Spring 2016
Assessment of Student Work from Across the Core:
Scoring & Results

Report prepared by the Office of Student Learning & Institutional Assessment
October 13, 2017

STEPHEN F. AUSTIN STATE UNIVERSITY

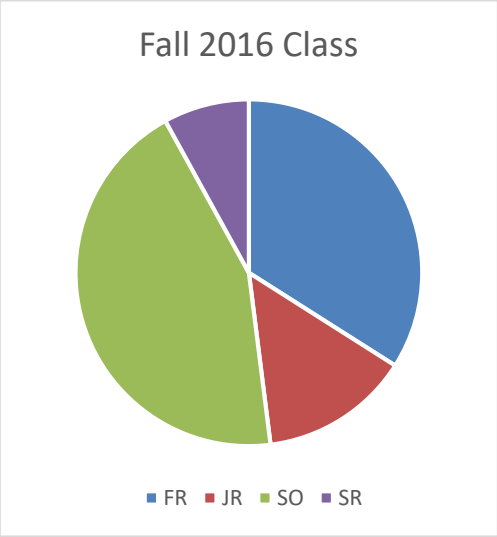
To assess the core objectives mandated by the Texas Higher Education Coordinating Board, Stephen F. Austin State University collected student work samples in core courses using LiveText. This report considers the Oral and Visual Communication samples, which were collected in Spring 2016 and Fall 2016.

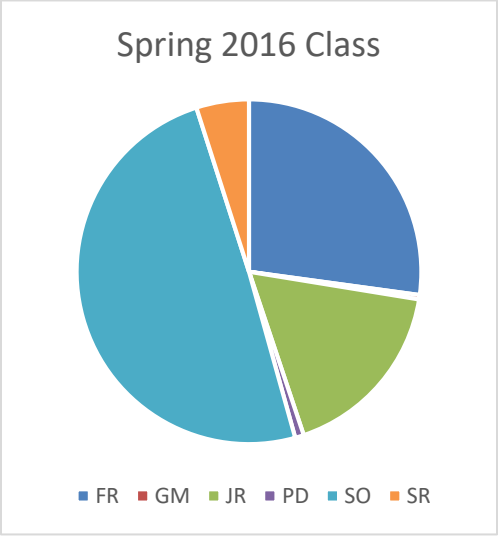
Method

Faculty members designed specific assignments for all related sections of courses designated "Core." Students then uploaded these assignments into the LiveText system online. From these collections, random samples were selected for review by a Core Curriculum Scoring Team.

Participants

The generated sample was similar to the overall SFA student population in terms of race and gender. The plurality of participants in the Fall 2016 semester were Sophomores, while the Spring 2016 class held a plurality of Freshmen. This may infer the plurality from *both* semesters emanated from the same entering class.





Section enrollments for the participating courses were larger in Spring 2016 when compared to Fall 2016. However, submission rates increased from one semester to the next, as indicated in Table 1.

Table 1: Course Enrollment and Submission Rates

	Spring 2016	Fall 2016
Enrollment	999	549
Submission Count	660	395
Submission Rate/Percentage	66.1%	71.9%

Scoring Team and Sampling

Student work was scored by teams of faculty who were nominated by their respective departments and then selected by the Core Curriculum Assessment Committee (CCAC). The team consisted of ten members drawn from departments teaching core courses in which core objectives were assessed.

Scoring Team members were asked to report any artifacts that did not match the assignment, were plagiarized, or contained no content. These artifacts were eliminated from the scoring sample. Because of the unique nature of these artifacts (student self-made video), a higher percentage of artifacts were unusable at first. Overall, 47 samples were deemed unusable in the Spring 2016 sample. Through improvements in communication and infrastructure, the situation improved. Only three samples were unusable in the Fall 2016 semester.

Rubric

The rubrics to assess each component of the core were developed by faculty teams who modified the Association of American Colleges and Universities (AAC&U) VALUES Rubrics. The AAC&U rubrics were adapted to best fit the objectives of the SFA core. The rubric for Oral

& Visual Communications can be found in Appendix A. Each rubric measures specific criteria using a 5-category continuum, labeled 0 - 4. For purposes of this report, the data has remained consistent with the rubric's scoring system. Benchmark labels are listed in Table 2.

Table 2: Rubric Category Scores and Corresponding Descriptions

Score	Correlation
0	Unacceptable
1	Beginning
2	Developing
3	Accomplished
4	Capstone

Rubric Calibration

In Fall 2016, each scoring team met for two rubric calibration sessions facilitated by the Office of Student Learning and Institutional Assessment. During these sessions, the team discussed the rubric extensively and developed rules for scoring student work. The calibration sessions were used to familiarize the faculty with the rubric that they would be using for scoring, allowing them to develop shared understanding of the language used on the rubric, and to become familiar with the process of scoring using LiveText. During the session, non-sample student artifacts were scored and discussed by the team. Further scoring rules were developed if needed following the scoring of each artifact.

Scoring

The LiveText sampling tool was used to draw a random sample of student work from each objective. The Spring 2016 sample (n = 223) was drawn with the intention of having a minimum of 200 pieces of scorable student work. This was keeping with previous practice. The Southern Association of Colleges and School Commission on Colleges (SACSCOC) recently imposed numerous sanctions on institutions based on sample size calculations. SACSCOC requires definitive reasoning behind any sampling presented to the Commission. Thus, changes were made to sampling procedures. Sample sizes were calculated with a confidence level of 80% and a margin of error of 10% using the following formula $Z^2*(p)*(1-p)/c^2$ where Z represents the Z value (in this case, 1.28), p is the population of submitted work in a specific core area, and c is the confidence interval (.1). This resulted in a sample size of 50 artifacts in the Fall of 2016.

Each artifact of student work in the sample was sent to two raters. Raters evaluated the paper in LiveText using an online copy of the rubric and following the rules developed in the calibration sessions. If the two raters had disagreement on a criterion, the artifact was then sent to a third rater to score only the criteria for which there was disagreement. A complete list of the rules for agreement/disagreement can be found in Appendix B. Faculty on the scoring teams were given two weeks to complete their first scoring round and then an additional week to finish their second round of scoring.

Results

Inter-rater agreement (within one point in each rating) was 91.5% for the Spring 2016 semester and 96.6% for the Fall 2016 semester. For those requiring a third rater, 59.5% needed a third rater for only one of the six elements being evaluated in the Spring. The same is true for 57.9% of the Fall scores.

Mean and mode are reported below for each rubric criterion (See Table 3 and Table 4). Frequency counts are illustrated through bar charts to assist with visualization and understanding. This is in keeping with admonishments from the Association of American Colleges & Universities:

> Do not, to the extent possible, show means in the absence of descriptive context as that reinforces the false notion of scale. As part of scorer training on the VALUE rubrics, individuals are "forced" to select a single performance level for each dimension. They must assign a student work product to a single, albeit ordered category of performance, not assign placement on a continuum or scale. Such ordinal data may be better described by medians, frequency distributions, and bar charts. Furthermore, this also implies that some statistical procedures may be more appropriate for analyzing the data generated from VALUE rubrics (e.g., analysis of variance, etc.) than others.
>
> Do not average the scores assigned to each dimension on a VALUE rubric to create a total score for the rubric. The power of the VALUE rubrics rests in the ability to focus attention on the specific learning addressed within each dimension; a total score for the rubric provides little diagnostic assistance to students or faculty. Furthermore, averaging across rubric dimensions makes methodological assumptions that are inappropriate when treating the VALUE data as ordinal.[1]

Tables 3 and 4: Oral and Visual Communication Means and Modes by Semester

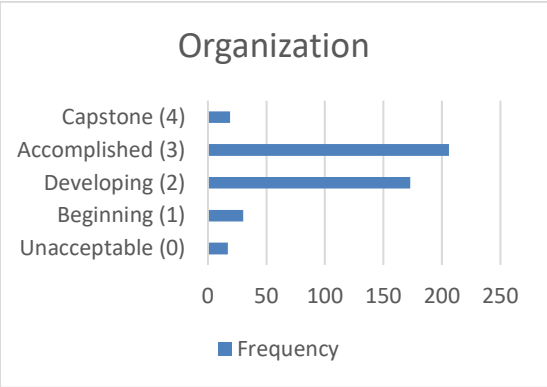
Spring 2016	Mean	Mode
Organization	2.40	3
Language	2.27	2
Delivery (oral/visual)	2.00	2
Evidence-based support	2.24	3
General purpose	2.48	3
Visual aids	1.85	2

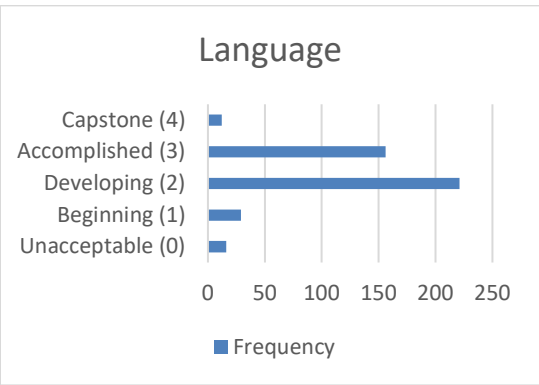
Fall 2016	Mean	Mode
Organization	2.42	3
Language	2.41	3
Delivery (oral/visual)	1.99	2
Evidence-based support	2.39	3
General purpose	2.75	3
Visual aids	2.09	3

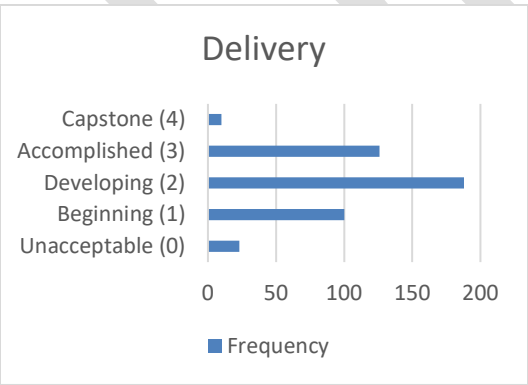
[1] On Solid Ground: VALUE Report 2017. Report. Association of American Colleges & Universities. Washington, DC, 2017. 28.

Frequency Counts: Oral and Visual Communication

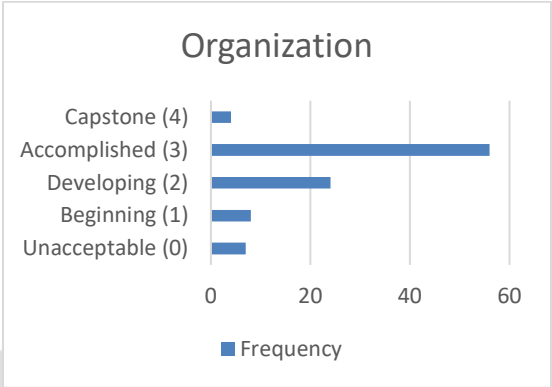
Spring 2016

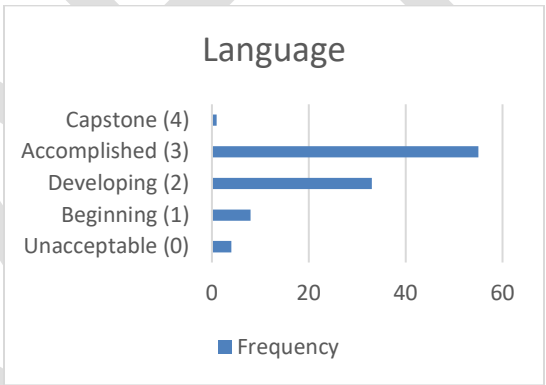


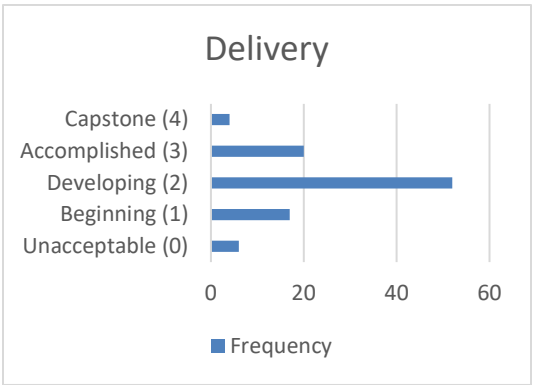


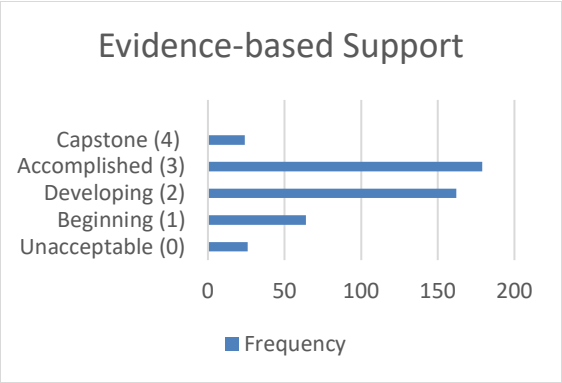


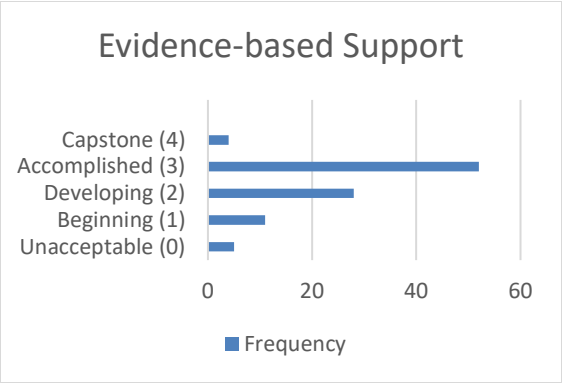
Fall 2016

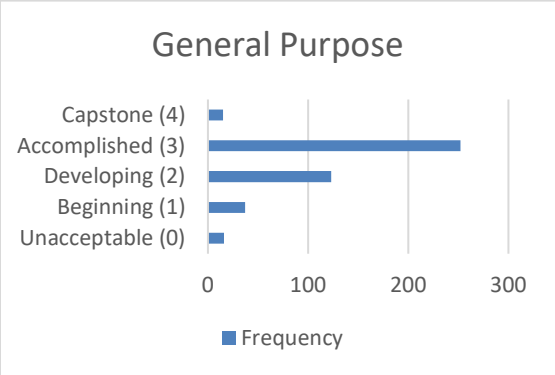


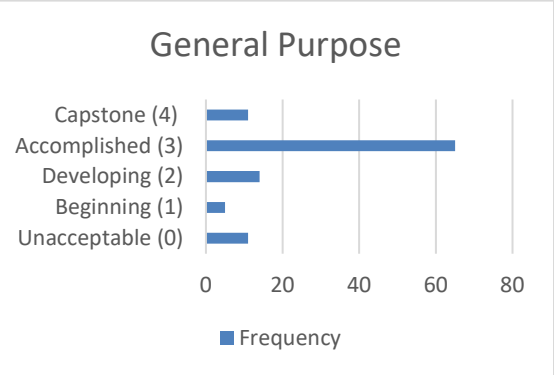


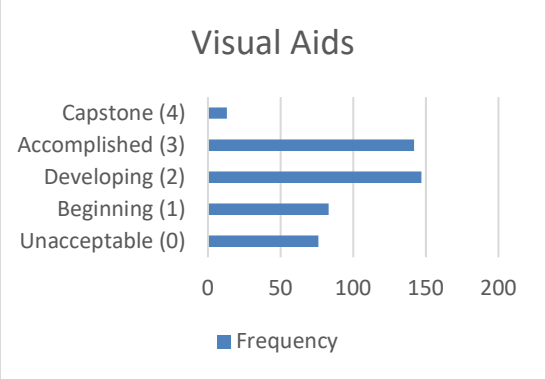


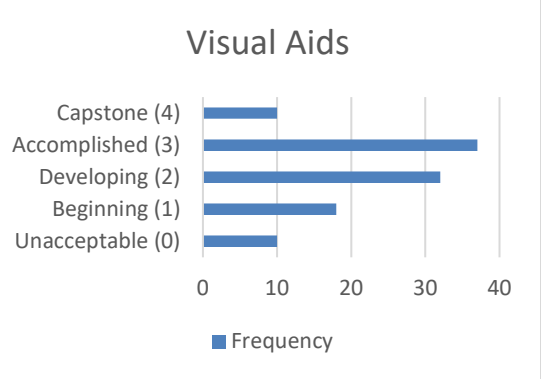












Scoring Team ratings generally followed similar patterns from one semester to the next. It should be noted that five of the six elements considered were highly correlated with each other, while the Visual Aids category showed moderate correlations. This may be due to the fact that a large number of Visual Aids ratings were zero (0), based on the lack of any visual aid,

whatsoever. The overall Cronbach's Alpha was .89. Table 5 indicates correlations between specific pairs of rubric elements.

Table 5

Inter-Item Correlation Matrix

	Organization	Language	Delivery	Evidence	Gen.Purp	VisAids
Organization	1.000	0.751	0.678	0.622	0.758	0.462
Language	0.751	1.000	0.695	0.626	0.702	0.399
Delivery	0.678	0.695	1.000	0.530	0.653	0.511
Evidence	0.622	0.626	0.530	1.000	0.646	0.487
General Purpose	0.758	0.702	0.653	0.646	1.000	0.488
Visual Aids	0.462	0.399	0.511	0.487	0.488	1.000

Although VALUE rubrics create ordinal and categorical data, mean averages of each element indicated an *increase in scores* from 2014 to 2016 (refer to Table 3 and Table 4). Mann-Whitney U analysis of scores is documented in Table 6. Analysis indicated statistically significant differences between semesters for two of the six elements. Language and General Purpose.

Table 6: Mann-Whitney U Comparison (Oral and Visual Communication)

	Organization	Language	Delivery	Evidence	Gen. Purpose	Visual Aids
Mann-Whitney U	20705.500	19078.500	21842.000	20479.000	17629.000	20865.500
Wilcoxon W	119940.500	113473.500	26792.000	124219.000	115975.000	127356.500
Z	-1.020	-2.225	-.213	-1.663	-3.449	-1.849
Asymp. Sig. (2-tailed)	.308	.026	.831	.096	.001	.064

Grouping Variable: Semester

One interesting change could be the Language element. Spring 2016 students were listed primarily as Developing (2); Fall 2016 students tended to be rated as Accomplished (3).

Visual Aid usage also was rated higher in the Fall, with a smaller percentage being rated as Unacceptable (a drop from 16.5% to 9.8%). There were slightly more students rated as Accomplished in the Fall, while the Spring sample indicated more students at the Developing level. The drop in Unacceptable markings likely accounts for the difference in ratings between Fall and Spring of 2016. The Fall semester used video artifacts from only one course, while the previous scoring sample included scores from multiple courses. One potential effect could be that General Purpose may have been easier to ascertain by Scoring Team members. This singular structural change may answer most of the score increase in this element.

These three elements indicate statistically significant changes; however, the *real* change in mean scores for the three elements ranged from .14 to .27. As Hilda Bastian wrote for the *Scientific American*,

> Statistical significance testing can easily sound as though it sorts the wheat from the chaff, telling you what's "true" and what isn't. But it can't do that on its own. What's more, "significant" doesn't mean it's important either. A sliver of an effect can reach the less-than-5% threshold.[2]

Moving Forward

Following each semester's artifact assessment, a debrief meeting was held with the Oral and Visual Scoring Team. At the end of the spring semester, team members noted their overall feelings on SFA students' oral and visual communication capabilities. The consensus was four words, "We're in good shape."

[2] Hilda Bastian, "Statistical significance and its part in science downfalls," *Absolute Maybe*, Scientific American, November 11, 2013, https://blogs.scientificamerican.com/absolutely-maybe/statistical-significance-and-its-part-in-science-downfalls/

As Linda Suskie recently posted, "Decisions are made with some level of uncertainty. Assessment results should reduce uncertainty but won't eliminate it."[3] While these rubric data are more descriptive in nature, some general concepts can be considered:

1. Students who begin their core are typically rated as at least Developing in their level of work.
2. When the Visual Aid factor is removed, Delivery seems to be the most challenging Oral Communication element for SFA students.
3. All assignments used in scoring likely need to be graded assignments. This keeps the spirit and effectiveness of the SFA VALUE rubrics.
4. SFA students may need more specific instruction on the use of Visual Aids.

[3] Linda Suskie, How to Assess Anything without Killing Yourself…Really, *Linda Suskie Blog*, May 30, 2017, http://www.lindasuskie.com/apps/blog/show/44560748-how-to-assess-anything-without-killing-yourself-really-

Appendix A: Oral and Visual Communication Rubric

	Capstone 4	Accomplished 3	Developing 2	Beginning 1	Unacceptable 0
Organization	Organizational development is clearly and consistently observable; skillfully makes content and expression of ideas in the presentation cohesive.	Organizational development and expression of ideas are clearly and consistently observable within the presentation; content is expressed reasonably well as a result.	Organizational development and expression of ideas are observable within the presentation.	Organizational development and expression of ideas are occasionally observable	Organizational development and/or expression of ideas are not observable within the presentation; lack of coherence and unity exist.
Language	Language choices are imaginative, memorable, and compelling; choices enhance presentation effectiveness. Language is appropriate to audience and aids the clear expression of ideas.	Language choices are thoughtful and generally support the effectiveness of the presentation. Language is appropriate to audience and is useful to the expression of ideas.	Language choices are mundane and commonplace and partially support the effectiveness of the presentation and the expression of ideas.	Language choices are sometimes unclear and minimally support the effectiveness of the presentation. Language appropriateness is inconsistent. Expression of ideas is hindered.	Language choices are unclear and fail to support the effectiveness of the presentation. Language is not appropriate to audience; ideas are not expressed clearly.
Delivery (Oral/Visual)	Delivery techniques make the presentation compelling; speaker appears polished and confident; speaker energy and emphases foster interpretation of ideas expressed. Dependency upon notes, if applicable, is not evident or intrusive. Non-verbal cues aid significantly.	Delivery techniques make the presentation interesting, and speaker appears comfortable; speaker tends toward conversational tone, and dependency upon notes is minimally noticeable. Non-verbal cues are appropriate and useful.	Delivery techniques make the presentation understandable; speaker appears tentative; speaker tends to be a bit casual, as evidenced in word choices; non-verbal cues do not particularly elevate audience's level of understanding or interpretation.	Delivery techniques sometimes detract from audience comprehension; speaker appears uncomfortable; speaker seems unenthusiastic, monotonic, or hesitancies suggest unpreparedness. Verbal cues include unnecessary gestures and purposeless body language.	Delivery techniques are either distracting from understandability of the presentation or fail to be effective; the speaker is clearly uncomfortable or unprepared.
Evidence-Based Support	Supporting materials make appropriate reference to information or analysis and significantly enhance development; materials establish presenter's credibility/authority.	Supporting materials make appropriate reference to information or analysis and generally supports development; presenter's credibility/authority is clear but evidence-based support could be stronger.	Supporting materials make appropriate reference to information or analysis but only partially fosters development and presentation of ideas. Presenter's credibility/authority could benefit from more careful exploration of evidence.	Insufficient supporting materials provide minimal information or analysis; presenter's credibility/authority on the topic is not particularly clear.	Supporting materials are virtually non-existent, or the supporting materials are not credible.
General Purpose	Purpose is compelling, precisely stated, appropriately repeated, memorable, and strongly supported. Purpose and evidence are aligned well.	Purpose is clear and consistent; purpose and evidence are appropriately aligned.	Purpose is understandable but is neither reinforced nor memorable; purpose and evidence are generally aligned.	Purpose can be deduced, but is not explicitly stated in the presentation. Alignment of purpose and evidence is not always clear.	Purpose is absent; the presentation does not seem to know what it is about. Unifying principles do not exist.
Visual Aids	Visual aids effectively support the communication of purposes and ideas; aids are integrated into the presentation seamlessly, thus fostering a full understanding of the message's content.	Visual aids generally support communication of the student's ideas and purposes; the aids effectively amplify or resonate the presentation of ideas and foster understanding.	Visual aids support the communication of the student's ideas and purposes but are only partially useful or informative.	Visual aids do not particularly support the communication of the student's ideas and purpose; they are insufficient to be of much use as they do little to elevate understanding.	Visual aids are virtually non-existent, serve no purpose, or are not credible

Appendix B: Rules for Scoring Student Work

Procedures for assessment of student work:

1. Each piece of student work will be initially assessed by two raters.

2. If the two raters agree on their rating on any element/criterion of a rubric then there is no need for a third rater on that element/criterion.

3. If the first two raters are no more than one integer apart on their ratings on an element/criterion of a rubric, then there is no need for a third rater on that element/criterion.

   For example, if Rater A gives a piece of student work a 2 on element/criterion of Audience, Context, and Purpose, and Rater B gives the piece of student work a 3 on Audience, Context, and Purpose, then the two ratings are averaged together to give a 2.5 on the Audience, Context, and Purpose element/criterion. If the two raters are more than one integer apart on their ratings on any element/criterion of a rubric, a third rater is asked to rate only the element(s)/criteria where there was disagreement.

For example, if Rater A gives a piece of student work a 1 on the element/criterion Audience, Context, and Purpose, and Rater B gives the piece of student work a 3 on Audience, Context, and Purpose. Also, rater A also gives the same piece of student work a 4 on Sources and Evidence, and Rater B gives that same piece of student work a 2. Then a third rater (Rater C) is asked to rate the student work only on the elements/criteria of Audience, Context, and Purpose and Sources and Evidence.

4. If Rater C's rating agrees with one of the other two ratings, then that rating is used and the rating that is not in agreement is discarded.

   For example, if Rater C and Rater A each rate a piece of student work a 2 on Content Development, but Rater B rates the work a 4, then Rater B's rating is discarded and the student work received a rating of 2 on Content Development.

5. If Rater C's rating does not agree with one of the other two ratings, and is no more than one integer from only one of the other ratings, then the rating that is more than one integer from the other ratings is discarded, and the two ratings that are no more than one integer apart are averaged.

For example, if Rater C rates a piece of student work 2, Rater A rated the work a 1, and Rater B rated the work 4 on Content Development. Rater B's rating of 4 is discarded and the ratings of Rater C and Rater A are averaged to get a rating of 1.5.

6. If Rater C's rating is no more than one integer from the other two ratings, then all of the ratings are averaged.

For example, if Rater C rates a piece of student work 3, Rater A rated the work a 2, and Rater B rated the work 4 on Content Development. All of the ratings are averaged for a rating of 3.

7. If Rater C's rating does not agree with one of the other two ratings and is more than one integer apart from the other two ratings, then Rater C's rating is discarded, and the other two ratings are averaged.

   For example, if Rater C rates a piece of student work 4, Rater A rated the work a 0, and Rater B rated the work a 2 on Content Development. Rater C's rating of 4 is discarded, and the other two ratings are averaged to get a rating of 1.