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 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 sample of student artifacts is selected and reviewed by a number of faculty raters. In previous years, each student artifact received two ratings by two different raters; however, in spring 2018, due to a large increase in the number of faculty raters, it was decided that each student artifact receive three rating by three different raters for each corresponding rubric (core objective).

In previous reports, where two raters submitted scores for randomly selected artifacts, a measure of percent of agreement between raters was utilized to assess inter-rater reliability. While this approach is commonly utilized, it has been criticized because it does not take into account agreement between raters that would occur by chance, and thus it is argued that it *over estimates* reliability measures between raters. Additionally, when more than two raters are involved, percent agreement is not recommended to assess reliability. In 2018, with the introduction of a third rater, a more robust measure of inter-rater reliability using Intraclass Correlation (ICC) was recommended. ICC measures the level of agreement or correlation between rater's measurements. This type of measurement tells the researcher how much conformity or agreement there is between raters who rate the same item. However, it is noted that this method also has drawbacks. According to the literature, under the current conditions, Fleiss kappa, an alternate form of inter-rater reliability is indicated. However, the literature also indicated that Fleiss kappa may be difficult to interpret and results obtained can vary widely dependent on the type of study being conducted. With the dearth of information being available on appropriate interpretations of Fleiss' kappa, and its associated pitfalls, the report writer opted to use a less ideal (under these circumstances) inter-rater reliability measure, the ICC. ICC is recommended to assess reliability for categorical data with two or more raters.

Methodology

To assess the inter-rater reliability for the annual COAT review of student artifacts, we utilized version 25 of SPSS, Intra Class Correlation. The confidence interval was set for 95%, analyzing the level of agreement for the average ratings by the three raters for each artifact. We specified testing for absolute agreement between the raters (how close the raters were to assigning the same scores), using a two-way mixed effects model as the raters were fixed (in this case fixed refers to the fact that the results are not meant to be generalized to the population of raters as a

whole) and random sampling was used to select artifacts for measurement. Koo and Li (2016) suggested that any time inter-rater reliability is 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 all three average ratings from faculty were utilized in the analysis. All data utilized in this report were provided by COAT.

There were approximately 200 artifacts rated in each of four different categories (for a total over 800 artifacts): Social Responsibility students who had earned between 12-15 credit hours, and Social Responsibility for those who had earned 30 or more credit hours, Critical Thinking, students earning 12-15 credit hours and Critical Thinking for students who had earned 30 plus credit hours. The 12-15 credit hour groups were thought to best approximate workforce education concentrators, whereas the 30 plus groups were thought to be representative of the academic degree concentrators and transfer group. The raters utilized rubrics where artifact scores could range from a low of 1 (does not meet expectations) to a high of 4 (exceeds expectations) across five categories for critical thinking and three categories for social responsibility. The scores were averaged out across each rater and then each artifact, with an average overall score across all four groups at or near 2.00 (reported in Tables 1-4 below).

Analysis

Because inter-rater reliability is an estimate of the true or actual reliability measure, we utilize the intraclass correlation coefficient along with the confidence interval to assess how reliable or consistent the ratings were overall. In this case, correlations between raters could range from 0 to 1, with 1 meaning there is perfect reliability or correlation (meaning all of the rater's scores were in absolute agreement) and 0 meaning the rater's 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 correlation coefficient is representative of the true results. The smaller the range between the confidence interval upper and lower bounds, the more confidence we can have in the results, as this indicates higher levels of agreement between raters. In the tables below, the correlation coefficient's strength (meaning the correlation between raters) in the row entitled "Average Measures" and 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 1 below, with a correlation of .701, and a confidence interval ranging between .624 and .764, we would say that the correlation/agreement between raters was from moderate to good based on the confidence interval range for that correlation coefficient.

We can see that the ICC for Table 1 Critical Thinking, 12 to 15 Credits, Table 3 Social Responsibility 12 to 15 Credits, and Table 4 Social Responsibility, 30 plus credit hours and their associated confidence intervals all range between a low of .612 and a high of .772. Therefore, for these three groups of ratings we would rate the inter-rater reliability as moderate to good. In Table 2 Critical Thinking 30 plus credit hours, the confidence interval for the coefficient ranges between .581 and .735, indicating a rating of moderate strength.

Table 1

Intraclass Correlation Coefficient for Critical Thinking, 12 to 15 Credits

		95% Confidence Interval	
	Intraclass Correlation	Lower Bound	Upper Bound
Average Measures	.701	.624	.764

n=213, mean=2.36

Table 2

Intraclass Correlation Coefficient, Critical Thinking 30 plus Credits

		95% Confidence Interval	
	Intraclass Correlation	Lower Bound	Upper Bound
Average Measures	.665	.581	.735

n=220, mean=2.37

Table 3

Intraclass Correlation Coefficient, Social Responsibility, 12 to 15 credits

		95% Confidence Interval	
	Intraclass Correlation	Lower Bound	Upper Bound
Average Measures	.712	.638	.772

n=219, mean=2.00

Table 4

Intraclass Correlation Coefficient, Social Responsibility, 30 plus Credits

		95% Confidence Interval	
	Intraclass Correlation	Lower Bound	Upper Bound
Average Measures	.693	.612	.760

n=202, mean=2.05

Conclusion

Overall, ICC estimates demonstrate that inter-rater reliability measurements between faculty raters range between moderate and good. Utilizing the average score each rater assigned to an artifact is appropriate in this research design versus utilizing the raw scores of the five portions of the rubric. As the most conservative values were used for assessing the strength of the ICC measures, overall raters have demonstrated solid reliability across multiple artifacts. With the ICC for Critical Thinking, 30 plus credit hours demonstrating the lowest correlation at .665, with a confidence interval between .581 and .735, it is suggested that additional training be provided for those raters to allow for a stronger reliability index in the next COAT artifact review cycle.

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.