Evolving from Quantity to Quality: A New Yardstick for Assessment

Keston H. Fulcher and Chris D. Orem
James Madison University

Author Note

Keston H. Fulcher, Center for Assessment and Research Studies, James Madison University; Chris D. Orem, Center for Assessment and Research Studies, James Madison University Correspondence concerning this article should be addressed to Keston H. Fulcher, Center for Assessment and Research Studies, James Madison University, Harrisonburg, VA 22807.
E-mail: fulchekh@jmu.edu

Abstract

Higher education experts tout learning outcomes assessment as a vehicle for program improvement. To this end the authors share a rubric designed explicitly to evaluate the quality of assessment and how it leads to program improvement. The rubric contains six general assessment areas, which are further broken down into 14 elements. Embedded within the article are links to the full rubric, an example of an exemplary assessment report, and a how-to guide for conducting and reporting quality assessment.

Introduction

As assessment practice in higher education evolves so too do the questions institutions and accreditors pose about assessment. Until recently the questions focused on participation and could be answered with statements like, "Ninety-seven percent of our academic degree programs submitted assessment reports in the current academic year." Although certainly important and an indicator of compliance, this information reveals little regarding the *quality* of assessment. If, as we believe, assessment's primary purpose is to guide programs toward improvement, then quality must be considered. Examples of legitimate questions include: Are objectives stated appropriately? Is there a clear link between the objectives and the methodology? Is the methodology sound? Is the interpretation of the program's strengths and weaknesses justified by the results? Do the program's plans for improvement logically fit with the results and interpretation? However, conveying information about quality is more challenging than conveying information about quantity.

Nonetheless, like Suskie (2009), we believe evaluating the integrity of assessment is a worthwhile endeavor. To this end James Madison University has developed a rubric to provide constructive feedback on the quality of assessment that can be used diagnostically at the academic program level and higher organizational levels. In this article we highlight the (a) focus of this rubric, (b) the assessment elements that are evaluated, (c) possible uses of resulting information, and (d) further considerations.

Focus of Rubric

To clarify our conceptual position, consider a scenario where a provost is reading two year-end assessment reports. Reviewing these documents, she discovers that the first program's report includes exceptionally positive results. On closer inspection, however, the results are based exclusively on indirect measures, course experiences are not mapped to learning outcomes, and information regarding the veracity of the assessment instruments or data collection design is absent. Further, the program provides no record of using results for improvement.

The second program's assessment report differs drastically from the first. It does not boast the same glowing results, but it clearly walks the reader through its assessment process. Specifically, the second program provides a convincing argument that the results are trustworthy and directly answer questions related to its objectives. Furthermore, the report clearly outlines how these results will be used to make improvements to both the program and the assessment process. If you were the provost, with which program would you be most satisfied?

This hypothetical scenario illustrates two contrasting perspectives when evaluating assessment reports. One approach concentrates primarily on the results; the other focuses on the trustworthiness of the results and how a program responds to its findings. From our perspective, we hope that administrators and faculty embrace the second. If assessment's primary role is for program improvement then assess-

ment should be evaluated on the quality of information it provides and the logic of the decisions that are derived from it.

Elements of the Rubric

From this perspective, James Madison University created a rubric that guides evaluative feedback on assessment. It is most directly applicable for academic degree programs. You can examine this rubric by going here: http://www.jmu.edu/assessment/JMUAssess/APT_Help_Package_4_15_2010.pdf The link also leads to several other related documents including a hypothetical exemplary report and a how-to-guide for conducting assessment. The interested reader will find that these documents provide much more detail than this article.

The rubric consists of six general areas that are further broken down into 14 elements (see Figure 1). The selection of elements was based upon several common models of assessment including Erwin's (1991) and Suskie's (2009). Although other rubrics have been developed for this purpose (e.g., Christopher Newport University: http://assessment.cnu.edu/docs/uaec_review_form.pdf), this rubric most clearly articulates the expectations for sound methodology, the area where many assessments break down.

<table>
<tr><th>Objectives</th><th>Learning Experiences</th><th>Methodology</th><th>Results</th><th>Results Shared</th><th>Program Improvement</th></tr>
<tr><td>Clarity & Specificity</td><td rowspan="5">Learning Experiences</td><td>Relationship Objs & Measures</td><td>Presentation</td><td rowspan="5">Results Shared</td><td>Program Improvement</td></tr>
<tr><td rowspan="4">Orientation</td><td>Types of Measures</td><td>History</td><td rowspan="4">Assessment Improvement</td></tr>
<tr><td>Specification</td><td rowspan="3">Interpretation</td></tr>
<tr><td>Data Collection</td></tr>
<tr><td>Additional Validity Evidencer</td></tr>
</table>

Figure 1. Organization of Rubric

Each of these elements is evaluated on a four point scale where 1 = Absent; 2 = Needs Improvement; 3 = Meets Expectations; and 4 = Exemplar. For each element the rubric provides a behavioral description associated with each level of performance. See Figure 2 for examples the verbs describing the desired actions of the students, and the content and skills to be exhibited – leads to the highest scores.

Learning Experiences

The rubric's second area targets the degree to which a program's courses/learning experiences are mapped to its objectives. Exemplary scores represent programs that have matched all of their objectives to curricular and sometimes co-curricular learning experiences. Note that a good curriculum map itself is not evidence of student learning. Rather, it represents where students should theoretically gain knowledge and skills.

Methodology

The rubric's third area covers methodology, the critical process that occurs between objectives and results. We find this is the area where faculty feel least comfortable and need the most feedback. Therefore, this section is divided granularly into five elements. The first element gauges the relationship between the measures (such as tests, essays, portfolios) used by a program and its objectives. Programs that score well not only provide a list of their measures, but they describe in detail *why* the measure is a good fit for assessing one or more objectives. To this end, faculty subject experts can specify exactly what component of a test corresponds to the objective. For example, a biology program could indicate that an entire rubric on oral communication corresponds to how it specified its objective on oral communication, which included eye contact, a good hook, clear organization, etc. Similarly, for a multiple choice test, the faculty would need to specify which items correspond to which objective(s). The main idea here is that faculty should choose a test or rubric that represents the skills and content outlined by one or more objectives.

3A. Data collection & research design integrity			
Absent	Needs Improvement	Meets Expectations	Exemplary
No information is provided about data collection process or data not collected.	Limited information is provided about data collection such as who and how many took the assessment, but not enough to judge the veracity of the process (e.g., thirty-five seniors took the test).	Enough information is provided to understand the data collection process, such as a description of the sample, testing protocol, testing conditions, and student motivation. Nevertheless, several methodological flaws are evident such as unrepresentative sampling, inappropriate testing conditions, one rater for ratings, or mismatch with specification of desired results.	The data collection process is clearly explained and is appropriate to the specification of desired results (e.g., representative sampling, adequate motivation, two or more trained raters for performance assessment, pre-post design to measure gain, cutoff defended for performance vs. a criterion)
6A. Improvement of programs regarding student learning and development			
No mention of any improvements.	Examples of improvements documented but the link between them and the assessment findings is not clear.	Examples of improvements (or plans to improve) documented and directly related to findings of assessment. However, the improvements lack specificity.	Examples of improvements (or plans to improve) documented and directly related to findings of assessment. These improvements are very specific (e.g., approximate dates of implementation and where in curriculum they will occur).

Figure 2. Examples of Behavioral Anchors Associated with Two Elements of the Rubric

The type of measure being used is also reviewed. Compared to essays, portfolios, or multiple choice tests, surveys are considered indirect and less objective. Correspondingly, the rubric rewards programs for using direct measures associated with each of its objectives. Note that it is good practice to include indirect measures but only as supplements to the direct measures.

Programs are also evaluated on whether they specify desired results for their objectives. The purpose of this element is to provide context for assessment results. Too often faculty will look at their assessment results and have little context for interpretation. If, at the outset (i.e., a priori), they indicate what results would indicate success, then the findings become more interpretable. Exactly what these results should look like depends on the type(s) of questions asked. What percentage of students meets a standard? How do students compare to similar programs across the country? To what degree did students change regarding their skills and knowledge? How does this cohort compare to the previous cohort? The rubric rewards specificity and rationale. As opposed to - "We intend for this cohort to perform better than last year's students." - a statement like this is much more powerful: "For the current cohort, our desired result is an average score of 83 on the exit exam. This score would connote a ½ standard deviation improvement from the previous year. We chose this moderate level of improvement because the current cohort is the first to undergo a modified curriculum where core content was emphasized more heavily." Articulating the desired results in such a fashion not only makes the results more interpretable but will likely entice faculty to engage with the findings; results are always more interesting when they address a question.

The next element under methodology is data collection. The most common problem we see in this area is insufficient information. At a minimum, an evaluator would need to know which students are targeted (i.e., population of interest, which should be specified in the objective), how students were sampled, the conditions under which students took the assessment, and their effort level. In addition, for a performance assessment, one would need to know about the raters and how they were trained. As an example, a program may report that "40 out of 41 seniors took the assessment during a set day in their senior seminar class in the spring semester; the test was proctored by faculty members, and was a graduation requirement. Consequently, proctors observed that students gave a good effort."

The fifth and final element under methodology refers to additional validity evidence. One may note that all six of the rubric's areas relate to validation of results and interpretations, or as Benson (1998) puts it, "…the process by which scores take on meaning" (p. 10). This element focuses on a particular part of validation: the psychometric properties of data like reliability. Note, we realize that some practitioners may be unfamiliar with these concepts. Nevertheless, they are necessary conditions of trustworthy results. We therefore strongly encourage faculty to consult with their institution's assessment consultants. Reliability estimates like coefficient alpha, inter-rater reliability, and other measures of consistency are all appropriate to report. The highest ratings are awarded to those programs whose assessment data have decent reliability and additional validity evidence. For example, if students who take more general education courses in mathematics score higher on a quantitative reasoning test, then such a result lends validity evidence to the test scores . Of all 14 elements on the rubric, this is likely the most difficult. Onlythe most mature programs who have worked with assessment experts (internal or external to the program) will receive exemplar marks.

Results

The fourth area of the rubric corresponds to assessment results, which is broken down into three elements: (a) presence of results—to what extent do they correspond to objectives? (b) history of results—in order to demonstrate trends, do programs report more than one year of data for some or all of their objectives? (c) interpretation of the results—does a program make reasonable inferences about the scores based on the methodology used? It is important to reiterate that the rubric does not directly evaluate whether or not desired results are achieved, but instead evaluates whether programs address the veracity of the results and how the program interprets and responds to them. In other words, a program can fall short of reaching their desired results, but still receive a high score. They can do so by providing a logical interpretation of the findings and reasons it believes the results fell short of expectations.

Sharing Results

Area five covers the ways in which a program disseminates its results to various stakeholders. Programs that do not share their results, or only provide data to a limited number of faculty members will score lower than ones that make their scores widely available to both internal and external audiences. The idea here is that assessment should be a collaborative enterprise among all faculty within a program and, ideally, external stakeholders such as an advisory board. Conversely, an assessment report viewed only by the eyes of the author rarely has bearing on a program.

Using Results

Making thoughtful programmatic changes to improve student learning is the very impetus of assessment, and it is the focus of the rubric's sixth area. The best assessments guide stakeholders in decision making, whether it be curricular, co-curricular, pedagogical, budgetary, etc. One may note that to make sound data-driven decisions, one needs to trust the assessment results first. Thus the emphasis on good objectives, methodology and the reporting of results noted in previous areas of the rubric. Exemplary assessment reports follow a clear logic from the assessment results to improvements that have been (or will be) implemented; as always, the more detailed the better.

In addition to evaluating the presence of results-driven improvements, the rubric also reviews whether programs address shortcomings to the assessment process itself. This element emphasizes that assessment is an ongoing process. As already stated, trustworthy results are a pre-requisite to using results for improvement. Therefore, by improving one's assessment, the likelihood that good decisions will be made about the program also increases. Recognizing that programs with strong assessment practices may not need to make drastic improvements to their assessment process, those who receive exemplary marks on the majority of the first five areas automatically receive a high score on this final element.

Using Information Obtained from Rubric

This article describes a process for evaluating assessment of academic programs via a rubric with six areas and fourteen elements. As with all assessments, it is essential to consider how the results will be used. We recommend two uses: (a) as a vehicle that provides diagnostic feedback about individual program's assessment and (b) as a mechanism to convey the quality of assessment across programs (i.e., at the department, college, and university levels). Regarding individual feedback, it informs faculty within programs about the strengths and weaknesses of their assessment. For example, perhaps a program's objectives are well articulated but concerns about methodology (e.g., absence of data collection procedures) cast doubt about the meaningfulness of the results. Consequently, in the next year the program can focus its efforts on improving the data collection process.

Additionally, feedback from the rubric can be diagnostic at the larger university level. The scores can be aggregated to identify strengths and weaknesses in the assessment process across programs, departments, and colleges. This information provides a university insight into how it can most efficiently support programs by creating or adapting services to bolster common needs. For example, the Office of Assessment could host a workshop on articulating desired results. Additionally, aggregated scores from across the university provide a gauge of where an institution stands regarding overall quality of academic program assessment. This information is easily interpreted by stakeholders and accrediting bodies. In essence, this aggregated data could be used to answer the quality-of-assessment questions at the macro level posed at the beginning of this article.

Further Considerations and Conclusions

While the primary focus of this article is on the rubric itself, there are several other important questions to consider when instituting an evaluation system of assessment reports. Will the assessment reports be collected electronically, or will they be turned in via hardcopy to a central location? Is there a common format required to make the reports easier to read, or are programs granted "creative discretion?" Who will rate the reports: faculty, students, or professional staff? How will raters be recruited and trained? Will feedback be provided for every section of the rubric or will general suggestions be made? What resources are available to programs that do not score well? Will the results of the rubric be used for high-stakes decision making, or simply for program improvement?

We acknowledge that assessment is a resource-intensive endeavor requiring money and, particularly, the time of faculty and staff. As such, this process needs to bear fruit in the form of enhanced student learning from improved degree programs. We hope that this rubric can be a resource toward that end. Regardless of whether this particular tool is appropriate for your institution, we recommend that every university incorporates some process of evaluating assessment. Too often this aspect of the assessment cycle is overlooked, an ironic fate for an endeavor rooted in reflection.

References

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17*, 10-22.

Erwin, T. D. (1991). *Assessing student learning and development.* San Francisco, CA: Jossey-Bass.

Suskie, L. (2009). *Assessing student learning: A common sense guide* (2nd Edition). San Francisco, CA: Jossey-Bass.