Association of American Colleges & Universities
A Voice and a Force for Liberal Education in the 21st Century

The Dependability of VALUE Scores: Lessons Learned and Future Directions

By: Gary Pike and Kathryne Drezek McConnell

In the decade since their release to the broader higher education community, the VALUE rubrics have been downloaded more than 70,000 times by individuals representing more than 5,895 organizations, including more than 2,188 colleges and universities. As part of their dissemination, institutions were encouraged to take the VALUE rubrics and make sense of them within their own unique culture and context. In this way, the original nomenclature—the VALUE meta-rubrics—provided an apt description of the rationale and appropriate use of these new assessment tools. Beginning in 2014, in addition to their use for locally based institutional assessment of student learning, the VALUE rubrics were used in the first-of-its-kind national scoring initiative (see McConnell and Rhodes 2017), which ultimately became the VALUE Institute. As the "intellectual and logistical stewards" of the VALUE rubrics (McConnell et al. 2019, 2), the Association of American Colleges and Universities (AAC&U) finds itself in a rather unique position vis-à-vis the VALUE approach, in that allowing (and even encouraging) local modification of the VALUE rubrics "signals a loosening of control—from modification and implementation to scoring and interpretation of data—that appears to be unique to the VALUE approach to assessment and stands in particular contrast to protocols associated with commercially available national standardized tests. As an approach to assessing student learning, VALUE must balance local pedagogical needs with methodological control" (McConnell et al. 2019, 2).

The VALUE approach to assessment is methodologically, epistemologically, and pedagogically complex, and as such, comparing and contrasting the VALUE approach with standardized tests will always represent an "apples to oranges" proposition (McConnell and Rhodes 2017; McConnell et al. 2019). That said, AAC&U recognizes that to fully realize their promise and achieve credibility commensurate with that enjoyed by standardized tests, the reliability and validity of the VALUE rubrics must be clearly established (Rhodes 2012b). Faculty from across the country have been involved in efforts to evaluate the content, convergent, and face validity of the VALUE rubrics (McConnell and Rhodes 2017; Pusecker et al. 2011; Rhodes and Finley 2013) and to assess levels of inter-rater agreement (Finley 2012; McConnell and Rhodes 2017; Rhodes 2012a). Yet more remained to be done. Establishing the credibility of the VALUE rubrics requires that the dependability of VALUE scores be evaluated consistent with the ways in which the scores are used for student-, institution-, and state-level assessment of student learning. This article briefly describes research that was designed to evaluate the dependability of VALUE scores (Pike 2018) and addresses the implications of this work for local and nationwide scoring efforts.

Framework for Evaluating the Dependability of VALUE Scores

First, the technical explanation. Generalizability theory represents the most appropriate method for assessing the dependability of scores obtained using the VALUE rubrics because it can be tailored to represent the assessment methods being used to make judgments about student learning (Pike 1995). Generalizability theory assumes that measures, whether they are questions on a standardized test or raters scoring student artifacts, are random samples from a larger universe of all possible observations (Haertel 2006). Ultimately, questions about the dependability of measures focus on whether the samples of test questions or raters allow for consistent generalizations about the universe of observations (Brennan 2006). Importantly for our consideration of the VALUE rubrics, generalizability theory allows us to account for multiple sources of error, which in turn allows assessment researchers to obtain more appropriate reliability indices and to identify how changes in an assessment design can influence the dependability of measurement (Erwin 1988; Webb, Rowley, and Shavelson 1988). While the "ideal" would be to base decisions on the average score over all possible measures (Cronbach et al. 1972)—such as an average score across all of the pieces of work a student generated in any given class or program—this ideal is seldom attainable. Instead, we must generalize from limited samples to the universe of all possible observations. The generalizability coefficient provides us with information about the dependability of generalizing from an observed score, based on our sample, to the mean score for all possible observations (Cronbach et al. 1972).

But what does all this mean in practical terms? How are we to ascertain and communicate the generalizability and dependability of the VALUE rubrics to faculty, faculty developers, assessment professionals, and ultimately perhaps even students, so that they are informed and empowered to make changes to enhance student learning?

To answer these important questions, it may be helpful to return to the imagery first evoked by the AAC&U report, *On Solid Ground*, of a "landscape of student learning" (McConnell and Rhodes 2017, 3). A landscape is more than simply a collection of topographic features; it is the natural expanse or scenery that one can see in a single view, from a single vantage point. What often matters most when taking in a breathtaking view is the overall effect, the patterns illuminated, the collective power of the panorama, with individual features—peaks and valleys, rivers and coastlines, forests and mountains—retreating to the background. Such landscapes can be found in the work of nineteenth-century painters like J. M. W. Turner, Robert Duncanson, and Claude Monet. However, the ability to see the full, complete picture is also dependent upon viewing these constituent parts of the landscape in relationship to one another. Artists of another kind, such as the eighteenth-century surveyors Charles Mason and Jeremiah Dixon, focused on accurately detailing and mapping the landscape, rather than capturing the broad expanse. In their case, the goal was to depict exactness, such as the "true" border between two American colonies, not breadth. The two approaches do not necessarily need to stand in contrast or in conflict, as both views—the forest and the trees—enhance our understanding of the world we see.

Extending the landscape metaphor to the VALUE work, we are reminded that assessment—as well as teaching and learning writ large—is both art and science. We aim to paint a picture of

learning and create a narrative of student success that is compelling and readily understandable to a host of critical audiences, while at the same time ensuring the accuracy of the picture we paint. Our work on generalizability and dependability is not unlike the work of surveyors trying to measure and map out the features of a given landscape. Generalizability helps us to map, like Mason and Dixon, the precision of our measurement, the "trueness" of our picture of learning, by depicting its constituent parts statistically. This, in turn, allows us to take a step back and, like Monet and Turner, see the emerging landscape more clearly.

Key Findings, Lessons Learned, and Future Directions

The data for the generalizability research were drawn from the data used in the AAC&U report, On Solid Ground (McConnell and Rhodes 2017). Specifically, the research uses the data from the subset (approximately 20 percent) of student work that was double scored (scored by two raters). These data came from the Multi-State Collaborative assessment project, as well as from the Great Lakes Colleges Association (GLCA) Collaborative and the Minnesota Collaborative. Details on the data collection and the institutions participating in the study are presented in On Solid Ground. Data for the student-level analyses of critical thinking scores included 1,572 student work products evaluated by two raters, and data for Written Communication included 1,683 student work products that were scored by two raters. The data for Quantitative Literacy included the work products from 1,496 students scored by two raters. Both G- (generalizability) and D- (decision) study models were generated to assess the dependability of VALUE rubric scores at the student, institution, and state levels.

It is important to note that the findings of the present research are limited in at least two important ways. First, the sampling and scoring protocols ensured that student artifacts would be randomly selected and that raters would be randomly assigned to score the artifacts. The assignments used to elicit student work, however, were not randomly selected. The Multi-State Collaborative recruited willing faculty on each campus to participate, who then in turn submitted an assignment and corresponding student work from their course. While not a random sample, this volunteer approach was a preferable to requiring institutions to force or require faculty to submit assignment prompts and student work. Institutions and states also self-selected into the project, thereby restricting variability across the states and institutions. Last, student work products connected with any one outcome (i.e., Written Communication, Critical Thinking, or Quantitative Literacy) were the products of multiple assignments, not a single, standardized assignment.

Despite the limitations of the present research, it is possible to draw some conclusions about the dependability of assessments using the VALUE rubrics. Based on our first foray into evaluating the generalizability of the VALUE approach, the dependability of the Critical Thinking, Written Communication, and Quantitative Literacy VALUE rubrics does not yet rise to the levels expected of standardized tests. Not surprisingly, the greatest source of variance at the student, institutional, and state levels of assessment is in raters' scores, which can reduce the dependability of students' scores (Pike 2018). We approach these results as a baseline understanding of the psychometric properties of the VALUE rubrics that, when triangulated

with other sources of data, confirm areas of relative strength and suggest areas for further refinement and improvement of the VALUE approach.

Improving raters' scores has potential implications for three constituent components of the VALUE approach—the scorers and the training they receive, the assignments that generate the student work that gets scored, and the VALUE rubrics themselves. While this research identified several possible avenues for improving inter-rater reliability, selecting among the range of strategies for enhancing dependability must balance methodological concerns with maintaining the core tenets of the VALUE approach to assessment. For example, one possible strategy for enhancing the dependability of the VALUE approach would be to simply increase the number of raters scoring each piece of student work from two to four, five, or even six raters. However, the resources required to achieve that level of scorer participation, either locally on a single campus or as part of the VALUE Institute, would be cost prohibitive. By way of a second example, the research revealed that variance across assignments was also an important source of error in institutional mean scores. Establishing whether this variance was attributable to differences in the difficulty of the assignments, or whether it was due to a poor match between some assignments and the rubrics themselves, was beyond the scope of the present investigation. One possible solution to this issue would be to develop and require the administration of standardized assignments. This solution, however, runs counter to VALUE's longstanding principle that faculty-designed and administered assignments from existing courses represent the most authentic learning of students at our institutions.

The findings of this research support several enhancements for each of the three constituent components of the VALUE approach to assessment:

1. Enhanced scorer calibration training. One possible method of improving inter-rater agreement is through better training of raters. Working with experts in performance-based assessment, AAC&U is revising its VALUE rubric training protocols to move to a more robust and rigorous protocol for training scorers, particularly for those scoring work as part of the VALUE Institute. Resulting protocol guidelines will be made available for local campus use, recognizing that individual institutions may choose to modify the protocols to meet local needs.
2. Improved assignments. AAC&U will continue to support assignment alignment with the VALUE rubrics through assignment (re)design. Drawing on the excellent work of and in partnership with organizations like the National Institute for Learning Outcomes Assessment, AAC&U will continue to work to help faculty and other higher education professionals find and/or (re)design assignments to ensure alignment between what is asked of students and the VALUE rubrics. Developing specifications for the types of assignments used to elicit products representing particular learning outcomes (e.g., Critical Thinking, Written Communication, or Quantitative Literacy) may help to improve the dependability of assessments. These specifications would almost certainly better ensure a match between the assignments and the dimensions of the scoring rubrics. Furthermore, the research suggests that increasing the number of assignments that each student completes—thereby increasing the number of artifacts of work generated per student—may prove helpful in reducing the error attributable to differences in assignments.

However, it may also require each student to submit as many as four or five products for scoring. While some practitioners may counter that this approach would be burdensome for students, faculty members who design the assignments, and raters, it actually aligns with one of the original design principles that informed the creation of the VALUE rubrics, namely:

> that good practice in assessment requires multiple assessments over time: well-planned electronic portfolios (ePortfolios) provide opportunities to utilize college data from multiple assessments across a broad range of learning outcomes and modes for expressing learning, while guiding student learning and building reflective self-assessment capabilities; and that assessment of student work in ePortfolios can inform programs and institutions on their progress in achieving expected goals for external reporting and at the same time, provide faculty with information necessary to improve courses and pedagogy. (Rhodes 2010)

3. Revisiting and revising the VALUE rubrics themselves. AAC&U will spearhead the revision of all sixteen VALUE rubrics beginning in 2019. This research will play a critical role in the revision process. For example, the research revealed that achieving acceptable levels of generalizability is easier for some dimensions of the VALUE rubrics than for others. As such, one possible avenue for improving inter-rater agreement is to carefully review the descriptive statements associated with score-points on the VALUE rubrics' dimensions. Dimensions with low levels of generalizability should be a starting point for reviewing and modifying these descriptive statements. Additionally, AAC&U will engage faculty and—for the first time—students through focus groups and campus vetting of revised versions of the VALUE rubrics. This work has the potential to improve not only the content and design of the VALUE rubrics but also the reliable and accurate application of the VALUE rubrics to student work.

AAC&U takes its role as steward for the VALUE approach seriously and is committed to addressing the methodological gaps identified by this research, starting with the recommendations delineated above. We believe the lessons learned and future directions described above do just that and welcome the continued efforts of others in the academy to help us refine and improve the VALUE approach to bring the emerging landscape of learning into full relief.

https://www.aacu.org/peerreview/2018/Fall/Research