored a 1. What can be inferred from this is that according to the assessment, not only are students weaker in these two dimensions but also that most students are at a similar level in both Grammar and Mechanics. In other words, level 3 is reached in Grammar and Mechanics as commonly as all other dimensions. The only difference from these dimensions and others is level 4 achievement.

In the Research dimension, 33/58, or 56% of all 4s come from the upper quartile of artifacts. In contrast, the upper quartile in all other dimensions encompasses nearly all of the 4s scored, ranging from 88% in Organization to 100% in both Grammar and Mechanics.

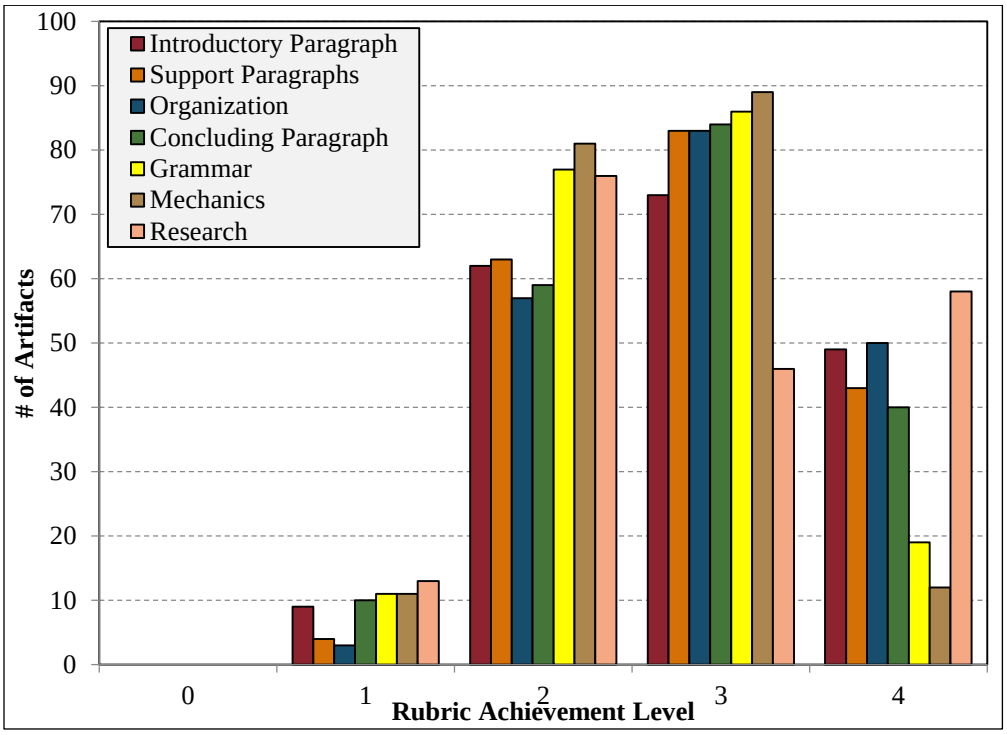

Figure 4. Histogram of Fall 2014 ENC0022 data distribution across achievement levels.

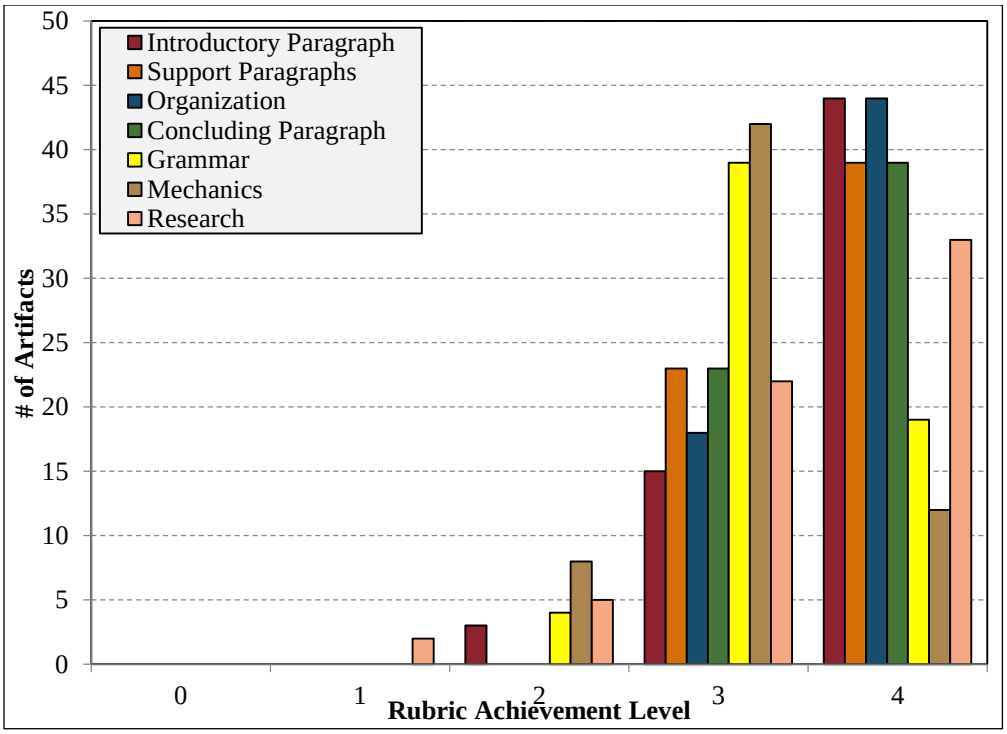

Figure 5. Histogram of Fall 2014 ENC0022 data distribution across achievement levels for upper 25% of possible overall scores (Combined rubric scores of 22-28).

One possible cause of the Research Dimension distribution differences might be in scoring. Figure 6 depicts a box-whisker plot showing the distribution of mean scores of each rubric dimension given by the nine instructors who scored artifacts. For example, to create the first box above "Intro Paragraph," the mean scores given by each of the nine instructors for that rubric dimension are pooled as one. The median of these nine means (one mean score for each instructor) is the red line, which for Intro Paragraph is 3.1. The box represents the 75th percentile and 25th percentile. The vertical lines, or whiskers, represent the full spread of the data excluding outliers. So the distribution of the mean scores given for Introductory Paragraph ranges from 2.0 to 3.9.

Whiskers or boxes that are fairly even both above and below the median (red line) can be interpreted as a normal or pseudo-normal distribution with an actual central tendency somewhere within the box. Whiskers that are uneven above or below the box can be interpreted as a small group of scores behaving somewhat differently from the larger group.

From Figure 6, the following interpretations can be made with reasonable certainty: 1) Rubric dimensions Introductory Paragraph and Research are not well agreed upon across instructors and exhibit a wide range (Introductory Paragraph range of 1.9, Research range of 2.1). 2) Support Paragraphs, Organization, and Research exhibit an uneven distribution of mean scores with respect to the median (red line). 3) Grammar and Mechanics are quite well agreed upon across instructors with a distribution at 1.3 and 1.4, respectively. 4) Median scores (red lines) by dimension give a rough estimate of typical success; the weakest areas, Grammar and Mechanics, are noticeably lower than the other five dimensions.

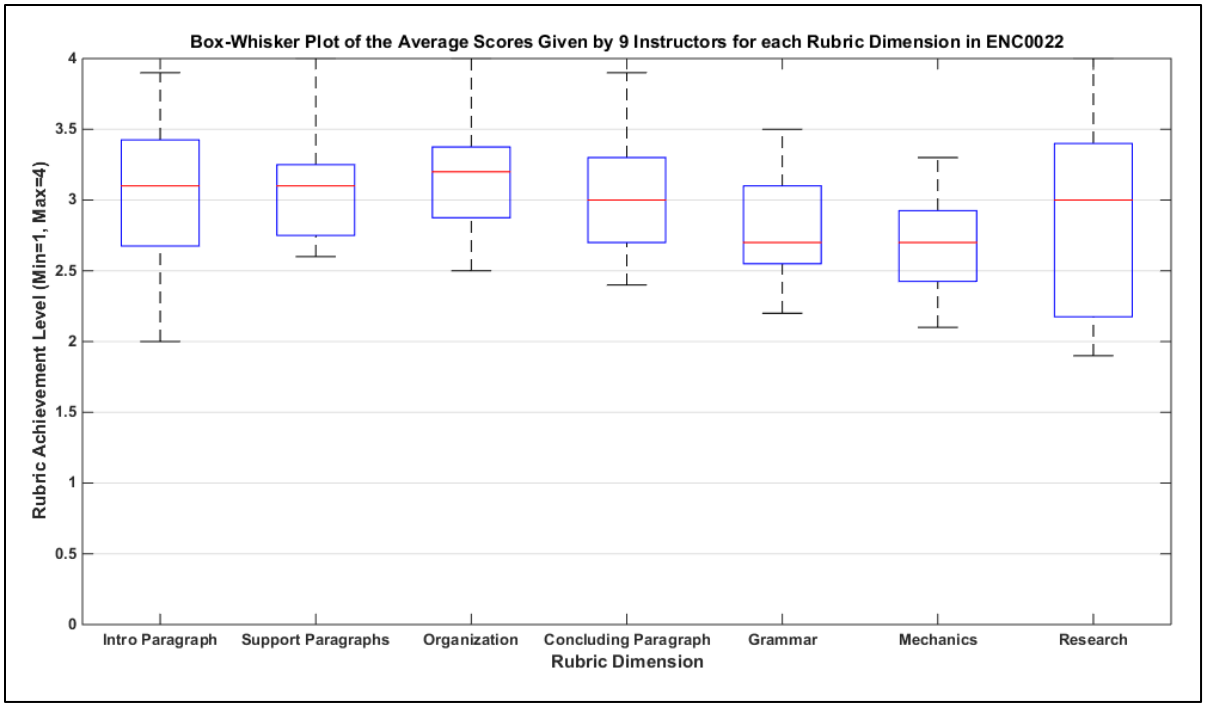

Figure 6. Box-Whisker plot of mean scores given by the 9 instructors scoring papers for ENC0022 for each rubric dimension. Red line depicts median score. Upper and lower box boundaries indicate 75% quartile and 25% quartile (box represents central 50% of the scores). Vertical lines represent remaining scores outside central 50% that are not outliers. Red '+'s denote outliers.

One of the nine instructors had just six artifacts with which their mean score was derived while a second instructor had just 10. With only 9 instructors, this can hinder interpretation. In the cases of Introductory Paragraph and Organization, this hindrance is contributing to the larger range and in the case of Organization, uneven distribution. This means that the dimensions of Support Paragraphs and Research still exhibit uneven distributions which is a result of a bimodal distribution. What it has no effect on, however, is that there is a cohort of mean scores that are clustered near the median (red line) and there is a smaller cohort of scores clustered in the upper quartile whisker causing both the larger range and uneven distribution. This, in turn, explains why artifacts with lower overall scores still attaining 4s in Research at a higher rate than other dimensions. The two cohorts are scoring Research sufficiently differently as to create the anomaly.

3.2 ENC1101

3.2.1 Comparison by Site, Format, or Student type

3.2.1.1 Dual Enrollment to non-Dual Enrollment Comparison

During the Fall 2014 semester, 1318 total Dual Enrollment (DE) students were enrolled in ENC1101 and 2172 non-DE students were enrolled in ENC1101. Of those, 236 DE artifacts were sampled along with 488 nonDE artifacts, for a representative sampling percentage of 17.9% and 22.4%, respectively.

Each rubric dimension and the overall score was tested for significance using a Welch's t-test according to standard methods (Davis, 1973; McDonald, 2009; Wilkinson, 1999). All rubric dimensions including the overall score exhibit a statistically significant difference in mean scores (see Table 6). Therefore, we

must reject the null hypothesis that the differences in the means of the DE and nonDE artifacts are equal to 0, and we can conclude with a 95% confidence that the differences in scores are not solely due to chance.

Effect size was calculated using the Rosenthal and Rosnow (1991) for meta-analytical purposes to serve as a common thread across institutions (Lipsey and Wilson, 1993). The statistically significant results exhibit what Cohen (1988) would consider medium to large effect sizes ranging from 0.43 to 0.69. In other words, non-overlap from DE artifacts to nonDE artifacts range from approximately 28% to 43%.

	Overall	Thesis	Evidence	Organization/Style	Grammar/Mechanics	Documentation
DE mean	16.8	3.5	3.4	3.4	3.3	3.2
nonDE mean	14.5	3.1	2.9	2.9	2.8	2.8
Effect size	*-0.69*	*-0.47*	*-0.60*	*-0.61*	*-0.64*	*-0.43*

Table 6. Difference in means for each rubric dimension and combined score (overall) of ENC1101 for DE and nonDE. Statistically significant results indicated by shaded cell. Parameters for significance tests are: Overall: $t(722)=9.239$, $p=2.46x10^{-20}$, Thesis: $t(722)=6.324$, $p=5.43x10^{-10}$, Evidence: $t(722)=8.015$, $p=6.40x10^{-15}$, Organization/Style: $t(722)=8.139$, $p=2.74x10^{-15}$, Grammar/Mechanics: $t(722)=8.647$, $p=6.05x10^{-17}$, Documentation: $t(722)=5.800$, $p=1.13x10^{-8}$. Positive effect sizes indicate a higher mean score for nonDE artifacts.

In order to determine if differences between DE and non-DE may be associated with site, a separate study was conducted comparing DE artifacts originating from onsite (students attended campus college course) and DE artifacts originating from offsite (conducted in a high school setting at the college level). Each rubric dimension and the overall score was tested for significance using a Welch's t-test according to standard methods (Davis, 1973; McDonald, 2009; Wilkinson, 1999) and there were no significant differences between the means of any rubric dimension or overall score.

	Overall	Thesis	Evidence	Organization/Style	Grammar/Mechanics	Documentation
DE Onsite mean	16.6	3.5	3.4	3.3	3.2	3.2
DE Offsite mean	16.9	3.5	3.4	3.5	3.3	3.2
Effect size	0.12	0.03	0.10	0.17	0.13	0.07

Table 7. Difference in means for each rubric dimension and combined score (overall) of ENC1101 for DE onsite and DE offsite. Statistically significant results indicated by shaded cell. Positive effect sizes indicate a higher mean score for DE offsite artifacts.

3.2.1.2 Online to Traditional Comparison

During the Fall 2014 semester, 350 total Online (OnL) students were enrolled in ENC1101 and 3140 Traditional (TD) students were enrolled in ENC1101. Of those, 56 OnL artifacts were sampled along with 668 TD artifacts, for a representative sampling percentage of 16.0% and 21.3%, respectively.

Each rubric dimension and the overall score was tested for significance using a Welch's t-test according to standard methods (Davis, 1973; McDonald, 2009; Wilkinson, 1999). None of the rubric dimensions including the overall score exhibit a statistically significant difference in mean scores (see Table 8). Therefore, we cannot reject the null hypothesis that the difference in the means of the OnL and TD artifacts are equal to 0, and we cannot rule out the possibility that the differences in scores are not solely due to chance.

Effect size was calculated using the Rosenthal and Rosnow (1991) for meta-analytical purposes to serve as a common thread across institutions (Lipsey and Wilson, 1993). The results exhibit what Cohen (1988) would consider small effect sizes ranging from 0.03 to 0.1. In other words, non-overlap from OnL artifacts to TD artifacts is less than 7% in all cases.

	Overall	Thesis	Evidence	Organization/Style	Grammar/Mechanics	Documentation
OnL mean	15.4	3.1	3.2	3.3	3.0	2.9
TD mean	15.2	3.2	3.1	3.1	2.9	2.9
Effect size	*-0.03*	*0.03*	*-0.06*	*-0.10*	*0.06*	*0.03*

Table 8. Difference in means for each rubric dimension and combined score (overall) of ENC1101 for OnL and TD. Statistically significant results indicated by shaded cell. Positive effect sizes indicate a higher mean score for TD artifacts.

3.2.1.3 Onsite to Offsite Comparison

During the Fall 2014 semester, 502 total Offsite (OFF) students were enrolled in ENC1101 and 2988 Onsite (ON) students were enrolled in ENC1101. Of those, 155 Off artifacts were sampled along with 570 ON artifacts, for a representative sampling percentage of 30.9% and 19.1%, respectively.

Each rubric dimension and the overall score was tested for significance using a Welch's t-test according to standard methods (Davis, 1973; McDonald, 2009; Wilkinson, 1999). All rubric dimensions for OFF including the overall score exhibit a statistically significant positive difference in mean scores (see Table 9). Therefore, we must reject the null hypothesis that the difference in the means of the OFF and ON artifacts are equal to 0, and we can conclude with a 95% confidence that the differences in scores are not solely due to chance.

Effect size was calculated using the Rosenthal and Rosnow (1991) for meta-analytical purposes to serve as a common thread across institutions (Lipsey and Wilson, 1993). The statistically significant results exhibit what Cohen (1988) would consider medium to large effect sizes ranging from 0.36 to 0.47. In other words, non-overlap from OFF artifacts to ON artifacts range from approximately 25% to 30%.

	Overall	Thesis	Evidence	Organization/Style	Grammar/Mechanics	Documentation
OFF mean	16.7	3.5	3.4	3.4	3.3	3.2
ON mean	14.8	3.1	3.0	3.0	2.9	2.8
Effect size	*-0.45*	*-0.36*	*-0.42*	*-0.46*	*-0.47*	*-0.37*

Table 9. Difference in means for each rubric dimension and combined score (overall) of ENC1101 for OFF and ON. Statistically significant results indicated by shaded cell. Parameters for significance tests are: Overall: $t(722)=6.096$, $p=3.02\times10^{-7}$, Thesis: $t(722)=4.830$, $p=2.23\times10^{-6}$, Evidence: $t(722)=5.607$, $p=4.64\times10^{-8}$, Organization/Style: $t(722)=6.200$, $p=1.84\times10^{-9}$, Grammar/Mechanics: $t(722)=6.299$, $p=1.56\times10^{-6}$,Documentation: $t(722)=4.900$, $p=1.56\times10^{-6}$. Positive effect sizes indicate a higher mean score for ON artifacts.

3.2.1.4 Full term to Mini-term Comparison

During the Fall 2014 semester, 124 students were enrolled in a 8-week section of ENC1101 and 3366 students were enrolled in a full term section of ENC1101. Of those, 116 8-week artifacts were sampled along with 608 full artifacts, for a representative sampling percentage of 93.5% and 18.1%, respectively. To be considered a full term artifact for the purposes of this study, the artifact must also have originated from an onsite location. Because no mini-term schedule courses exist offsite, this parameter is applied to avoid any differences as a result of onsite/offsite as opposed to the target study.

Each rubric dimension 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 Organization/Style and Documentation dimensions exhibit a statistically significant difference in mean scores (see Table 10); therefore, we must reject the null hypothesis that the difference in the means of the 8-week and full artifacts are equal to 0, and we can conclude with a 95% confidence that the differences in scores are not solely due to chance. For the remaining rubric dimensions and the overall score, we cannot reject the null hypothesis, meaning the differences in mean scores for those artifacts can be a result of chance.

Based on the work of Johnson (2013), there is a 17-25% chance that the marginally significant results depicted in Table 3 may be false positives (i.e. Type I errors). These marginal results, defined as those within the 95-99% confidence level, include the Documentation dimension.

Effect size was calculated using the Rosenthal and Rosnow (1991) for meta-analytical purposes to serve as a common thread across institutions (Lipsey and Wilson, 1993). The statistically significant results exhibit what Cohen (1988) would consider small effect sizes ranging from 0.15 to 0.18. In other words, non-overlap from 8-week artifacts to full artifacts range from approximately 12%.

	Overall	Thesis	Evidence	Organization/Style	Grammar/Mechanics	Documentation
8-week mean	15.8	3.2	3.2	3.2	3.0	3.1*
full mean	15.1	3.2	3.0	3.1	2.9	2.9*
Effect size	*-0.13*	*-0.05*	*-0.09*	*-0.15*	*-0.09*	*-0.18*

*Table 10. Difference in means for each rubric dimension and combined score (overall) of ENC1101 for 8-week and full. Statistically significant results indicated by shaded cell. Parameters for significance tests are: Organization/Style: t(722)=6.200, $p=1.84x10^{-9}$, Documentation: t(722)=2.422, p=0.016. Positive effect sizes indicate a higher mean score for full artifacts. *Denote marginal significance as defined by Johnson (2013).*

3.2.1.5 Comparison of Full-time and Part-time

Of the artifacts sampled for the full ENC1101 study, 594 originated from courses taught by adjuncts while 133 originated from courses taught by full-time faculty. A comparison of the means for each rubric dimension was conducted as well as overall score. Each rubric dimension 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 Documentation dimension and overall score exhibit a statistically significant difference in mean scores (see Table 11). Therefore, we must reject the null hypothesis that the differences in the means of the adjunct and full-time artifacts for these areas are equal to 0, and we can conclude with a 95% confidence that the differences in scores are not solely due to chance.

Effect size was calculated using the Rosenthal and Rosnow (1991) for meta-analytical purposes to serve as a common thread across institutions (Lipsey and Wilson, 1993). The statistically significant results exhibit what Cohen (1988) would consider medium to large effect sizes ranging from 0.04 to 0.27. In other words, non-overlap from adjunct faculty artifacts to full-time faculty artifacts range from approximately 4% to 19%.

	Overall	Thesis	Evidence	Organization/Style	Grammar/Mechanics	Documentation
Adjunct	12.8	3.2	3.0	3.1	2.9	2.9
Full-time	14.2	3.0	3.1	3.1	2.9	3.2
Effect size	*0.19*	*-0.14*	*0.04*	*0.07*	*0.01*	*0.27*

Table 11. Difference in means for each rubric dimension and combined (overall) of ENC1101 for Adjunct and Full-time faculty. Statistically significant results indicated by shaded cell. Parameters for significance tests are: Documentation: t(725)=3.627, $p=3.51x10^{-4}$, and Overall: t(725)=2.766, p= p=0.006. Positive effect sizes indicate a higher mean score for Full-time faculty artifacts.

3.2.2 Data Distribution

Results from section 2.1 briefly described the distribution in scores among rubric dimensions. With 33 instructors providing 33 elements of variability towards assessment, it becomes increasingly important to understand that variability, particularly as it relates to any abnormal score distribution (see ENC0022 above). Figure 7 depicts the distribution of mean scores of each rubric dimension given by those 33 instructors.

Figure 7 depicts the distribution of mean scores of each rubric dimension given by the 33 instructors who scored artifacts. From the figure, the following interpretations can be made with reasonable certainty: 1) Rubric dimension scoring is quite variable, with mean scores ranging as low as 1.7 to as high as 3.7 in the Documentation dimension. 2) The Thesis dimension exhibits an uneven distribution of mean scores with respect to the median (red line). 3) Grammar/Mechanics and Documentation have the tightest distribution with some outliers. 4) Median scores (red lines) by dimension give a rough estimate of typical success; with the exception of Thesis which is noticeably higher, all dimensions are fairly consistent (somewhat visible in Table 2).

It is reasonable to expect variability to a degree as there are a number of factors going into the data including differences between DE/nonDE, OnL/TD, 8-week/full, and ON/OFF as mentioned above. The largest differences in mean scores existed between DE/nonDE and ON/OFF site samples at slightly less than 0.5. Further, these differences are present in both the Grammar and Documentation dimensions, which mean it is reasonable to conclude the box-whisker spread visible in these two dimensions is reflective of the sample variability. With these two dimensions serving as a baseline, this means Thesis, Evidence, and Organization/Style exhibit a nearly doubled range for the central 50%.

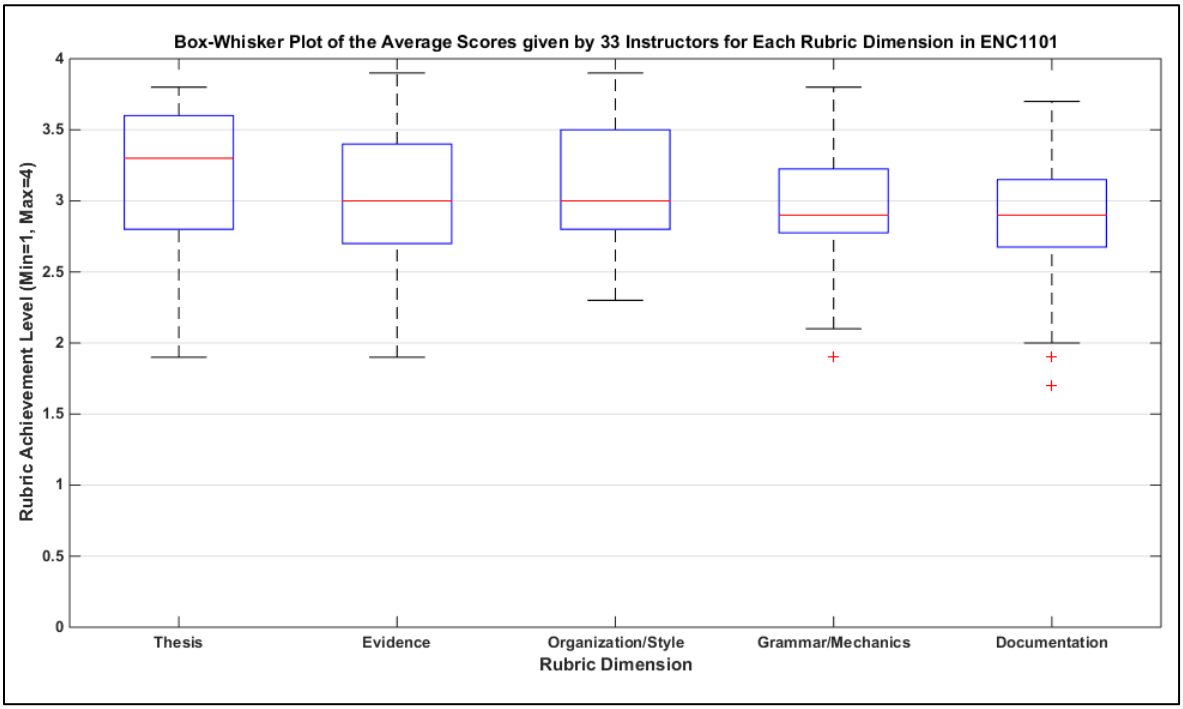

Figure 7. Box-Whisker plot of mean scores given by the 33 instructors scoring papers for ENC1101 for each rubric dimension. Red line depicts median score. Upper and lower box boundaries indicate 75% quartile and 25% quartile (box represents central 50% of the scores). Vertical lines represent remaining scores outside central 50% that are not outliers. Red '+'s denote outliers.

A histogram of actual artifact scores for the entire 722 samples is shown in Figure 8. For comparison, the histogram of the upper quartile of possible scores (highest 25% of possible overall scores; rubric score: 16-20), the 2nd tier quartile of possible scores (rubric score: 11-15), and the 3rd tier quartile of possible scores (rubric score: 6-10) are provided in Figures 9, 10, and 11. Note that this is not the upper quartile of scored artifacts, but the upper quartile of possible scores.

From a review of the histogram of all artifact scores compared with that of the quartile histograms several observations can be made. First, the median score for all rubric dimensions in the upper quartile of possible scores (16-20) is level 4 for all except Grammar/Mechanics. Second tier (11-15) quartile exhibits medians of 3s, and 3rd tier (6-10) exhibit medians of 2s. In both of these cases, Grammar/Mechanics does not deviate from the distribution frequency as it did with the upper quartile. In the 3rd tier, Grammar/Mechanics artifacts show a wider distribution to include both an increased number of level 3s and level 1s when compared with other dimensions.

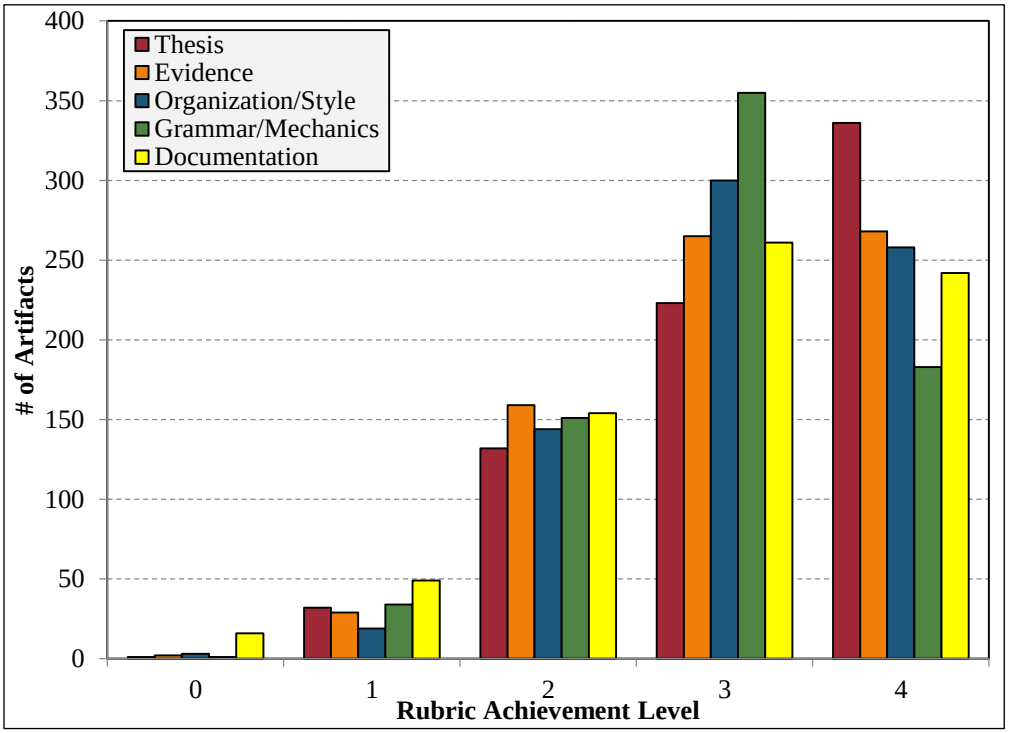

Figure 8. Histogram of Fall 2014 ENC1101 data distribution across achievement levels.

Each of the three observations above depicts artifact scores in the Grammar/Mechanics dimension behaving slightly differently than the other dimensions overall. The interpretation is that upper quartile artifacts exhibit weaker Grammar/Mechanics than other dimensions while 2nd tier artifacts exhibit no discernible difference. In other words, high achieving students consistently struggle with Grammar/Mechanics while 2nd tier students do not struggle with this dimension any more than any other dimension. Additionally, under-achieving students display a wider variety of success in the Grammar/Mechanics dimension than in any other. These attributes together create a circumstance where, as a group, students do not actually perform any poorer in Grammar/Mechanics than any other dimension, but rather the performance distribution among students is markedly different.

To clearly distinguish the properties of the five rubric dimensions based on overall achievement, a color map or binary raster image was created by calculating the average scores for each dimension by overall combined score (Figure 12). Line 'A' depicts the trend of scores by dimension for the upper quartile of lower Grammar/Mechanics dimension scores at higher overall scores compared with other dimensions. By the 2nd tier quartile (line 'B'), this trend has given way to a more variable achievement across dimensions.

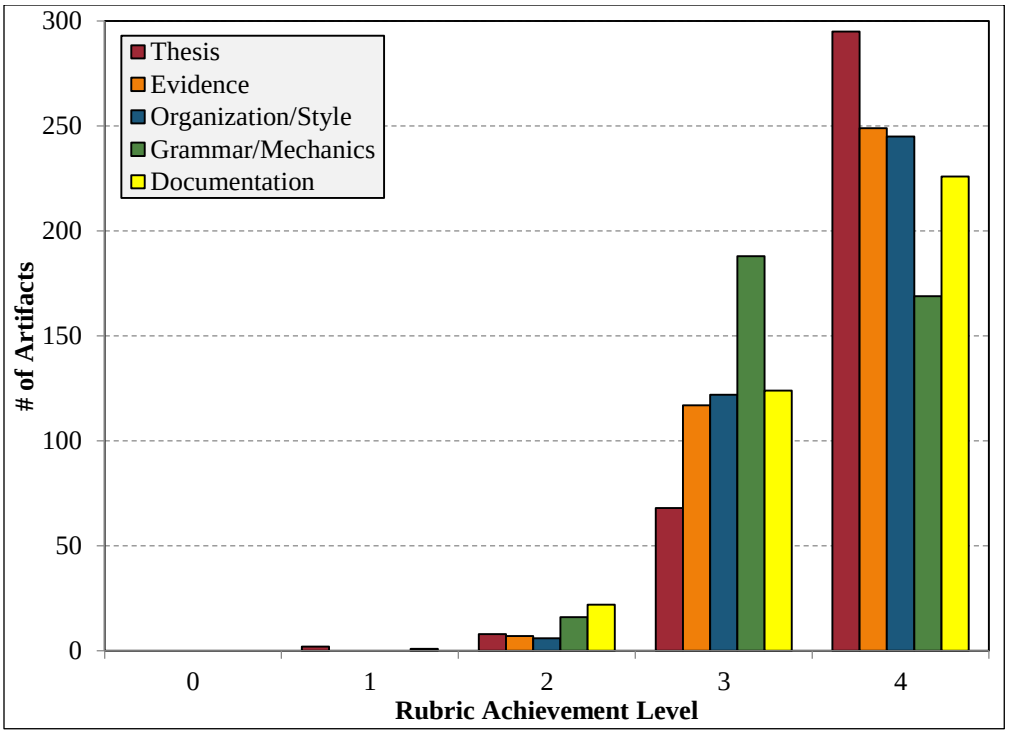

Figure 9. Histogram of Fall 2014 ENC1101 data distribution across achievement levels for upper 25% of overall scores (Combined rubric scores of 16-20).

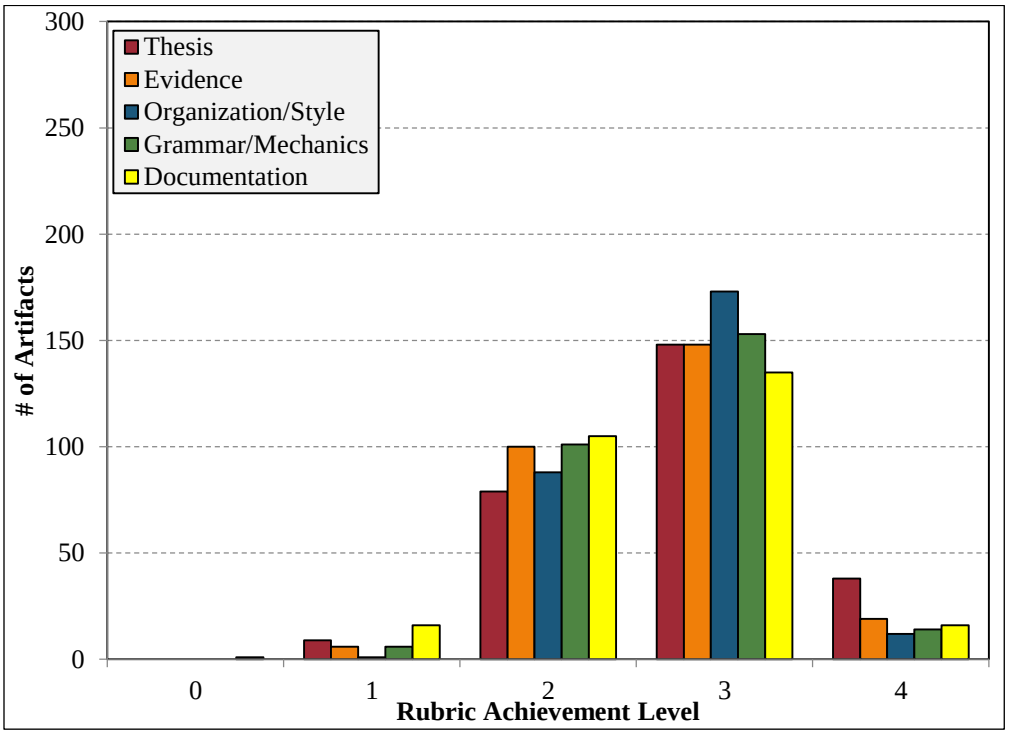

Figure 10. Histogram of Fall 2014 ENC1101 data distribution across achievement levels for 2nd tier quartile of overall scores (Combined rubric scores of 11-15).

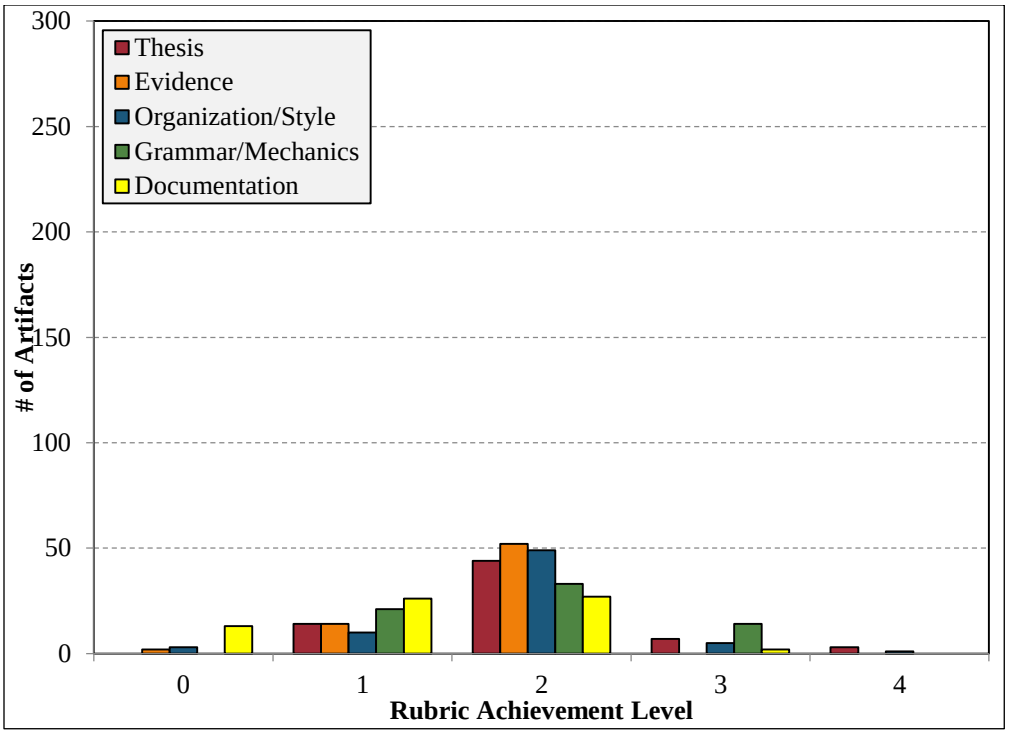

Figure 11. Histogram of Fall 2014 ENC1101 data distribution across achievement levels for 3rd tier quartile of overall scores (Combined rubric scores of 6-10).

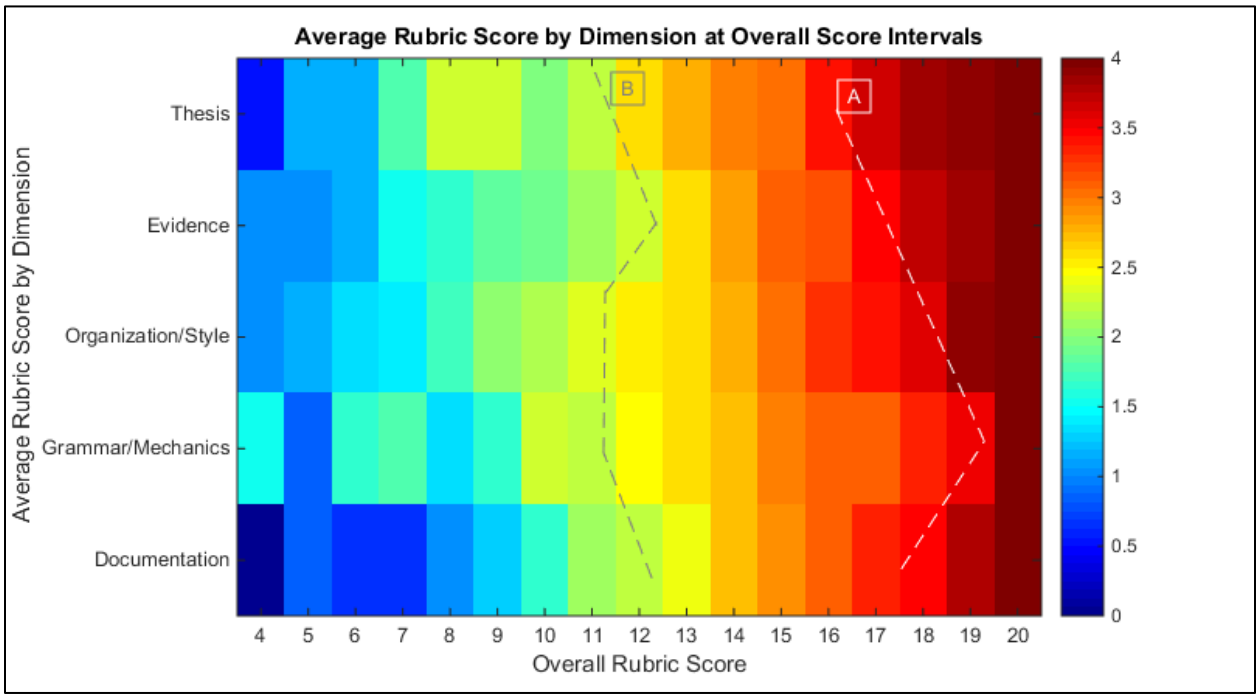

Figure 12. Color map of average achievement in each rubric dimension tied to overall (combined) assessment score.

3.3 ENC1102

3.3.1 Comparison by Site, Format, or Student type

3.3.1.1 Dual Enrollment to non-Dual Enrollment Comparison

During the Fall 2014 semester, 176 total Dual Enrollment (DE) students were enrolled in ENC1102 and 1133 non-DE students were enrolled in ENC1102. Of those, 31 DE artifacts were sampled along with 239 nonDE artifacts, for a representative sampling percentage of 17.6% and 21.1%, respectively.

Each rubric dimension 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 overall score, Thesis, Evidence, and Organization/Style dimensions exhibit a statistically significant difference in mean scores (see Table 12); therefore, we must reject the null hypothesis that the difference in the means of the DE and nonDE artifacts are equal to 0, and we can conclude with a 95% confidence that the differences in scores are not solely due to chance. For the remaining rubric dimensions, we cannot reject the null hypothesis, meaning the differences in mean scores for those artifacts can be a result of chance.

Effect size was calculated using the Rosenthal and Rosnow (1991) for meta-analytical purposes to serve as a common thread across institutions (Lipsey and Wilson, 1993). The statistically significant results exhibit what Cohen (1988) would consider medium effect sizes ranging from 0.34 to 0.48. In other words, non-overlap from DE artifacts to nonDE artifacts range from approximately 24% to 32%.

	Overall	Thesis	Evidence	Organization/Style	Grammar/Mechanics	Documentation
DE mean	16.4	3.5	3.3	3.4	3.1	3.0
nonDE mean	14.7	3.1	2.9	3.1	2.8	2.8
Effect size	*-0.48*	*-0.37*	*-0.35*	*-0.34*	*-0.24*	*-0.19*

Table 12. Difference in means for each rubric dimension and combined score (overall) of ENC1102 for DE and nonDE. Statistically significant results indicated by shaded cell. Parameters for significance tests are: Overall: t(268)=3.914, p=1.64x10^{-4}, Thesis: t(268)=3.046, p=0.004, Evidence: t(268)=2.826, p=0.007, Organization/Style: t(268)=2.760, p=0.008. Positive effect sizes indicate a higher mean score for nonDE artifacts.

A study to determine if Dual Enrollment and non-Dual Enrollment may be associated with site could not be completed since only seven offsite samples were collected. A sample size this small is marginal in terms of completing a reliable comparison statistical significance study (de Winter, 2013) and therefore was not completed.

3.3.1.2 Online to Traditional Comparison

During the Fall 2014 semester, 324 total Online (Onl) students were enrolled in ENC1102 and 985 Traditional (TD) students were enrolled in ENC1102. Of those, 44 OnL artifacts were sampled along with 226 TD artifacts, for a representative sampling percentage of 13.6% and 22.9%, respectively.

Each rubric dimension 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 Documentation dimension exhibits a statistically significant difference in mean scores (see Table 13); therefore, we must reject the null hypothesis that the difference in the means of the OnL and TD artifacts are equal to 0, and we can conclude with a 95% confidence that the differences in scores are not solely due to chance. For the remaining rubric dimensions, we cannot reject the null hypothesis, meaning the differences in mean scores for those artifacts can be a result of chance.

Based on the work of Johnson (2013), there is a 17-25% chance that the marginally significant results depicted in Table 13 may be false positives (i.e. Type I errors). These marginal results are those that are within the 95-99% confidence level and so the result of statistically significant differences of the mean may not be true of the population.

Effect size was calculated using the Rosenthal and Rosnow (1991) for meta-analytical purposes to serve as a common thread across institutions (Lipsey and Wilson, 1993). The results exhibit what Cohen (1988) would consider small to medium effect sizes ranging from 0.08 to 0.64. In other words, non-overlap from TD artifacts to OnL artifacts range from approximately 7% to 41%.

	Overall	Thesis	Evidence	Organization/Style	Grammar/Mechanics	Documentation
OnL mean	14.3	3.3	2.7	3.0	2.8	2.6*
TD mean	15.0	3.1	3.0	3.1	2.9	2.9*
Effect size	*0.17*	*-0.08*	*0.22*	*0.11*	*0.64*	*0.29*

*Table 13. Difference in means for each rubric dimension and combined score (overall) of ENC1102 for OnL and TD. Statistically significant results indicated by shaded cell. Parameters for significance tests are: Documentation: t(268)=2.356, p=0.022. *Denote marginal significance as defined by Johnson (2013). Positive effect sizes indicate a higher mean score for TD artifacts.*

3.3.1.3 Onsite to Offsite Comparison

Only seven offsite (OFF) samples were collected in the study for ENC1102. A sample size this small is marginal in terms of completing a reliable comparison statistical significance study (de Winter, 2013) and therefore was not completed.

3.3.1.4 Full term to Mini-term Comparison

During the Fall 2014 semester, 175 students were enrolled in an 8-week section of ENC1102 and 1134 students were enrolled in a full term section of ENC1102. Of those, 13 8-week artifacts were sampled along with 257 full artifacts, for a representative sampling percentage of 7.4% and 22.7%, respectively.

Each rubric dimension 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 overall rubric score exhibits a statistically significant difference in mean scores (see Table 14); therefore, we must reject the null hypothesis that the difference in the means of the 8-week and full artifacts are equal to 0, and we can conclude with a 95% confidence that the differences in scores are not solely due to chance. For the remaining rubric dimensions, we cannot reject the null hypothesis, meaning the differences in mean scores for those artifacts can be a result of chance.

Based on the work of Johnson (2013), there is a 17-25% chance that the marginally significant results depicted in Table 14 may be false positives (i.e. Type I errors). These marginal results are those that are within the 95-99% confidence level and so the result of statistically significant differences of the mean may not be true of the population.

	Overall	Thesis	Evidence	Organization/Style	Grammar/Mechanics	Documentation
8-week mean	14.3*	3.0	2.9	2.8	2.8	2.8
full mean	14.9*	3.2	2.9	3.1	2.9	2.9
Effect size	*0.07*	*0.09*	*-0.01*	*0.14*	*0.05*	*0.05*

*Table 14. Difference in means for each rubric dimension and combined score (overall) of ENC1102 for 8-week and full. Statistically significant results indicated by shaded cell. Parameters for significance tests are: Overall: t(268)=0.544, p=0.031. *Denote marginal significance as defined by Johnson (2013). Positive effect sizes indicate a higher mean score for full artifacts.*

Effect size was calculated using the Rosenthal and Rosnow (1991) for meta-analytical purposes to serve as a common thread across institutions (Lipsey and Wilson, 1993). The results exhibit what Cohen (1988) would consider small effect sizes ranging from 0.01 to 0.14. In other words, non-overlap from 8-week artifacts to full artifacts is less than 9%.

3.3.1.5 Comparison of Full-time and Part-time

Of the artifacts sampled for the full ENC1102 study, 130 originated from courses taught by adjuncts while 140 originated from courses taught by full-time faculty. A comparison of the means for each rubric dimension was conducted as well as overall score. Each rubric dimension and the overall score was tested for significance using a Welch's t-test according to standard methods (Davis, 1973; McDonald, 2009; Wilkinson, 1999). There was no significant difference in any rubric dimension or in the overall score. Documentation dimension and overall score exhibit a statistically significant difference in mean scores (see Table 15). Therefore, we cannot reject the null hypothesis, meaning the differences in mean scores for those artifacts can reasonably be the result of chance.

	Overall	Thesis	Evidence	Organization/Style	Grammar/Mechanics	Documentation
Adjunct	15.0	3.2	2.9	3.1	2.9	2.9
Full-time	14.8	3.0	3.1	2.8	2.8	3.0
Effect size	*0.06*	-0.04	0.12	-0.02	-0.21	-0.11

Table 15. Difference in means for each rubric dimension and combined (overall) of ENC1102 for Adjunct and Full-time faculty. Statistically significant results indicated by shaded cell. Positive effect sizes indicate a higher mean score for Full-time faculty artifacts.

3.3.2 Data Distribution

ENC1102 included 15 instructors providing 15 elements of variability towards assessment. To better to understand this variability, particularly as it relates to any abnormal score distribution (see ENC0022 above), a box-whisker plot of the distribution of mean scores of each rubric dimension given by those 15 instructors is presented in Figure 13.

Scoring averages across instructors is much less variable than both ENC0022 and ENC1101. As with ENC1101, Grammar/Mechanics has the highest agreement among instructors in terms of average scoring with the exception of one outlier. While the Thesis dimension has somewhat higher scoring, the remaining four dimensions are scored fairly similarly across instructors.

A histogram of actual artifact scores for the entire 270 samples is shown in Figure 14. All dimensions are fairly similarly distributed across rubric achievement level. The Grammar/Mechanics, as with ENC1101, has the lowest number of exemplar (4s) achieved, although it doesn't appear as dramatic as ENC1101 because the range of exemplar achievement by the four other dimensions is a bit more varied. When comparing the percent difference in 4s scored in Grammar/Mechanics with the highest dimension, in both cases Thesis, the differences are -54.4% for ENC1101 and -55.4% for ENC1102.

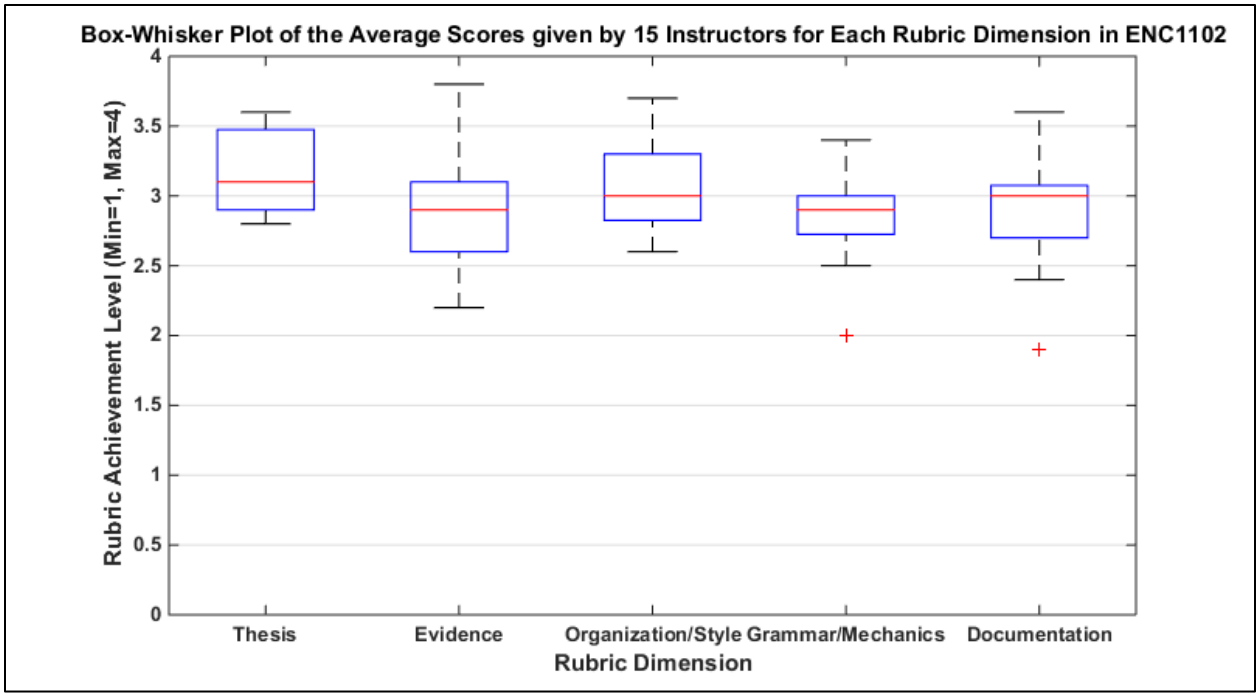

Figure 13. Box-Whisker plot of mean scores given by the 15 instructors scoring papers for ENC1102 for each rubric dimension. Red line depicts median score. Upper and lower box boundaries indicate 75% quartile and 25% quartile (box represents central 50% of the scores). Vertical lines represent remaining scores outside central 50% that are not outliers. Red '+'s denote outliers.

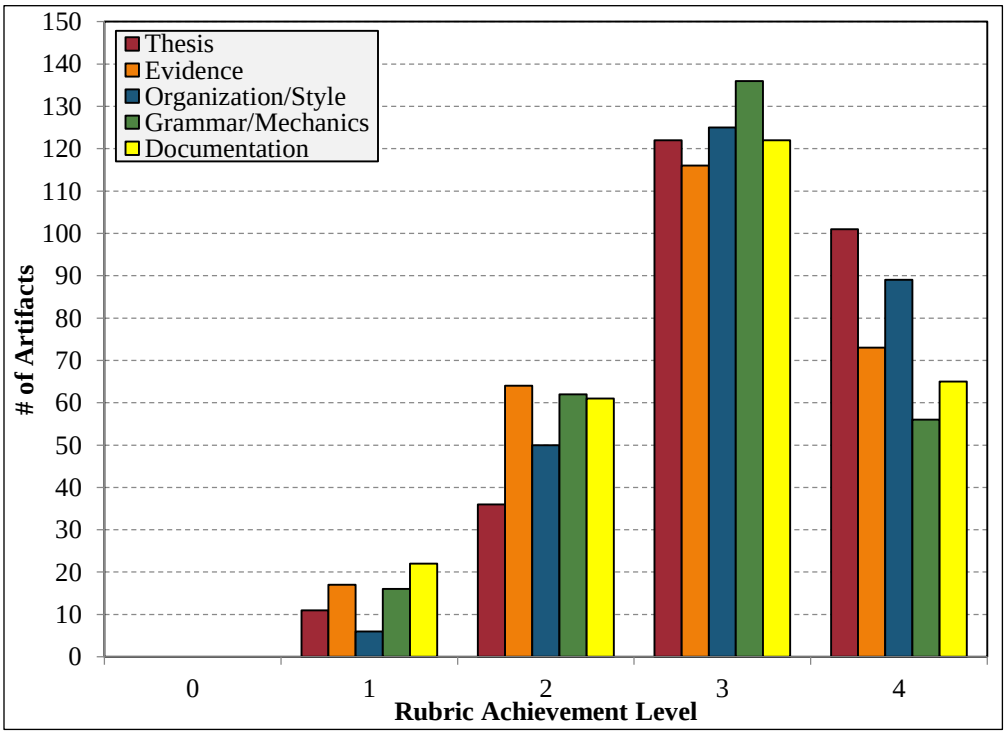

Figure 14. Histogram of Fall 2014 ENC1102 data distribution across achievement levels.

4 CONCLUSIONS

The English Department of FSW outlined an initial plan for assessment in three courses: English for College Success (ENC0022), Composition I (ENC1101), and Composition II (ENC1102). In each course instructors use a common rubric with seven identified rubric dimensions in the case of ENC0022, and five dimensions for both ENC1101 and ENC1102 to evaluate achievement with a goal towards informing faculty upon establishing Student Learning Objectives (SLOs) for future assessment plans.

A drilldown of ENC0022 results are as follows:

1. All seven rubric dimensions had > 90% achievement at level 2 or higher.
2. Introductory Paragraph, Support Paragraphs, Organization, and Concluding Paragraph dimensions had > 60% of achievement at level 3 or higher.
3. In the historically lowest-performing dimensions (Grammar and Mechanics), scores of 3 are earned in Grammar and Mechanics as commonly as all other dimensions. Only for high-achieving students do these dimensions exhibit lower averages than other dimensions causing the <60% achievement at level 3 or higher.
4. In a comparison of full term to mini-term (8-week) courses, there was a statistically significantly higher mean score for full-term artifacts in both Introductory Paragraph and Research dimensions.
5. In a comparison of full-time faculty to adjunct faculty, there was a statistically significantly higher mean score for adjunct faculty artifacts in both Support Paragraphs, Research, and the overall score.
6. In the Research dimension, 56% of level 4s come from high achieving students which is substantially lower than all other dimensions which range from 88%-100%. There is disagreement among instructor scoring which appears to be the cause.

A drilldown of ENC1101 results are as follows:

1. All five rubric dimensions had > 70% achievement at level 3 or higher.
2. In a comparison of dual enrollment to traditional students, there was a statistically significantly higher mean score for all dual enrollment rubric dimensions including the overall combined rubric score.
3. A follow-up study of dual enrollment students determined there was no statistically significant differences in the means of dual enrollment students on campus and dual enrollment students receiving instruction offsite.
4. In a comparison of online to traditional or face-to-face students, there was no statistically significant difference across all rubric dimensions including the overall combined rubric score.
5. In a comparison of on campus students to offsite students, there was a statistically significantly higher mean score for all offsite dimensions including the overall combined rubric score.
6. In a comparison of mini-term (8-week) students to full term, both the Organization/Style and Documentation dimensions exhibit a statistically significantly higher mean score for 8-week students.
7. In a comparison of full-time faculty to adjunct faculty, there was a statistically significantly higher mean score for full-time faculty artifacts in both Documentation and the overall score.

8. Additionally, in a study comparing scoring across instructors, it was determined that rubric dimension scoring across instructors was found to be quite variable with mean rubric scores ranging from 1.7 to 3.7.
9. In the same instructor scoring study, the Grammar/Mechanics and Documentation dimensions have the highest agreement.
10. The median score for all rubric dimensions in the upper quartile of possible scores (overall combined rubric score of 16-20) is level 4 for all except Grammar/Mechanics.
11. For Grammar/Mechanics, scores of 3 are earned in Grammar and Mechanics as commonly as all other dimensions. Only for high-achieving students do these dimensions exhibit lower averages than other dimensions. As a group, students do not actually perform any poorer in Grammar/Mechanics than any other dimension, but rather the performance distribution among students is markedly different.

A drilldown of ENC1102 results are as follows:

1. Four of five rubric dimensions had > 70% achievement at level 3 or higher. The dimension that did not score greater than 70% at 3 or higher was Documentation (69.3%).
2. In a comparison of dual enrollment to traditional students, there was a statistically significantly higher mean score for the Thesis, Evidence, and Organization/Style dimensions as well as the overall combined rubric score. No follow up study comparing dual enrollment students by location could be completed due to a small sample size for offsite artifacts.
3. In a comparison of online to traditional or face-to-face students, there was a statistically significantly higher mean score for the traditional artifacts.
4. In a comparison of mini-term (8-week) students to full term, the overall rubric score exhibits a statistically significantly higher mean score for full artifacts.
5. In a comparison of full-time faculty to adjunct faculty, there was no statistically significant differences in mean scores for any rubric dimension or the overall score.
6. In a study comparing scoring across instructors, it was determined that rubric dimension scoring across instructors was found to be fairly consistent with the exception of the Thesis dimension, which fairs slightly higher than other dimensions.
7. In the same instructor scoring study, the Grammar/Mechanics and Documentation dimensions have the strongest agreement.
8. For Grammar/Mechanics, achievement level 3 is reached as commonly as all other dimensions. Only for high achieving students do these dimensions exhibit lower averages than other dimensions. As a group, students do not actually perform any poorer in Grammar/Mechanics than any other dimension, but rather the performance distribution among students is markedly different.

5 REFERENCES

Cohen, J. 1988. Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Earlbaum Associates, Hillsdale, NJ.

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.

de Winter, J.C.F. 2013. Using the Student's T-Test with Extremely Small Sample Sizes. Practical Assessment, Research, and Evaluation, 18(10), 1-12.

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.

Johnson, V. 2013. Revised Standards for Statistical Evidence. Proceedings of the National Academy of Science, 110(48), 19313-19317.

Lipsey, M.W. and Wilson, D.B. 1993. The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. American Psychologist, 48, 1181-1209.

McDonald, J.H. 2009. Handbook of Biological Statistics (2nd ed.). Sparky House Publishing, Baltimore, Maryland.

Rosenthal, R. and Rosnow, R.L. 1991. Essentials of behavioral research: Methods and data analysis (2nd ed.). McGraw Hill, New York, NY.

Wilkinson, L. 1999. APA Task Force on Statistical Inference. Statistical Methods in Psychology Journals: Guidelines and Explanations. American Psychologist 54 (8), 594–604.