Karnas-Haines, C. (2019) Making assessment actionable through assessor training: A tool for building trust through moderation and calibration. *Intersection: A Journal at the Intersection of Assessment and Learning, 2(3).*

Making Assessment Actionable through Assessor Training: A Tool for Building Trust through Moderation and Calibration

Colleen Karnas-Haines, Ph.D.

Intersection: A Journal at the Intersection of Assessment and Learning
Vol 2, Issue 3, 2021

Abstract: In order to use assessment findings for learning improvement, it is essential to improve the validity of these data. When an assessment is distributed across numerous course sections with multiple faculty members acting as assessors, issues arise concerning validity and calibration. To address these concerns, we created a training program to better transmit student learning outcome definitions, provide practice activities aimed at calibration, and promote social moderation among our assessment community. This training, hosted on our university's learning management system, explicitly defines outcomes, establishes common indicators to inform the rating of student proficiency levels, and provides quizzes with sample student work to practice and refine assessment calibration. This training was developed by the assessment team, is given to an assessor shared across all sections, and is modified annually by faculty and program leadership.

Keywords: *calibration, assessor training, learning management system, written communication, assessment*

Introduction

The field of assessment has shifted from conducting assessment that demonstrates compliance towards conducting assessment that produces actionable findings for learning improvement (Blaich & Wise, 2018; Pasquerella, 2018; Singer-Freeman & Robinson, 2020a, 2020b). To effectively drive learning improvement, it is essential to improve assessment methodology so that assessment findings inform our understanding of the outcomes associated with pedagogical practices. Before a program can identify the causes of gaps in student learning and success, identify evidence-based solutions, determine whether selected interventions are implemented correctly, and measure the extent to which the interventions drive improvements (Eubanks, 2017; Fulcher et al., 2017), the program needs to communicate about data validity. Currently, many assessment processes result in reporting primarily pedagogical successes. This approach to assessment results from past views of assessment as an exercise in demonstrating compliance. For assessment data to inform learning improvement, there is a need to create assessment processes that illuminate successes, failures, and all the levels of proficiency between the two. Effective calibration is an essential first step towards the creation of assessment findings that can serve as an impetus for improvements.

There are many reasons why a program should reevaluate its assessment procedures: adherence to a system of cyclical check-ups, the introduction of new programs or

participants, substandard assessment results, confounding assessment results, etc. Each scenario suggests different interventions with differing levels of intensity. The scenario that brought about the following intervention was one where the results may have been "too good to be true." This inspired a complete reimagining of the student learning outcome assessment process with a renewed interest in defining the outcome, determining levels of proficiency, proposing indicators of proficiency, creating a new system of assessment team communication using a learning management system (LMS), and empowering the process of moderation.

Background

The scenario that precipitated the development of the assessment training tool involved a written communication student learning outcome (SLO) that consistently passed at an extremely high rate, the last year producing a 99% pass rate. Not only did nearly all of the assessed students pass, the vast majority passed with the highest "exemplary" rating in all categories. The impetus for reimagining the student learning outcome assessment was not to reduce the passing rate, but to investigate the validity of the assessment and discover areas that could be targeted for learning improvement; "Validity relates to the notion of 'truth' in the sense that we want to know how closely assessment results reflect reality" (Holzman et al., 2020). In other words, we needed, and still need, confidence that the "exemplary" rating in the written communication SLO indicates evidence of exemplary writing communication skills.

The Bachelor of Science and the Bachelor of Arts in Computer Science both contain a written communication SLO measured in a third-year course titled, "Computers and Their Impact on Society." Being a writing intensive course that is required for all 2,000+ BA and BS Computer Science undergraduates, there are a large number of small sections offered each semester with many different faculty teaching those sections. Fall 2020 had six different faculty teaching over 300 students in 13 sections. The Spring 2021 semester added a seventh faculty member across 15 sections. There has been a recent drive to create a team mentality among the course faculty to help unify curriculum by employing two full-time lecturers to act as team leads for the course. These full-time faculty have also worked with university-wide programs such as *Communication Across Curriculum* to find points of agreement across colleges and degree programs about written communication standards. Despite these activities, the curriculum and assignments across the various sections have varying degrees of similarity, which can create challenges in generating comparable assessments across sections.

While there are challenges, the benefit of having multiple small sections is the small faculty-to-student ratio and the resultant positive learning environment. This course structure builds strong relationships between faculty and students, encourages personalized learning with small groups that promote engaging discussions, and leverages a wider variety of faculty experience and expertise for a richer learning experience. The specific assessment challenges of this course structure include trying to calibrate scores across numerous assessors, identifying and honoring different faculty data collection preferences, and ensuring that assessments function separately from grading.

Old Process

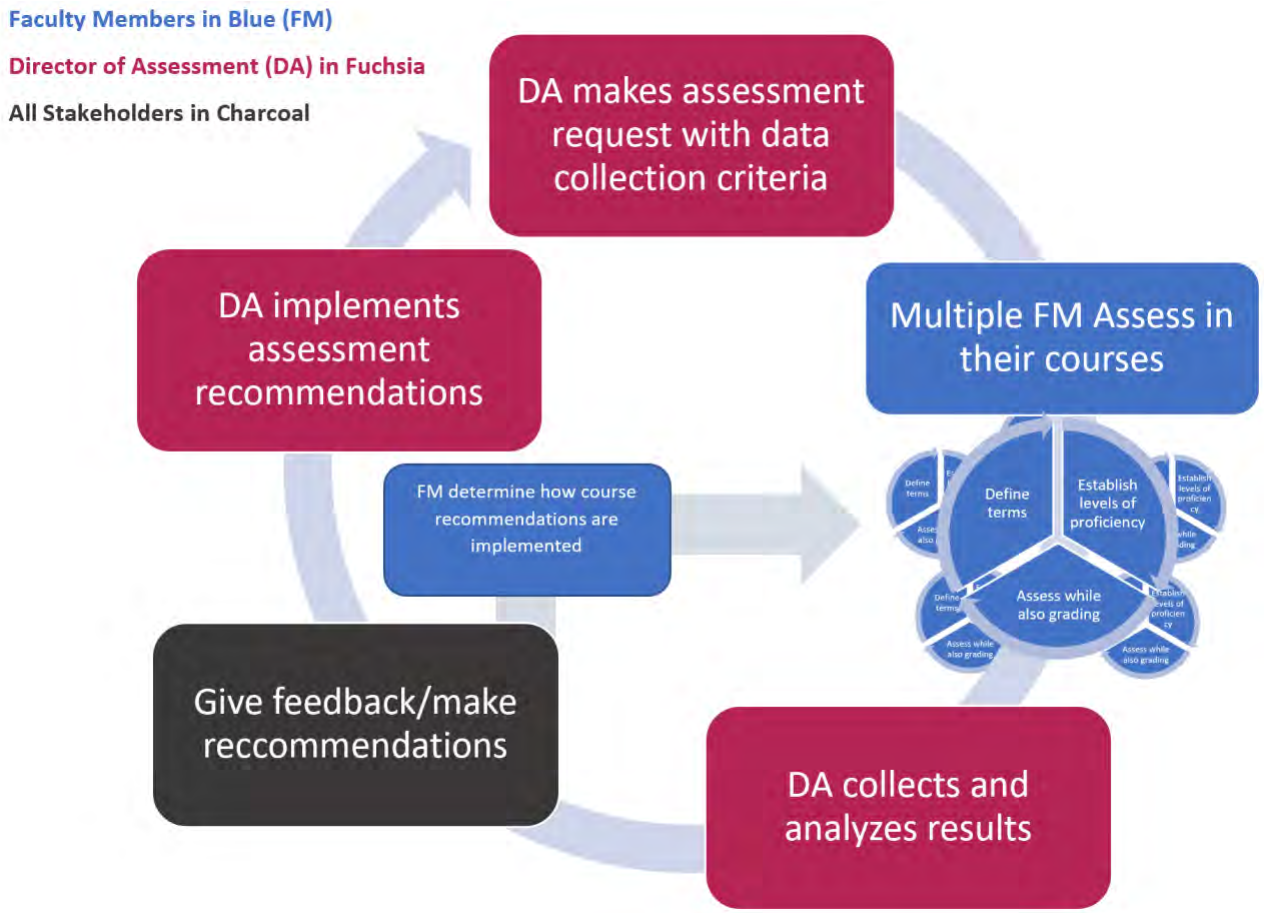
Our former assessment process is depicted in Figure 1. As can be seen in Figure 1, prior to the changes described later in this paper, our assessment process for this multi-section course began with the Director of Assessment (DA). The DA distributed previously developed rubrics to all faculty members and reminded faculty of the expectations regarding the assignment to be evaluated, how to collect the data, and when to report the results back to the DA. Faculty members interpreted rubrics on their own and were required to simultaneously grade and assess the written assignments – two related but different processes. At the completion of the semester, the DA analyzed the collected data and prepared a report for faculty and program leaders to explore. The report was distributed to faculty members and program leaders in order to inform their assessment close-the-loop meetings. While feedback from all stakeholders was encouraged, some were unable to attend meetings and chose not to offer their feedback in alternative ways, i.e. an emailed response. From the feedback collected, the DA synthesized the suggested improvements and distributed them to the faculty scheduled to teach the class during the next semester. Implementation of suggestions was often not coordinated among all faculty members with competing priorities sometimes dropping assessment-based improvements lower on the faculty's list of demands than they intended.

Calibration

With so many different faculty teaching different sections, it was difficult to calibrate assessment scores across the different faculty members and different semesters. Assessing non-quantitative written communication skills will always involve some degree of subjectivity, but as O'Connell et al. (2015) explain, the solution is not to narrow the measurement or find a way to objectify the outcome, but rather "realiz[e] that both subjective and objective knowledge are involved in this very human enterprise, and that through facilitation and collaboration this volatility can be significantly reduced." When we speak of calibration, we do not imply a goal of matching our ratings to some objective, universally accepted scale; instead we understand that "calibration is based on the assumption that standards are socially constructed, and it therefore involves dialogue, negotiation and joint decision making" (Advance HE, 2018). Many programs encounter this problem not only with multi-section courses, but also in courses that use adjuncts or graduate students as instructors and assessors. One faculty member interviewed by Crisp (2017) for an article on the benefits of calibration explained it thusly, "'adjuncts are not institutional insiders with the creation and deployment of rubrics for assessment. Therefore, many variations in communication and interpretation open up even with a very clear rubric'" (pg. 9). Even when well-written rubrics are used, the descriptions of ratings on the rubrics cannot replace the necessary discussions, negotiations, and explanations that are required to capture all the nuanced differences in ratings. It is not, for example, enough to assess proper punctuation in citations. One has to ask if mastery of proper punctuation is considered an essential component of basic acceptable citation creation or a skill that when mastered moves the rating from acceptable to exemplary. Is the citation punctuation question universally agreed upon by faculty across all sections? Does a single faculty member answer the citation punctuation question the same way across all semesters and all papers, or does the faculty member demand more proficiency as the semester proceeds? Yearly assessment feedback meetings can address these types of

questions when noticeable issues arise, but it can be prohibitively time consuming to ask every faculty member to review all such questions of calibration to discover, negotiate, and correct hidden discrepancies. In addition to providing more valid ratings of proficiency, calibration exercises can shift faculty members' views around assessment away from compliance and towards learning improvement. Taking the time to consider what constitutes exemplary written communication can be a starting place for considerations of ways to respond to student areas of continuing struggle with targeted changes to instruction.

Figure 1. The Old Assessment Process



Note: Notice how each faculty member independently defines SLO-related terms, establishes their own levels of proficiency, and has to wear two hats, grader and assessor.

Data Collection Preferences

Our previous system allowed faculty to collect data using whatever system they preferred. As a result, some assessment results were ready before the semester was over and some after faculty completed semester grades. Some assessors preferred to use the university's LMS to automatically report assessment data while others preferred manual data reports. While all preferences were valid and could be accommodated, integrating the various methods of data collection was unnecessarily complicated and time consuming and differing preferences may have negatively impacted the

comparability of the scores across sections. For example, scores that were assigned at the end of the semester may have been more likely to reflect students' progress over the course of the semester than scores that were given mid-semester.

Assessment, Not Grading

Lastly, without explicit training on the purpose of assessment and the ways in which assessment differs from grading, different faculty approached the task quite differently. Some faculty expressed concern over the emotional impact of assessing students in a visible way through the LMS. Some faculty wished to grade students in a way that acknowledged improvement, but worried that pairing grades with more static assessment proficiency levels might discourage students and hamper their progress. In other words, they wanted a student who had poor written communication skills but who greatly improved, to focus on their grade-reflected improvement, rather than their static assessment rating. Some faculty were able to successfully pair visible grades with hidden-from-student-view assessment results. This strategy allowed the visible grades to have a formative function for the student, while the hidden programmatic assessment data of student abilities functioned as a summative measure of program performance. As mentioned with other processes, coordinated implementation of this strategy was not successful. Some faculty were visibly grading immediately, but procrastinating for the hidden assessments creating two temporally distinct points of assessment. Some faculty were simply using grades as assessment ratings so as not to have two points of assessment. Others were allowing grades and assessments to be visible while acknowledging their proficiency ratings were artificial higher than they would have been otherwise.

These challenges resulted in assessments that invite validity issues. Every faculty must be very clear concerning the two hats they must wear: formative grader of the student and summative assessor of the program. Faculty must distinguish between a student showing evidence of an "exemplary" ability to organize a paper and a student showing an "exemplary" ability to improve their paper organization. Student visibility of ratings threatened to turn what was meant to be a summative programmatic assessment into the faculty's formative assessment tool. In order to effectively use assessments of student learning to guide program improvements, it is essential to resolve these issues that threaten the validity and usefulness of the measurements.

New Process

Rethinking an assessment process allows each organization to discover bottlenecks, resolve questions that participants were previously hesitant to bring forward, unify goals, and establish a culture of progress and improvement. The simple act of declaring that a process can and will change breaks participants free from assumptions born of convention. If process change is pursued in a manner that builds in cyclical moments of reevaluation, participants are more likely to experiment and view mishaps as fodder for feedback, not as failures. Therefore, while the process shown below is our "new" process, it is not our "final" process. The primary goals for our new process were: defining outcomes,

levels of proficiency, indicators of proficiency, and assessed assignment requirements; communicating expectations, processes, results, and suggestions; and empowering our assessment community to embrace discussions and negotiations that lead towards learning improvement.

Development

Over the summer of 2020 a new process for the assessment of written communication assessment data was developed in preparation for implementation in the Fall 2020 semester. This was accomplished in four steps: identification of subtopics that contribute to successful written communication skills, clear definitions of measures, establishment of indicators of proficiency levels, and the creation of samples for practice assessments. The products of these steps are housed in a Canvas (our LMS) Assessment Training Course.

To identify the subtopics that contribute to successful written communication skills, we reviewed the assignment that was evaluated and the accompanying subtopics used in previous cycles. Traditionally, we evaluated a formal research paper. Although the College is interested in eventually exploring varied modes of written communication, including writing in the context of project proposals and coding comments, for this assessment project we chose to maintain our traditional assignment. We previously evaluated these assignments using five subtopics: supported conclusions, appropriate content, well-deployed writing mechanics (grammar and spelling), logical organization, and ethically cited references. After a review and discussion of these sub-topics, we decided that they were appropriate and elected to continue to use them in the new process.

Having decided to continue using the same subtopics and artifact for evaluation, we created a page in our training course that defined and delineated each of the five subtopics. The subtopic definitions focus on what constitutes a successful outcome. The delineation guidelines were designed to provide clear ways to respond to common subtopic overlaps faculty had encountered in student papers. For example, can well-supported conclusions be present in a paper with irrelevant content? Does a conclusion that is contained in the second paragraph of a seven-paragraph essay indicate a problematic conclusion or problematic organization? We chose to house the definitions and delineations of subtopics in a Canvas training page so faculty could revisit the information easily when new and unique situations arose. Housing the information in our Canvas page also allowed us to easily disseminate adjustments based on ongoing feedback.

To link the high-level definitions of subtopics to concrete examples of varying proficiency we identified common indicators that an assessor might encounter when rating students' success for each subtopic. Common indicators are elements of student writing that often influence the assessed level of proficiency. It is not enough to simply say that a student must have "good" writing mechanics; instead, we break down common mistakes and assign a level of severity. Examples in the writing mechanics category include: unresolved spellcheck notifications, passive voice, subject-verb disagreement, misuse of homonyms, etc.

These were presented as, “Things to look for” in each subtopic. As with everything else, the levels of proficiency are dynamic and can be easily renegotiated now that they are transparently documented and open for communal inspection.

Having created all of the instructive materials, we wanted to create opportunities for assessors to practice applying the guidelines in order to calibrate ratings. We provided sample excerpts from mock student papers and assessment practice quizzes for each subtopic. Authentic examples are integral in illuminating the differences between levels (Crisp, 2017, pg. 9). Some questions were set up as matching-types where there is one example for each level of achievement: beginner, needs improvement, acceptable, accomplished, and exemplary. Other questions simply contained a sample written excerpt with instructions to assign it one of the five possible ratings: Beginner (1 pt) to Exemplary (5 pts). In all cases, when the question is answered incorrectly, the LMS quiz provides an explanation. This explanation refers back to the subtopic definitions and/or common indicators. The quizzes can be taken multiple times as the quizzes are not intended to evaluate the assessor’s knowledge, but to train (what to look for) and calibrate (match our established scale).

Implementation

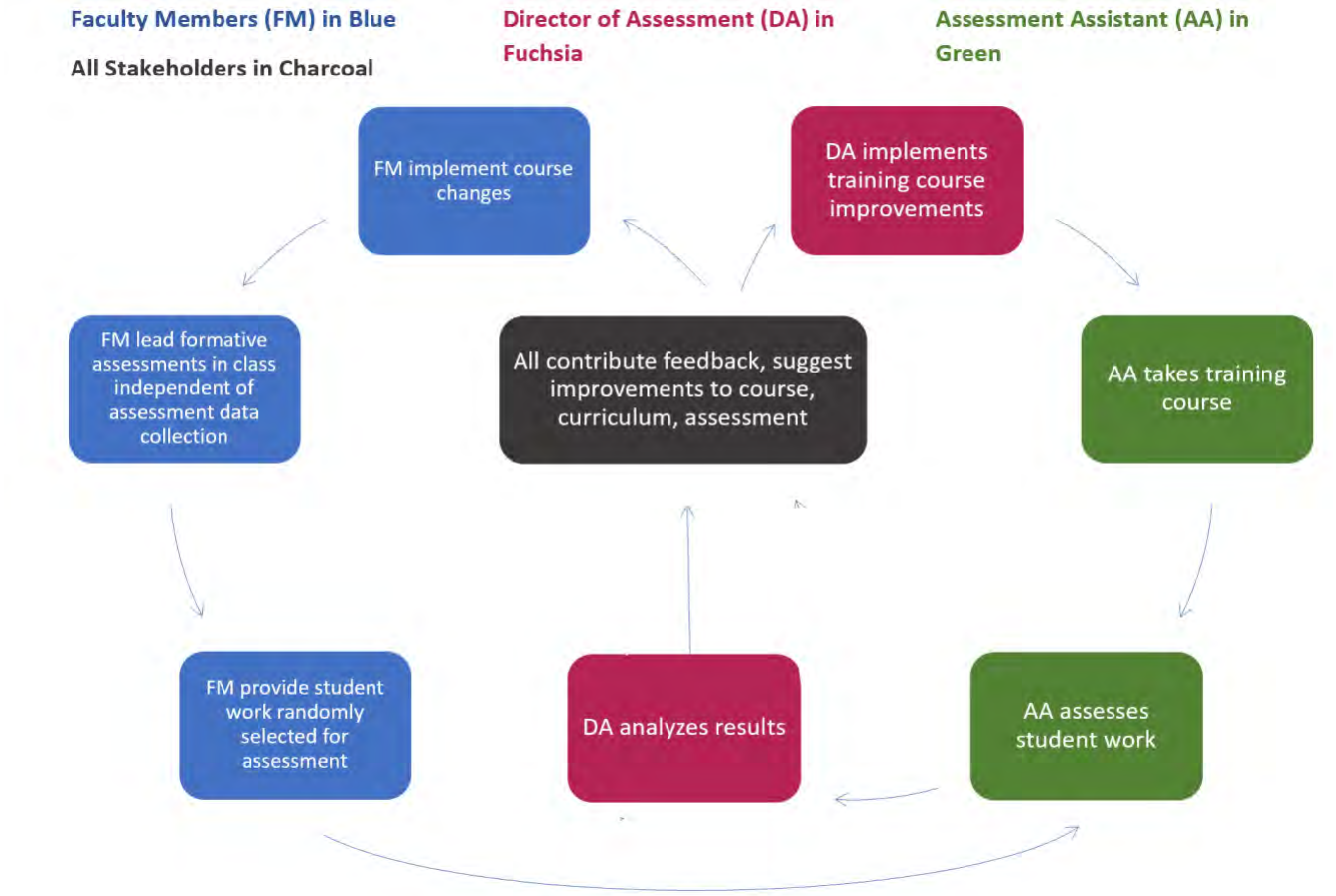
The assessment team for the College of Computing and Informatics (CCI) implemented this new method by training, sampling, assessing, comparing, and modifying. Notice in Figure 2 that many of the same actors are involved in the new method as in the old method, but there is better delineation of tasks leading to more constructive and efficient communication.

Training. With the Canvas training page complete, CCI hired a student Assessment Assistant (AA) with Teaching Assistant experience to be shared among all course sections. Outside assessors of anonymized student papers not only address issues of calibration, but also eliminate the possibility of faculty bias that occur when faculty assess their own students (Kim & Helms, 2016). Not only was the AA introduced to the entire assessment process, they were also required to take the Canvas training course at the beginning of the semester. The AA was required to pass all the subtopic calibration quizzes before conducting any assessments. We believed the use of a single evaluator who is not directly involved in instruction or grading student work will reduce some of the challenges faced by faculty assessors such as feelings that assessment of learning would discourage students or reflect poorly on themselves. Additionally, the use of a single evaluator across many sections and faculty helped us to focus on areas of curricular weakness that will result in reduced proficiency across many sections. By freeing faculty from the work of assigning scores we hoped to engage them more fully in the process of making meaning from patterns of strength and weakness that emerge.

Sampling. Faculty submitted a random sampling of student papers across all sections of the “Computers and Their Impact on Society” course from Spring 2020 (the previous semester) and Fall 2020 (the semester of the new process’s implementation) to CCI’s Assessment Team. A random number generator

was used to identify which students' papers would be submitted based on their position on the class roster.

Figure 2. The New Assessment Process



Note: The addition of the AA role helps to separate grading from assessment tasks. Assessment is unified across all sections by the shared AA assessor, trained by an assessment training course modified by faculty and program leader feedback. This reduces the process of calibration from a many-to-many process (multiple faculty giving feedback in order to recalibrate multiple faculty assessors) to a many-to-one process (multiple faculty negotiating changes to the single training course).

Assessing. Once the papers were submitted for assessment, the Director of Assessment removed all personally identifiable information from the papers, unless already removed by the faculty, before submitting them to the AA. The AA read each paper and rated the performance in each subtopic. Ratings for each subtopic ranged from Beginner (1 pt) to Exemplary (5 pts) and a final score for the assessment was calculated by averaging all five subtopics.

Comparing. Once the AA completed the assessments, the Director of Assessment analyzed the results to determine any strengths or persistent weaknesses. These results were compared to previous years' results and with a limited number of sections who maintained the

old method of assessment data collection. The faculty saw the aggregated results--not only were students' names removed, but section identification was also removed to reduce any desire to defend, justify, or explain weak areas of performance and instead focus attention on broader programmatic assessment.

Modifying. Currently, CCI is in the process of evaluating the written communication performance data and the new (not final) process of assessment. CCI Assessment is sharing the results with program leadership and the faculty in the courses through our college's Canvas Undergraduate Assessment project page. The Undergraduate Assessment project page presents the written communication SLO in the context of all the other programmatic SLOs.

Assessment meetings are held every Spring semester. Participating stakeholders include faculty teaching the course being assessed, program leadership, the assessment director, and the AA student. A possible future direction is to expand student participation and feedback beyond the AA. These assessment meetings, run asynchronously through our LMS system, provide an opportunity to discuss areas of strength and weakness in written communication as revealed by the new assessment process. Are we satisfied with the results? Are we surprised by any results? Are there interventions that can remedy what we view as weaknesses in written communication skills? During these meetings we will also evaluate the assessment process**.** Initial analysis shows that the old methods produce far higher assessment scores than the new process: scores that left little room for improvement. While the assessment team believes the new process produces more objective results, a consensus is sought among multiple stakeholders: assessment team, faculty, program leadership, and the student representative. Finally, during these yearly meetings we will consider whether there is a need for modifications to the training activities. While the training is the first step in the more objective assessment process, one of the benefits is the ease with which it can be recalibrated. Instead of retraining all the faculty members to adjust their scoring criteria, only the Canvas Assessment training page needs to be adjusted for next year's AA training. Faculty input and consensus is sought for those modifications.

Benefits

We want assessment to remain a dynamic, engaging process that promotes conversation among stakeholders and results in learning improvements for students, not a process that, "impos[es] standardisation of curriculum, teaching or assessment" (Sadler, 2012). Sadler identifies a benefit of engaging in calibration processes beyond the resultant scores, "through engagement with certain calibration procedures, assessors become able to tune their judgement-making ability." While it is true that moderation is often pursued to, "ensure that the mark a particular student is awarded is independent of which marker does the marking" (Sadler, 2012), it can do more. We hope that our process of moderation ensures "that an assessment outcome is valid, fair and reliable and that marking criteria have been applied consistently" (Bloxham et al., 2016), but simply pursuing moderation can have the effect

of improving the culture of assessment. “Seeing others’ marking and discussing marking decisions can have an important role in staff development and the creation of an assessment community amongst marking teams (Swann & Ecclestone, 1999)” (Bloxham, 2019). In our case this expands beyond the marking teams to those developing the training, the markers taking the training, and the faculty leading the direction and modification of the training materials.

While reimagining the assessment process improves our culture of assessment, it also has a direct impact on assessment validity. As Holzman states, “In reality, validity is a process, not an outcome . . . There are no numerical indices or benchmarks to determine whether validity has been ‘reached.’ . . . decisions about validity . . . require critical review of a body of evidence” (2020, pg. 8). The fact that we now have a system for training an assessor to both understand the definition of the SLOs and to understand what evidence indicates the level of proficiency, does not magically make the results “valid.” What it does is give the stakeholders, the aforementioned assessment personnel, faculty members, program leadership, and student representative, a better body of evidence to examine questions of validity, “a forum in which assessors establish a common vocabulary and set of meanings in relation to the mark to be awarded in that assessment event” (Sadler, 2012). Additionally, by addressing issues of calibration alongside SLO subtopic definitions our examinations of validity not only address the topics being assessed, but also create, “a shared understanding of what constitutes proficiency” (Crisp, 2017). The separate assessor helps ensure that summative assessment criteria is independent from formative grading criteria. This delineation of purpose reduces accidental crossover of summative and formative assessment goals. Faculty, in addition to recognizing established levels of proficiency, may reward student improvement, student effort, timeliness of submissions, and other measures that may serve a formative purpose without having to change hats to support programmatic summative assessments. Faculty are still integral to the programmatic assessment process through the discussion and modification of SLO definitions and proficiency levels, but those discussions benefit from temporal delineation from the other grading duties.

While the aforementioned delineation of grading and assessment tasks is a current benefit, this setup is dependent upon additional resources, namely, hiring a third-party assessor. While the hours the Assessment Assistant submitted for training and assessing were far less than a typical Teaching Assistant appointment, budgetary concerns may put this position in danger of elimination. A beneficial product of this process still remains—an assessment training procedure that defines the SLO, identifies its components, and calibrates levels of proficiency. This single training system can transfer from Assessment Assistant to multiple faculty assessors to generate discussions regarding outcomes and indicators of proficiency, periodically refresh faculty calibration, and ultimately create a body of faculty assessors who are accustomed to analyzing assessment *practices*, not just assessment *results*.

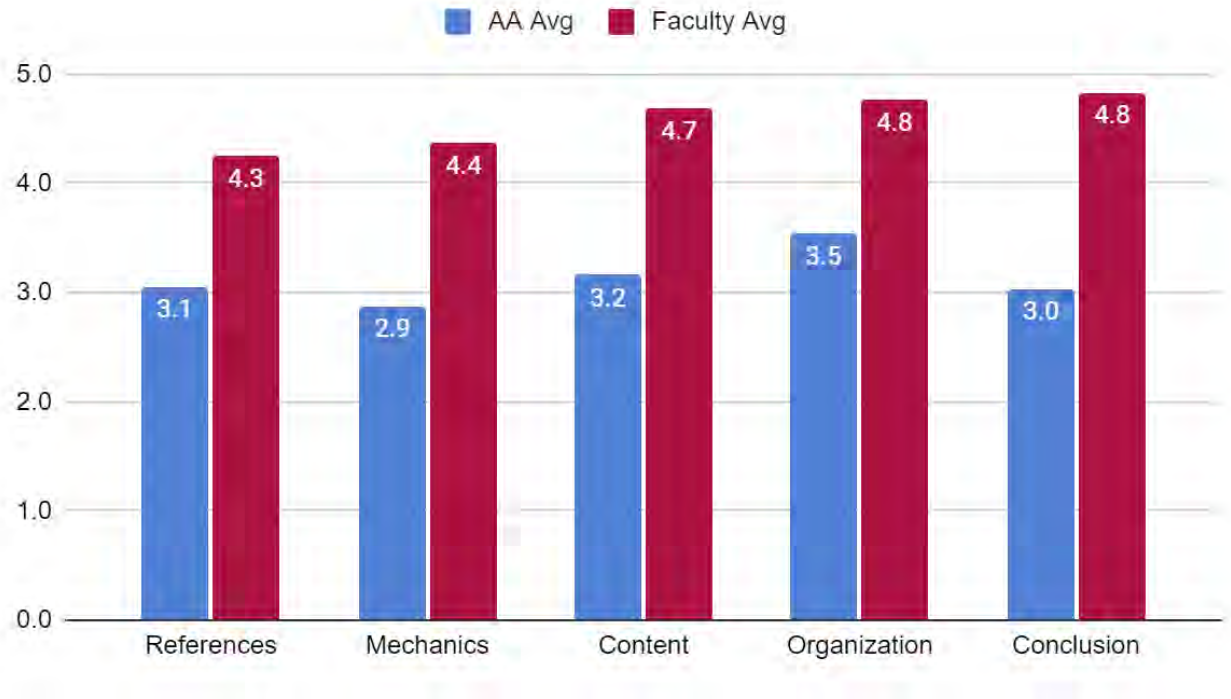
Results

The results of our first cycle of assessment reveal three items of note, 1. faculty regularly assign higher assessment scores than the Assessment Assistant (AA) with less variation, 2. the effect of the Covid-19 emergency remote teaching/learning environment is reflected to a greater extent in the AA scores and to a lesser extent in the faculty scores, and 3. the AA and faculty agree on student performance strengths and weakness in four out of five written communication skill categories.

Averages

Across the board, the average faculty scores are higher than the AA scores with far less variation. Figure 3 shows the average scores in each category with faculty rating student performance, on average, 1.45 points higher than the AA. With the exception of the Conclusion category, the ranking of category performance is rather consistent between the two groups of assessors (References and Mechanics as lowest ranking versus Content and Organization as highest ranking). This is discussed further in the *Category Performance* section.

Figure 3. Average Assessment Scores for Each Category, by Assessor Group



Note: AA (n=51) vs. faculty (n=51).

Additionally, we also found that the AA has more variation in scores. Figure 4 shows a breakdown of average scores and standard deviations for each faculty who participated in assessment using a common rubric. The higher average scores typically correspond to lower standard deviations. This indicates that while faculty may note performance weaknesses, there are some who do not rate those weaknesses as severely as others. Additionally, the AA, removed from grading responsibilities, is free to highlight writing communication weaknesses through lower scores without worrying that the students are being "punished" through poor grades. An acceptable rating (3 points) does not have to correlate to an "C" grade. This is especially true in this measurement since this assessment point is not located in a capstone course. We expect students to continue to improve in their written communication skills after completing this course; however, that improvement is difficult to measure if the majority are already earning exemplary 5-point ratings.

Figure 4. Overall Average SLO Scores and Standard Deviation of Assessors who Assessed Student Written Communication During the Spring 2020 and/or Fall 2020 Semesters A Using Common Rubric

	Mean	SD
AA	3.14	0.89
Faculty A	4.36	0.71
Faculty B	4.63	0.58
Faculty C	4.77	0.4
Faculty D	4.78	0.41

Covid-19 Emergency Response

There is a noted difference in student performance between the Spring 2020 and Fall 2020 semesters which may be a result of the emergency remote measures taken in response to the Covid-19 global pandemic. After moving from an emergency remote teaching/learning environment in the spring to a better planned, blended, remote teaching/learning environment in the fall, the faculty found student performance increased from 4.51 points to 4.69 points on average: a 0.18-point difference. Compare this to the AA rating which increased from 2.99 points to 3.35 points on average: a 0.36-point difference.

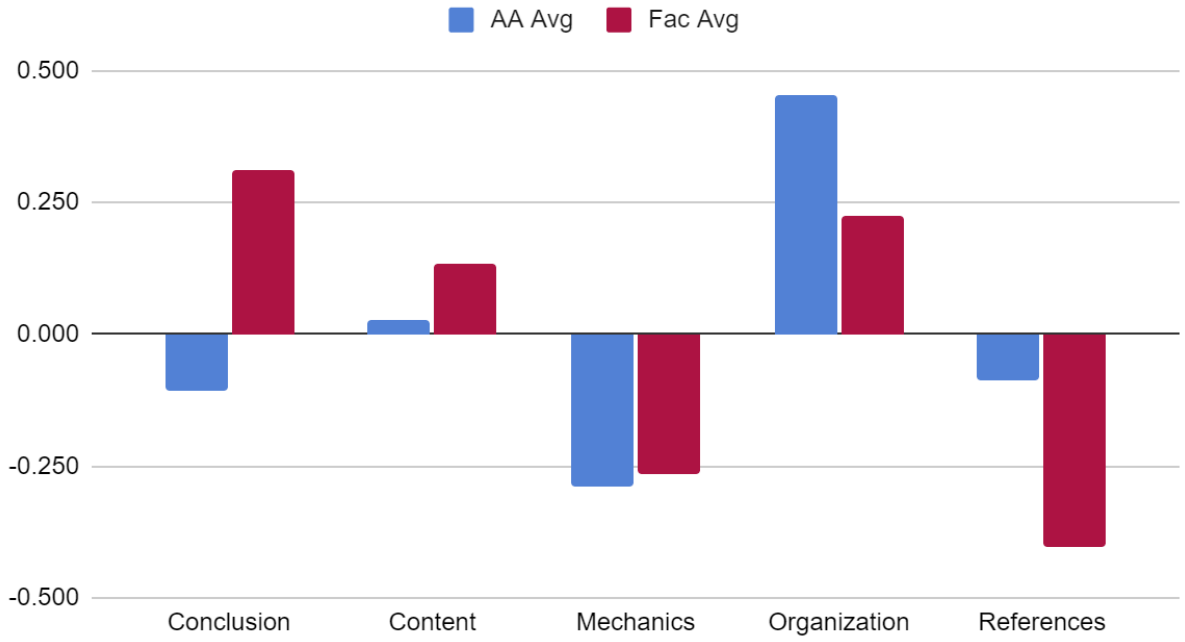
While it is completely reasonable to change expectations of students in response to a global pandemic and the resultant emergency response measures, it is also helpful to have a way to measure the resultant student performance changes outside of the grading context and away from student view. In other words, the quality of work students produce is expected to be different given the emergency context and we can measure that without penalizing students through grades or causing anxiety through visibly reduced assessment scores. Given that the AA ratings are not provided to students as judgements of their

individual performance, the AA may have felt more comfortable highlighting the impacts of the sudden disruption in teaching/learning through lower assessment scores.

Category Performance

While Figure 3 clearly shows the difference in AA vs. faculty assessment scores, the scoring differences indicate that the two types of assessors use different scales. In order to make the data comparable, we standardize the two scoring systems. This allows us to evaluate student over- and underperformance in the various written communication categories. Figure 5 shows the standardized scores—above zero indicates a relative strength, under zero indicates a relative weakness. In four out of five categories the AA and the faculty agree on student strengths and weaknesses. "Content" is a moderate strength. "Mechanics" is a significant weakness. "Organization" is a strength, although the degree of perceived proficiency differs. "References" is a weakness, although, again, the extent of the weakness differs.

Figure 5. Standardized Category Performance by Assessor Group



Note: Standardized scores account for the different scales used by the two groups of assessors. Above zero indicates a strength, below zero indicates a weakness. Four out of five categories show directional agreement between the AA and faculty.

A noticeable point of disagreement between the AA and the faculty concerns the Conclusion category. The faculty rated students' abilities to provide a justified and supported conclusion as a strength while the AA rated it as a weakness. Since the faculty are the subject matter experts, it is imperative to gain their input in correcting this discrepancy. The third-party assessor should be trained as an unbiased representation of faculty expectations not as an

agent of contradiction. This point of disagreement provides our assessment team with an opportunity for fruitful discussion concerning the expectations for a well-supported conclusion. This has identified, not an issue of calibration, but of definition—the AA and faculty do not agree on the components of an acceptable conclusion. Before the Fall 2021 semester, faculty will more precisely define the subtopic, review the stated expectations, and suggest how to adjust the training materials.

Conclusion

Assessment data that suggests further exploration is needed presents the opportunity for organizations to not only reexamine data points and adjust analyses, but to rethink entire assessment processes. Rethinking processes not only gives stakeholders the opportunity to redesign systems of data collection, redefine outcomes, and redetermine definitions of outcomes of proficiency, it also reasserts that assessment is not static activity, but one that is subject to discussion, negotiation, and modification. Developing an assessment training course forced our organization to abandon assumptions and old conventions, and declare that change is not only possible but desirable. While we are reaping the benefits of better-defined outcomes and levels of proficiency, more efficient uses of the learning management system, and cleaner delineations of grading and assessing tasks, the lasting benefit may prove to be an improved culture of assessment in which each stakeholder understands their role and how their voice can positively influence programmatic decisions and pedagogical directions. In this new landscape, assessment findings will more directly guide learning improvements.

References

Advance HE. (2018). *What is calibration?* https://www.heacademy.ac.uk/project-section/what-is-calibration

Blaich, C. F., & Wise, K. S. (2011). From gathering to using assessment results: Lessons from the Wabash National Study (NILOA Occasional Paper No.8). National Institute for Learning Outcomes Assessment. https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/OccasionalPaper8.pdf

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment and Evaluation in Higher Education*, *34*(2), 209–220. https://doi.org/10.1080/02602930801955978

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education*, *41*(3), 466–481. https://doi.org/1080/02602938.2015.1024607

Crisp, E. (2017). Calibration: Are you seeing what I'm seeing? *Intersection: A Journal at the Intersection of Assessment and Learning*, *Winter*, 7-13.

Eubanks, D. (2017). A guide for the perplexed. *Intersection: A Journal at the Intersection of Assessment and Learning*, *Fall*, 4-14. https://www.aalhe.org/assets/docs/AAHLE_Fall_2017_Intersection.pdf

Fulcher, K., H., Smith, K. L., Sanchez, E. R. H., Ames, A. J., & Meixner, C. (2017). Return of the pig: Standards for learning improvement. *Research & Practice in Assessment, 11*, 10-40. http://www.rpajournal.com/dev/wp-content/uploads/2017/03/A2.pdf

Holzman, M., Pope, A. M., & Horst, J. (2020). Reliability and validity 101: A primer for student affairs assessment. *Intersection: A Journal at the Intersection of Assessment and Learning*, *Winter*, 5-10.

Kim, D., & Helms, M. M. (2016). Assessing faculty bias in rating embedded assurance of learning assignments. *Journal of Education for Business, 91*(3), 159-165.

O'Connell, B., De Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B., & Watty, K. (2015). Does calibration reduce variability in the assessment of accounting learning outcomes? *Assessment & Evaluation in Higher Education*, *41*(3), 331-349. https://doi.org/10.1080/02602938.2015.1008398

Pasquerella, L. (2018). Planning for America's future: Educating for democracy. *Liberal Education*, *104*. https://www.aacu.org/liberaleducation/2018/winter/pasquerella

Sadler, D. R. (2012). Assuring academic achievement standards: From moderation to calibration. *Assessment in Education: Principles, Policy & Practice, 20*(1), 5-19. https://doi.org/10.1080/0969594x.2012.714742

Singer-Freeman, K. E., & Robinson, C. (2020a). Grand challenges for assessment in higher education. *Journal of Research and Practice in Assessment*, *15*(2), 1-20. https://www.rpajournal.com/dev/wp-content/uploads/2020/11/Grand-Challenges-for-Assessment-in-Higher-Education.pdf

Singer-Freeman, K. E., & Robinson, C. (2020b). Grand challenges in assessment: Collective issues in need of solutions (NILOA Occasional Paper 47). National Institute for Learning Outcomes Assessment. https://www.learningoutcomesassessment.org/wp-content/uploads/2020/11/GrandChallenges.pdf

About the Authors

Dr. Colleen Karnas-Haines is the Director of Assessment, Planning, and Accreditation in the College of Computing and Informatics at the University of North Carolina at Charlotte. She can be reached at: ckarnash@uncc.edu