Foreign Language 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 Foreign Language Department has employed common rubric elements used by all faculty as a means to evaluate student progress and make informed comparisons between Dual Enrollment (DE) and non-Dual Enrollment (nonDE) students, as well as Online (OnL) and Traditional (TD) students as highlighted in the QEP course level assessment plan.

The results of common course assessments for both introductory Spanish (SPN1120, SPN1121) and French (FRE1120, FRE1121) courses are herein detailed. 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 assist in gauging student progress and provide support toward instructive improvement, therefore allowing assessment to drive instruction as defined by Elder and Paul (2007).

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

2 OUTCOMES AND RESULTS

2.1 ASSESSMENT ANALYSIS & SIGNIFICANCE TESTS

2.1.1 Norming Analysis

2.1.1.1 French

No norming exercise was conducted by the French course faculty.

2.1.1.2 Spanish

Five Spanish 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 five faculty scored three unique 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 is shown in Figure 1.

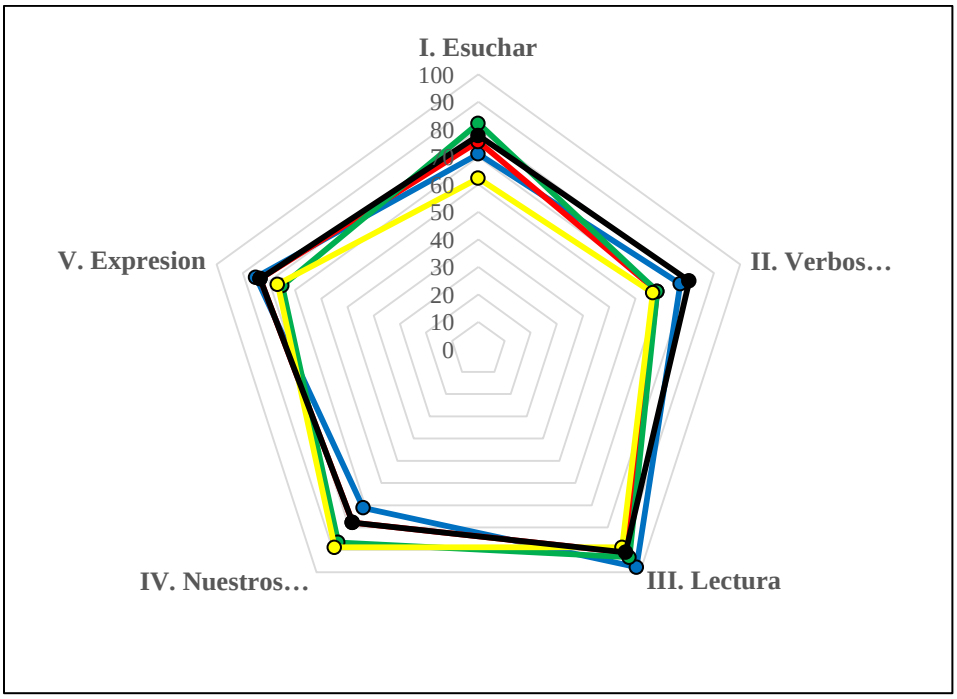

Figure 1. Radar plot of norming depicting average rater score for each rubric criteria of three common artifacts. Blue is Rater 1, red is Rater 2, green is Rater 3, yellow is Rater 4, and black is Rater 5. Each rubric criterion had a unique maximum number of points. Scores were normalized to 100-scale for clarity and interpretation.

Raters must assign each artifact a variable score based on rubric criterion. Each rubric criterion is weighted differently. Topics I, III, and IV are determined using a 15-point scale, Topic II is determined using a 60-point scale, and Topic V is determined using a 20-point scale. The maximum possible overall score is 125 points. These ratings are graphed here in percent of their individual rubric maximum (i.e. 15/15 = 100%, or 60/60 = 100%, depending on rubric). This was done for comparison purposes and possible links to rubric interpretation by providing information to determine if a rater is more/less inclined to deduct greater/fewer points if the total is 15 as opposed to 20 or 60. An example of this would be that if an artifact has a near perfect Topic 1, raters may be inclined to score 14/15, but a similar near perfect Topic III might result in a 59/60 for some raters and a 56/60 if raters interpret the measurement of perfection across topics should be equalized. In short, with weighted rubrics, rater interpretation as piecewise or holistic becomes paramount.

Topics III and V show moderate agreement with scoring ranges of 8.89% and 10.0% respectively. By rubric weight, this translates to disagreement among raters of 1.4/15 and 1.5/20 respectively. In contrast, Topic I exhibits the largest range at 20.0%, or 3/15.

When comparing raters overall scores, there is no statistically significant difference in mean scores between them based on a single factor ANOVA (Analysis of Variance). However, this may be largely due

to the extensive variety of the topics (e.g. Rater 1 might inflate scores in Topic 1 compared to Rater 2, but the situation is reversed for Topic 2, thus negating mean score differences). To effectively make a rater-to-rater comparison by topic, each artifact was plotted on a radar plot as a function of rater in Figures 2-6.

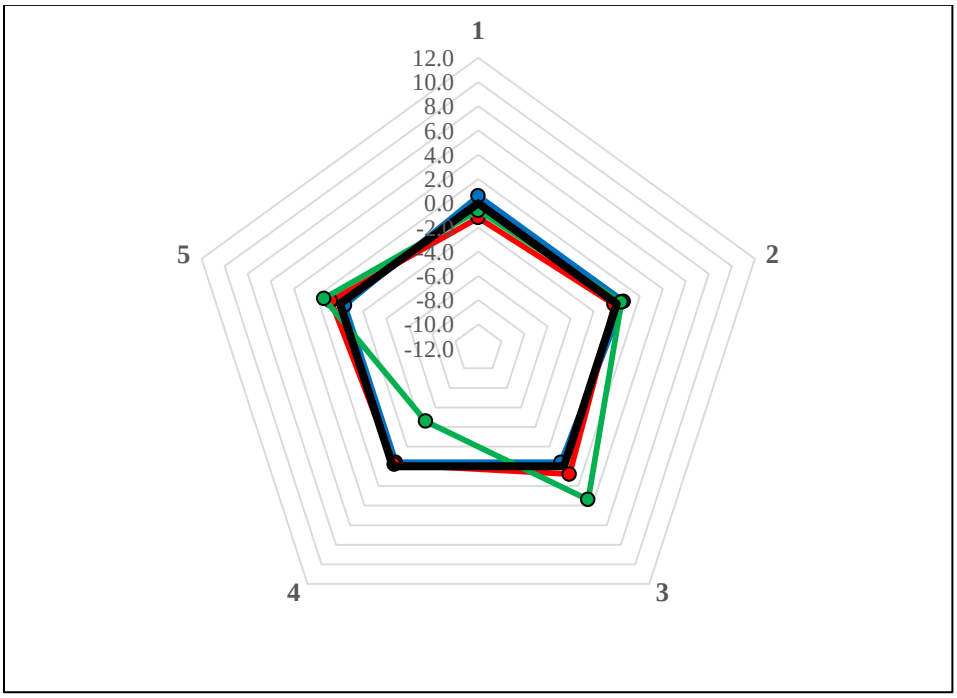

Figure 2. Difference in artifact scores by rater for Topic I: Escuchar. Highest score (artifact 1) in blue, mid-range score (artifact 2) in red, lowest score (artifact 3) in green. Bold black line denotes mean score for each of the five raters plotted as the zero line while each rater score is plotted as the difference from the mean.

In Figure 2, each of the three artifacts for Topic I: Escuchar are plotted on a radar plot corresponding to each of the five raters. The black line represents the average score of the five raters displayed here as a zero line, while each artifact (blue, red, and green) is plotted as the difference from the mean at each rater.

For example, corner 3 on the radar plot represents Rater 3, who scored an 11/15 for the lowest scoring artifact (green line). This score is a difference of 3.4 rubric points over the mean rater score (depicted as the bold black line). Rater 3's difference from the mean for the other two artifacts is minimal (0.4 below mean for blue, and 0.8 for red). This is of note, since this representation reflects an inflated grade for Rater 3 on the lowest artifact, but fairly uniform otherwise. Similarly, Rater 4 deflates substantially with a lowest scoring artifact compared with the mean (4.6 rubric points below mean) but is otherwise uniform with higher scoring artifacts.

Inflations and deflations of this magnitude on a 100-point scale (as seen in Figure 1) translate to 22.6%, and -30.7%, respectively. Rater 3's inflation and Rater 4's deflation for this lowest scoring artifact is clearly visible in Figure 1, where Rater 3's mean score is highest of the five raters for topic I and Rater 4's is lowest.

Figure 3 depicts each of the three artifacts for Topic II: Verbos plotted on a radar plot corresponding to each of the five raters. In Topic II, the scoring is determined using a 60-point scale so it isn't

unreasonable to see larger variability in the results (e.g. Topic II 33% increase = 20 points compared with 5 points in Topics I, III, and IV).

Figure 3 exhibits a trend towards increasing disagreement among raters with decreasing artifact score. The highest scored artifact (blue, mean: 122/125) exhibits very good agreement across all raters. The moderate artifact score (red, mean: 98/125) shows deflation from the mean from Raters 3 and 4 and inflation from Raters 1 and 5. The lowest scored artifact (green, mean: 68/125) shows increased deflation from Raters 3 and 4 and increased inflation from Raters 1 and 5. Moreover, Rater 2 also exhibits deflation from the mean at the lowest scoring artifact.

By example, the artifact with the greatest disagreement among raters (green), with all other topics scored identically, would be scored 16 points higher by Rater 5 than Rater 3. A 16-point difference on the exam's 125-point scale amounts to 13% of the total score, more than one full letter grade. The results for Topic II support two possible interpretations: A) raters disagree on poor performance rubric definitions, or B) the rubric at the lower end of the scoring spectrum is not clear enough to effectively serve as a common assessment tool. These possibilities may also play a role to a lesser extent in the differences among raters with Topic I as well.

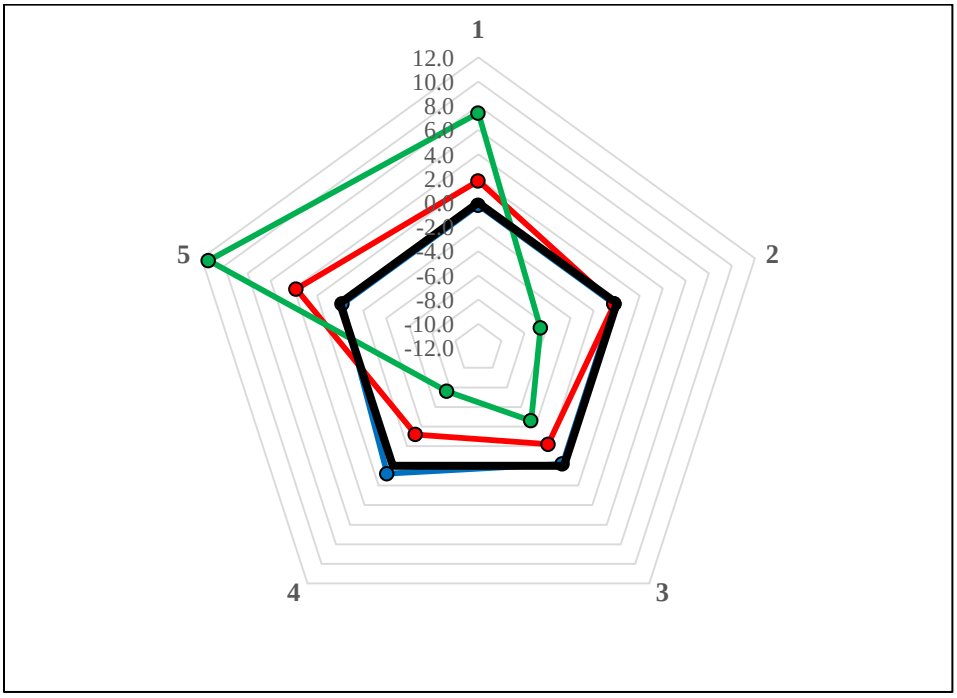

Figure 3. Difference in artifact scores by rater for Topic II: Verbos. Highest score (artifact 1) in blue, mid-range score (artifact 2) in red, lowest score (artifact 3) in green. Bold black line denotes mean score for each of the five raters plotted as the zero line while each rater score is plotted as the difference from the mean.

Figure 4 depicts each of the three artifacts for Topic III: Lectura plotted on a radar plot corresponding to each of the five raters. In Topic III the scoring is determined using a 15-point scale. Topic III exhibits strong agreement across all raters with the largest variation again reflected in the lowest scoring artifact (green) where Rater 4 scores an 11/15 while the mean rater score is 12.4/15, a difference of 1.1% of the overall score.

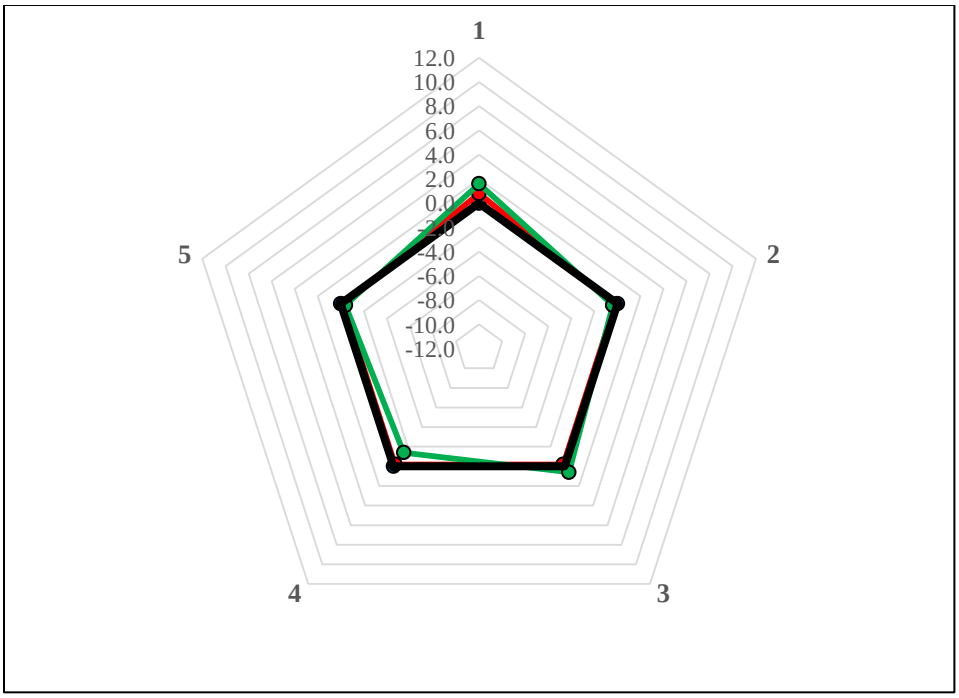

Figure 4. Difference in artifact scores by rater for Topic III: Lectura. Highest score (artifact 1) in blue, mid-range score (artifact 2) in red, lowest score (artifact 3) in green. Bold black line denotes mean score for each of the five raters plotted as the zero line while each rater score is plotted as the difference from the mean.

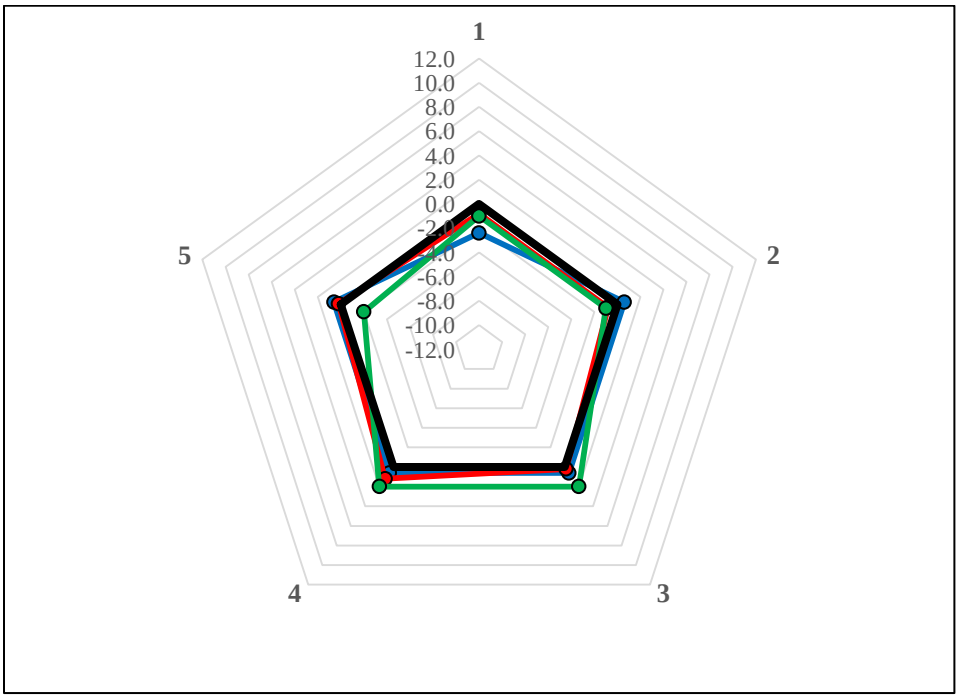

Figure 5. Difference in artifact scores by rater for Topic IV: Nuestros. Highest score (artifact 1) in blue, mid-range score (artifact 2) in red, lowest score (artifact 3) in green. Bold black line denotes mean score for each of the five raters plotted as the zero line while each rater score is plotted as the difference from the mean.

Figure 5 depicts each of the three artifacts for Topic IV: Nuestros plotted on a radar plot corresponding to each of the five raters. In Topic IV, the scoring is determined using a 15-point scale. Topic IV exhibits strong agreement across all raters with the largest variation again reflected in the lowest scoring artifact (green) where Rater 5 scores an 7/15 while the mean rater score is 9/15, a difference of 1.6% of the overall score.

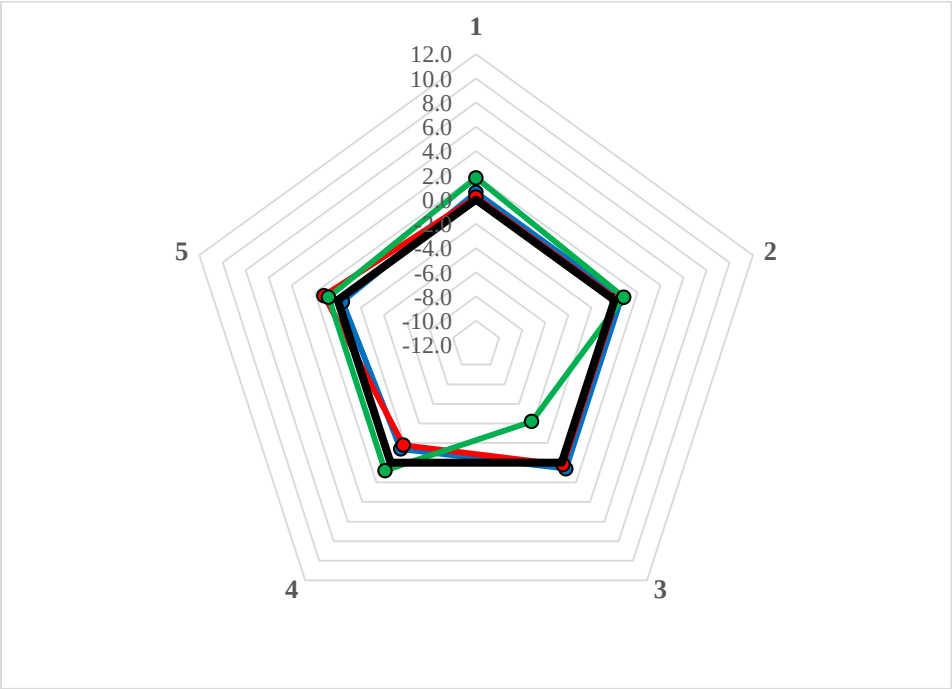

Figure 6. Difference in artifact scores by rater for Topic V: Expresion. Highest score (artifact 1) in blue, mid-range score (artifact 2) in red, lowest score (artifact 3) in green. Bold black line denotes mean score for each of the five raters plotted as the zero line while each rater score is plotted as the difference from the mean.

Figure 6 depicts each of the three artifacts for Topic V: Expresion plotted on a radar plot corresponding to each of the five raters. In Topic V, the scoring is determined using a 20-point scale. Topic V exhibits strong agreement across all raters, with the exception of the lowest scoring artifact (green), where Rater 3 scores a 10/15, 4.2 rubric points below the mean rater score of 14.2/15, a difference of 3.36% of the overall score.

2.1.2 Descriptive Statistics

2.1.2.1 French

2.1.2.1.1 FRE1120

During the Fall 2013 semester, 42 total artifacts were recorded for FRE1120. Of those, the ratio of non-Dual Enrolment (nonDE) to Dual Enrollment (DE) (nonDE/DE) students was 34/8 and all were Traditional students (TD) students with no Online (OnL) students enrolled. No sections of FRE1120 were offered for Spring 2014.

FRE1120 is scored using a rubric with 10 different sections, each with a different maximum. The maximum scores for those sections (Sections A, B, C, etc.) are 10, 6, 20, 6, 20, 8, 10, 10, 14, and 20, respectively, for a maximum score of 124 points. The average overall score for the 42 artifacts is

111.0/124 or 89.5% (Table 1). Section A had the highest mean score (9.09/10, 90.9%) and Section I had the lowest mean score (10.01/14, 71.5%). No sections of FRE1120 were offered for Spring 2014.

The average overall score for DE students (111.00) was substantially higher than that of nonDE (97.18) (see Section 2.1.3 for details on significance) (Table 2). Nine of ten rubric criteria for DE artifacts exhibit higher means than their nonDE counterparts, although again, see 2.1.3 for significance.

N = 42	A	B	C	D	E	F	G	H	I	J	Total
max. possible pts	***10***	***6***	***20***	***6***	***20***	***8***	***10***	***10***	***14***	***20***	***124***
mean	9.09	5.08	15.60	4.96	16.33	6.52	7.98	8.31	10.01	15.93	99.82
median	9	5	18	5.5	18	7	8.5	8.5	11.25	16	104.75
mode	9.5	5	20	6	20	8	9	10	13	16	115.5
standard deviation	0.80	0.81	5.24	1.43	4.52	1.72	1.56	1.41	3.55	2.53	16.40
Kurtosis	1.90	0.92	-0.42	2.87	0.46	-0.06	5.03	-0.87	-1.19	0.45	-0.53
Skewness	-1.21	-0.79	-0.96	-1.75	-1.20	-0.99	-2.13	-0.36	-0.54	-0.73	-0.72

Table 1. Basic descriptive statistics of Fall 2013 FRE1120 artifacts (42 samples).

	nonDE: N = 34 DE: N = 8	A	B	C	D	E	F	G	H	I	J	Total
nonDE	***max. possible pts***	***10***	***6***	***20***	***6***	***20***	***8***	***10***	***10***	***14***	***20***	***124***
	mean	9.08	5.01	15.12	4.93	15.59	6.60	7.88	8.09	9.53	15.35	97.18
	median	9	5	17	5	17.5	7	8.5	8	10.5	16	101
	mode	9	5	20	6	20	8	9	7	7	16	115.5
	standard deviation	0.85	0.81	5.28	1.40	4.71	1.55	1.52	1.42	3.63	2.39	16.39
	Kurtosis	1.71	1.31	-0.66	4.07	-0.19	-0.09	6.10	-0.97	-1.41	0.47	-0.74
	Skewness	-1.17	-0.82	-0.79	-1.93	-0.91	-0.92	-2.19	-0.16	-0.34	-0.80	-0.59
DE	**mean**	9.13	5.38	17.63	5.13	19.50	6.19	8.38	9.25	12.06	18.38	111.00
	median	9.5	5.75	20	6	20	7.75	9	9.5	13	18.5	113.5
	mode	9.5	6	20	6	20	8	9	10	13	19	n/a
	standard deviation	0.58	0.79	4.84	1.64	0.93	2.45	1.77	0.89	2.37	1.41	11.51
	Kurtosis	0.62	-0.69	6.62	0.72	0.00	-1.09	8.00	-1.48	2.67	-0.56	3.84
	Skewness	-1.36	-0.90	-2.53	-1.55	-1.44	-0.89	-2.83	-0.62	-1.62	-0.48	-1.79

Table 2. Basic descriptive statistics of FRE1120 artifacts for Fall 2013 with respect to nonDE vs. DE students (N=34, N=8, respectively). Higher values for DE over nonDE denoted with shaded cell.

2.1.2.1.2 FRE1121

No sections of FRE1121 were offered for Fall 2013. During the Spring 2014 semester, 26 total artifacts were recorded for FRE1121. Of those, the ratio of nonDE/DE was 16/10 and all were TD with no OnL students enrolled.

FRE1121 uses 13 common scoring elements (Sections I, II, III, etc.) and maximum scores for those sections are 5, 6, 12, 10, 6, 13, 12, 12, 12, 6, 8, 26, 20, for a maximum score of 148 points. The average overall score is 119.04/148 or 80.43% (Table 3). Section III has the highest mean score (9.13/10, 91.1%) and Section XI has the lowest mean score (5.69/8, 71.1%).

N = 26	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	Total
max. possible pts	***5***	***6***	***12***	***10***	***6***	***13***	***12***	***12***	***12***	***6***	***8***	***26***	***20***	***148***
mean	4.15	4.33	8.60	9.13	4.02	11.27	10.02	10.48	9.21	4.67	5.69	20.54	16.92	119.04
median	4.5	5	9.75	9.5	4.5	12	10.5	10.75	9.5	5	6	21.5	17	121.75
mode	5	6	10	10	4.5	13	11.5	12	9.5	5.5	5	23	17	136.5
standard deviation	1.10	1.79	3.01	1.52	1.48	2.65	1.80	1.76	1.77	1.33	1.44	4.20	2.23	17.45
Kurtosis	1.80	0.69	0.12	7.96	-0.78	13.35	2.69	2.66	0.16	5.63	0.29	0.17	5.80	3.15
Skewness	-1.56	-1.10	-1.03	-2.74	-0.55	-3.34	-1.53	-1.58	-0.72	-2.15	-0.58	-0.91	-1.88	-1.36

Table 3. Basic descriptive statistics of Spring 2014 FRE1121 artifacts (26 samples).

The average overall score for DE students (122.85) was higher than that of nonDE (116.66) (see Section 2.1.3 for details on significance) (Table 4). Ten of thirteen rubric criteria for DE artifacts exhibit higher means than their nonDE counterparts, although again, see 2.1.3 for significance. Nine of thirteen rubric criteria for DE artifacts exhibit a more negative skewness, meaning scores are tending towards higher scores than nonDE artifacts with a tail towards lower scores (see Figure 7 for example).

nonDE N = 16 DE N = 10		I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	Total
nonDE	*max. possible pts*	*5*	*6*	*12*	*10*	*6*	*13*	*12*	*12*	*12*	*3*	*8*	*26*	*20*	*148*
	mean	4.50	3.72	7.69	8.88	3.56	10.91	9.69	10.34	9.34	4.81	5.69	20.84	16.69	116.66
	median	4.8	4.0	9.0	9.8	3.3	12.0	10.3	10.5	9.5	5.3	6.3	21.0	17	114.75
	mode	5.0	5.0	10.0	10.0	3.0	12.0	11.5	10.5	9.5	5.5	5.0	23.0	16	140
	standard deviation	0.80	1.91	3.38	1.88	1.45	3.12	2.10	1.76	1.43	1.41	1.70	3.82	2.77	20.48
	Kurtosis	6.38	-0.06	-1.06	4.18	-0.80	11.23	1.39	5.22	0.81	9.85	-0.24	1.66	3.01	2.05
	Skewness	-2.38	-0.75	-0.55	-2.12	-0.05	-3.18	-1.27	-1.99	-0.57	-2.89	-0.59	-1.01	-1.45	-1.16
DE	mean	3.60	5.30	10.05	9.55	4.75	11.85	10.55	10.70	9.00	4.45	5.70	20.05	17.30	122.85
	median	4	6	10	9.5	5	12.5	11	11.75	9.75	5	6	22.5	17	124.5
	mode	4	6	9	9.5	5.5	13	11.5	12	10.5	5	6	24	17	136.5
	standard deviation	1.33	1.06	1.54	0.50	1.27	1.65	1.09	1.84	2.29	1.21	0.95	4.92	0.82	10.97
	Kurtosis	-0.03	1.26	0.35	0.91	5.19	2.68	-1.74	0.48	-0.73	0.50	-0.35	-0.83	1.24	-0.31
	Skewness	-0.89	-1.44	-0.55	-1.08	-2.06	-1.75	-0.48	-1.31	-0.61	-0.97	-0.23	-0.81	0.81	-0.49

Table 4. Basic descriptive statistics of FRE1121 artifacts for Spring 2014 with respect to nonDE vs. DE students (N=34, N=8, respectively). Higher values for DE over nonDE denoted with shaded cell.

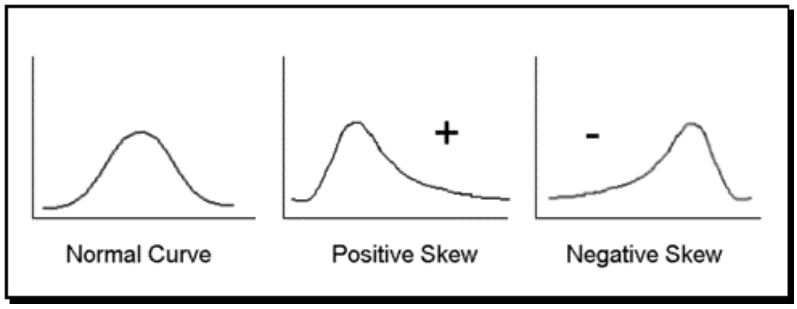

Figure 7. Example of skewness. The normal curve (left) has a skewness of 0.0. A positive value skewness (center) and negative value skewness (right) depict an ideal scenario (Starkweather, 2010).

2.1.2.2 Spanish

2.1.2.2.1 SPN1120

During the Fall 2013 semester, 58 total artifacts were recorded for SPN1120. Of those, the ratio of nonDE to DE students was 55/3 and all were TD students with no OnL students enrolled. During the Spring 2014 semester, 90 total artifacts were recorded for SPN1120. Of those, the ratio of nonDE to DE students was 82/8 and all were TD with no OnL students enrolled.

SPN1120 is scored using a rubric with five different sections, each with a different maximum. The maximum scores for sections (Sections I, II, III, etc.) are 15, 60, 15, 15, and 20 respectively, for a maximum score of 125. The average overall score for the 58 artifacts in Fall 2013 is 91.27/125, or 73.0% (Table 5). The average overall score for Spring 2014 is 99.17/125 or 79.3%. All rubric criteria exhibited increases in means from Fall to Spring although for significance tests see Section 2.1.3. In most cases,

the Spring 2014 artifacts reflect an increase in standard deviation (spread of data distribution) and kurtosis that is more leptokurtic (Figure 8).

Fall 2013: N = 58 Spring 2014: N = 90		I. Escuchar	II. Verbos	III. Lectura	IV. Nuestros	V. Expresion	Total
Fall 2013	max. possible score	*15*	*60*	*15*	*15*	*20*	*125*
	mean	12.12	40.40	11.95	10.64	16.17	12.12
	median	12.5	42	12	11	17	12.5
	mode	14	35	12	11	17	14
	standard deviation	2.31	10.18	2.07	3.44	2.98	2.31
	Kurtosis	0.88	-0.93	0.99	3.38	2.24	0.88
	Skewness	-1.01	-0.13	-1.00	-1.73	-1.56	-1.01
Spring 2014	mean	12.38	45.14	13.23	11.67	16.76	12.38
	median	14	48.5	14	12	17.5	14
	mode	15	58	15	15	20	15
	standard deviation	3.33	11.72	2.21	3.24	3.18	3.33
	Kurtosis	1.91	-0.34	1.31	2.23	2.22	1.91
	Skewness	-1.51	-0.76	-1.38	-1.31	-1.32	-1.51

Table 5. Basic descriptive statistics of SPN1120 artifacts for Fall 2013 (58 samples) and Spring 2014 (90 samples). Measured increases from Fall 2013 to Spring 2014 are denoted with shaded cell.

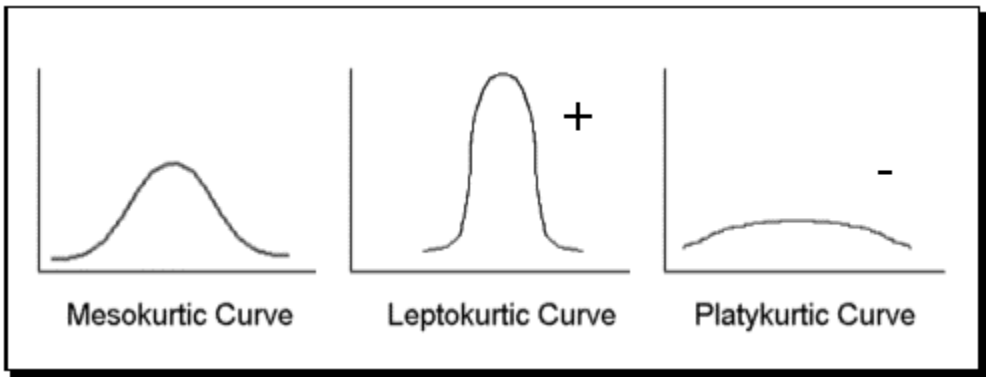

Figure 8. Example of kurtosis. The normal curve (left) has a kurtosis of 0.0. A positive value, or leptokurtic distribution (center), and negative value, or platykurtic distribution (right), are depicted here in an ideal scenario (Starkweather, 2010).

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.

In the Fall 2013 semester the average overall score for DE students (103.83/125) is substantially higher than that of nonDE students (90.59/125) (see Section 2.1.3 for details on significance) (Table 6). Most descriptive statistics for the DE artifacts are not useful as the sample size is only 3.

In the Spring 2014 semester the average overall score for DE students (105.13/125) is higher than that of nonDE students (99.17/125) (see Section 2.1.3 for details on significance) (Table 6). The increased standard deviation in this case may be a result of a small sample size for DE that includes outliers at an abnormally high percentage of the overall sample number.

Fall 2013: nonDE N = 55, DE = 3 Spring 2014: nonDE N = 82, DE = 8		I. Escuchar	II. Verbos	III. Lectura	IV. Nuestros	V. Expresion	Total
Fall 2013 nonDE/DE	max. possible score	*15*	*60*	*15*	*15*	*20*	*125*
	mean	12.06 / 13.17	39.96 / 48.33	11.92 / 12.50	10.59 / 11.50	16.05 / 18.33	90.59 / 103.83
	median	12.5 / 13	42 / 50	12 / 12	11 / 11	17 / 19	90.5 / 102
	mode	14 / na	35 / na	12 / na	10 / 11	17 / 19	96 / na
	standard deviation	2.35 / 1.26	10.14 / 8.62	2.11 / 1.32	3.52 / 0.87	3.01 / 1.15	17.15 / 9.88
	Kurtosis	0.73 / na	-0.93 / na	0.85 / na	3.04 / na	2.06 / na	-0.58 / na
	Skewness	-0.97 / 0.59	-0.09 / -0.84	-0.97 / 1.46	-1.67 / 1.73	-1.52 / -1.73	-0.30 / 0.81
Spring 2014 nonDE/DE	mean	12.34 / 12.81	44.74 / 49.25	13.27 / 12.81	11.57 / 12.69	16.68 / 17.56	98.59 / 105.13
	median	14 / 15	48.5 / 51.5	14 / 14.5	12 / 14	17.5 / 19	103.25 / 115.5
	mode	15 / 15	58 / 58	15 / 15	15 / 12	20 / 19	117 / 122
	standard deviation	3.28 / 4.05	11.76 / 11.25	2.08 / 3.46	3.24 / 3.33	3.18 / 3.29	19.94 / 24.32
	Kurtosis	2.35 / 0.03	-0.36 / 1.07	1.38 / -0.05	2.36 / 5.08	2.45 / 1.28	0.50 / 1.47
	Skewness	-1.57 / -1.44	-0.75 / -1.21	-1.33 / -1.38	-1.29 / -2.17	-1.34 / -1.53	-0.91 / -1.55

Table 6. Basic descriptive statistics of SPN1120 artifacts for Fall 2013 and Spring 2014 with respect to nonDE vs. DE students (N=55/3, N=82/8, respectively). Higher values for DE over nonDE denoted with shaded cell.

2.1.2.2.2 SPN1121

During the Fall 2013 semester 10 total artifacts were recorded for SPN1121. Of those the ratio of nonDE to DE students was 9/1 and all were TD students with no OnL students enrolled. During the Spring 2014 semester 115 total artifacts were recorded for SPN1121. Of those, the ratio of nonDE to DE students was 67/48 and all were TD with no OnL students enrolled.

SPN1121 is scored using a rubric with seven different sections, each with a different maximum. The maximum scores for sections (Sections I, II, III, etc.) are 15, 15, 40, 15, 12, 15, and 20 respectively for a maximum score of 132. The average overall score for the 10 artifacts in Fall 2013 is 95.95/132, or 72.7% (Table 7). The average overall score for Spring 2014 is 96.99/132 or 73.5%. Five of seven rubric criteria exhibited increases in means from Fall to Spring, although for significance tests see Section 2.1.3. In most cases the Spring 2014 artifacts reflect an increase in standard deviation (spread of data distribution) and kurtosis is more leptokurtic (see Figure 8). 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.

In the Fall 2013 semester only one DE artifact was recorded and so basic descriptive statistics are not calculated. The single artifact scored 115.5/132, which is distributed over the seven rubric criteria I-VII, as 13, 14, 38, 13, 9, 10, and 18.5.

Fall 2013: N = 10 Spring 2014: N = 115		I. Comprension Oral	II. Situaciones	III. Una Cita	IV. Preguntas	V. Comparaciones	VI. Lectura	VII. Expresion	Total
Fall 2013	**max. possible score**	***15***	***15***	***40***	***15***	***12***	***15***	***20***	***132***
	mean	11.50	9.45	34.20	9.50	7.50	9.60	14.20	95.95
	median	11.75	10.25	35.75	11.5	8	10	14.5	101
	mode	13	10.5	32	12	8	10	N/A	N/A
	standard deviation	1.41	3.36	5.26	4.25	1.27	2.72	4.32	18.82
	Kurtosis	-1.60	0.31	2.48	-0.23	0.25	-0.64	0.96	-1.65
	Skewness	-0.33	-0.78	-1.46	-0.98	-0.82	-0.39	-0.97	-0.30
Spring 2014	**mean**	12.27	9.62	32.32	11.38	5.70	10.34	15.37	96.99
	median	14	10.5	35	12	6	12	17	104
	mode	15	13	39	15	7	15	18	120
	standard deviation	3.14	4.84	7.41	3.50	3.01	4.56	5.39	23.57
	Kurtosis	1.39	0.51	3.54	2.03	0.19	-0.14	2.82	0.61
	Skewness	-1.38	-0.29	-1.66	-1.47	-0.05	-0.98	-1.92	-1.11

Table 7. Basic descriptive statistics of SPN1121 artifacts for Fall 2013 (10 samples) and Spring 2014 (115 samples). Measured increases from Fall 2013 to Spring 2014 denoted with shaded cell.

Spring 2014: nonDE N = 67 DE N = 48		I. Comprension Oral	II. Situaciones	III. Una Cita	IV. Preguntas	V. Comparaciones	VI. Lectura	VII. Expresion	Total
Spring 2014	**mean**	11.89 / 12.79	7.85 / 12.08	30.02 / 35.54	10.46 / 12.66	5.40 / 6.11	8.42 / 13.02	15.34 / 15.40	89.38 / 107.60
	median	13 / 14	8 / 13	31 / 37	11.5 / 13	6 / 7	8.5 / 14	16.5 / 17.5	91.5 / 112.5
	mode	14 / 15	0 / 14	31 / 39	12 / 15	8 / 7	12 / 15	18 / 20	85.5 / 120
	standard deviation	3.13 / 3.10	5.40 / 2.29	8.23 / 4.45	3.97 / 2.19	3.54 / 2.00	4.68 / 2.67	4.45 / 6.53	25.35 / 15.74
	Kurtosis	0.86 / 3.03	0.74 / -0.75	2.24 / 3.19	0.67 / -0.35	-0.41 / 1.77	-0.99 / 11.26	3.38 / 1.94	-0.22 / 3.84
	Skewness	-1.21 / -1.79	0.46 / -0.57	-1.34 / -1.67	-1.20 / -0.73	0.20 / -0.54	-0.42 / -2.83	-1.82 / -1.87	-0.71 / -1.78

Table 8. Basic descriptive statistics of SPN1121 artifacts for Spring 2014 with respect to nonDE vs. DE students (N=67/48, respectively). Higher values for DE over nonDE denoted with shaded cell.

In the Spring 2014 semester the average overall score for DE students (107.60/132) was substantially higher than that of nonDE students (89.38/132) (see Section 2.1.3 for details on significance) (Table 8). DE artifacts recorded higher mean scores for all rubric criteria. Additionally, DE artifacts recorded lower standard deviations and kurtosis for six of seven criteria which mean scores are narrowly distributed and largely fall into the same rubric scoring level consistently.

2.1.3 Significance Testing

Study goals demand that 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

course are shown in Tables 9 thru 12. 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.

2.1.3.1 French

2.1.3.1.1 FRE1120

With no FRE1120 course offered during Spring 2014 and no OnL artifacts, the significance tests were only conducted on nonDE vs. DE difference in mean scores of Fall 2013. The Welch's t-test results indicate that when comparing nonDE to DE students of the Fall 2013 semester Sections E, H, I, and J are significantly different (Table 9). However, the small sample size for DE artifacts has been shown to result in Type I errors (false positives) approximately 30% of the time for all statistically significant results (de Winter, 2013). Type II errors (false negatives) can also be of concern here. Therefore, we must bear this in mind when rejecting the null hypothesis that the difference in the means of nonDE and DE artifacts are equal to 0, and concluding this with a 95% confidence that the differences in scores are not solely due to chance. 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.

nonDE / DE: df = 15	A	B	C	D	E	F	G	H	I	J	Total
mean diff	0.04	0.36	2.51	0.20	3.91	-0.42	0.49	1.16	2.53	3.02	13.82
t_{crit}	2.13	2.13	2.13	2.13	2.13	2.13	2.13	2.13	2.13	2.13	2.13
t_{obs}	0.17	1.15	1.30	0.32	4.49	-0.46	0.73	2.93	2.43	4.68	2.79
p-value	0.864	0.274	0.221	0.759	$6.03x10^{-5}$*	0.658	0.484	0.010*	0.027*	$1.85x10^{-4}$*	0.014

*Table 9. Significance test of the difference in means of FRE1120 for nonDE vs. DE. Positive mean scores indicate DE > nonDE. *Denote marginal significance as defined by Johnson (2013).*

2.1.3.1.2 FRE1121

With no FRE1121 course offered during Fall 2013 and no OnL artifacts, the significance tests were only conducted on nonDE vs. DE difference in mean scores of Spring 2014. The Welch's t-test results indicate that when comparing nonDE to DE students of the Spring 2014 semester Sections II, III, and V are significantly different (Table 10). However, the small sample size for DE artifacts have been shown to result in Type I errors (false positives) approximately 20% of the time for all statistically significant results (de Winter, 2013). Type II errors (false negatives) can also be of concern here. Therefore, we must bear in mind when rejecting the null hypothesis that the difference in the means of nonDE and DE artifacts are equal to 0, concluding with a 95% confidence that the differences in scores are not solely due to chance. In 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.

nonDE / DE: df = 13	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	Total
Mean diff	-0.90	1.58	2.36	0.68	1.19	0.94	0.86	0.36	-0.34	-0.36	0.01	-0.79	0.61	6.19
t_{crit}	2.16	2.16	2.16	2.16	2.16	2.16	2.16	2.16	2.16	2.16	2.16	2.16	2.16	2.16
t_{obs}	-1.94	2.71	2.43	1.37	2.19	1.01	1.37	0.49	-0.43	-0.70	0.02	-0.44	0.83	1.00
p-value	*0.075*	*0.012**	*0.024**	*0.189*	*0.040**	*0.325*	*0.182*	*0.631*	*0.677*	*0.494*	*0.981*	*0.669*	*0.419*	*0.327*

*Table 10. Significance test of the difference in means of FRE1121 for nonDE vs. DE. Positive mean scores indicate DE > nonDE. *Denote marginal significance as defined by Johnson (2013) and de Winter (2013).*

2.1.3.2 Spanish

2.1.3.2.1 SPN1120

Significance tests were only conducted on the difference in mean scores of Fall 2013 to Spring 2014, and Spring 2014 nonDE-to-DE. There were no SPN1120 OnL artifacts in either Fall 2013 or Spring 2014, and while there were DE samples for Fall 2013, recent studies suggest significance testing may not be sufficiently accurate for this study (de Winter, 2013).

The Welch's t-test results indicate that when comparing nonDE to DE students of the Spring 2014 there is no significant difference in any rubric criterion or the overall score. (Table 11). 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 difference in means of the Fall 2013 to Spring 2014 artifacts indicate that for Sections II, III, and the overall score, we must reject the null hypothesis that the difference in the means of the two semesters' artifacts is 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 criteria 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 4 may be false positives (i.e. Type I errors). These marginal results, defined as those within the 95-99% confidence level, include only the overall score.

nonDE / DE: df = 8 F '13 / Sp '14: df = 145		I. Escuchar	II. Verbos	III. Lectura	IV. Nuestros	V. Expresion	Total
Spring 2014 nonDE / DE	**mean diff**	0.48	4.51	-0.46	1.11	0.89	6.53
	t_{crit}	2.30	2.30	2.30	2.30	2.30	2.30
	t_{obs}	0.32	1.08	-0.37	0.91	0.73	0.74
	p-value	0.755	0.310	0.725	0.390	0.486	0.483
F '13 / Sp '14	**mean diff**	0.26	4.74	1.28	1.03	0.58	7.90
	t_{crit}	1.98	1.98	1.98	1.98	1.98	1.98
	t_{obs}	0.56	2.61	3.57	1.83	1.13	2.55
	p-value	0.574	0.010	0.001	0.070	0.260	0.012*

*Table 11. Significance test of the difference in means of SPN1120 for nonDE vs. DE and Fall 2013 vs. Spring 2014. Positive mean scores indicate DE > nonDE, and Spring 2014 > Fall 2013. *Denote marginal significance as defined by Johnson (2013).*

2.1.3.2.2 SPN1121

Significance tests were only conducted on the difference in mean scores of Fall 2013 to Spring 2014, and Spring 2014 nonDE-to-DE. There were no SPN1121 OnL artifacts in either Fall 2013 or Spring 2014 and with only one DE sample for Fall 2013. Significance testing will not be reliable (de Winter, 2013).

The Welch's t-test results indicate that when comparing nonDE to DE students of the Spring 2014 semester for Sections II, III, IV, VI, and the overall score, we must reject the null hypothesis that the difference in the means of the two semesters' 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 criteria we cannot reject the null hypothesis, meaning differences in mean scores can be a result of chance.

nonDE / DE: df = 102 F '13 / Sp '14: df = 18		I. Comprension Oral	II. Situaciones	III. Una Cita	IV. Preguntas	V. Comparaciones	VI. Lectura	VII. Expresion	Total
Spring 2014 nonDE / DE	mean diff	0.90	4.23	5.52	2.19	0.72	4.60	0.05	18.21
	t_{crit}	1.98	1.98	1.98	1.98	1.98	1.98	1.98	1.98
	t_{obs}	1.53	5.73	4.62	3.79	1.38	6.67	0.05	4.74
	p-value	0.130	$1.17x10^{-7}$	$1.07 x10^{-5}$	$2.47 x10^{-4}$	0.170	$1.10 x10^{-9}$	0.962	$6.33 x10^{-6}$
F '13 / Sp '14	mean diff	0.77	0.17	-1.88	1.88	-1.80	0.74	1.17	1.04
	t_{crit}	2.10	2.10	2.10	2.10	2.10	2.10	2.10	2.10
	t_{obs}	1.44	0.15	-1.04	1.36	-3.69	0.77	0.80	0.16
	p-value	0.168	0.886	0.317	0.204	0.002	0.454	0.440	0.873

*Table 12. Significance test of the difference in means of SPN1121 for nonDE vs. DE and Fall 2013 vs. Spring 2014. Positive mean scores indicate DE > nonDE, and Spring 2014 > Fall 2013. *Denote marginal significance as defined by Johnson (2013) and de Winter (2013).*

The Welch's t-test results of the difference in means of the Fall 2013 to Spring 2014 artifacts indicate that for Sections V we must reject the null hypothesis that the difference in the means of the two semesters' 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 criteria we cannot reject the null hypothesis, meaning the differences in mean scores for those artifacts can be a result of chance.

2.2 Supporting Statistical Analyses

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, allowing assessment to drive instruction as defined by Elder and Paul (2007).

2.2.1 French

2.2.1.1 FRE1120

Figure 9 depicts the distribution of artifact scores for Fall 2013 based on 10 percentage point scoring bins down to less than 30. All except for one rubric criteria (a.k.a. topics or sections) exhibit modality (peak of the distribution) in either the 80-89% or >90% scoring bins. The exception is Section H which is trimodal (three peaks), centered in bins 30-39%, 50-59% and >90%. Based on descriptive statistics (see Section 2.1.2) this does not appear to be nonDE/DE related. The author suggests a norming session to examine any possible relationship to rater differences.

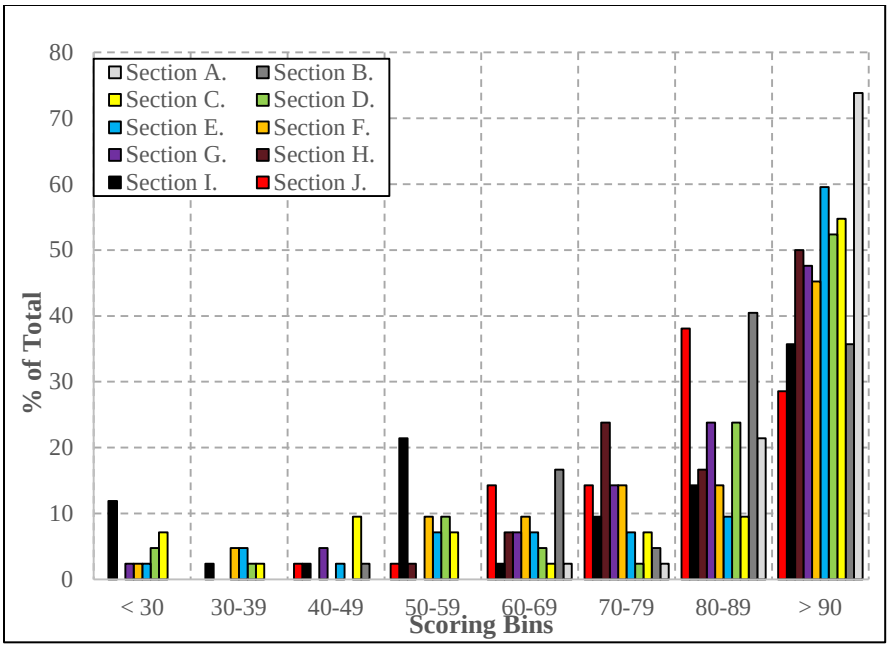

Figure 9. Histogram of Fall 2013 FRE1120 data distribution across 10% scoring bins.

2.2.1.2 FRE1121

Figure 10 depicts the distribution of artifact scores for Spring 2014 based on 10 percentage point scoring bins down to less than 30. All except for two rubric criteria (a.k.a. topics or sections) exhibit modality (peak of the distribution) in either the 80-89% or >90% scoring bins. The exceptions are Section IX, which is centered in bin 70-79%, and Section XI, which is bimodal and centered on 70-79% and >90%. Based on descriptive statistics it appears with Section IX there may be a weak correlation between nonDE and DE students although more study is needed.

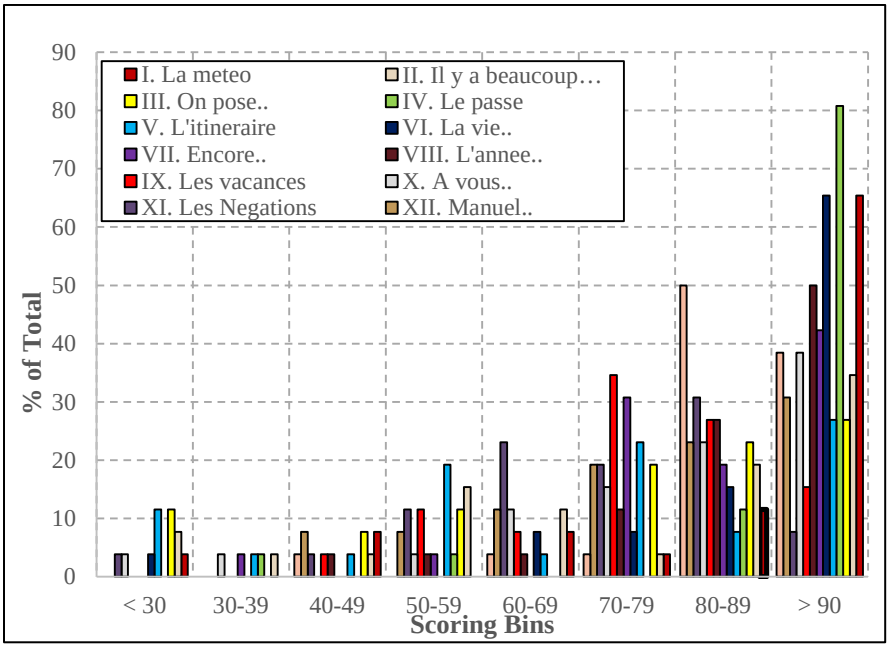

Figure 10. Histogram of Spring 2014 FRE1121 data distribution across 10% scoring bins.

2.2.2 Spanish

2.2.2.1 SPN1120

Figure 11 depicts the distribution of scores based on 10 percentage point scoring bins down to less than 30 comparing overall scores for artifacts from Fall 2013 and Spring 2014. Recall from Section 2.1.3 that the increase from Fall to Spring was marginally statistically significant and those results are evident in the shift of artifact distribution. Fall 2013 mode (peak) is centered in the 70-79% scoring bin, while Spring 2014 is centered in the >90% scoring bin.

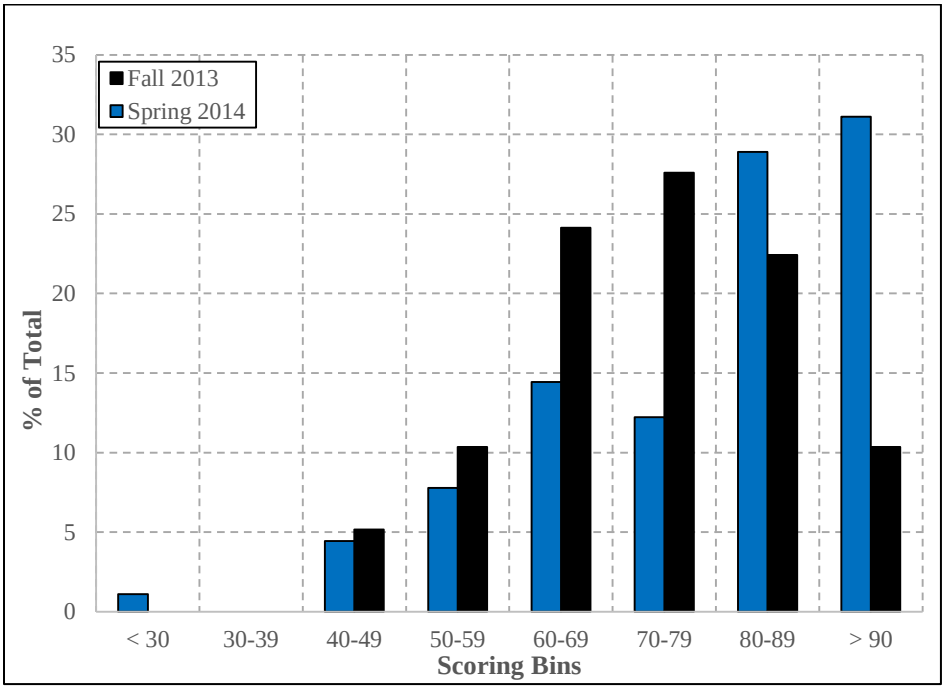

Figure 11. Histogram of SPN1120 for Fall 2013 (black) and Spring 2014 (blue) data distribution across 10% scoring bins.

Figure 12 depicts the distribution of artifact scores for Fall 2013 based on 10 percentage point scoring bins down to less than 30. All except for one rubric criteria (a.k.a. topics or sections) exhibit modality (peak of the distribution) in either the 80-89% or >90% scoring bins. The exception is Section II which is centered on the 70-79% bin.

Figure 13 depicts the distribution of artifact scores for Spring 2014 based on 10 percentage point scoring bins down to less than 30. All rubric criteria (a.k.a. topics or sections) exhibit modality (peak of the distribution) in the >90% scoring bin, which is a marked improvement over Fall 2013 artifacts. Note the negative skewness of the Spring 2014 data when compared with Fall 2013.

Section II, Verbos, remains the poorest scoring among all sections. Section II: Verbos includes more artifacts in bins 30-39%, 40-49%, 50-59%, and 60-69% than any other section in the artifact. Recall from Section 2.1.1 that Topic II: Verbos was the most disagreed upon by raters in the norming session. Raters exhibited potential for a scoring differential of a full letter grade or more when scoring this topic. This increased rater variability for Topic II may negate some of the results depicted.

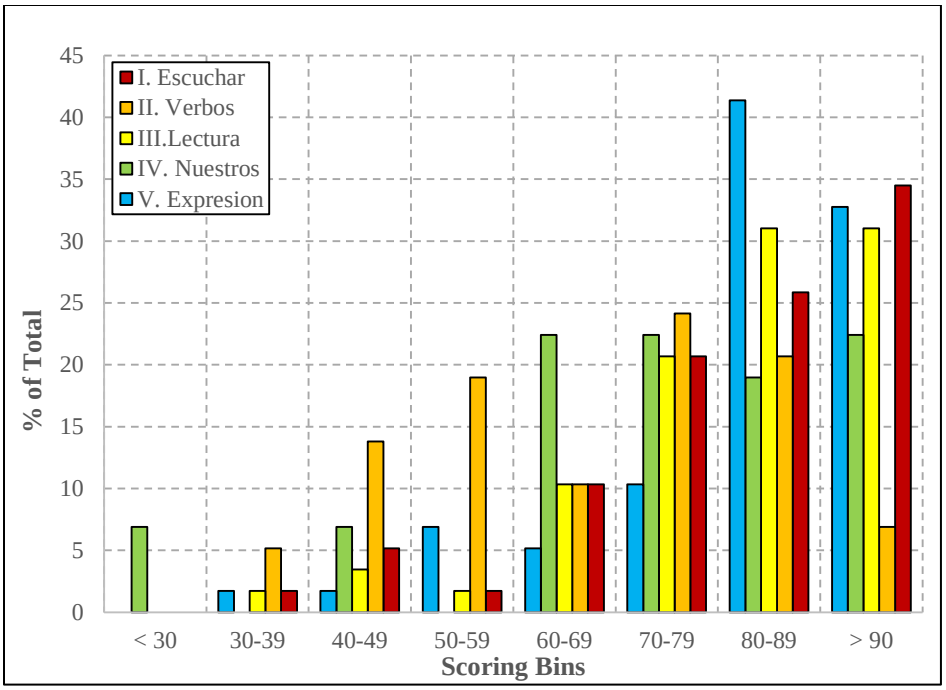

Figure 12. Histogram of Fall 2013 SPN1120 data distribution across 10% scoring bins.

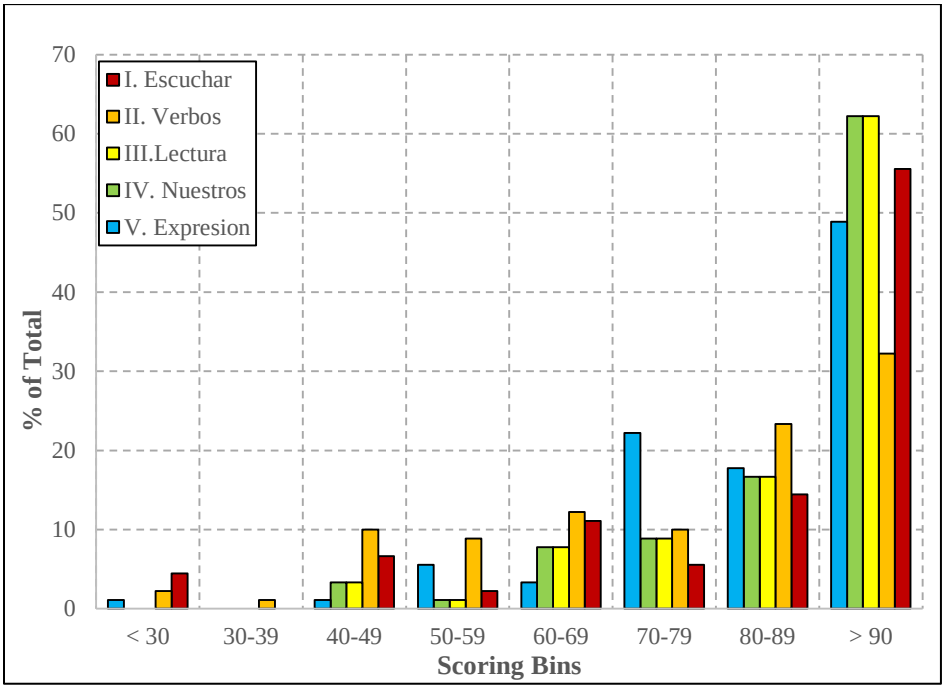

Figure 13. Histogram of Spring 2014 SPN1120 data distribution across 10% scoring bins.

2.2.2.2 SPN1121

Figure 14 depicts the distribution of scores based on 10 percentage point scoring bins down to less than 30 comparing overall scores for artifacts from Fall 2013 and Spring 2014. Recall from Section 2.1.2 that Fall 2013 data was limited to 10 artifacts. Any comparisons made from the distribution, like significance testing, are suspect.

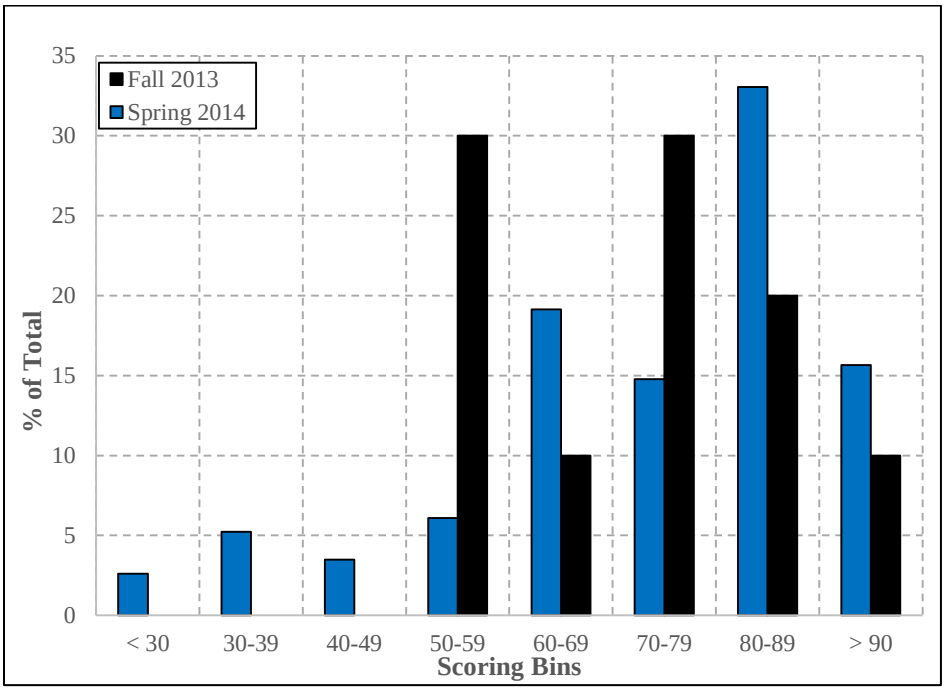

Figure 14. Histogram of SPN1121 for Fall 2013 (black) and Spring 2014 (blue) data distribution across 10% scoring bins.

Figure 15 depicts the distribution of artifact scores for Fall 2013 based on 10 percentage point scoring bins down to less than 30. Note the blocky appearance of data distribution within the scoring bins. This is the result of the paucity of artifacts for the Fall 2013 semester (only 10 samples). As a result, it is difficult to distinguish patterns but is included here in keeping with the comprehensiveness of the report.

Figure 16 depicts the distribution of artifact scores for Spring 2014 based on 10 percentage point scoring bins down to less than 30. Both Sections V and VI exhibit modalities centered on the 60-69% bin while the remaining sections are 80-89% and above. These topics standing apart from the rest are similar to the differences seen in SPN1120 with II: Verbos. With this in mind the author suggests a norming session for SPN1121 to examine any possible relationship to rater differences.

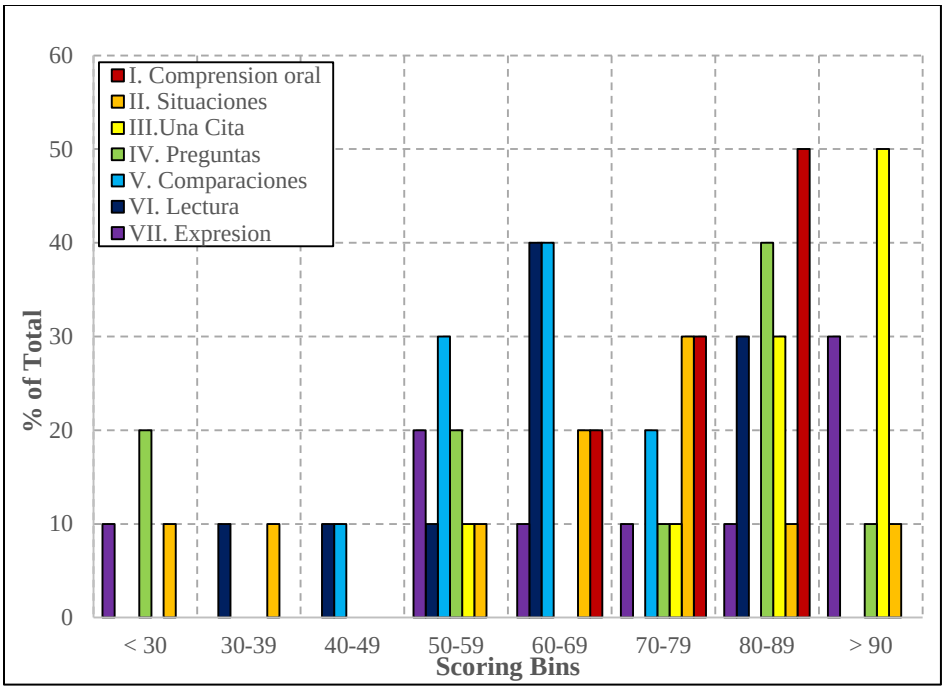

Figure 15. Histogram of Fall 2013 SPN1121 data distribution across 10% scoring bins.

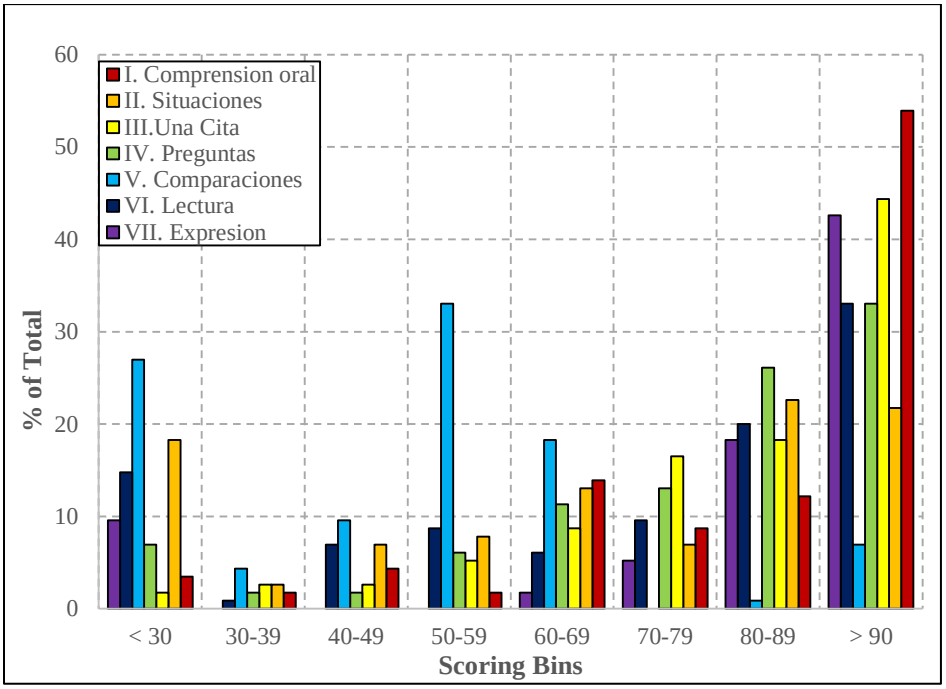

Figure 16. Histogram of Spring 2014 SPN1121 data distribution across 10% scoring bins.

3 CONCLUSIONS

Florida SouthWestern's Foreign Language Department employed common rubric elements used by all faculty for both Spanish and French introductory courses as a means to evaluate student progress and make informed comparisons between Dual Enrollment (DE) and non-Dual Enrollment (nonDE) students as well as Online (OnL) and Traditional (TD). Additionally, a norming session was conducted for Spanish only, to assess variation among scorers using common criteria.

3.1 FRENCH

No analysis on student progress across semesters could be conducted on FRE1120 or FRE1121 as each only were offered in Fall 2013, and Spring 2014 respectively. Descriptive statistics exhibit consistently higher scores for DE artifacts compared with nonDE artifacts in both FRE1120 and FRE1121.

The results of significance testing for FRE1120 find 4/10 criteria exhibited significantly higher mean scores among DE artifacts compared with nonDE. In FRE1121, 3/13 criteria exhibited significantly higher mean scores among DE artifacts compared with nonDE. However, in both cases significance tests are within a substantial margin of error due to the small sample size and minimal p-value (probability of significance) (de Witter, 2013; Johnson, 2013). When comparing artifact scores by rubric criteria exploratory analysis exhibits possible variability determined by the artifact status as DE or nonDE; although due to paucity of data, more study is needed.

3.2 SPANISH

Before assessment was engaged in Fall 2013, a norming analysis was conducted on SPN1120 artifacts by a subset of the Spanish faculty (Raters 1-5). Variability across the five rubric criteria ranged from 9% to 20% of individual criterion score. In rubric criteria Topic I, Raters 3 and 4 showed disagreement from the rater mean of 23% and -30% respectively. Topic II exhibited less agreement with lower artifact scores. The results support two possible interpretations: A) raters disagree on poor performance rubric definitions; or B) the rubric at the lower end of the scoring spectrum is not clear enough to effectively serve as a common assessment tool. These possibilities may also play a role in the differences among raters with Topic I as well. Topics III through V exhibited less distinction between raters. Rater 3 exhibited the largest difference from the rater mean over the five criteria, with inflated scores for one criteria, and deflated for two criteria.

The results of significance testing for SPN1120 find 2/5 criteria and the overall artifact score exhibited significantly higher mean scores for Spring 2014 compared with Fall 2013. No significant differences among mean scores were found when comparing DE with nonDE in Fall 2013. The results of significance testing for SPN1121 find only 1/7 criteria exhibited significantly higher mean scores for Spring 2014 compared with Fall 2013. Five of seven criteria and the overall artifact score exhibited significantly higher mean scores for DE compared with nonDE in Spring 2014

When comparing artifact scores by rubric criteria for SPN1120, exploratory analysis exhibits Section II: Verbos includes more artifacts in lower scoring bins compared with other rubric criteria which may be related to rater reliability. Similar distributions exist for both Sections V and VI in SPN1121. The author suggests a norming session for SPN1121 to examine any possible relationship to rater differences.

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.

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.

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.