Abstract

When evaluating student learning, educators often employ scoring rubrics, for which quality can be determined through evaluating validity and reliability. This article discusses the norming process utilized in a graduate organizational leadership program for a capstone scoring rubric. Concepts of validity and reliability are discussed, as is the development of a scoring rubric. Various statistical measures of inter-rater reliability are presented and effectiveness of those measures are discussed. Our findings indicated that inter-rater reliability can be achieved in graduate scoring rubrics, though the strength of reliability varies substantially based on the selected statistical measure. Recommendations for determining validity and measuring inter-rater reliability among multiple raters and rater pairs in assessment practices, among other considerations in rubric development, are provided.

AUTHORS

Brent J. Goertzen, Ph.D.
Fort Hays State University

Kaley Klaus, Ed.D.
Fort Hays State University

Is it actually reliable? Examining Statistical Methods for Inter-rater Reliability of a Rubric in Graduate Education

Faculty in graduate education utilize a variety of activities to measure student learning—case studies, discussions, essays, or even high-impact practices such as research projects or capstones. For graduate education in particular, high-impact summative activities are commonly utilized at the end of the students' program experience; however, one cannot assume that "high-impact" guarantees students are achieving the program learning goals (Finley, 2019), and one must still competently measure student performance. When evaluating student learning, educators often employ scoring rubrics, but how does one know if a rubric is of sound quality? Is it objective? Does it measure what one wants it to? Does it provide good data? Whether one uses a holistic or analytic rubric (Moskal, 2000) to evaluate student performance, educators must ask these essential questions, especially in contexts involving several raters.

To determine the quality of the scoring rubric used by multiple evaluators for a graduate capstone project in organizational leadership, faculty at [redacted] University participated in a rubric norming process which utilized research-based best practices to determine the inter-rater reliability. This norming process can be employed across academic disciplines to ensure quality evaluations are utilized when measuring student learning. During this process, we discovered varying strengths of inter-rater reliability, depending on the statistical formula used to calculate it. In this article, we outline the statistical methods used

CORRESPONDENCE
Email
bjgoertzen@fhsu.edu

to calculate inter-rater reliability and recommend how educators should measure inter-rater reliability in their assessment practices, among other considerations in rubric development.

Literature

Scoring rubrics are among the most popular forms of direct assessment in the academy (Kuh and Ikenberry, 2009; Gallardo, 2020), and multiple studies have shown that scoring rubrics positively influence students' effort and learning (Charamba and Dlamini-Nxumalo, 2022; Panadero and Romero, 2014). Rubrics provide two important benefits. First, they provide specified criteria and the extent to which the criteria had been reached. Second, they provide important student feedback concerning performance improvement (Moskal, 2000). The authors of this article have been utilizing scoring rubrics for nearly all student assignments for over fifteen years. Anecdotally, students express appreciation for the scoring rubric when shared in concert with general instructions for each assignment, and if designed well, rubrics provide a clear expectation of performance for students and aid instructors in evaluating that performance.

Validity and Reliability

Validity and reliability are essential psychometric properties in survey design; however, these principles are rarely applied to the development and implementation of scoring rubrics. If faculty, directors, and administrators of graduate education programs are using scoring rubrics to inform decisions regarding quality improvement, we must design these rubrics to ensure they yield both valid information and reliable data.

Validity seeks to answer the question, "Does it measure what it was intended to measure?" Validity refers to the "degree to which the evidence supports that these interpretations are correct and that the manner in which the interpretations are used is appropriate" (Moskal and Leydens, 2000). There are three common types of evidence that support validity of an instrument: content, construct, and criterion. Content-related evidence is concerned with the extent to which the assessment instrument adequately samples students' knowledge of the content domain. Construct-related evidence refers to processes that are internal to the individual. While construct-related evidence occurs internally to the student, the performance task and corresponding rubric ought to address not only the product but also provide convincing evidence of the students' underlying processes. Criterion-related evidence describes the extent to which the results of the assessment are related to current or future performance and may be generalized to other, perhaps more relevant, activities.

If faculty, directors, and administrators of graduate education programs are using scoring rubrics to inform decisions regarding quality improvement, we must design these rubrics to ensure they yield both valid information and reliable data.

Reliability refers to the consistency in the assessment scores. A reliable scale is one whereby a student would expect "to attain the same score regardless of when the student completed the assessment, when the response was scored, and who scored the response" (Moskal and Leydens, 2000, p. 1). There are typically two forms of reliability in assessment: inter-rater and intra-rater (McHugh, 2012). Inter-rater reliability concerns the potential variance of scores between multiple raters. Intra-rater reliability refers to any situation in which the scoring process of a single rater may change over time. These inconsistencies result from influences internal to the rater rather than factors associated with differences in student performance.

Three of the most reported strategies for reporting inter-rater reliability are: consensus estimates, consistency estimates, and measurement estimates (Stemler, 2004). Consensus estimates presume that reasonable observers should achieve precise agreement about applying various levels of a scoring rubric. Consistency estimates assume that it is not necessary for raters to share common meaning of the rating scale so long as each rater is consistent in evaluating each dimension of the scale. Measurement estimates presume one should use all available information from all judges, including discrepant ratings, when creating a summary score for each respondent.

Statistical Methods of Inter-rater Reliability

Several statistical methods are common to determine the level of agreement between raters when they review the same product of student performance. One common method

involves a calculation of the intraclass correlation coefficient (ICC) (Gray et al., 2017; Khan et al., 2012). The ICC measures the proportion of variance explained by the objects of measurement (Kahn, et al., 2012). It is advantageous over other types of bivariate correlations (e.g., Pearson *r*) as it accounts for the variance across multiple raters.

Other methods recommended to tabulate consensus estimates of inter-rater reliability include Cohen's kappa statistic, simple percent agreement, and percent adjacent scoring (Stemler, 2004). Cohen's kappa statistic estimates "the degree of consensus between two judges after correcting the percent agreement figure for the amount of agreement that could be expected by chance alone" (Stemler, 2004, p. 2). The kappa statistic assumes that: (1) the phenomenon being rated are independent of one another; (2) the rating categories are mutually exclusive and independent from one another; and (3) the two raters operate independently (Cohen, 1960). It is a robust statistic to compare reliability between rater pairs. Kappa, similar to a correlation coefficient, is a standardized value, ranging from -1 to +1, where 0 represents agreement due to chance and 1 represents perfect agreement (McHugh, 2012). Weighted kappa is an extension of Cohen's kappa. Whereas Cohen's kappa is most suitable for categorical data, weighted kappa can be used for ordinal variables such as scales of a grading rubric (Gisev et al., 2013).

While various other statistical methods exist to evaluate inter-rater reliability, the present study focused on four commonly cited approaches: intraclass correlation coefficient, Cohen's kappa, percent agreement, and percent adjacent.

Percent agreement and percent adjacent are also common methods for calculating interrater reliability, perhaps because of their strong intuitive appeal and that they are easy to calculate and explain (Stemler, 2004). In contrast to ICC or Cohen's kappa, percent agreement and adjacent scoring do not consider chance agreement (Graham et al., 2012). Percent agreement is tabulated by adding up the number of cases that received the same score between rater pairs and dividing by the total number of cases. Percent adjacent assumes that raters to not need to come to exact agreement but can differ by no more than one point above or below the other judge; therefore, adjacent scores are tabulated by adding up the number of cases that received no more than one point differential between raters on a case and dividing by the total number of cases. While various other statistical methods exist to evaluate inter-rater reliability (see McHugh, 2012), the present study focused on four commonly cited approaches: intraclass correlation coefficient, Cohen's kappa, percent agreement, and percent adjacent.

Examining Rubric Validity and Reliability

It is important to understand the context of the assignment and scoring rubric utilized in the organizational leadership graduate program and for this study. In lieu of a traditional comprehensive exam, the [redacted] Department adopted a comprehensive e-portfolio project and associated scoring rubric to measure student mastery of the program competencies.

The e-portfolio is the primary pathway for graduate students to demonstrate mastery of the program's six learning goals. They do this by critically reflecting on selected "artifacts" that provide evidence of their learning for each program goal (e.g., papers, group projects, interviews, discussion postings, journals, peer assessments). Artifacts are mostly comprised of assignments completed in their coursework; however, students can also make use of artifacts from their professional experience, if such work was accomplished during their graduate experience (e.g., team and individual projects, professional development activities). While artifact selection is a key step in developing the e-portfolio, the critical reflection component of the portfolio is what truly demonstrates students' learning and achievement of the program learning goals.

Students are assessed using an analytic rubric with a four-point scale for two categories for each learning goal. The first category is *Selection of Artifacts* and measures whether a student's selected artifacts clearly and directly relate to the corresponding learning goal. For the second category, *Reflection*, students must articulate important learning experienced while creating the artifact and express how they are applying these insights in other contexts in which they engage in leadership. Further, students are to envision new contexts in which they will continue to develop and grow in the future. A distinguished critical reflection meets the following criteria:

- All reflections clearly describe the growth, achievement, and accomplishments, and include goals for continued learning (long- and short-term).

- All reflections illustrate the ability to effectively critique work and provide suggestions for constructive practical alternatives.
- A variety of connections are made between coursework and other parts of the student's life; expressiveness of personality is clearly apparent in the content, and creativity is evident through writing, pictures, media, etc.
- The student superbly incorporates Kolb's experiential model and the DRAG-IT structure for reflective writing (Luzynski and Hamilton, 2017).
- The student accurately connects examples with experience and describes relevant related experiences from other situations.
- The student includes a detailed understanding of their cultural/personal lens and plans for future development.

Valid Judgments of Student Performance: Assignment and Rubric Design

We applied Moskal and Leyden's (2000) framework for creating scoring rubrics by intentionally considering content-related, construct-related and criterion-related evidence in the design of the e-portfolio project and corresponding grading rubric. Students are expected to provide content-related evidence of their mastery for each of the six program learning goals within the e-portfolio project. We intentionally developed the e-portfolio instructional guidelines to assist students in identifying appropriate artifacts representing their learning, in part by suggesting several artifacts commonly used by prior students. The expectations are expressed via the *Selection of Artifacts* domain of the scoring rubric.

The *Reflection* domain of the rubric addresses construct-related and criterion-related evidence by inviting students to reflect on their artifacts; convey what they could have done better; and express how they will improve in future contexts. This reflection requires students to articulate their 'internal reasoning,' an essential pathway to achieve construct validity. Because we also invite students to envision future context in which they will apply their knowledge and insights, the rubric integrates criterion-related evidence as a key feature of student reflection. Additionally, the e-portfolio instructional guidelines and other supporting materials further detail performance expectations by inviting students to relate their experiences to Kolb's Experiential Learning Model and to model their reflective writing with the DRAG-IT structure (Luzynski and Hamilton, 2017). These resources provided students a framework for quality reflection and enhance raters' ability to make valid judgments of student performance.

Conducting both the ICC and the subsequent tests for rater-pair agreement provided insight into how raters might approach evaluating e-portfolios of the growing program in the future.

Improving Inter-rater Reliability

Maki (2004) described a norming process that establishes inter-rater reliability in scoring students' performance. This iterative process requiring successive applications of the scoring rubric ensures consistency in raters' responses, whereby: (1) raters independently score a set of student samples; (2) raters are brought together to review responses and discuss patterns of consistent and inconsistent responses; (3) raters deliberate and resolve inconsistent responses; (4) raters repeat the process of independent scoring for a new set of student work; and (5) again, raters are brought together to discuss consistent and inconsistent patterns in their responses, and raters deliberate and resolve responses.

We employed Maki's (2004) process to include multiple debrief sessions and inter-rater analysis. For the purposes of this study, we performed statistical analysis to test inter-rater reliability of rater responses, including the ICC for overall inter-rater reliability, as well as tests for inter-reliability among rater pairs (i.e., Cohen's weighted kappa, percent-agreement, and percent-adjacent) between the first round of review (see Maki, 2004 stages 1 and 2) and the second round of review (see Maki, 2004 stages 4 and 5). Conducting both the ICC and the subsequent tests for rater-pair agreement provided insight into how raters might approach evaluating e-portfolios of the growing program in the future, as faculty participating in the present study envision continuously increasing program enrollments. As student numbers and e-portfolio submissions increase, teams of three or more raters per e-portfolio will become impractical; therefore, planning for rater-pairs is the preferred level of analysis.

Moreover, rater-pair agreement can lead to greater consensus estimates as they imply judges are providing the same information (Stemler, 2004). Consensus estimates of inter-rater reliability assume that observers should be able to come "to exact agreement about how to apply the various levels of a scoring rubric to the observed behaviors" (Stemler, 2004, p. 2). Consensus estimates are particularly useful for dealing with nominal variables on a rating scale that represent qualitatively different categories and they are beneficial for diagnosing challenges in differing interpretations of how raters apply the rating scale. As a result of our calculations, we observed an increase in inter-rater reliability consensus estimates (Stemler, 2004) over the first several iterations of review; however, the degree to which inter-rater reliability was high was dependent on the statistical method used to calculate it.

Consensus estimates are particularly useful for dealing with nominal variables on a rating scale that represent qualitatively different categories and they are beneficial for diagnosing challenges in differing interpretations of how raters apply the rating scale.

Results

Intraclass Correlation Coefficient (ICC)

There are multiple types of intraclass correlation coefficients. Decisions for identifying the appropriate form of ICC are based on: (1) the model, (2) the type, and (3) the definition (Koo and Li, 2016). Because (1) the selected reviewers are the only reviewers of interest (the model); (2) since we used measurement from a single rater as the unit of analysis (the type); and (3) we were interested in absolute agreement between different raters, we selected to use the absolute agreement of a single measure "two-way mixed" approach to calculate the ICC (Koo and Li, 2016) for both domains (*Selection of Artifacts* and *Reflection*) of the rubric scoring for round-one review and again for the second-round review.

All rater scores for both the first and second round evaluation of e-portfolios were entered into SPSS and the ICC test was run using the *absolute agreement* of a *single measure* "*two-way mixed*" method. Results indicated "poor" and "moderate" reliability, with coefficient scores ranging between .368 and .669 on the first round while the second round yielded "moderate" to "good" with coefficient scores between .546 and .766 (see Table 1).

Table 1
Single Measures of ICC (Absolute Agreement)

	Intraclass Correlation Coefficients (ICC)
FIRST ROUND	
Selection of Artifacts	.368
Reflection	.669*
SECOND ROUND	
Selection of Artifacts	.546*
Reflection	.766⁂

Note. * .5 - .75 Moderate Reliability; ⁑.75 - .9 Good Reliability; ⁂ > .9 Excellent Reliability (Koo and Li, 2016)

Cohen's Weighted Kappa

Individual responses between each rater-pair were dummy coded (agreement = 1; non-agreement = 0) and the weighted kappa statistic was calculated using SPSS. The first round of scoring achieved a weighted kappa range between .166 to .521 on *Selection of Artifacts*; whereas the *Reflection* scores ranged between .206 to .591 (see Table 2). Landis and Koch (1977) recommended a framework for interpreting the statistic (e.g., .21-.40 Fair; .41-.60 Moderate; .61-80 Substantial; .81-1.00 Almost perfect). Further interpretation of the results indicated four of the items achieved a fair level of agreement while five items achieved moderate agreement. The weighted kappa results for the second round of scoring improved, ranging between .320 and

.605 on the *Selection of Artifacts* dimension, and the *Reflection* dimension ranged between .452 and .701. Two of the items achieved at least a fair level of agreement and the remaining five items achieved a moderate level of agreement; five other items achieved a substantial level of agreement.

Table 2
Cohen's Weighted Kappa statistic

	Rater #01 & #02	Rater #01 & #03	Rater #01 & #04	Rater #02 & #03	Rater #02 & #04	Rater #03 & #04
FIRST ROUND						
Selection of Artifacts	.322*	.166	.170	.521⁑	.308*	.318*
Reflection	.586⁑	.545⁑	.206	.571⁑	.591⁑	.373*
SECOND ROUND						
Selection of Artifacts	.587⁑	.490⁑	.320*	.605⁂	.487⁑	.248*
Reflection	.701⁂	.609⁂	.452⁑	.699⁂	.627⁂	.577⁑

Note. * .21 - .40 Fair Agreement; ⁑.41 - .60 Moderate Agreement; ⁂.61 - .80 Substantial Agreement (Landis and Koch, 1977)

Percent Agreement and Percent Adjacent

Individual responses between each dyad pair of raters were dummy coded (agreement = 1; non-agreement = 0) and percent-agreement was tabulated. For the first round, percent-agreement ranged from 33.33 to 72.22% with an overall average of 50% on the *Selection of Artifacts* category, and 33.33 to 72.22% with an overall average of 49.07% on the *Reflection* category (see Table 3). General agreement increased in the second round of scoring as the percent-agreement ranged from 46.67 to 73.33% with an overall average of 62.22% on *Selection of Artifacts*, and a range of 40 to 70% with an overall average of 57.78% on *Reflection* (see Table 3). Only one dyad pair achieved the desired percent-agreement threshold of 70% (Stemler, 2004) for both the *Selection of Artifacts* and *Reflection* elements for the first round of scoring. The results yielded modest improvement for the second round of scoring as two dyad pairs met the threshold for each of the *Selection of Artifacts* and *Reflection* categories.

The results of the present study suggest it was inadequate as a sole means of inter-rater reliability as it cannot detect between which rater-pairs' agreement (or disagreement) was experienced.

Adjacent scoring was also tabulated by first dummy coding individual responses between each dyad pair of raters (agreement or adjacent = 1; non-adjacent = 0). The first round of adjacent averages ranged from 88.89 to 100% with an overall average of 95.37% on *Selection of Artifacts*, and a range of 94.44 to 100% with an overall average of 99.07% on *Reflection* (see Table 4). The second round of scoring yielded similarly high results with a range of 88.33 and 100% with an overall average of 93.89% on *Selection of Artifacts*, and a range of 86.67 and 100% with an overall average of 94.44% on the *Reflection* category. Many adjacent averages among the dyad pairs, including the overall averages for both the first round and second round of scoring, achieved the desired threshold of 90% (Stemler, 2004).

Discussion

The consensus estimates produced mixed results (see Table 5) regarding inter-rater reliability. The Intraclass Correlation Coefficient (ICC) is a common method to evaluate inter-rater reliability and is frequently used in norming grading rubrics (Gray et al., 2017), as it provides a single, holistic metric for each dimension across multiple raters. The ICC has been argued as a preferred method over other methods such as percent agreement (Bryer, 2019). If the ICC was used as the sole measure in the present study, we would conclude that we achieved a sufficient level of reliability, particularly at the conclusion of the second-round review; however, while the ICC may provide important insight, the results of the present

Table 3
AGREEMENT: Average Per Rater Combination

	Rater #01 & #02 (%)	Rater #01 & #03 (%)	Rater #01 & #04 (%)	Rater #02 & #03 (%)	Rater #02 & #04 (%)	Rater #03 & #04 (%)	Composite Average (%)
FIRST ROUND							
Selection of Artifacts	50.00	33.33	33.33	44.44	66.67	72.22*	50.00
Reflection	61.11	50.00	38.89	33.33	72.22*	38.89	49.07
SECOND ROUND							
Selection of Artifacts	73.33*	68.33	50.00	73.33*	61.67	46.67	62.22
Reflection	70.00*	60.00	40.00	70.00*	56.67	50.00	57.78

Note. * >70%, recommended minimum threshold for Rater Pair Agreement

Table 4
ADJACENT: Average Per Rater Combination

FIRST ROUND							
Selection of Artifacts	88.89	88.89	94.44*	100.00*	100.00*	100.00*	95.37*
Reflection	100.00*	100.00*	94.44*	100.00*	100.00*	100.00*	99.07*
SECOND ROUND							
Selection of Artifacts	100.00*	98.33*	88.33	100.00*	90.00*	86.67	93.89*
Reflection	100.00*	100.00*	86.67	100.00*	90.00*	90.00*	94.44*

Note. * >90%, recommended minimum threshold for Rater Pair Adjacent

study suggest it was inadequate as a sole means of inter-rater reliability as it cannot detect between which rater-pairs' agreement (or disagreement) was experienced.

The additional consensus estimates (e.g., Cohen's weighted kappa, percent agreement, percent adjacent) are analogous to post hoc tests affording a refined examination of the data to precisely detect patterns of agreement (or disagreement) between rater-pairs. When examining the collective results of the additional consensus estimates, there was substantial agreement between rater-pairs of reviewers 1 and 2 and reviewers 2 and 3, especially from the second round of evaluation. The percent agreement tests, however, yielded disappointing results. Nearly all individual results were stronger in the second-round review when compared to the first-round findings. Only one item between two different rater pairs, however, achieved the desirable threshold in the first round, and two other rater pairs (Raters 1-2; and Raters 2-3) achieved the desirable threshold in the second round. The results illuminate one of the disadvantages of using consensus estimates like percent agreement as it can take substantial time and energy to train raters to come to an exact agreement (Stemler an Tsai, 2008).

The first round of review of the percent adjacent scores were strong, while the Cohen's weighted kappa results most frequently achieved a moderate-level of consistency; however, the percent agreement results were quite disappointing with only one of the six rater-pair combinations achieving a satisfactory level.

The first round of review of the percent adjacent scores were strong, while the Cohen's weighted kappa results most frequently achieved a moderate-level of consistency; however, the percent agreement results were quite disappointing with only one of the six rater-pair combinations achieving a satisfactory level. These results were not surprising as the reviewers evaluated student performance independently before engaging in a debriefing session.

Table 5
Cohen's Weighted Kappa statistic

	Absolute Agreement	Cohen's Weighted Kappa	Percent Agreement	Percent Adjacent
FIRST ROUND				
Selection of Artifacts	Poor	3 of 6 items *fair* agreement; 1 of 6 items *moderate* agreement 0 of 6 items *substantial* agreement	1 of 6 items meets minimum threshold	4 of 6 items meet minimum threshold
Reflection	Moderate	1 of 6 items *fair* agreement; 4 of 6 items *moderate* agreement 0 of 6 items *substantial* agreement	1 of 6 items meets minimum threshold	5 of 6 items meet minimum threshold
SECOND ROUND				
Selection of Artifacts	Moderate	2 of 6 items *fair* agreement; 3 of 6 items *moderate* agreement 1 of 6 items *substantial* agreement	2 of 6 items meet minimum threshold	4 of 6 items meet minimum threshold
Reflection	Good	0 of 6 items *fair* agreement; 2 of 6 items *moderate* agreement 4 of 6 items *substantial* agreement	2 of 6 items meet minimum threshold	5 of 6 items meet minimum threshold

We recommend educators regularly engage in the norming process to enhance inter-rater reliability among reviewers.

We expected and observed appreciable improvement across all consensus estimates between the first and second rounds of scoring. Notably, the ICC test produced moderate to good levels of reliability and the Cohen's weighted kappa yielded moderate to substantial reliability. Similarly, the percent adjacent calculations were strong as ten of the 12 items achieved desirable reliability. One explanation for the percent adjacent results is the findings may be artificially inflated due to the limited number of categories from which to choose (e.g., 1 to 4) (Stemler, 2004). Scholars noted it is often possible to get artificially inflated percent agreement because values can frequently fall under one category of the rating scale (Hayes and Hatch, 1999); however, of the various statistical models in the present study, we observed percent agreement as the weakest reliability measure.

Recommendations

Capstone assessment methods in graduate education, such as the e-portfolio and rubric discussed in this article, often serve as a central feature of program-level assessment; therefore, if we are to make data-informed decisions for program improvement, it is paramount we develop accurate and reliable evaluation of student learning and performance. Based on the results and experiences evaluating our rubric, we offer recommendations for practice.

First, we recommend educators regularly engage in the norming process to enhance inter-rater reliability among reviewers. In our case, this will require regular, ongoing conversations to develop a shared understanding for both sets of dimensions associated with artifact selection and reflection quality. Given we have used the scoring rubric in its present form for several years, individual raters may have experienced "construct drift" when rating student performance on the performance levels. We will need to re-examine aspects of both content and construct validity (Moskal and Leydens, 2000) to ensure the scoring rubric accurately addresses all important and relevant aspects related to the intended content. Refining the definition for each performance level will help raters evaluate student performance and increase rater-pair agreement.

Second, we recommend educators utilize multiple statistical tests for determining inter-rater reliability of scoring rubrics. While one may provide desired results, our study demonstrates that not all measures of inter-rater reliability are equal. While the ICC provides a holistic view of inter-rater reliability, it does not account for differences between individual raters. Utilizing post hoc measures such as Cohen's weighted kappa and percent agreement and percent adjacent further delineate patterns of agreement (or disagreement) between rater-pairs.

In addition to ensuring inter-rater reliability of the scoring rubric as discussed above, it is important to continuously improve and monitor the raters' ability to make valid judgments of student performance related to the scoring rubric.

In addition to ensuring inter-rater reliability of the scoring rubric as discussed above, it is important to continuously improve and monitor the raters' ability to make valid judgments of student performance related to the scoring rubric. As academic programs evolve and adjust to the needs of student learning, so should the evaluation methods. While we applied principles related to content-, construct-, and criterion-related evidence (Moskal and Leydens, 2000) to assist us in making valid inferences of student performance at the present, that may not always be the case in the future. Thus, when faculty make changes at the assignment, course, or program levels, we should ensure our instructional guidelines and scoring rubric correspondingly aligned. While in some instances the changes may enhance valid judgments of student performance, however, it is not always guaranteed.

Conclusion

Many benefits can be achieved by having valid and reliable assessment instruments, especially for projects that serve as critical summative assessments of student learning. As our graduate program continues to experience growth in student enrollment, it will become impractical for all reviewers to evaluate every student's e-portfolio. Through this study, we sought greater consistency between and across raters so we may possess greater confidence that student performance will be evaluated fairly and equitably, regardless of which combination of raters are assigned to judge each student. Our findings indicate that inter-rater reliability can be achieved in graduate scoring rubrics. To do so, faculty must be willing to conduct a comprehensive norming process and select the appropriate measures for inter-rater reliability when conducting statistical analysis.

References

Bryer, J. (2019, October 7). *Relationship between intraclass correlation (ICC) and percent agreement*. IRRsim: An R package for simulating inter-rater reliability. http://irrsim.bryer.org/articles/IRRsim.html

Charamba, E., & Dlamini-Nxumalo, N. (2022). Same yardstick, different results: Efficacy of rubrics in science education assessment. *EUREKA: Social and Humanities*, (4), 82-90. https://doi.org/10.21303/2504-5571.2022.002455

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. http://dx.doi.org/10.1177/001316446002000104

Finley, A. (2019, November). *A comprehensive approach to assessment of high-impact practices* (Occasional Paper No. 41). University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NIOLA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/11/Occasional-Paper-41.pdf

Gallardo, K. (2020). Competency-based assessment and the use of performance-based evaluation rubrics in higher education: Challenges towards the next decade. *Problems of Education in the 21st Century*, *78*(1), 61-79. https://doi.org/10.33225/pec/20.78.61

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330-338. http://dx.doi.org/10.1016/j.sapharm.2012.04.004

Graham, M., Milanowski, A., & Miller, J. (2012, February). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. The Center for Educator Compensation Reform. https://files.eric.ed.gov/fulltext/ED532068.pdf

Gray, J. S., Brown, M. A., & Connolly, J. P. (2017). Examining construct validity of the quantitative literacy VALUE rubric in college-level STEM assignments. *Research & Practice in Assessment*, *12*, 20-31. http://www.rpajournal.com/examining-construct-validity-of-the-quantitative-literacy-value-rubric-in-college-level-stem-assignments/

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication*, *16*(3), 354-367. http://dx.doi.org/10.1177/0741088399016003004

Khan, R., Khalsa, D. K., Klose, K., & Cooksey, Y. Z. (2012). Assessing graduate student learning in four competencies: Use of a common assignment and a combined rubric. *Research & Practice in Assessment*, *7*, 29-41. https://www.rpajournal.com/assessing-graduate-student-learning-in-four-competencies-use-of-a-common-assignment-and-a-combined-rubric/

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. https://doi.org/10.1016/j.jcm.2016.02.012

Kuh, G. D., & Ikenberry, S. O. (2009). *More than you think, less than we need: Learning outcomes assessment in American higher education*. University of Illinois and Indiana University, National Institute of Learning Outcomes Assessment (NIOLA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/2009NILOASurveyReportAbridged.pdf

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174. http://dx.doi.org/10.2307/2529310

Luzynski, C., & Hamilton, C. (2017). *DRAG-IT: A guide to critical reflection for enhancing college student learning and leadership development* [Conference session]. Association of Leadership Educators 27th Annual Conference, Charleston, SC, United States. https://www.leadershipeducators.org/resources/Documents/ALE%202017%20Conference%20Proceedings.pdf

Maki, P. L. (2004). *Assessing for learning: Building a sustainable commitment across the institution*. Stylus.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276-282. http://dx.doiorg/10.11613/BM.2012.031

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation*, *7*(3). https://doi.org/10.7275/a5vq-7q66

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment Research & Evaluation*, *7*(10). https://doi.org/10.7275/q7rm-gg74

Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy and Practice*, *21*(2), 133-148. https://doi.org/10.1080/0969594X.2013.877872

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). https://doi.org/10.7275/96jp-xz07

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Sage. https://doi.org/10.4135/9781412995627