Abstract

Meta-assessment, or the assessment of the outcomes assessment process, is a useful strategy to communicate, guide, document, and provide feedback on assessment practices. We share the creation of a rubric that aligns with assessment processes aimed to improve student learning and development in higher education. More specifically, the rubric was created to align with the student affairs professional standards and explicitly evaluates equity-related aspects of each component of the outcomes assessment process. Additionally, we highlight the rubric as a necessary step in a broader change management effort. We then share procedures of the initial use of the rubric to evaluate assessment reports. After analyzing rubric scores (i.e., G-study) and qualitative feedback from several raters, we then describe additional support resources intentionally created to facilitate more reliable ratings. We freely share the rubric, training reports, and support materials to enable a culture of improvement in higher education institutions.

AUTHORS

Sara J. Finney, Ph.D.
James Madison University

Jonathan P. Stewart, Ph.D.
James Madison University

Autumn N. Wild, M.A.
James Madison University

Riley K. Herr, M.A.
James Madison University

Katarina E. Schaefer, M.A.
James Madison University

A Meta-Assessment Rubric to Guide Professional Development and Practice in Equitable Outcomes Assessment

The outcomes assessment process facilitates the evaluation of higher education programming to meet intended student learning and development outcomes. The information gathered via the assessment process can uncover needed changes to programming to better achieve intended outcomes. However, the usefulness of assessment data to guide program improvement depends on its quality and the process to gather outcomes assessment data.

What is Meta-Assessment and its Utility?

Meta-assessment is the evaluation of the quality of the assessment process (e.g., Fulcher & Good, 2013). Stemming from the evaluation literature (Ory, 1992), the practice of meta-assessment developed to meet the increased engagement in outcomes assessment in higher education. That is, guidelines and standards associated with outcomes assessment needed to be developed to guide high-quality processes and data. Meta-assessment served this need for both summative (for accountability) and formative (for improvement) purposes (McDonald, 2010).

To structure a meta-assessment process, a rubric or checklist is typically developed that specifies characteristics of high-quality assessment practice. Ratings from the rubric or checklist can serve as evidence of engagement in the assessment process for accreditation

CORRESPONDENCE
Email
finneysj@jmu.edu

purposes. Likewise, ratings allow administrators to efficiently identify programs that have evidence of impact versus programs where the level of effectiveness is unknown.

A second, and equally important, use of a meta-assessment rubric is to clearly communicate expectations related to assessment. Bichsel et al. (2023) reported that approximately 39% of student affairs professionals were either likely or very likely to seek other employment opportunities within the next year. Thus, when professionals experienced with assessment practice leave the university, we must offer and facilitate assessment-related training for those new to both the institution and assessment practice. Our meta-assessment rubric is an efficient way to introduce the assessment process and describe quality practice. We can then offer training aligned with specific aspects of the assessment process outlined in the rubric.

A third use of meta-assessment rubrics is for continuous improvement (Fulcher et al., 2016). Rubrics can be intentionally designed to prompt improvements to assessment processes or programming. Ratings can direct action and identify if support is needed for improvement work.

Why Another Meta-Assessment Rubric?

Given their benefits, meta-assessment rubrics have been developed at numerous institutions. Thus, you may question why we believed another rubric was needed. Before sharing our new rubric and training materials, we want to acknowledge existing rubrics and how they influenced the creation of our rubric. We then discuss the unique qualities of our rubric.

Existing Meta-Assessment Rubrics

When reviewing existing meta-assessment rubrics, we were fortunate to find many exemplars. Our goal when reviewing these existing rubrics was to determine what rating structure and content fit best with the needs of student affairs professionals.

Our first and primary inspiration was the *Assessment Progress Template (APT) Rubric* from James Madison University (Fulcher & Orem, 2010; James Madison University, 2015). Used to assess programs in the division of academic affairs (e.g., B.A. in Psychology, B.S. in Mathematics), the APT rubric consists of four developmental levels (beginning, developing, good, exemplary) spanning six criteria (e.g., improvement of assessment process). The rubric structure, use of developmental levels, and detailed criteria enhanced the utility of the APT rubric for pedagogical purposes.

A third use of meta-assessment rubrics is for continuous improvement. Rubrics can be intentionally designed to prompt improvements to assessment processes or programming.

Other inspirations for our rubric included the University of the District of Columbia's (UDC) *Meta-Rubric for Evaluating Institutional Assessment Reports* (2023). UDC's rubric includes space for rater comments and a glossary of terms, both of which we decided to include in our rubric. Including these aspects in our rubric would enhance the quality of feedback provided and the didactic nature of the rubric. We structured our developmental levels as "Exemplary," "Proficient," and "Developing" based off the UDC rubric. The labeling of the "Missing" category was inspired by Wayne State University's *Assessment Practices Feedback Rubric* (2023), which includes a "Not Submitted" category rather than a "Not Developed" or "Needs Attention" category. We used the label "Missing" to encourage submission of assessment reports for review even if processes were incomplete (e.g., learning outcomes are complete but measures are "missing"). We were also inspired by Texas A&M International University's *Assessment Plan Rubric* (2021), which assigned numeric values to levels of development. We felt these scores would facilitate interpretation of improvement over time, something leadership in the division values.

Need for a New Rubric

We evaluated existing rubrics to identify criteria for quality assessment that were relevant for structuring, evaluating, and providing guidance on improving student affairs programs. Moreover, we determined which criteria for quality assessment were missing from existing rubrics. Below we discuss two key features of our rubric: alignment with the professional standards of student affairs and an explicit and consistent focus on equity.

Alignment with Student Affairs Professional Standards. Professional standards provide one means to communicate best practice for programs and personnel (Finney & Horst, 2019b). Regarding outcomes assessment, three sets of professional standards have been mapped directly to the outcomes assessment cycle (Finney & Horst, 2019a): two personal competency standards (ASK Standards, American College Personnel Association, 2006; ACPA-NASPA Competencies, American College Personnel Association & National Association of Student Personnel Administrators, 2015) and one program-related set of standards (CAS, Council for the Advancement of Standards in Higher Education, 2023). It was imperative that our rubric aligned with this mapping.

However, training in assessment and training in equity and inclusion have historically been separate rather than integrated. Thus, our rubric purposefully intertwines assessment and equity.

Unlike existing meta-assessment rubrics, we incorporated the following aspects of outcomes assessment into our rubric to align with current professional standards: program theory (Finney et al., 2021; Pope et al., 2019, 2023; Smith & Finney, 2020), evidence-informed programming (Finney & Buchanan, 2021; Horst et al., 2021), and implementation fidelity (Fisher et al., 2014; Gerstner & Finney, 2013; Smith et al., 2019). For example, the current CAS Standards (2023), which are often used for program reviews, make explicit calls for articulating program theory, employing evidence-informed programming, and gathering implementation fidelity data. As just a few examples, CAS (2023) states that *each* functional area *must*:

- "Provide a research-informed, theory-informed, or evidence-based rationale for designing programs and services, strategies, and tactics intended to influence student learning, development, and success goals" (p. 44).
- "Use theory, research, and evidence to develop and implement its programs and services to achieve stated mission, goals, and outcomes" (p. 46).
- "Document the extent to which intentionally designed programming, strategies, and tactics are implemented as planned" (p. 45).

Moreover, student affairs educators are expected to demonstrate competency in equity and inclusion (West & Henning, 2023). However, training in assessment and training in equity and inclusion have historically been separate rather than integrated (Henning & Lundquist, 2018). Thus, our rubric purposefully intertwines assessment and equity, as described below.

Explicitly Addressing Equity in our Rubric

> Given the central focus of students in the work of student affairs practitioners, student affairs professionals and our close partners in higher education are uniquely positioned to engage in assessment practices which center the lived experiences of historically underserved students, to challenge policies and processes which foster inequities, and to champion a better future for students by leveraging data to advance equity (Heiser, Schnelle, & Tullier, 2023, p. 4).

We agreed with Heiser and colleagues and thus purposefully integrated equity considerations within each criterion of our rubric. "Equity" is not simply the last rubric criterion which could be ignored. Instead, equity considerations are ever-present sub-criteria for each rubric criterion, priming professionals to consider equity throughout the assessment process.

The equity sub-criteria were informed by scholars in higher education assessment (e.g., Henning & Lundquist, 2022; Montenegro & Jankowski, 2017, 2020), educational measurement (e.g., Randall, 2021; Randall et al., 2022; Russell, 2023, 2024), and culturally-responsive evaluation (e.g., Hood et al., 2015). These scholars offered definitions and frameworks for equity-centered assessment, characteristics of assessment for social justice, and strategies for culturally-responsive use of results for improvement. Our goal was to overtly infuse these ideas into common assessment practice to offer explicit equity-focused "moves" via the rubric. Thus, our rubric addresses barriers to infusing equity in assessment that we hear often: no time to read equity-related scholarship and translate it into practice; definitions

and frameworks of equity in assessment do not facilitate specific assessment-related action; and unclear how to improve current efforts to infuse equity in assessment. These are not the only barriers (Heiser et al., 2023a, 2023b), but we believed a meta-assessment rubric would address these barriers. We are not the first to suggest the utility of a meta-assessment rubric to center equity in assessment. When discussing ways to cultivate equitable assessment practice, Levy and Heiser (2018) suggested "Institutions may want to create a meta-assessment rubric or checklist to help ensure assessment practice is following proper process as intended by the institution in accordance with institutional goals and values" (p. 3). They suggested sharing a clear vision of high-quality assessment (i.e., a rubric) to combat unchecked biases.

The Student Affairs Assessment Improvement Rubric

The Student Affairs Assessment Improvement Rubric was created to define what we mean by "high-quality" assessment practice and to provide constructive feedback to educators (Finney et al., 2024). The goal of our rubric is continuous improvement, as reflected in its name and contents. Multiple iterations of the rubric were created over two years by colleagues in the division of student affairs and our office for assessment. Moreover, the rubric was written to be accessible to those new to the assessment process; hence, the didactic presentation of terms (i.e., glossary at end) and steps.

Given previous meta-assessment rubrics and student affairs professional standards related to assessment, it was easy to articulate the general criteria of the rubric: student learning and development outcomes, program theory, selecting or designing measures, implementation fidelity, gathering data, analyzing data and reporting findings, and using results for improvement. Articulating the more specific sub-criteria of the general criteria took more thought and time. Those experienced in outcomes assessment may feel some sub-criteria are basic and thus unnecessary (e.g., "outcomes are student-focused"). However, we believe including these sub-criteria can guide novices and be easily achieved when first starting the assessment process. By far the most difficult and time-intensive task was articulating characteristics that differentiated "exemplary," "proficient," and "developing" practice. These descriptions needed to be concise yet detailed enough to distinguish the three levels. We had experts in meta-assessment and equity in assessment review these descriptions and provide feedback prior to our training on the rubric.

Regardless of why change is needed, leading and maintaining change in higher education can be difficult due to complex organizational and power structures.

Our Rubric Development Process and its Relation to Change Management in Our Division

Early conversations on developing a meta-assessment process for our student affairs division began in fall of 2019 after new leadership highlighted a significant need for changing the culture around assessment and evidence within the division. Regardless of why change is needed, leading and maintaining change in higher education can be difficult due to complex organizational and power structures (Buller, 2014; Clark, 2003; Strine-Patterson, 2022). Within a division, one unit might operate in a more hierarchical top-down manner, whereas another operates with a flatter distribution of power, influence, and decision making. Thus, regardless of whether you are exploring meta-assessment as a tool for broad cultural change or for leading assessment practices on your campus, it is important to understand your institution's unique politics and power structures to lead effectively (Henning & Roberts, 2024; Roberts, 2024).

Steps to establish a culture of evidence at our institution reflect Kotter's (2012) eight-step model for managing change. The model starts with establishing a sense of urgency and ends with incorporating change into the culture. To establish a sense of urgency, people need to understand why change is needed. Change within an institution typically happens for three reasons: 1) change is forced on the institution from external forces (reactive change); 2) the institution knows change will eventually be forced on them from external forces (proactive change); or 3) change is required due to internal rather than external factors (interactive change) (Buller, 2014).

In our case, all three reasons were occurring simultaneously. Thus, we were able to leverage multiple angles to help build a sense of urgency for changing from what Culp and

Dungy (2012) refer to as a "culture of good intentions" toward a "culture of evidence." First, some units had not yet responded effectively to expectations from higher education regarding assessment (reactive change). Second, external expectations for evidence of effectiveness will only increase; thus, the ability of units to "hide" behind the excellence of other units will end (proactive change). Third, a needs assessment examining behaviors, perceptions, and barriers related to the use of theory and research in program development indicated several areas of concern and recommendations for improvement (interactive change). Also, concerns were expressed to leadership from discouraged staff that no one was reading their assessment reports. Thus, a sense of urgency toward changing the culture of assessment was appearing at all levels.

With leadership and partners across the university serving as a guiding coalition and the "change vision" of moving toward a culture of evidence underway (Kotter's next two steps for managing change), we began planning. The curricular approach to student learning (Kerr et al., 2020) was a way to communicate our vision of a culture of assessment, secure buy-in, and create opportunities for broad-based action (steps three and four of managing change). For example, by agreeing to a curricular approach to student learning, we committed to a cycle of assessment.

When discussing how to build a cycle of assessment, divisional leadership decided that framing within a meta-assessment process would be ideal. In June 2021, directors were introduced to meta-assessment via professional development. Directors were informed that intentional planning would begin with development of a meta-assessment rubric. In fall of 2021, division leadership and staff serving on the division's assessment council were invited to participate in a rubric development workshop. From there, staff that expressed continued interest were invited to a working group to develop the rubric. Through these actions, we were moving along Kotter's (2012) sixth ("short term wins") and seventh ("never let up") steps.

The rubric development group was comprised of five student affairs professionals from different departments, one faculty member from academic affairs, and three graduate students. This group met weekly for one-hour working sessions from February through July 2022 and bi-weekly from then on. The consistency of these sessions allowed the team to develop the rubric and rater training over two years.

Development of Rater Training Materials and Process

Planning for rater training began simultaneously with rubric development. Essential questions such as who would be invited to participate, how many reports we would like to rate for the pilot, and how many raters we would like for each report were discussed before the initial draft of the rubric began. However, more specific planning could only begin after drafting the rubric. Once the criteria and sub-criteria were established, a report template was created and distributed to staff across the division. Multiple communications were sent regarding the rater training and rating pilot that included reinforcement of our change vision, encouragement from our vice president, and multiple avenues to participate in the process.

The curricular approach to student learning was a way to communicate our vision of a culture of assessment, secure buy-in, and create opportunities for broad-based action (steps three and four of managing change).

Ten student affairs professionals, representing many offices, engaged in a three-day rater training to examine the rubric and learn how it could inform assessment practice. Of these 10 professionals, four were paid ($500) raters of actual Assessment Improvement Reports submitted by offices for feedback on assessment practice (additional two days). During training, we reinforced why the division was implementing a meta-assessment process, guided a detailed review of the rubric, and provided the opportunity, time, and resources needed to rate multiple mock reports. A detailed schedule of the training is in the Appendix.

With the implementation of a new meta-assessment process in student affairs, it was necessary to create tools that demonstrated different levels of assessment practice. Three mock reports were created based on hypothetical but realistic programs. Reports were designed to have an average rating corresponding with a level of the rubric: exemplary, proficient, or developing. Rating keys were generated to demonstrate how each reports' scores reflected

the specific level. Keys included detailed comments to provide explanations of how each report differed in quality.

Prior to the rater training, assessment consultants participated in rating the mock reports to identify issues with the reports themselves or their keys. When raters did not align, the mock reports were edited to increase consistency. Additionally, applying the rubric to the mock reports highlighted issues with the rubric itself (e.g., unclear transitions, confusing wording). Thus, amendments were made to the rubric. The rubric and mock reports were then used during rater training to guide raters on how to evaluate assessment reports submitted for review.

The G-coefficient, a reliability index analogous to classical test theory reliability, equaled 0.98. Thus, we have evidence that reports (exemplary, proficient, developing) can be rated consistently.

Evaluating the Quality of the Ratings from the Rubric

Generalizability Theory (G-theory) was used to gather evidence of the quality of ratings when piloting the rubric. A G-study provides reliability-like coefficients (e.g., G-coefficient). It also partitions the variability of ratings. For our purpose, we wanted to ensure a large amount of variance in ratings was due to three different reports reflecting three different levels of development (developing, proficient, exemplary). Next, we wanted to determine whether there was rating variation due to rater harshness (some raters consistently rating reports very high or very low, regardless of the report or sub-criteria being rated, which is not desirable). Finally, we wanted to investigate how much variability was due to different sub-criteria (did some sub-criteria receive consistently lower ratings, regardless of report or rater, which is not desirable). For example, we expected some variability to be due to sub-criteria, given the equity sub-criteria were designed with some lower ratings, regardless of report. Although G-theory provides some validity evidence, further evidence was gathered by the match between professionals' ratings and each report key. We wanted ratings within a half point (.50 on the 0 to 3 scale, where 0 = missing, 1 = developing, 2 = proficient, and 3 = exemplary) of the key.

Data Collection Design and Method

The exemplary report was rated first, followed by the proficient report, and then the developing report. We expected a better match between the key and the ratings as raters completed additional reports (i.e., raters might have more difficulty rating the first report). The G-study design was fully crossed (all 10 raters rated all three mock reports using all criteria). The variance of ratings was partitioned into variation due to the object of measurement (three reports of varying quality), raters, sub-criteria, their interactions, and error. All effects were treated as random. For raters, random means that raters have been theoretically sampled from a universe of possible raters. For sub-criteria, random means the sub-criteria were sampled from a universe of sub-criteria that reflect parts of the assessment process. Sub-criteria purposefully designed to be "Missing" were not included in the G-study.

Results and Interpretations

The G-coefficient, a reliability index analogous to classical test theory reliability (Shavelson & Webb, 1991), equaled 0.98. Thus, we have evidence that reports (exemplary, proficient, developing) can be rated consistently. The relative standard error of measurement (SEM) was 0.113, which is useful in gauging rating precision. For example, the SEM can be used to create a confidence interval around an average mock report rating of 2 (rating of proficient). Multiplying the SEM by a critical z-value (e.g., 1.96) results in a plausible range of values for the proficient report between 1.76 and 2.24 (on the rubric scale of 0 to 3), indicating the amount of uncertainty in ratings. The high number of raters and low number of reports influenced the size of the G-coefficient and SEM. Without the extensive rater training, the G-coefficient likely would have been lower and the SEM would have been higher.

To further understand the consistency in ratings, the percent of variance associated with each variance component (variance due to report quality, raters, sub-criteria) and interpretations are included in Table 1. Much of the variance in ratings was due to the object of measurement, which is desirable. That is, 46.5% of the variance in ratings was due to systematic differences in the quality of the mock reports. Fortunately, there was no variance in ratings due to rater harshness (0%) and very little variance was due to systematic rater

harshness by sub-criteria (0.5%). In other words, raters were successfully trained to avoid being systematically too harsh or lenient across mock reports or on different sub-criteria.

Some of the variation in ratings (12.8%) was due to various sub-criteria receiving systematically higher or lower ratings, regardless of rater or mock report quality. We suspected the ratings of the equity sub-criteria were underlying this variation. The mock reports attempted to prompt very low ratings (developing report), moderate ratings (proficient report), and high ratings (exemplary report). However, some of the equity components had lower scores by design. For example, in the exemplary report, the text was designed to elicit the highest rating (i.e., 3) for most sub-criteria, with the exception of the equity sub-criteria where the text reflected slightly lower quality. It proved difficult to craft text that reflected exemplary levels of the equity sub-criteria. Thus, we were not surprised that some variability in rating was due to sub-criteria.

Table 1
Variance Components & Descriptions from G-Study of Ratings of Mock Reports

Variance Component	Variance	% of Variance	Interpretation
Mock Report	0.703	46.5%	Variance in ratings due to differences in quality of the reports (developing, proficient, exemplary). *Interpretation*: 46.5% of the variance in ratings was due to differences in report quality. Fortunately, raters consistently rank ordered different reports.
Rater	-0.023	0.0%	Variance due to rater harshness. *Interpretation*: 0% of the variance was due to rater variance. Fortunately, raters were not systematically harsh or systematically lenient.
Sub-Criteria	0.193	12.8%	Variance in ratings due to sub-criteria. *Interpretation*: 12.8% of the variance was due to the sub-criteria being rated. For example, the equity sub-criteria were rated lower, on average, than other sub-criteria, regardless of rater or mock report.
Mock Report x Rater	0.088	5.8%	Variance due to interaction between mock report and rater. *Interpretation*: Fortunately, not much variation in ratings (only 5.8%) was due to different raters rating reports in different ways.
Mock Report x Sub-Criteria	0.077	5.1%	Variance due to interaction between mock report and sub-criteria. *Interpretation*: Fortunately, not much variation in ratings (only 5.1%) was due to different mock reports having sub-criteria rated in different ways.
Rater x Sub-Criteria	0.007	0.5%	Variance in ratings due to interaction between raters and sub-criteria. *Interpretation*: Fortunately, only 0.5% of variation in ratings was due to different raters systematically rating sub-criteria differently.
Mock Report x Rater x Sub-Criteria	0.445	29.4%	Error

Note: GENOVA (v3.1) uses ordinary least squares estimation. Thus, small negative variances are interpreted as 0.

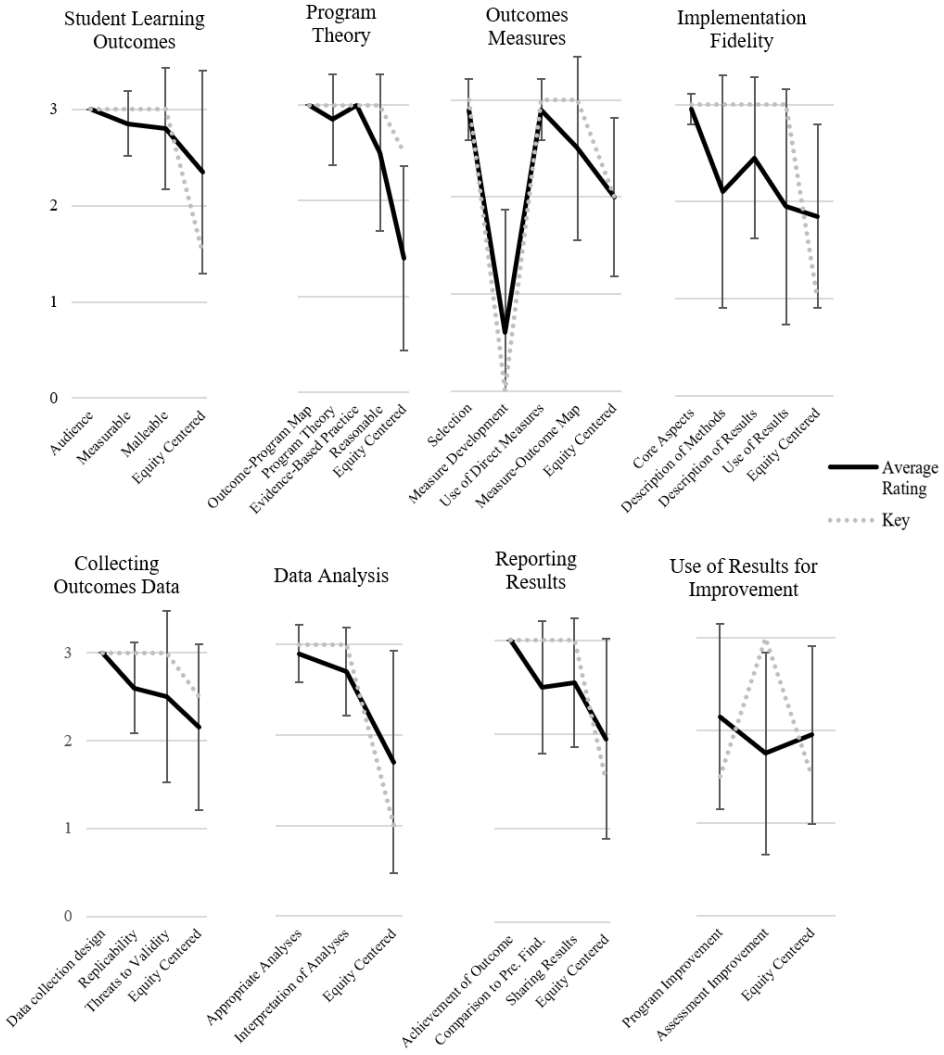
The G-study does not uncover which sub-criteria were causing variation in ratings. Thus, we employed a key matching analysis (average sub-criteria rating compared to correct rating) to identify areas of concern (see Table 2 and Figure 1). Regarding report quality, raters had the most difficulty matching the key for the proficient mock report. Although there was some variability in proportion of matches between the exemplary, proficient, and developing mock reports, half of the raters (58.5%, 50.3%, 54.8%, respectively) exactly matched the keys. More raters (68.8%, 62.1%, 66.1%) were within 0.5 of the keys. Finally, most raters (81.5%, 87.9%, 93.9%) were within a maximum of 1 point of the keys.

Table 2
Key Matching Percentages by Mock Report Level

	Mean (SD)		
	Exemplary	Proficient	Developing
Exact match with key	58.5% (11.5%)	50.3% (8.2%)	54.8% (10.7%)
Within half a point of key	68.8% (8.6%)	62.1% (8.1%)	66.1% (10.7%)
Within one point of key	81.5% (6.1%)	87.9% (4.0%)	93.9% (4.0%)

Note: For each report, percentages shown are averages across 10 raters and 31 sub-criteria.

Figure 1
Exemplary Mock Report Average Ratings and Key by Sub-Criteria



Note: Error bars indicate one standard deviation above and below the average rating.

Regarding sub-criteria, Figure 1 displays the average ratings by sub-criteria for the exemplary mock report in comparison to the rating key. For the equity sub-criteria, ratings were generally more variable across raters (larger error bars) than the other sub-criteria.

We were satisfied with the reliability of the ratings as well as how close most of the ratings were to the mock report keys. These results provided evidence of the success of the rubric training. The lack of variability due to raters (i.e., raters rating consistently harsh or lenient) and the rater by sub-criteria interaction justifies the multiday training required to calibrate the raters and speaks to the clarity of the mock reports.

Table 3
Qualitative Themes from Rubric Training by Question

Question Posed to Raters	Qualitative Themes
What was the best aspect of this PD?	• order of rater training process (i.e., rating individually, with partner, then with whole group) • enthusiasm of the facilitators • using the rubric/training resources to learn how to write a high-quality assessment report
What did you learn?	• improved understanding of assessment • expectations for assessment reports
What was the worst aspect of PD?	• first two days of training were too lecture-heavy • information presented first two days was too dense
Do you have suggestions for improving this PD?	• more interactive activities during first two days • more time spent on equity-centered assessment • more emphasis on the rubric being a means to guide improvement (not penalize programs)
How can you use what you learned (resources you gathered) in your office?	• can speak about the value of the training at future within-office and division-wide meetings • can promote partnership between student affairs division and assessment professionals on campus
What additional resources do you need to facilitate rating reports?	• more examples of how to write comments
Are there any changes to the rubric you'd like to suggest?	• providing more clarity in text that describe each exemplary/proficient/developing rating • adding a general comment box to the end of rubric

Note: PD = Professional Development.

Qualitative Feedback from Training

The G-study and key matching analyses provided insight regarding the success of our training. To supplement the quantitative G-study results, we solicited qualitative feedback from raters to identify strengths and weaknesses of our week-long training. Specifically, at the conclusion of our workshop, we posed a series of open-ended questions to elicit rich insight from raters. Questions pertained to the quality of the training and rubric. Facilitators transcribed notes while raters provided feedback, which were later compiled into one summary of comments (see Table 3). This feedback illuminated areas of improvement for future rater training.

Generally, participants requested alterations to the structure of the training. Their primary concern was the first two days of the training were too dense (e.g., too many materials, too much content, lecture heavy, lack of interactive activities). We addressed this feedback in two ways. First, we incorporated more engaging discussion-based activities into the training and distributed the training across two weeks instead of one week. Second, we created a nine-month professional development session that introduced each step of the assessment cycle and provided time to work on assessment-related activities. The latter facilitated engagement in assessment throughout the year and the creation of new assessment reports to rate during training.

We were satisfied with the reliability of the ratings as well as how close most of the ratings were to the mock report keys. These results provided evidence of the success of the rubric training.

Notably, one theme that emerged from qualitative feedback aligned with results from the G-study. Specifically, raters mentioned that increased equity resources would strengthen our rater training workshop, suggesting their knowledge was insufficient to accurately rate equity-centered sub-criteria and provide feedback. Recall the high variability across raters

Our new equity-centered resource has two main purposes: 1) increase equity-centered assessment practice on campus via examples of high-quality practice and 2) produce more accurate ratings and feedback on equity sub-criteria of our rubric.

and notable deviation from the key with respect to equity sub-criteria ratings, regardless of mock report or rubric criterion (Figure 1). Thus, it was necessary to create an additional equity-centered resource that aligned with our rubric.

A Resource to Model Equity in Assessment

Experts of culturally responsive assessment (e.g., Montenegro & Jankowski, 2017, 2020) emphasize the importance of: 1) embedding equity-centered assessment into professional development and 2) using assessment results to make evidence-based changes to improve equity. Thus, using hypothetical student affairs programs, we created a new resource that depicted equity practices that would reflect an "exemplary" rating. Text was created for each equity-centered sub-criterion (e.g., student learning outcomes, implementation fidelity, using results for improvement). We then deliberately altered the "exemplary" text to reflect "proficient" equity-centered practices and "developing" equity-centered practices. We purposely highlighted aspects of the text that would prompt the "exemplary," "proficient," or "developing" rating.

Our new equity-centered resource has two main purposes: 1) increase equity-centered assessment practice on campus via examples of high-quality practice and 2) produce more accurate ratings and feedback on equity sub-criteria of our rubric. We consulted various sources (Brocato et al., 2021; Cerna et al., 2021; Montenegro & Jankowski, 2020) to inform this equity-centered resource. The resource is freely available at our Open Educational Resources website.

Conclusion

We share our meta-assessment rubric, mock reports, training information, and equity resource in the spirit of advancing high-quality programming and outcomes assessment. We believe this package of assessment-related support materials can help to build or reinforce an institution's quality improvement process. Most importantly, we hope our colleagues find these resources useful for considering how to apply an equity frame when engaging in continuous improvement efforts. In turn, we should increase the odds that all students will benefit from high-quality, equitable programming on our campuses.

Appendix

Day 1	
Time	**Topic**
9:00 – 9:30am	Introductions and Overview
9:30 – 10:00am	The What, Why, and How of Improvement Reports
10:00 – 10:30am	Questions that Matter to SA Work: Utility of Assessment
10:30 – 10:40am	Break
10:40 – 12:30pm	Detailed Review of the Assessment Improvement Rubric
12:30 – 1:30pm	Lunch
1:30 – 2:00pm	Resources Overview
2:00 - 2:30pm	Finding Evidence to Inform Rater Feedback
2:30 - 3:00pm	Rater Adjudication
3:00 – 3:10pm	Break
3:10 - 3:40pm	System to House Ratings and Comments
3:40 – 4:00pm	Questions & Answers
Day 2	
9:00 – 9:15am	Overview of the Day
9:15 – 12:15pm	Rate Mock "Exemplary" Report 9:15 – 11:00am • Read report on own • Rate report on own & provide written feedback 11:00 – 11:45am • Adjudicate with partner 11:45 – 12:15pm • Full group debrief
12:15 – 1:15pm	Lunch
1:15 – 3:45pm	Rate Mock "Proficient" Report
3:45 – 4:00pm	Questions & Answers
Day 3	
9:00 – 9:15am	Overview of the Day
9:15 – 12:15pm	Rate Mock "Developing" Report

References

American College Personnel Association. (2006). *ASK standards: Assessment skills and knowledge content standards for student affairs practitioners and scholars*. Washington, DC: Author.

American College Personnel Association & National Association of Student Personnel Administrators. (2015). *ACPA/NASPA professional competency areas for student affairs educators*. Washington, DC: Authors.

Bichsel, J., Fuesting, M., Tubbs, D., & Schneider, J. (2023, September). *The CUPA-HR 2023 higher education employee retention survey*. College and University Professional Association for Human Resources. https://www.cupahr.org/surveys/research-briefs/higher-ed-employee-retention-survey-findings-september-2023/

Brocato, N., Clifford, M., Brunsting, N., & Villalba, J. (2021, February). *Wake Forest University: Campus life and equitable assessment* (Case Study Example). National Institute for Learning Outcomes Assessment, Council for the Advancement of Standards in Higher Education, and Anthology. https://www.learningoutcomesassessment.org/wp-content/uploads/2021/02/EquityCase-WFU-2.pdf

Buller, J. (2014). *Change leadership in higher education: A practical guide to academic transformation* (1st ed.). Jossey-Bass. http://doi.org/10.1002/9781119210825

Cerna, O., Condliffe, B., & Wilson, A. (2021). *Guiding questions for supporting culturally responsive evaluation practices and an equity-based perspective* [Issue Focus]. MDRC. https://www.mdrc.org/sites/default/files/Equity-Guiding_Questions.pdf

Clark, B. (2003). Sustaining change in universities: Continuities in case studies and concepts. *Tertiary Education and Management*, *9*(2), 99–116. https://doi.org/10.1080/13583883.2003.9967096

Council for the Advancement of Standards in Higher Education. (2023). *CAS professional standards for higher education* (11th ed.). Washington, DC: Author.

Culp, M., & Dungy, G. (2012). *Building a culture of evidence in student affairs: A guide for leaders and practitioners*. NASPA Student Affairs Administrators in Higher Education.

Finney, S., & Buchanan, H. (2021). A more efficient path to learning improvement: Using repositories of effectiveness studies to guide evidence-informed programming. *Research & Practice in Assessment*, *16*(1), 36–48. https://eric.ed.gov/?id=EJ1307022

Finney, S., & Horst, S. J. (2019a). Standards, standards, standards: Mapping professional standards for outcomes assessment to assessment practice. *Journal of Student Affairs Research and Practice*, *56*(3), 310—325. https://doi.org/10.1080/19496591.2018.1559171

Finney, S. & Horst, S. J. (2019b). The status of assessment, evaluation, and research in student affairs. In V. L. Wise & Z. Davenport (Eds.), *Student affairs assessment, evaluation, and research: A guidebook for graduate students and new professionals* (pp. 3–19). Charles Thomas Publisher. https://psycnet.apa.org/record/2019-16536-000

Finney, S., Stewart, J., & Wild, A. (2024). *The Student Affairs Assessment Improvement Rubric*. OER Commons. Retrieved May 15, 2024, from https://oercommons.org/courseware/lesson/112329/overview

Finney, S., Wells, J., & Henning, G. (2021). *The need for program theory and implementation fidelity in assessment practice and standards* (Occasional Paper No. 51). National Institute for Learning Outcomes Assessment (NILOA). https://eric.ed.gov/?id=ED612091

Fisher, R., Smith, K., Finney, S., & Pinder, K. (2014). The importance of implementation fidelity data for evaluating program effectiveness. *About Campus*, *19*(5), 28–32. https://doi.org/10.1002/abc.21171

Fulcher, K., Coleman, C., & Sundre, D. (2016). Twelve tips: Building high-quality assessment through peer review. *Assessment Update*, *28*(4), 1–2, 14–16. https://doi.org/10.1002/au.30062

Fulcher, K. & Good, M. (2013, November). *The surprisingly useful practice of meta-assessment*. National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/08/Viewpoint-FulcherGood.pdf

Fulcher, K., & Orem, C. (2010). Evolving from quantity to quality: A new yardstick for assessment. *Research & Practice in Assessment*, *5*, 13–17. https://eric.ed.gov/?id=EJ1062646

Gerstner, J., & Finney, S. (2013). Measuring the implementation fidelity of student affairs programs: A critical component of the outcomes assessment cycle. *Research & Practice in Assessment*, *8*, 15–28. https://eric.ed.gov/?id=EJ1062846

Heiser, C., Coleman, J., Yngve, K., Dixon, K., Henning, G., Lundquist, A., & Rice, A. (2023a). *Exploring barriers to equity-centered assessment in higher education* (Equity Report No. 2). National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2023/08/Equity-Report_Barriers-to-Equity_Centered-Assessment.pdf

Heiser, C., Coleman, J., Yngve, K., Dixon, K., Henning, G., Lundquist, A., & Rice, A. (2023b). *Exploring what is needed to support equity-centered assessment in higher education* (Equity Report No. 1). National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2023/08/Equity-Report_Supports-Equity_Centered-Assessment.pdf

Heiser, C., Schnelle, T., & Tullier, S. (2023). Equity-centered assessment: Leveraging assessment to advance equity and justice. *Journal of Student Affairs Inquiry*, *6*(1), 4–17. https://doi.org/10.18060/27921

Henning, G., & Lundquist, A. (2018). *Moving towards socially just assessment* (Equity Response). National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/08/EquityResponse-HenningLundquist.pdf

Henning, G., & Lundquist, A. (2022). Using assessment to advance equity. *New Directions for Student Services*, 2022, 185–194. https://doi.org/10.1002/ss.20439

Henning, G., & Roberts, D. (2024). *Student affairs assessment: Theory to practice* (2nd ed.). Routledge. https://doi.org/10.4324/9781003447207

Hood, S., Hopson, R., & Kirkhart, K. (2015). Culturally responsive evaluation: Theory, practice, and future implications. In K. Newcomer, H. Hatery and J. Wholey (Eds.), *Handbook of Practical Program Evaluation* (4th ed., pp. 281–317). John Wiley & Sons, Inc. https://nasaa-arts.org/wp-content/uploads/2017/11/CRE-Reading-1-Culturally-Responsive-Evaluation.pdf

Horst, S. J., Finney, S., Prendergast, C., Pope, A., & Crewe, M. (2021). The credibility of inferences from program effectiveness studies published in student affairs journals: Potential impact on programming and assessment. *Research & Practice in Assessment*, *16*(2), 17–32. https://files.eric.ed.gov/fulltext/EJ1348828.pdf

James Madison University. (2015). *Assessment Progress Template (APT)* Rubric. https://www.jmu.edu/assessment/_files/pdf/apt_rubric_revised.pdf

Kerr, K., Edwards, K., Tweedy, J., Lichterman, H., & Knerr, A. (2020). *The curricular approach to student affairs: A revolutionary shift for learning beyond the classroom* (1st ed.). Routledge. https://doi.org/10.4324/9781003447740

Kotter, J. (2012). Leading change. Boston, MA: Harvard Business Review Press.

Levy, J., & Heiser, C. (2018, March). *Inclusive assessment practice* (Equity Response). National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/08/EquityResponse_LevyHeiser.pdf

McDonald, B. (2010). Improving learning through meta assessment. *Active Learning in Higher Education*, *11*(2), 119-129. https://journals.sagepub.com/doi/10.1177/1469787410365651

Montenegro, E., & Jankowski, N. (2017). *Equity and assessment: Moving towards culturally responsive assessment* (Occasional Paper No. 29). National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/OccasionalPaper29.pdf

Montenegro, E., & Jankowski, N. (2020). *A new decade for assessment: Embedding equity into assessment praxis* (Occasional Paper No. 42). National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2020/01/A-New-Decade-for-Assessment.pdf

Ory, J. (1992). Meta-assessment: Evaluating assessment activities. *Research in Higher Education*, *33*(4), 467–481. https://www.jstor.org/stable/40196044

Pope, A., Finney, S., & Bare, A. (2019). The essential role of program theory: Fostering theory-driven practice and high-quality outcomes assessment in student affairs. *Research & Practice in Assessment*, *14*, 5–17. https://eric.ed.gov/?id=EJ1223397

Pope, A., Finney, S., & Crewe, M. (2023). Evaluating the effectiveness of an academic success program: Showcasing the importance of theory to practice. *Journal of Student Affairs Inquiry*, *6*(1), 35–50. http://doi.org/10.18060/27924

Randall, J. (2021). Color-neutral is not a thing: Redefining construct definition and representation through a justice oriented critical antiracist lens. *Educational Measurement: Issues & Practice*, *40*(4), 82–90. http://doi.org/10.1111/emip.12429

Randall, J., Slomp, D., Poe, M., & Oliveri, M. (2022). Disrupting white supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, *27*(2), 170–178. https://doi.org/10.1080/10627197.2022.2042682

Roberts, D. (2024). Politics in student affairs assessment. In G. Henning, E. Bentrim & K. Yousey-Elsener (Eds.). *Coordinating divisional and departmental student affairs assessment* (2nd ed. pp. 123–137). Routledge. https://doi.org/10.4324/9781003460695

Russell, M. (2023). Shifting educational measurement from an agent of systemic racism to an anti-racist endeavor. *Applied Measurement in Education*, *36*(3), 216–241. https://doi.org/10.1080/08957347.2023.2217555

Russell, M. (2024). *Systemic racism and educational measurement: Confronting injustice in testing, assessment, and beyond. Routledge.*

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Sage.

Smith, K., & Finney, S. (2020). Elevating program theory and implementation fidelity in higher education: Modeling the process via an ethical reasoning curriculum. *Research & Practice in Assessment*, *15*(2), 5–17. https://eric.ed.gov/?id=EJ1293385

Smith, K., Finney, S., & Fulcher, K. (2019). Connecting assessment practices with curricula and pedagogy via implementation fidelity data. *Assessment and Evaluation in Higher Education*, *44*(2), 263–282. https://doi.org/10.1080/02602938.2018.1496321

Strine-Patterson, H. (2022). Assessment is a leadership process: The multilevel assessment process. *New Directions for Student Services*, 2022, 61–76. https://doi.org/10.1002/ss.20429

Texas A&M International University. (2021). *Assessment Plan Rubric*. https://www.tamiu.edu/adminis/ie/documents/assessment/academic-assessment-plan-rubric-.pdf

University of the District of Columbia. (2023). *Meta-Rubric for Evaluating Institutional Assessment Reports*. https://docs.udc.edu/assessment/Meta-Rubric-August-2023.pdf

Wayne State University. (2023). *Assessment Practices Feedback Rubric*. https://wayne.edu/assessment/rubrics_asmt_practices/assessment_practices_feedback_rubric.pdf

West, A., & Henning, G. (2023). Global implications of student affairs competencies and standards. *New Directions for Student Services*, 2023(183), 31–39. https://doi.org/10.1002/ss.20476