Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 22 Number 11, December 2017 ISSN 1531-7714

The Miscalculation of Interrater Reliability: A Case Study Involving the AAC&U VALUE Rubrics

Robert F. Szafran, *Stephen F. Austin State University*

Institutional assessment of student learning objectives has become a fact-of-life in American higher education and the Association of American Colleges and Universities' (AAC&U) VALUE Rubrics have become a widely adopted evaluation and scoring tool for student work. As faculty from a variety of disciplines, some less familiar with the psychometric literature, are drawn into assessment roles, it is important to point out two easily made but serious errors in what might appear to be one of the more straightforward assessments of measurement quality—interrater reliability. The first error which can occur when a third rater is brought in to adjudicate a discrepancy in the scores reported by an initial two raters has been well-documented in the literature but never before illustrated with AAC&U rubrics. The second error is to cease training before the raters have demonstrated a satisfactory level of interrater reliability. This research note describes an actual case study in which the interrater reliability of the AAC&U rubrics was incorrectly reported and when correctly reported found to be inadequate. The note concludes with recommendations for the correct measurement of interrater reliability.

Mandated by state governments and regional accrediting agencies, institutional assessment of student learning outcomes has become a fact of life for institutions of higher learning in the US (Ikenberry and Kuh 2015) and the Association of American Colleges and Universities' (AAC&U) "Valid Assessment of Learning in Undergraduate Education" (VALUE) rubrics have become a widely used metric for evaluating and scoring student work. These VALUE rubrics, first developed in 2007-09, and currently available for 16 learning outcomes have been viewed by over 3,300 colleges and universities (Association of American Colleges & Universities 2017). Each rubric includes a definition of the learning outcome, its component dimensions, and, for each dimension, descriptions of four levels of performance.

As faculty from a variety of disciplines, some less familiar with the psychometric literature, are drawn into assessment roles, it is important to point out two easily made but serious errors in what might appear to be one of the more straightforward assessments of measurement quality—interrater reliability. The first error which can occur when a third rater is brought in to adjudicate a discrepancy in the scores reported by an initial two raters has been well-documented in the literature but never before illustrated with AAC&U rubrics. The second error may occur when rater training ceases before a satisfactory level of interrater reliability has been demonstrated on practice objects. This research note describes an actual case study in which the interrater reliability of the AAC&U rubrics was incorrectly reported and when correctly reported found to be inadequate. The note concludes with recommendations for the correct measurement of interrater reliability. The case study will also highlight the fact that reliability is ultimately a feature of a set of scores and that the same measurement instrument (or rubric)

can provide good reliability in some circumstances and poor reliability in others.

Reliability can be simply defined as consistency of measurement (Nunnally 1978). Measurement reliability is different from but a necessary condition for measurement validity which is when what is being measured corresponds closely to what was intended to be measured (Nunnally 1978). An easy to understand method of assessing reliability is interrater reliability where multiple raters judging the same objects and following the same measurement procedures are in general agreement as to the scores to be assigned to individual objects (Vogt and Johnson 2011). So, for example, scores generated by a rubric designed to assess content development in a written essay would have interrater reliability if the content development scores given to essays by one rater matched those given by a second rater. The scores would have validity if the content development scores actually reflect the extent of content development evident in the essays as content development was theoretically defined by the rubric designers.

A more formal description of reliability can be found in classical test theory which treats any observed test score as the combination of the true score plus some degree of measurement error; and the reliability of a set of scores is the ratio of the variance in true scores to the variance in observed scores. Reliability approaches zero when observed scores contain much measurement error and it approaches one when observed scores contain little measurement error. (Nunnally 1978)

While more reliability is clearly better than less reliability, psychometric theory is mute on what constitutes "acceptable" reliability. That is largely a matter of expert judgment and disciplinary convention. Schmitt (1996) cautions against the notion of any enshrined level of minimum reliability. Nunnally (1978) suggests a minimum of .70 but notes that minimally acceptable reliability depends on the purpose to which the test is intended with those having individual high-stakes consequences demanding the highest reliability. Consistent with Nunnally's observation, Cherry and Meyer (1993) suggest the minimum threshold for acceptable reliability may be lower when scores are used to draw conclusions about group behavior than when scores are used to draw conclusions about individual behavior. In practice, 0.70 has come to be the most often cited minimum acceptable level of reliability in the social sciences (for example, Nuendorf 2002; Finley 2011; Krippendorff 2013) although some authors would allow tentative conclusions based on reliabilities as low as .50 (Koo and Li 2016) or even .40 (Cicchetti 1994).

While interrater reliability is often calculated based on multiple raters evaluating just a sample of objects and, once adequate reliability is demonstrated, leaving the remainder of objects to be scored by just single raters; in high-stakes evaluations in which outcomes carry large consequences for individuals or institutions, it is common to have all objects scored by two independent raters. When the scores assigned by the two raters to a particular object differ (sometimes only when they differ substantially), a not uncommon practice is to bring in a third rater to resolve the discrepancy. This *tertium quid* method of discrepancy resolution can be implemented in a variety of ways (Johnson, Penny, and Gordon 2000; Johnson et al 2003; Penny and Johnson 2011). The following case study describes what would appear to be a straightforward but ultimately ill-advised implementation of the tertium quid procedure and calculation of reliability[1].

The Miscalculation of Interrater Reliability

Following the implementation of a new core curriculum, student mastery level of written communication was assessed at what we will call Grand Plains University (GPU). GPU chose to assess a sample of course writing assignments from core classes using a modified version of the Written Communication Rubric developed by the Association of American Colleges and Universities (AAC&U) as part of their VALUE Project. The rubric consists of five dimensions: (a) Audience, Context and Purpose; (b) Content Development; (c) Sources and Evidence; (d) Organization and Presentation; and (e) Control of Syntax and Mechanics. On each dimension, a student's level of mastery was scored on a five-point scale: (0) unacceptable, (1) beginning, (2) developing, (3) accomplished, and (4) capstone.

[1] The use of an additional rater in cases where the original raters substantially disagree may be justified in order to provide the individual test-taker with a less questionable score (Cherry and Meyer 1993). When this is done, however, the calculation of interrater reliability can become confusing. It is this confusion in the calculation of interrater reliability to which the present paper is addressed.

A sample of 113 student written assignments from a variety of courses across the core curriculum were randomly selected for assessment by a panel of faculty from departments teaching courses within the core. The assignments for which the students were submitting their work had all been previously reviewed by a university assessment committee and judged appropriate for showing students' mastery of written communication across the five dimensions. The panel of faculty raters all received training and practice in using the modified AAC&U scoring rubric. While some of the written assignments a faculty rater might score may have come from his or her own class or department, raters were asked to score a sample of assignments from across the core courses. While the raters might or might not be familiar with the subject matter of the assignment, they were asked to focus on and score only the five dimensions of written communication identified in the rubric.

The 113 written assignments were independently scored by two members of the scoring panel. If the two raters disagreed by more than one point in their score on an assignment on one or more of the dimensions, a third rater scored the assignment in question but only on the dimension(s) where the initial raters disagreed by more than a point. If just two raters scored a paper on a particular dimension, the score assigned to the paper on that dimension was the average of the two ratings. If three raters scored a paper on a particular dimension, the score assigned to the paper on that dimension was the average of the two closest ratings. The outlying score was discarded except in the case when the third rater's score is no more than one integer from the other two ratings, in which case all three scores were averaged.

As row 1 in Table 1 indicates, the number of writing assignments requiring a third rater did not vary dramatically across dimensions ranging from a low of 12 (11%) to a high of 18 (16%).

For each dimension, an intra-class correlation coefficient (ICC) was calculated and reported as a measure of reliability. As row 2 in Table 1 shows, these reported reliability coefficients ranged from 0.56 for "organization and presentation" to 0.75 for "sources of evidence."[2] However, it is essential to note that these reliability coefficients were calculated only after outlying original scores were replaced by a third rater's score.

It should be said that this was GPU's first attempt at reporting assessment results for the new core, the university did demonstrate an awareness of reliability as an important measurement characteristic, and it was calculating interrater reliability in a manner that had at one time been fairly common (Johnson et al 2000). However, it was a method of calculating interrater reliability that had subsequently been reviewed in the psychometric literature and criticized for presenting an inflated claim of reliability because it discarded outlying scores legitimately generated by the scoring rubric in favor of scores more conducive to claims of high reliability. If reliability is the ratio of true score variance to observed score variance as classical test theory assumes, replacing observed outlying scores with more

Table 1. Intra-class Correlation Coefficients After and Before Replacement

	Audience, Content and Purpose	Content Development	Sources and Evidence	Organization and Presentation	Control of Syntax and Mechanics
Assignments requiring a 3rd rater	18 (16%)	12 (11%)	13 (12%)	18 (16%)	13 (12%)
ICC* after replacement (reported by GPU)	.69	.68	.75	.56	.69
ICC* before replacement (not reported by GPU)	.09	.41	.47	.21	.36

* Intra-class Correlation Coefficient

[2] Interested readers may contact the author for a copy of the institutional report in which these coefficients appear.

consistent observed scores falsifies the calculation of observed score variance. Cherry and Meyer (1993:122) describe the resulting reliability coefficient as "vastly inflated and largely meaningless" and "a kind of fraud." Others who reviewed the technique are more reserved in their language but no less clear in their criticism. Johnson et al (2000:136) note that "choosing the score of the rater that is in closest agreement with the expert introduces an artificial inflation to the reliability of domain scores." McCormick (2009) also decries the inflation of reliability coefficients following this tertium quid substitution and demonstrates that the lower the initial reliability of the original two raters, the greater the inflation introduced. Krippendorff (2013:275) states "The only publishable reliability is the one measured before the reconciliation of disagreements. The reliability of the data after this reconciliation effort is merely arguable."

McCormick (2009) demonstrated that even randomly generated pairs of numbers would produce minimally satisfactory reliability coefficients if this tertium quid method of calculating reliability only after discarding outliers were applied. Following McCormick's lead, a simulation was run to estimate the inflation present in the published GPU reliability coefficients. 113 pairs of random integers between 0 and 4 were generated using a normal distribution with the mean and standard deviation from the original GPU ratings.[3] Only when the original pair of integers differed by more than 1 was a third random integer generated. For those cases with now three randomly generated scores, the outlying score was discarded and the intra-class correlation coefficient was calculated. This simulation was run a total of 10 times—each time beginning with a new 113 pairs of random numbers. These 10 simulations using random numbers produced an average reliability coefficient of 0.64—nearly equaling the 0.67 average reliability reported by GPU for the five written communication dimensions.

Had GPU calculated and reported the more appropriate intra-class correlation coefficient based solely on the scores from the original two raters, the reliability coefficients would have been those appearing in the third row of Table 1. These correct intra-class correlation coefficients range from just .09 for "audience, content and purpose" to .47 for "sources and evidence" and would generally have been judged unacceptable.

The Contextual Nature of Reliability

Cherry and Meyer (1993), and more recently Thompson (2003), note that reliability is correctly described as an attribute of a set of scores produced by some measurement instrument and not an attribute of the instrument itself. While some measurement instruments might be unable to produce scores with good reliability under any circumstances, other instruments might produce scores with good reliability under some circumstances but not under others. That fact is demonstrated by a further description of the assessment process at GPU.

At the same time that the six faculty member panel of judges described above were scoring 113 student objects on each of five dimensions of written communication using the VALUE rubric, a second panel of six different faculty judges was scoring a different set of 132 student objects on the same five dimensions of written communication along with a sixth dimension pertaining to visual communication, specifically the use of visual aids to supplement the written text. The two sets of student products came from different core curriculum courses: the first set of 113 student works came from courses charged with teaching and assessing student mastery of written communication while the second set of 132 student works came from courses charged with teaching and assessing student mastery of written communication and visual communication.

Just like for the first panel, the interrater reliabilities which were initially calculated and reported by GPU for the second panel were inflated due to the replacement of outlying ratings by ratings provided by a third scorer. The second row of Table 2 shows the inflated scores as reported by GPU and the third row shows the corrected reliability coefficients based on the original two raters. The difference between the inflated and corrected reliability coefficients are much less for the second panel because the second panel's actual reliability levels were much higher than were the first panel's. This is consistent with McCormick's (2009) observation that the lower the initial reliability of the original two raters, the greater the inflation introduced

[3] The random integers were generated using the SPSS numerical expression rnd(rv.normal(mean, stdev)).

by the tertium quid procedure. Of greater interest, however, is a comparison of the reliability levels achieved by the second panel compared to those achieved by the first panel once the reliability coefficients for both panels have been corrected to reflect the ratings of the original raters (row 3 in Table 1 compared to row 3 in Table 2). The second panel achieved much higher reliability levels than did the first panel. On the five dimensions of written communication which both panels rated, the average corrected reliability for panel one was 0.31 but 0.75 for panel two.

Table 2. Second Panel's Intra-class Correlation Coefficients After and Before Replacement

	Audience, Content and Purpose	Content Development	Sources and Evidence	Organization and Presentation	Control of Syntax and Mechanics	Visual Aids
Assignments requiring a 3rd rater	6 (4%)	5 (4%)	8 (6%)	6 (4%)	5 (4%)	9 (7%)
ICC* after replacement (reported by GPU)	.84	.84	.91	.84	.72	.91
ICC* before replacement (not reported by GPU)	.73	.78	.84	.78	.61	.86

* Intra-class Correlation Coefficient

Both panels were using the same scoring rubric for the five dimensions of written communication. Both panels received similar training. And reports from both panels suggest that all panel members diligently performed their tasks. Nevertheless, the ratings from the second panel showed so much better reliability than did those of the first panel!

A comparison of demographic characteristics of the students who produced the essays rated by the 1st panel and by the 2nd panel found no significant differences in age, year in college, major, ACT scores, SAT scores, or college grade point average.

Since the assignments that produced the student essays were course-specific, it is possible that the various assignments that produced the 132 student works rated by the 2nd panel were of a nature that made differences in the quality of student work clearer and, as a result, resulted in higher interrater reliability than the various assignments that produced the 113 works rated by the 1st panel. However, systematic differences in assignments that resulted in differences in reliability would seem unlikely. The works rated by both panels came from a large number of different courses (19 courses for the 1st panel and 24 for the 2nd), and both panels had an almost identical 80/20 split of papers coming from 1st year courses versus 2nd year courses.

Perhaps the best explanation for the 2nd panel's much higher levels of reliability came from conversations with faculty who served on the two panels when they described their training in the use of the VALUE rubrics. While both panels had similar training which consisted of two sessions each lasting approximately two hours during which the rubrics were discussed and four sample papers were individually rated and the ratings then discussed, participants in the 2nd panel noted that some of their panel members were particularly forceful in their explanation of their ratings of the practice papers and this led to a productive discussion of rating criteria by the panel members. In contrast, members of the 1st panel described the discussions following the practice grading as not being particularly noteworthy. The fortuitous selection of two particularly verbal faculty to the 2nd panel may account for that panel's superior rating reliability. That more thorough discussion of rubrics during training may be a contributing factor to greater interrater reliability cannot be rigorously tested in the post hoc investigation recounted here but certainly seems plausible and deserving of further study.

Recommendations

I conclude with some recommendations for calculating and reporting interrater reliability that are not new but deserve repeating as more and more faculty find

themselves analyzing and reporting on institutional assessment:

- If a tertium quid procedure is to be used, there is an obligation to report the initial interrater reliability between the two original raters. (for example, Stevens, Lyles and Berke 2014). Reporting reliability coefficients after replacement lacks a basis in psychometric theory and provides readers with a false sense of confidence in the measurement instrument.
- If time and resources permit, all objects should be scored by three raters. In this case, reliability can be calculated based on all three raters. This will typically improve reliability and minimize the effect of outlying ratings (for example, Krippendorff 2013).
- Preliminary assessment of interrater reliability should occur within the training sessions. When reliability is found to be inadequate, additional or improved training of raters is needed (for example, Johnson, Penny, and Gordon 2009). Training of raters needs to continue to the point that reliability in grading practice papers achieves acceptable levels. Some panels may achieve that sooner than others, but quality assessment should not depend on the fortuitous selection of particularly verbal and assertive raters.
- Where reliability still remains inadequate, the difficulty may lie not with the raters but with the rubric which may have great face validity but be impossible to implement in a consistent, reliable fashion. In such cases, reconceptualization of the grading rubric may be necessary (for example, Moskal and Leydens 2000). In the present case study, had both panels registered very low reliabilities, suspicion might have turned to the VALUE rubric for written communication; but the acceptable levels of reliability recorded by the 2nd panel suggests the VALUE written communication rubric could produce acceptably reliable scores.

References

Association of American Colleges & Universities. (2017). VALUE Rubric Development Project. Washington, DC: Association of American Colleges & Universities. Retrieved July 18, 2017 https://www.aacu.org/value/rubrics.

Cherry, R. D. & Meyer P. R. (1993). Reliability issues in holistic assessment. (pp.109-141) in Validating holistic scoring for writing assessment, edited by M.W. Williamson and B.A. Huot. Cresskill, NJ: Hampton Press.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* (6) 284-290.

Finley, A. P. (2011). How reliable are the VALUE rubrics? Peer Review Fall2011/Winter2012, Volume 13/14, Issue 4/1: 31-33.

Ikenberry, S. O. &. Kuh, G.D. (2015). From compliance to ownership: why and how colleges and universities assess student learning. (pp. 1-23) in Using Evidence of Student Learning to Improve Higher Education, edited by G.D. Kuh, S.O. Ikenberry, N.A. Jankowski, T.R. Cain, P.T. Ewell, P. Hutchings, and J. Kinzie. San Francisco, CA: Jossey-Bass.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education* 13(2), 121-138.

Johnson, R. L., Penny, J., Steve Fisher, & Kuhs, T. (2003). Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement in Education* 16(4), 299-322.

Krippendorff, K. (2013). Content Analysis: An introduction to its methodology. 3rd ed. Thousand Oaks, CA: Sage.

McCormick, C. M. (2009). The potential for interrater reliability inflation with tertium quid rater Adjudication: A Simulation. Unpublished Master's Thesis. University of Nebraska, Lincoln, NE.

Moskal, B. M. & Leydens J.A. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation* 7(10). Available online: http://PAREonline.net/getvn.asp?v=7&n=10 .

Nuendorf, K. A. (2002). The Content Analysis Guidebook. Thousand Oaks, CA: Sage.

Nunnally, J. C. (1978). Psychometric Theory. 2nd ed. New York, NY: McGraw-Hill.

Penny, J. A. & Johnson, R.L. (2011). The accuracy of performance task scores after resolution of rater disagreement: A monte carlo study. *Assessing Writing* 16:221-236.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment* 8(4), 350-353.

Stevens, M. R., Ward L., & Berke, P.R. (2014). Measuring and reporting intercoder reliability in plan quality evaluation research. *Journal of Planning Education and Research* 34(1), 77-93.

Thompson, B. (2003). Understanding reliability and coefficient alpha, Really. (pp. 3-23). in Score Reliability: Contemporary Thinking on Reliability Issues, edited by B. Thompson. Thousand Oaks, CA: Sage.

Vogt, W. P. & Johnson, R. B. (2011). Dictionary of Statistics & Methodology: A nontechnical guide for the social sciences. 4th ed. Los Angeles, CA: Sage.

Citation:

Szafran, Robert. (2017). The Miscalculation of Interrater Reliability: A Case Study Involving the AAC&U VALUE Rubrics. *Practical Assessment, Research & Evaluation*, 22(11). Available online: http://pareonline.net/getvn.asp?v=22&n=11

Corresponding Author

Robert F. Szafran
P.O. Box 13047 SFA Station
Nacogdoches, TX 75962

email: rszafran [at] sfasu.edu