Report on COAT Interrater Reliability

August 10, 2020

Prepared by Beenah Moshay

Background

The Core Objective Assessment Team (COAT) at Collin College has been meeting for several years to address the Southern Association of Colleges and Schools Commission on Colleges (SACSCOC) and Texas Higher Education Coordinating Board (THECB) continuous quality improvement guidelines. The guidelines state that General Education or Core educational courses should conduct regular assessments of student performance on selected student learning outcomes (SLOs) to ascertain the extent to which students have gained knowledge of certain topics and to establish targets for continuous improvement.

Based on a three-year cyclical schedule, all Core courses offered at the College assess targeted Student Learning Outcomes or SLOs on a variety of student artifacts. At the end of the fall and spring semesters, all sections of the slated courses submit student artifacts to be assessed by a team of faculty raters. A random sample of student artifacts is selected and reviewed by a fixed number of faculty members. In academic year 2019-20, each student artifact received two ratings by two different raters using the rubric corresponding to the relevant Core objective being evaluated. This year (2019-20), two learning domains, Communication and Teamwork were assessed. For the purpose of this report, only the Communication SLOs will be discussed as Teamwork assessments were conducted by the students themselves. For each Communication SLO students were categorized in two groups based on the number of credit hours completed. The two groups were students who had completed 12 to 15 credit hours (representing those in workforce programs who may only earn certificates) and students who had completed 30 or more credit hours (representing associate degree earners or transfer students). Thus, in total, there were two groups of student artifacts that were evaluated (page 2 lists the 2 groups).

It is important to note that in 2019-20, instead of faculty meeting in person and rating artifacts together in a large room over a two-day period, all faculty rated artifacts separately as well as remotely (due to the COVID-19 pandemic, large in-person meetings were discouraged). Additionally, a new mode of training for faculty assessors was adopted. Rather than an in-person inter-rater reliability training as had historically been done, faculty assessors were asked to watch a video and complete a training module where they practiced rating student artifacts. Faculty members who did not achieve a minimum score of 70% in the initial training module's rating assessment were given the opportunity to rate another set of artifacts. If they did not achieve a minimum score of 70% the second time, faculty were still allowed to participate in the actual assessment of student artifacts. At this time, it is unclear how the two changes noted above might impact the results of the study.

Methodology

To assess the interrater reliability for the 2020 COAT review of student artifacts, Intraclass Correlation Coefficient or ICC was utilized using SPSS Statistical version 25. The confidence interval was set for 95%, analyzing the level of agreement for the average ratings by the two raters for each artifact. The model specified testing for absolute agreement between the raters (how close the raters were to assigning the same scores based on the rubrics provided by COAT), using a two-way mixed effects model as the raters were fixed (consistent group of raters for all artifacts from the population of raters) and random sampling was used to select artifacts for measurement. Koo and Li (2016) suggested that any time ratings from two raters are being measured (as well as test-retest) the above methodology is the appropriate methodology. Listwise deletion was used in selecting cases, meaning only cases that had both average ratings from faculty were utilized in the analysis. All data used in this report were provided by COAT.

There were approximately 300 to just over 400 artifacts rated in each of the following two categories (for a total over 700 artifacts) based on credit hours earned by students in general education courses:

(1) Communication— students who had earned between 12-15 credit hours;
(2) Communication —students who had earned 30 or more credit hours;

Below are the three domains for which each group of the students were rated:

1. **Development:** The student organizes content support of a central idea.
2. **Expression:** The student shows appropriate awareness of intended audience, adjusting the subject matter, syntax, and mechanics of the product.
3. **Interpretation:** The student uses relevant content that conveys understanding of the subject matter.

Analysis

As part of the analysis, Cronbach’s alpha was computed and is presented in Tables 1 and 2 below. Cronbach’s alpha is a measure of internal consistency or reliability to determine how much the items on the scale measure the underlying dimension or construct. Its value ranges between 0 and 1. The closer the value is to 1, the higher the amount of consistency or reliability indicating that the average of the items (in this case the Learning Outcomes) together are measuring the same thing. For all of the items in this report Cronbach’s alpha is relatively high; .821 for Communication Skills (12 to 15 credits) in Table 1 and .855 in Table 2 Communication Skills (30-Plus credits) indicating that over 80% of the variability in scores is captured by the constructs. So overall, both SLOs demonstrated good levels of reliability for the average measures.

Table 1

Cronbach’s Alpha for Communication Skills: Students Who Completed 12 to 15 Credits

Cronbach's Alpha Based on Standardized Items	Number of Items
.821	6

Table 2

Cronbach’s Alpha for Communication Skills: Students Who Completed 30-Plus Credits

Reliability Statistics

Cronbach's Alpha Based on Standardized Items	Number of Items
.855	6

As with Cronbach’s Alpha, ICC between raters could range from 0 to 1, with 1 meaning there is perfect reliability or correlation (meaning all of the raters’ scores were in absolute agreement) and 0 meaning the raters’ scores showed no correlation or agreement between raters. The confidence interval simply tells us that if the correlation coefficient falls within the lower and upper bounds of the confidence interval, we can have, in this case, 95% confidence that the true correlation coefficient is likely to fall between the two values. The smaller the range between the confidence intervals' upper and lower bounds the more confidence we can have in the results, as this indicates higher levels of agreement between raters. In Tables 3 and 4 below, the correlation coefficient's strength (meaning the correlation between raters) in the row entitled "Average Measures" can be interpreted as follows: 0.5 or below is poor correlation, 0.5 to 0.75 is moderate, 0.75-0.9 is good, and 0.9 and above is considered excellent. The correlation coefficient should fall within the upper and lower bounds of the confidence interval. For example, in Table 3 below (Communication Skills 12-15 Credits), the ICC is .819 with a confidence interval ranging between .791 and .845 which means that there is 95% chance that the true ICC value is between .791 and .845. Therefore, we infer that the correlation/agreement between raters was "good" based on the confidence interval range for that correlation coefficient. In general, a lower bound confidence interval of at least .700 is preferable to indicate high interrater reliability. Additionally, the lower and upper confidence bounds had relatively low ranges, suggesting that there was high correlation between scores. In Table 4 (Communication Skills 30-Plus Credits) the average measures indicated good correlation (meaning the raters scores were similar) with an ICC of .855 and a confidence interval ranging between .826 and .876, we can conclude that the correlation or level of agreement between raters was "good" as well, and slightly stronger than those for Communication Skills 12-15 Credits.

Table 3

Intraclass Correlation Coefficient Communication Skills: Students Who Have 12 to 15 Credits

		95% Confidence Interval	
	Intraclass Correlation[b]	Lower Bound	Upper Bound
Single Measures	.431[a]	.387	.476
Average Measures	.819[c]	.791	.845

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Table 4

Intraclass Correlation Coefficient Communication Skills: Students Who Have 30-Plus Credits

		95% Confidence Interval	
	Intraclass Correlation[b]	Lower Bound	Upper Bound
Single Measures	.491[a]	.442	.541
Average Measures	.853[c]	.826	.876

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

In Tables 5 and 6 below, descriptive statistics of the learning outcomes are provided. They include the average scores for each domain (ratings ranged between 1 and 4, where 1 equals “Does Not Meet Expectations” and 4 equals “Exceeds Expectations”). Each mean was above 2.5 for the domains and the standard deviation or the average variation of scores from the mean ranged from a low of .779 in Table 5 to a high of .925 in Table 6. The standard deviation can be interpreted as relatively high, indicating that there is much variation in scores (on average the scores differ from the mean by almost 1 point). The lower the standard deviation, the more confidence we can have in our results.

Table 5

Item Statistics for Communication Skills: Students Who Completed 12-15 Credits

Learning Outcome*	Mean	Std. Deviation	Number of Artifacts
Development[1]	2.76	.890	416
Expression[1]	2.76	.833	416
Interpretation[1]	2.88	.813	416
Development[2]	2.78	.859	416
Expression[2]	2.77	.812	416
Interpretation[2]	2.88	.779	416

*Note. Superscript 1 denotes rater 1 and superscript 2 denotes rater 2.

Table 6

Item Statistics for Communication Skills: Students Who Completed 30-Plus Credits

Learning Outcome*	Mean	Std. Deviation	Number of Artifacts
Development[1]	2.73	.890	324
Expression[1]	2.69	.854	324
Interpretation[1]	2.85	.841	324
Development[2]	2.59	.925	324
Expression[2]	2.78	.890	324
Interpretation[2]	2.82	.874	324

*Note. Superscript 1 denotes rater 1 and superscript 2 denotes rater 2.

Conclusion

As noted earlier, while the average measures of the ICC for each learning objective was relatively high which meant that there was high level of agreement between the scoring of two raters, of interest were the **single measures**, which looked at the ICC of non-averaged scores or each score the raters gave, individually. The single measures numbers indicated low ICC, at .431 in Table 3 and .491 in Table 4. These results contradict the average scores result as they indicate weak ICC between the single measures. However, as indicated by the literature, it is perhaps best to compare, over time, the ICC results of COAT outcomes against themselves rather than to use correlation benchmarks found in the literature. Thus, we can cautiously conclude that interrater reliability in the academic year 2019-20 appears to be high. Due to the unusual nature of this academic year with faculty rating artifacts online rather than in person and the change in training for raters, it is difficult to ascertain the impact these circumstances may have had on this year's results.

References

Koo, T.K, & Li, M.Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine,* 15(2):155-163.

Kwet, K.L. (2014). Handbook of Interrater Reliability (Fourth Edition). Gaithersburg, Maryland: Advanced Analytics