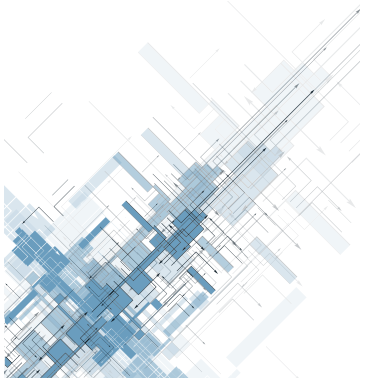
Abstract

Meta-assessment is a useful strategy to document assessment practices and guide efforts to improve the culture of assessment at an institution. In this study, a meta-assessment of undergraduate and graduate academic program assessment reports evaluated the maturity of assessment work. Assessment reports submitted in the first year (75 undergraduate and 35 graduate programs) provided baseline data. As part of implementation of revised reporting processes, the authors facilitated faculty workshops to promote effective assessment practices and increase the clarity of communication in assessment reports. Review of assessment reports submitted the following year (69 undergraduate and 41 graduate programs) evaluated the impact of institutional efforts to develop a more mature culture of assessment. Reviewers used a rubric to score assessment reports on reporting compliance, assessment maturity, and evidence of impact on student learning. Findings indicate reliable improvements in compliance and assessment maturity, but no evidence of efforts to evaluate impact on learning.



AUTHORS

Claudia J. Stanny, Ph.D.
University of West Florida

Angela A. Bryan, PhD.
University of West Florida

Meta-Assessment of the Assessment Culture: Using a Formal Review to Guide Improvement in Assessment Practices and Document Progress

Although higher education institutions have been engaged in the assessment of educational programs for several decades, they continue to struggle to meet the expectations for program-level assessment set by accreditors. Early standards for assessment emphasized sustained assessment efforts rather than episodic assessment (American Association for Higher Education, 1992, cited in Hutchings, Ewell, & Banta, 2012). However, institutional accreditors have shifted their focus to emphasize assessment work that "provides evidence of seeking improvement based on analysis of the results" (Southern Association of Colleges and Schools Commission on Colleges, 2020, p. 66). In addition, recent conversations around program-level assessment of student learning (in the United States) and evaluation of academic programs (in Europe and the UK) raise concerns about what impact (if any) these efforts have had on the quality of academic programs and student learning (e.g., Banta & Blaich, 2011; Blaich & Wise, 2011; Fulcher, Smith, Sanchez, & Sanders, 2017; Kuh, Jankowski, Ikenberry, & Kinzie, 2014).

The history of program-level assessment of student learning can be characterized by a continuing conflict between assessment for compliance and assessment for improvement (e.g., Blumberg, 2018; Stitt-Bergh, Kinzie, & Fulcher, 2018; Suskie, 2015, 2018; Walvoord, 2014). Assessment critics have argued that assessment processes represent little value beyond compliance with external mandates (Gilbert, 2018; Worthen, 2018). In contrast,

CORRESPONDENCE

Email
abryan@uwf.edu

professional organizations aligned with assessment advocate that mature assessment cultures should focus on the use of results to improve teaching, learning, and assessment (e.g., Association for the Assessment of Learning in Higher Education, Association of American Colleges & Universities, National Institute for Learning Outcomes Assessment). Although the requirements of external stakeholders such as government agencies and accrediting bodies can motivate efforts to assess student learning, external mandates tend to focus attention and effort on assessment for compliance (Stanny & Halonen, 2011; Suskie, 2015). However, institutions should nurture an assessment culture that focuses on improvement because this strategy can enhance the quality of academic programs (Isabella & McGovern, 2018; O'Neill, Slater, & Sapp, 2018; Lattuca, Terenzini, & Volkwein, 2006; Magruder, McManis, & Young, 1997).

That is, when assessment is done in the right way for the right reasons, accountability should take care of itself.

Fulcher and his colleagues describe a culture of assessment as one in which academic programs define learning outcomes, map outcomes to the curriculum, select an assessment instrument, collect assessment data, analyze and report the results, and communicate their findings to stakeholders (Fulcher, Good, Coleman, & Smith, 2014; Fulcher, Swain, & Orem, 2012). In a mature culture of assessment, the institution uses assessment processes and evidence as opportunities for self-reflection and identification of actions that might promote student learning (Fulcher et al., 2014; Fulcher et al., 2017; Lending, Fulcher, Ezell, May, & Dillon, 2018; Maki, 2010; Stanny, 2015, 2018, 2020; Suskie, 2015). Programs should assess and analyze student learning data, identify and implement changes to the curriculum and/or instructional methods (if needed), and then reassess to evaluate the impact of the implemented changes on student learning (Lending, et al., 2018; Stanny 2021). Discussions of changes implemented and how learning changed following implementation, grounded in an analysis of follow-up assessment findings, are the two most vital components of a culture of improvement and are often missing in assessment reports (Reder & Crimmins, 2018; Stitt-Bergh et al., 2018; Suskie, 2015, 2018).

How can institutions maintain accountability to external stakeholders yet still foster a culture of improvement? Wehlburg (2008, 2013) argues that programs can best meet accountability expectations when they assess with the goal of increasing program effectiveness. These programs focus on using assessment findings to identify promising areas to improve student learning and conduct follow-up assessments to determine whether implemented changes made a difference for student learning. When programs document these activities, they can meet expectations for accountability set by external stakeholders (Souza & Rose, 2021). That is, when assessment is done in the right way for the right reasons, accountability should take care of itself.

Meta-assessment, such as a formal review of assessment, can yield a wealth of useful information that serves multiple goals (Stanny, 2020). It can provide a broad description of the types of assessment practices in use, including evidence of efforts to use assessment results to improve programs. Findings can be used to guide future professional development efforts and campus interventions that promote the adoption of mature assessment practices. Dissemination of review findings communicates to faculty and administrators that assessment reports are read by multiple individuals. Because the findings provide formative feedback on both the quality of assessment processes and how well assessment reports communicate the program's assessment story to reviewers, a formal review can improve the quality of reporting. Walvoord (2014) offers general suggestions for how to "tell the story of how you are assessing and improving," (p. 45). Stitt-Bergh et al. (2018) identify five elements required to connect and align assessment activities and improvement initiatives to tell a compelling assessment story: clearly identify the learning targeted, specify the scope of the initiative (course, program, institution), identify specific changes and actions implemented, collect multiple types of evidence from two points in time to evaluate whether improvements occurred, and reflect on and interpret the assessment evidence.

Efforts to promote a mature culture of assessment have been guided by the framework of an assessment cycle, presented in guidelines for assessment (e.g., Maki, 2010; Suskie, 2018; Walvoord, 2014) and discussions of the characteristics of "mature" assessment, and expectations for documentation for institutional accreditation reports. Rubric elements articulate these goals in concrete language intended for the campus audience as part of

a pragmatic effort to transparently communicate expectations for assessment work and assessment reporting to chairs and faculty assessment committees.

The rubric tries to balance two points of view in language that will be understandable to the campus community. First, the compliance items reflect documentation needs established through prior experience preparing reports for external audiences (such as institutional and disciplinary accrediting bodies). Second, the maturity of assessment items reflect best practices in the literature and describe assessment processes that move beyond compliance and motivate efforts to improve student learning. The rubric is microscopic because we wanted to track the emergence of specific practices and document the number of departments that adopted each practice over time. Written in the spirit of rubrics for specifications grading advocated by Nilson (2015), these detailed, specific rubric items connect mature practices to unambiguous, concrete characteristics of assessment work that could appear in an assessment report. An added advantage of these concrete criteria is that the rubric elements can be scored as present or absent, which promoted more reliable scoring and simplified on-the-fly computation of inter-rater agreement.

When departments receive feedback from reviewers that describe problem areas and see a score that can be compared to a mean of their college or the university as a whole, we can refocus the conversation on improvement, even if initial changes are directed at improving the assessment report itself (Stanny, Stone, & Mitchell-Cook, 2018). Evaluations of the clarity of reporting can guide decisions about the design of report templates and guiding instructions prepared by an Assessment Office or Office of Institutional Effectiveness (IE). Together, the findings and follow-up interventions can both provide evidence that programs comply with accreditor expectations and shift the culture toward a focus on efforts to seek improvement.

Audit of the Assessment Reporting Process

As part of preparation for an impending compliance report to an institutional accreditor, the Office of IE conducted an audit of the assessment process and reviewed three years of programmatic assessment reports, the reporting template, and the submission process (Walvoord, 2014). The audit revealed strengths and weaknesses in institutional assessment processes. The good news was that the institution could document a systematic and ongoing culture of assessment. Nearly all departments had reported assessment activities annually for each of their educational programs, with few departments failing to participate in the process. The audit also revealed areas for improvement. Specifically, the report template included question prompts and instructions for several reporting fields that were vague, ambiguous, or did not elicit narratives that fully documented the assessment work completed by faculty. Because reports were submitted as responses to questions in a Qualtrics survey that had limited text fields, narratives frequently lacked the level of detail needed to understand the work reported. In addition, the submission process, which required completing a new form for each learning outcome, was awkward, repetitive, and cumbersome. As a result, most departments reported assessment work for only one or two student learning outcomes, although evidence from recent disciplinary accreditations and program reviews indicated that several programs engaged in more extensive assessment activity. In addition, few departments had created a multi-year assessment plan. There was scant evidence that any department had reassessed a student learning outcome to determine the impact of changes implemented in a prior year. Thus, the structure of the reporting process encouraged departments to treat each assessment cycle as a snapshot of work from the current reporting year, with no thought given to assessments that could evaluate the impact of changes made in a prior year (Suskie, 2018).

Evaluations of the clarity of reporting can guide decisions about the design of report templates and guiding instructions prepared by an Assessment Office or Office of Institutional Effectiveness (IE).

Information from the audit motivated us to modify assessment processes. First, IE staff designed a new report template based on an Excel spreadsheet with revised prompts that more clearly communicate expectations about the information requested. Second, each department was asked to develop a five-year assessment plan for each educational program that described how the department planned to conduct a full, multi-year cycle of assessment for each program-level student learning outcome within a five-year period. A full cycle of assessment was defined as a two to three year process. In the first year of an assessment

cycle, the program collects baseline assessment data. Then, the program should reflect on the findings and make decisions about possible implementation of an improvement initiative. In the final year of the cycle, the program conducts follow-up assessments to either evaluate the impact of the implemented change or document the stability of student performance on the targeted learning outcome.

Rubric elements describe "best practices" and hallmarks of a mature assessment process. These best practice elements contribute to assessment work that is likely to produce meaningful information and guide faculty decisions about curriculum and instruction.

In addition to changing the assessment reporting process, the Director of IE and the Director of the Center for Teaching and Learning (CTL) facilitated a series of workshops designed to educate faculty and administrators in assessment leadership positions on how to write assessment reports that would clearly document an actionable use of assessment results toward seeking improvement in student learning (Fulcher, et al., 2017; Walvoord, 2014). Workshops included a half-day mini-conference on assessment, an annual peer review of assessment events (described in Stanny, et al., 2018), workshops on effective assessment practices, targeted workshops on specific assessment skills (writing measurable learning outcomes, creating a five-year assessment plan), presentations to disseminate findings from the current formal review, and one-on-one consultations with chairs and members of assessment and curriculum committees.

The institution had adopted an annual formal review of assessment reports, in which trained reviewers used a rubric to evaluate the quality of assessment work described in assessment reports. Although the previous four formal reviews had documented improvements in assessment reporting (Stanny, 2020), the audit confirmed the need for extensive changes to the reporting process, which had emerged from a series of conversations during the peer review event and in one-on-one consultations. The formal review was extended to evaluate the impact of changes made to the new report template and other initiatives to promote a more mature assessment culture. The rubric was revised to reflect the new reporting fields and guiding language in the new Excel template. The review continued our evaluation of submitted assessment reports as a meta-assessment of the impact of these changes on the quality of assessment reporting. In addition, examination of the types of assessment practices documented in these reports enabled us to describe the ongoing evolution toward a more mature culture of assessment.

Method

Rubric

The rubric used for the review is comprised of three major sections: *Reporting Compliance Criteria*, *Maturity of Assessment*, and *Evidence of Impact*. A list of the rubric elements is presented in Table 1. Rubric elements were scored as a 0 (evidence is weak, missing, or the criterion is not applicable to the reporting program, as when no evidence is provided for an optional item) or 1 (evidence that a report meets expectations).

Scores for the *Reporting Compliance* and *Evidence of Impact* sections are based on the number of rubric elements that describe best practices for this section (two – six rubric elements). *Maturity of Assessment* produced scores on six dimensions of maturity, based on the number of rubric elements that described best practices for this dimension (two – five rubric elements). Summary findings report the scores for each dimension as diagnostic feedback and report an overall score for Maturity of Assessment (23 elements). Rubric elements describe "best practices" and hallmarks of a mature assessment process. These best practice elements contribute to assessment work that is likely to produce meaningful information and guide faculty decisions about curriculum and instruction. Composite scores, based on the rubric elements included in a section or dimension of a section, create global measures of the quality of reporting and maturity of the assessment culture.

Reporting Compliance Criteria. This score was based on the sum of 6 rubric elements that evaluate key elements that should appear in every assessment report to adequately document the program's compliance with expectations for reporting assessment processes with clear and compelling narratives. The elements evaluated the following characteristics: (1) report documents assessment on at least 20% of program student learning outcomes (SLOs), (2) completion of the summary tab portion of the Excel template for assessment reports, (3) clear description of program delivery, including locations and modalities of

Table 1
Rubric used for Scoring Annual Assessment Reports

Each rubric element scored a criterion as present/met (1) or absent/not met (0).
Department reports assessment for at least 20% of identified SLOs for the program
Summary narrative of assessment activity
Clear description of the delivery mode of the program
Evidence of faculty engagement and reflection on the assessment findings
Curriculum Map is available (posted on the IE website)
5-Year Assessment Plan is available (posted on the IE website)
Maturity of Assessment (6 dimensions)
Quality of Measures (4 criteria)
At least one measure aligns with the SLO(s) assessed
Assessments include at least one direct measure for each SLO
At least one SLO was assessed with multiple measures
Discussion of the reliability or validity of at least one measure used to assess an SLO
Credible Data Collection Processes and Representative Sampling (4 criteria)
Measures used for assessment have face validity for and align with the SLO assessed
Data analysis includes disaggregation by locations and delivery modes as appropriate
Report includes the number of course sections that provided data
Report includes the number of students assessed
Report of Results (5 criteria)
Report identifies a benchmark and description of criteria for meeting the benchmark
Report includes the number of students that meet or exceed expectations
Narrative compares current findings with evidence from previous assessments
Narrative summarizes results that appear in another document
Department provides the meeting date(s) where faculty discussed assessment findings
Department documents the attendance of faculty at the meeting
Department submitted the meeting minutes as supporting evidence
Narrative describes a logical relationship between decisions and assessment findings
Use of Results for Improvement (2 criteria)
Department describes an ***actionable*** use of results to improve student learning that is clearly related to the assessment evidence
Narrative provides convincing evidence of a concrete plan to implement
Faculty Engagement with Assessment Processes (4 criteria)
Evidence of broad faculty engagement
Narrative describes how assessment findings and decisions are communicated
Evidence that findings were disseminated to all appropriate faculty
Evidence that findings were disseminated to other relevant stakeholders
Evidence of Impact on Student Learning (2 criteria)
Narrative includes an evaluation of the impact of any changes implemented during a prior academic year on student learning
Evidence provided about the impact (either positive or negative) of a new initiative

instruction, (4) documentation of faculty engagement and reflection on assessment evidence for program improvement, (5) curriculum map posted to the IE website, and (6) five-year assessment plan posted to the IE website.

Maturity of Assessment. An overall score for maturity of assessment was based on the sum of 23 rubric elements, which described six dimensions or characteristics of a mature assessment process: (1) quality of measures (four rubric elements), (2) credible data collection processes and representative sampling (four rubric elements), (3) report of results (five rubric elements), (4) interpretation of findings (four rubric elements), (5) use of results for improvement (two rubric elements), and (6) faculty engagement with assessment processes (four rubric elements).

Evidence of Impact. This metric identifies programs that provide concrete examples of tangible changes in student learning that can be attributed to teaching and learning initiatives motivated by assessment findings. The metric was based on two rubric elements: (1) evidence that the program assessed and evaluated impact and (2) evidence presented for the impact of changes implemented was compelling.

Sample

The sample included assessment reports submitted to the Office of IE during two cycles of assessment reporting (ending in 2019 and 2020). The 2018-2019 assessment cycle included 75 reports for undergraduate programs and 35 reports for graduate programs. The 2019-2020 assessment cycle included 69 reports for undergraduate programs and 41 reports for graduate programs. Departments submitted assessment reports using an Excel spreadsheet template prepared by the Office of IE. Departments were encouraged to supplement information in their report narratives by uploading supporting documents (such as meeting minutes, examples of assignments or rubrics, and reports summarizing large data analyses). Reviewers examined the narratives and all supporting documents when they scored each report.

Procedure for training and maintaining inter-rater reliability

Serving as a reviewer of programmatic assessment reports is regarded as intensive professional development for faculty. Although faculty may serve as reviewers more than once, we encourage applications from new reviewers each year to increase assessment expertise across the university.

Reviewers. Each year, the CTL issues a call for faculty reviewers. Faculty are invited to submit letters of interest that include information regarding their full-time status, their department and college, and their availability to meet during the spring semester. The CTL and IE collaborate to review the applications. Four reviewers are selected based on their application responses and availability with the constraint that the four reviewers come from different colleges. This ensures that no reviewer scores assessment reports submitted by departments from the college in which they teach (except during initial training, when all reviewers score all reports in the training sample).

Serving as a reviewer of programmatic assessment reports is regarded as intensive professional development for faculty. Although faculty may serve as reviewers more than once, we encourage applications from new reviewers each year to increase assessment expertise across the university. For both years included in this study, four reviewers were selected for both 2019 and 2020, for a total of eight reviewers over the two-year period. Reviewers received formal training on how to apply the rubric to score the assessment reports. The reliability of scoring was evaluated and monitored continuously during the review.

Reviewer training and reliability. Reviewers completed an initial training and discussed how to score the assessment reports based on the rubric elements. Next, reviewers scored a training sample of assessment reports (six reports in 2019, seven reports in 2020). Reports were read and scored by all four reviewers. To compute inter-rater agreement, each reviewer was first paired with every other reviewer and we computed individual rater agreement scores (pair-wise) for each rubric element. We then computed the average agreement score across all possible pair-wise comparisons for each rubric element. Thus, agreement scores are the percentage of pair-wise comparisons that produced identical scores for a rubric element. We also computed the average percent agreement across all rubric elements.

After computing the initial reliability data, reviewers discussed areas of disagreement on individual rubric elements. Reviewers developed guidelines to help them apply the rubric consistently. Reviewers then independently rescored the reports in the training sample. The second calculation of reliability scores established acceptable levels of reliability (82% agreement, averaged over all rubric elements for the 2019 review and 81% agreement for the 2020 review).

Scoring procedures. After achieving an acceptable level of consensus (exceeding the target of 75% average agreement), reviewers scored the remaining reports. The review was completed as a series of assignments (three assignments for undergraduate reports, with 19-28 reports per assignment; three assignments for graduate reports, with 10-13 reports per assignment). Each reviewer was paired with every other reviewer for a subset of the reports included in an assignment. Two reviewers independently scored each report. Thus, percent agreement scores reflect the scoring consistency of each reviewer with every other reviewer and the average rater agreement score reflects the collective judgment of all four reviewers. No reviewer scored reports submitted by a department from his or her college.

The most problematic rubric elements entailed judgments about the maturity of assessment, especially practices that either did not apply to all programs … or did not have an obvious location or prompt in the reporting template…

Scoring consistency was maintained by computing the rater agreement metrics for scores submitted for each assignment (percent agreement for individual rubric elements, average agreement across all rubric elements). In addition, we computed cumulative percent agreement scores (individual rubric elements and average across rubric elements) for all reports scored to date. Reviewers discussed the reliability data and developed consensus about problem areas they encountered in the most recent assignment before they scored reports in the next assignment. Reviewers added notes to the scoring guidelines as needed to resolve emerging challenges and maintain consistency throughout the review. For the few instances when the scores submitted by two reviewers were not identical, differences were resolved by computing the average of the submitted scores.

Results and Discussion

Reliability

Reviewer agreement was monitored for each assignment and for the population of reports reviewed. We monitored agreement for individual rubric elements and for the agreement averaged across all rubric elements, with the goal of maintaining aggregate agreement above 75%. Final reliability metrics were based on the entire population of assessment reports in a given year, disaggregated by program (undergraduate or graduate).

The average percent agreement for the 2019 review was 90% for undergraduate reports (n = 75) and 87% for graduate reports (n = 35). Similarly, the average percent agreement for 2020 was 81% for undergraduate reports (n = 69) and 84% for graduate reports (n = 41). Agreement scores for individual rubric elements (31 elements) ranged from 58% to 100%. During the 2019 review, only 3 of the 31 rubric elements (10%) produced percent agreement scores that were less than 75% agreement (values were 63%, 69%, and 72%) when reviewing undergraduate reports. Among the graduate reports (when scoring pivoted to remote work), seven rubric elements (22.6%) fell below 75% agreement (four elements ranged between 70% and 74% agreement; the remaining three elements ranged between 66% and 69% agreement). The review of the 2020 reports, completed entirely through remote work, was a bit more variable: eight rubric elements (26%) for the undergraduate reports produced percent agreement scores that were less than 75% agreement (values ranged from 59% to 74% agreement) and eight rubric elements (26%) for the graduate reports fell below 75% agreement (values ranged from 58% to 74% agreement).

Examination of the rubric elements that posed the greatest challenges for reliable scoring reflected as much about the quality of the template and prompts as the judgment of reviewers. The most problematic rubric elements entailed judgments about the maturity of assessment, especially practices that either did not apply to all programs (e.g., *data analysis includes disaggregation by locations and delivery methods as appropriate*) or did not have an obvious location or prompt in the reporting template (e.g., *comparison of current findings with evidence from previous assessments*, *summaries of results in a supporting*

document, *discussions of how findings and decisions were communicated*, *evidence that findings were disseminated to all appropriate faculty*).

The most difficult rubric elements were two criteria that concerned the use of results for improvement (*description of actionable use of results* and *description of a concrete plan to implement*). Reviewers said they had difficulty seeing a distinction between these two aspects of use of results. Future reviews might merge these items because we found that when reviewers disagreed, they usually scored one element as present and the other as absent, but chose different elements to score as present (versus one reviewer scoring both elements as present while the other reviewer scored both elements as absent). The items became more reliable when rescored as a single item (scoring one if at least one of the original two elements had been scored one and zero only when both elements were scored as zero). In addition to the challenge of attempting to capture a nuanced characteristic of mature assessment, reliable scoring of these two elements was further hampered by ambiguities inherent in the way the template requested information about decisions and actions (either implemented or planned for the coming year). This illustrates the multi-layered value of a formal review. Difficulties establishing reliability for some rubric elements often surfaced problems with the reporting template and ambiguous communications from IE to faculty responsible for reporting assessment work.

An interesting observation during this review was related to the impact of COVID-19 and the pivot to remote work. In 2019, reviewers had completed their work on undergraduate program reports by the end of February. In March, we shifted to remote work and continued weekly meetings via web conference software. The following year, the entire review, including initial training and weekly meetings, was implemented via web conferences. The data on inter-rater agreement reflect the challenges associated with clear communication via web meetings to maintain calibration and consensus. These challenges were compounded by schedule conflicts that prevented all reviewers from meeting at the same time. Based on these observations, we conclude that although it is possible to maintain better than 75% agreement among reviewers under these conditions, reviewers will reach higher levels of consensus if they can meet in person at the same time. It is unclear whether meeting via web conferencing software or meeting as two groups contributed to the lower agreement values observed during remote work.

Analysis of rubric scores

Difficulties establishing reliability for some rubric elements often surfaced problems with the reporting template and ambiguous communications from IE to faculty responsible for reporting assessment work.

Reporting compliance. The sum of the first six rubric elements served as a global measure of compliance with reporting expectations. Values could range from 0 (no report filed, no documents posted to the IE website) to 6 (all reporting criteria met expectations). In 2019, the mean reporting compliance score was 1.94 for undergraduate reports (SD = 1.222) and 2.64 for graduate reports (SD = 1.579). In 2020, reporting compliance scores increased to 4.88 for undergraduate reports (SD = 1.192) and 4.65 for graduate reports (SD = 1.744). Analysis of the reporting compliance composite scores produced a significant main effect of year ($F(1, 216) = 158.090$, $MSe = 1.914$, $p < .001$, partial $\varepsilon^2 = .423$) as well as a significant interaction of year by type of program report (undergraduate, graduate) ($F(1, 216) = 5.681$, $MSe = 1.914$, $p < .02$, partial $\varepsilon^2 = .026$).

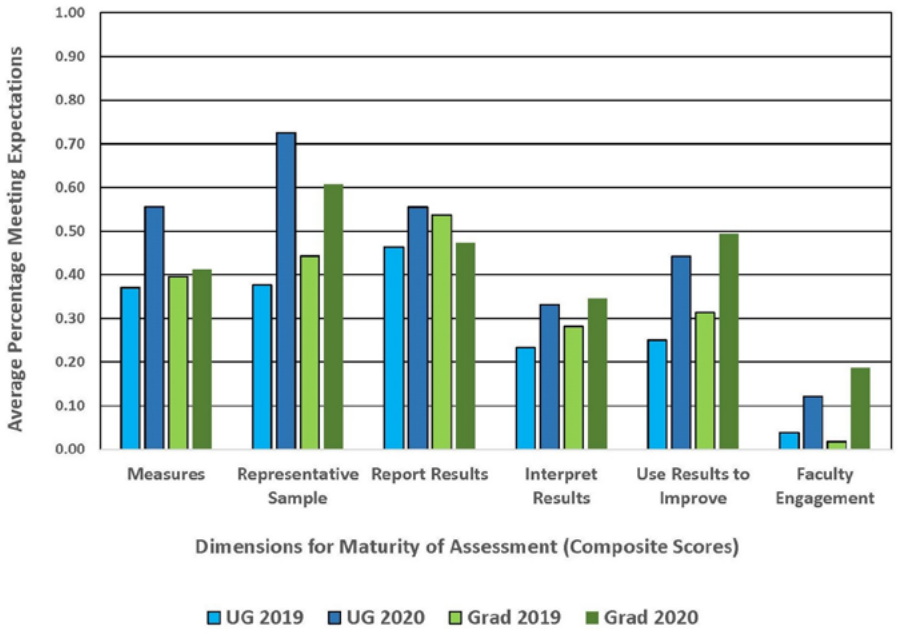
Maturity of assessment. Because each dimension of assessment maturity was based on two to five rubric elements, we computed an average of the contributing rubric elements instead of a sum to create composite scores with the same range of values (0 to 1, representing the average proportion of rubric elements in a dimension that met expectations). A 2 X 2 X 6 repeated measures analysis of variance was computed on composite scores in which report year (2019, 2020) and type of report (undergraduate, graduate) were between subjects factors and the six composite scores were repeated measures: *quality of measures* (four rubric elements), *credible data collection processes and representative sampling* (four rubric elements), *report of results* (five rubric elements), *interpretation of findings* (four rubric elements), *use of results* to improve student learning (two rubric elements), and faculty engagement with assessment processes (four rubric elements). A parallel statistical analysis, based on the raw scores produced by sums of rubric elements, produced the same pattern of findings. Only one analysis is reported here.

Average maturity of assessment improved from the first review (M = .31, SE = .020) to the second review (M = .437, SE = .019), producing a significant main effect of year of report ($F(1, 216) = 20.644$, MSe = .232, $p < .001$, partial ε^2 = .087). Although reports received significantly different scores for the six dimensions of assessment maturity ($F(5, 1080) = 153.899$, MSe = .035, $p < .001$, partial ε^2 = .416), this factor produced significant two-way interactions with the year of report and type of program as well as a significant three-way interaction between maturity scores, year of report, and type of report. As a result, this discussion focuses on the significant three-way interaction ($F(5, 1080) = 4.189$, MSe = .035, p = .001, partial ε^2 = .019). Other comparisons (the main effect of type of report and the interaction between year of review and type of report) were not statistically reliable.

Differences among dimensions of assessment maturity reflect strengths and weaknesses in the culture of assessment.

Mean composite scores (average proportion of rubric elements in a dimension that met expectations) are presented in Figure 1 as a function of the year of report (2019, 2020) and type of report (undergraduate and graduate program reports). Consistent with the significant main effect of year of report, with only a few exceptions, scores for both undergraduate and graduate reports improved from 2019 to 2020. The exceptions were that graduate programs showed no change on the *quality of measures* metric and showed a small decline on the *report of results metric*. Differences among dimensions of assessment maturity reflect strengths and weaknesses in the culture of assessment. Undergraduate programs showed pronounced improvements in the *quality of measures* gathered and the *collection of assessment evidence from a representative sample of student work*. The findings also suggest areas for further growth and maturation in the areas of *interpretation of findings* and *breadth of faculty engagement*.

Figure 1
Two-year comparison (2019 versus 2020) of the proportion of rubric elements within each of six dimensions of assessment maturity that met expectations for undergraduate and graduate programs.



Note: Number of rubric elements varied across dimensions: Measures (four elements), Representative Sample (four elements), Report of Results (five elements), Interpret Results (four elements), Use Results to Improve (two elements), Faculty Engagement (four elements).

Evidence of impact. Two rubric elements generate the composite score for evidence of impact. However, this metric produced no evidence for change across reports for either graduate or undergraduate programs, with few reports submitting documentation of the impact of an implemented change on assessments of student learning. Although few departments currently meet expectations on these rubric elements, they remain part of the

review to enable the institution to capture and document when departments reach this level of assessment maturity.

Conclusions

Our findings clearly indicate positive changes in the culture of assessment. Building on improvements documented in previous years (Stanny, 2020), data generated by the new rubric and reporting process document additional advances in both compliance with reporting expectations and adoption of assessment practices that characterize a more mature assessment culture. Strengths included widespread use of direct measures of student learning, improved alignment of assessment measures with targeted learning outcomes, collection of artifacts from a representative sample of students, more complete documentation of faculty discussions and reflections on assessment findings, and increases in the breadth of faculty engagement. Although the absolute value of scores for rubric elements related to mature assessment practices indicate substantial room for additional improvement, the changes from year one (baseline use of the new reporting template in 2019) to year two (2020) unambiguously document movement in the desired direction.

Institutional change often occurs at a glacial pace (Halonen, Ellenberg, Stanny, El-Sheikh, 2011). Assessment professionals charged with leading an initiative to promote a culture of assessment might feel they are making little progress from year to year. This project illustrates the value of meta-assessment to monitor progress on these large-scale efforts. Systematic monitoring of the maturity of assessment helped make incremental changes in the culture of assessment visible. The findings, along with informal observations from reviewers, suggested opportunities where small modifications could drive ongoing change. For example, during training, reviewers sometimes commented that they were unsure where in the assessment report they should look to find evidence for a given assessment practice. Reviewers also identified ambiguous language in report instructions. These observations identified shortcomings in template prompts and instructions that interfered with our ability to gather the information needed to document assessment activities. Revision of the reporting template was informed by the various observations gleaned from reviewer comments. Reviewer feedback also informed the design of professional development workshops to guide faculty charged with writing assessment reports and help them "tell their assessment story" to reviewers outside their discipline (Stitt-Berg, et al., 2018).

Assessment professionals charged with leading an initiative to promote a culture of assessment might feel they are making little progress from year to year. This project illustrates the value of meta-assessment to monitor progress on these large-scale efforts.

In summary, this formal review of assessment reports supported assessment efforts in several ways. The data provided tangible evidence of the quality of assessment work on campus, creating a year-to-year snapshot that proved useful as documentation of the institution's compliance with accreditation standards for assessment. The review provided formative feedback to the Office of IE and the CTL about the progress made toward achieving unit operational goals. The findings informed decisions about how to structure assessment reporting, such as the format of templates, how we framed requests for assessment information, and the logistics of reporting (timelines and interfaces with software and other reporting technology). These structural changes helped eliminate unintended obstacles to effective reporting. The data provided formative feedback to academic departments about their assessment practices and identified areas where small, realistic changes could produce tangible improvements in the quality of their assessment work. Dissemination of the findings helped allay a common misconception among faculty and critics of assessment: the belief that assessment reports are simply not read (Stanny, 2021).

An additional, serendipitous benefit emerged while the institution prepared a major accreditation report for its institutional accreditor. Scores on rubric elements for mature assessment practices served as an index to the population of assessment reports. When the authors of the accreditation report wanted to locate examples from assessment reports to include as evidence in the report narrative, they consulted the data file of rubric scores to identify programs that submitted relevant documentation with their assessment report. Rubric scores accurately identified relevant examples of departments that had disaggregated data, uploaded a rubric or description of an embedded assessment assignment, submitted minutes of a faculty meeting in which faculty reflected on assessment results and discussed curriculum changes or other initiatives intended to improve an aspect of student learning.

AUTHORS NOTE:
The authors thank the faculty reviewers for their contributions to scoring assessment reports for this project: Christopher L. Atkinson, Eric Bostwick, Jasara Norton, Pamela Meyers, Mizanoor Rahman, Bhuvaneswari Ramachandran, April Schwartz, and Jacqueline Thomas.

References

Association for the Assessment of Learning in Higher Education (AALHE). https://www.aalhe.org/

Association of American Colleges & Universities (AAC&U). https://www.aacu.org/

Banta, T. W., & Blaich, C. F. (2011, January). Closing the assessment loop. *Change: The Magazine of Higher Learning*, *43*, 22-27. https://doi.org/10.1080/00091383.2011.538642

Blaich, C. F., & Wise, K. S. (2011, January). *From gathering to using assessment results: Lessons from the Wabash National Study* (Occasional Paper No. 8). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). https://www.bu.edu/provost/files/2015/09/From-Gathering-to-Using-Assessment-Results_Lessons-from-the-Wabash-Study-C.-Blaich-K.-Wise1.pdf

Blumberg, P. (2018, Summer/Fall). Two underused best practices for improvement focused assessments. *Research & Practice in Assessment*, *13*, 78-84. http://www.rpajournal.com/dev/wp-content/uploads/2019/01/RPA_Summer_Fall_Issue_2018_NIB.pdf

Fulcher, K. H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014, December). *A simple model for learning improvement:weigh pig, feed pig, Weigh pig*. (Occasional Paper No. 23). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://in.ewu.edu/facultycommons/wp-content/uploads/sites/129/2016/12/A-Simple-Model-for-Learning-Improvement_Weigh-Pig-Feed-Pig-Re-Weigh-Pig.pdf

Fulcher, K. H., Smith, K. L., Sanchez, E. R. H., Sanders, C. B. (2017). Needle in a haystack: Finding learning improvement in assessment reports. *Professional File*, *Article 141*. http://doi.org/10.34315/apf1412017

Fulcher, K. H., Swain, M. S., & Orem, C. D. (2012, January-February). Expectations for assessment reports: A descriptive analysis. *Assessment Update*, *24* (1), 1-2, 14-16. https://uncw.edu/assessment/documents/fultcherswainandorem2012.pdf

Gilbert, E. (2018, January 12). An insider's take on assessment: It may be worse than you thought. *The Chronicle of Higher Education*. https://www.chronicle.com/sectin/Commentary/44

Halonen, J. S., Ellenberg, G. B., Stanny, C. J., & El-Sheikh, E. (2011). First things first: Attending to assessment issues, accountability, and accreditation. In D. S. Dunn, M. A. McCarthy, S. C. Baker, & J. S. Halonen, *Using quality benchmarks for assessing and developing undergraduate programs* (pp. 46-70). Jossey-Bass.

Hutchings, P., Ewell, P., & Banta, T. (2012, May). *AAHE principles of good practice: Aging nicely*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/08/Viewpoint-Hutchings-EwellBanta.pdf

Isabella, M., & McGovern, H. (2018). Identity, values, and reflection: Shaping (and being shaped) through assessment. *New Directions for Teaching and Learning*, *2018* (155), 89-96. https://doi.org/10.1002/tl.20307

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. (2014). *Knowing what students know and can do: Current state of student learning outcomes assessment in U.S. colleges and universities*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/2013AbridgedSurveyReport.pdf

Lattuca, L. R., Terenzini, P. T., & Volkwein, J. F. (2006). *Engineering change: A study of the impact of EC2000 - Executive summary*. Accreditation Board for Engineering and Technology. https://www.abet.org/wp-content/uploads/2015/04/EngineeringChange-executive-summary.pdf

Lending, D., Fulcher, K., Ezell, J. D., May, J. L., & Dillon, T. W. (2018, Winter). Example of a program-level learning improvement report. *Research & Practice in Assessment*, *13*, 34-50. http://www.rpajournal.com/dev/wp-content/uploads/2019/02/W18_A6.pdf

Magruder, J., McManis, M. A., & Young, C. C. (1997). The right idea at the right time: Development of a transformational assessment culture. *New Directions for Higher Education*, *100*, 17-29. https://doi.org/10.1002/he.100002

Maki, P. L. (2010). *Assessing for learning: Building a sustainable commitment across the institution* (2nd ed). Stylus.

National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/

Nilson, L.B. (2015). *Specifications grading: Restoring rigor, motivating students, and saving faculty time*. Stylus.

O'Neill, M., Slater, A., & Sapp, D. G. (2018). Writing and the undergraduate curriculum: Using assessment evidence to create a model for institutional change. *New Directions for Teaching and Learning*, *2018*(155), 97-104. https://doi.org/10.1002/tl.20308

Reder, M., Crimmins, C. (2018, Winter). Why assessment and faculty development need each other: Notes on using evidence to improve student learning. *Research & Practice in Assessment*, *13*, 15-19. http://www.rpajournal.com/dev/wp-content/uploads/2019/02/W18_A3.pdf

Southern Association of Colleges and Schools Commission on Colleges (2020). *Resource manual for the principles of accreditation: Foundations for quality enhancement (Third Edition)*. SACSCOC. https://sacscoc.org/app/uploads/2019/08/2018-POA-Resource-Manual.pdf

Souza, J.M. & Rose, T.A. (Eds.) (2021). *Exemplars of assessment in higher education: Diverse approaches to addressing accreditation standards*. Stylus Publishing, LLC.

Stanny, C. J. (2015). Assessing learning in psychology: A primer for faculty and administrators. In D. S. Dunn (Ed.), *The Oxford handbook of undergraduate psychology education* (pp. 813-831). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199933815.013.065

Stanny, C. J. (2018). Putting assessment into action: Evolving from a culture of assessment to a culture of improvement, *New Directions for Teaching and Learning*, *2018* (155), 113-116. https://doi.org/10.1002/tl.20310

Stanny, C. J. (2020, June). *Promote change by assessing the maturity of your assessment culture*. Single paper focus session presented at the Association for the Assessment of Learning in Higher Education (AALHE) conference. Conference proceedings: https://www.aalhe.org/2020-conference-proceedings

Stanny, C. J. (2021). Overcoming obstacles that stop student learning: The bottleneck model of structural reform. In S. A. Nolan, C. M. Hakala, & R. E. Landrum (Eds.), *Assessing undergraduate learning in psychology: Strategies for measuring and improving student performance*, (pp. 77-93). APA Books. https://doi.org/10.1037/0000183-007

Stanny, C. J., & Halonen, J. S. (2011). Accreditation, accountability, and assessment: Addressing multiple agendas. L. Stefani (Ed.), *Evaluating the effectiveness of academic development: A professional guide* (pp. 169-181). Routledge.

Stanny, C. J, Stone, E., & Mitchell-Cook, A. (2018). Evidence-based discussions of learning facilitated through a peer review of assessment. *New Directions for Teaching and Learning, 2018* (155), 31-38. https://doi.org/10.1002/tl.20300

Stitt-Bergh, M., Kinzie, J., Fulcher, K. (2018, Winter). Refining an approach to assessment for learning improvement. *Research & Practice in Assessment*, *13*, 27-33. http://www.rpajournal.com/dev/wp-content/uploads/2019/02/W18_A5.pdf

Suskie, L. (2015). *Five dimensions of quality: A common sense guide to accreditation and accountability*. Jossey-Bass.

Suskie, L. (2018). *Assessing student learning: A common sense guide* (3rd ed.). Jossey-Bass.

Walvoord, B. E. (2014). *Assessment clear and simple: A practical guide for institutions, departments, and general education (2nd ed)*. Jossey-Bass.

Wehlburg, C. M. (2008). *Promoting integrated and transformative assessment: A deeper focus on student learning*. Jossey-Bass.

Wehlburg, C. M. (2013). "Just right" outcomes assessment: A fable for higher education. *Assessment Update: Progress, Trends, and Practices in Higher Education*. *25* (2). https://doi.org/10.1002/au.252

Worthen, M. (2018, February 22). The misguided drive to measure 'learning outcomes.' *The New York Times*. https://www.nytimes.com/2018/02/23/opinion/sunday/colleges-measure-learning-outcomes.html