peerReview

EMERGING TRENDS AND KEY DEBATES IN UNDERGRADUATE EDUCATION

A Decade of VALUE

Inquiry and Analysis

Critical Thinking

Creative Thinking

Written Comm.

Oral Comm.

Reading

Quantitative Literacy

Information Literacy

Teamwork

Problem Solving

Civic Engagement

Intercultural Knowledge

Ethical Reasoning

Lifelong Learning

Global Learning

Integrative Learning

Essential Learning Outcomes

CONTENTS

peerReview
EMERGING TRENDS AND KEY DEBATES IN UNDERGRADUATE EDUCATION
VOL. 20, NO. 4 | FALL 2018

Welcome to the this special issue of *Peer Review*, in which we commemorate the tenth anniversary of the VALUE rubrics' release. Although it was not known at the time, 2009 began a decade of the public's questioning the worth of higher education for students' success. There was a decline in public funding and support for colleges and universities, and a litany of complaints about lack of student learning and faculty disinterest or bias.

The creation of the VALUE (Valid Assessment of Learning in Undergraduate Education) rubrics was an intentional response to that environment. VALUE was built on the premises that the best indicators of student accomplishment or attainment of Essential Learning Outcomes was the work students produced from assignments given by faculty and other educators both in formal classroom settings and outside activities; that professional educators did exercise practiced expertise and judgment about the quality of student performance; and that it was possible to articulate the shared core dimensions or elements of the Essential Learning Outcomes so students, faculty, employers, and others could judge for themselves the quality of learning that students' work demonstrates.

The assumptions underlying the VALUE rubrics are (1) that all students bring with them a set of skills and abilities that offer a basis upon which further growth and development of learning is built; (2) that the better higher education can articulate what outcomes and corresponding levels of learning are expected for successful completion of educational goals, the more likely it is that educators can design and deliver intentional practices that will enhance student ability to achieve quality performance; and (3) that learning is an iterative and complex set of experiences that need to be integrated and practiced over time, with educational settings designed to be progressively more challenging and assessed for high levels of attainment.

This issue's content—selected and guided by Kathryne (Kate) Drezek McConnell, AAC&U assistant vice president for research and assessment—includes a range of authors who have been involved with different aspects of the development of the VALUE rubric approach over the past ten years. Several authors were active through the Multi-State Collaborative that piloted the large-scale implementation of the VALUE approach (www.aacu.org/value/msc). Michael Ben-Avie, Kevin Kuna, and I examine how VALUE is a strategic approach to learning connected to the work of the campus rather than a test or isolated initiative. John Hathcoat explores the meaning of the rubric scores and how each score represents important learning opportunities rather than failures or inadequacies. Martha Stassen and Anne Harrington direct attention to the ways in which faculty are engaged by the VALUE assessment approach in pedagogical and learning improvement. Eric Vanover connects the curriculum to civic engagement and the historical conception of community colleges. Kimberly Filer and Gail Steehler dispel the myths that perfection in assessment is possible and that more good may come from the imperfect measurement of the messy reality that is learning in practice. Gary Pike and Kate McConnell discuss the ability to generalize from data generated by using the VALUE rubrics. Finally, David Eubanks argues that inter-rater reliability and rubric methodology have often masked the actual usefulness of rubric scores and results for learning and improvement. He also considers the potential and pitfalls of reconceptualizing assessment based on large scale examination of student work and generalizability around technical and social requirements for learning improvement.

This broad dive into the many dimensions of the VALUE approach to assessment continues to illustrate the strength of critical friends in enhancing assessment of student learning. A single number, score or grade is not sought through the VALUE approach, but rather multiple measures over time and circumstance that together make meaningful sense of essential learning and abilities for translating learning into practice to improve lives and societies for oneself and others.

—**TERREL RHODES**
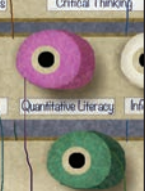
Assessment as a Strategy, Not a Stand-Alone Activity

▶ **Michael Ben-Avie,** Senior Director for Learning Assessment and Research, Quinnipiac University, formerly of Southern Connecticut State University

Kevin Kuna, Student, Southern Connecticut State University

Terrel Rhodes, Vice President, Office of Quality, Curriculum, and Assessment and Executive Director of VALUE, AAC&U

When the Association of American Colleges and Universities (AAC&U) developed the VALUE rubrics ten years ago, the rubrics began as a set of shared expectations around key learning outcomes and essential skills associated with student success in school and life. These rubrics were created and agreed upon by educators in the field who described what to look for in student work as learners progress through their educational pathways to acquire progressively more sophisticated knowledge and competence as they approach attainment levels associated with a baccalaureate degree. Intended as metarubrics, VALUE rubrics use general descriptors to represent and capture fundamental dimensions of each learning outcome while encouraging students to demonstrate learning in different contexts, media, and forms appropriate to their circumstances.

Through the VALUE approach to assessment, institutions collect direct evidence of student learning using performance-based assessments. VALUE is exemplary in this regard because the direct evidence (e.g., final term papers) is closely linked to students' learning experiences in college courses. This is in contrast to other direct evidence measures that are collected through standardized tests administered outside of the classroom and not aligned with the curriculum.

More than 70,000 individuals—affiliated with 5,895 organizations, including 2,188 colleges and universities—have downloaded the rubrics. In 2014, AAC&U was able to engage in a systematic scaling of the VALUE rubric approach in collaboration with the State Higher Education Executive Officers association's Multi-State Collaborative (MSC), eventually involving thirteen states and ninety-two two- and four-year institutions. Most recently, in 2017, AAC&U announced the creation of the VALUE Institute.

MUCH MORE THAN A SCORE

Paying attention to students' development does not detract from their learning. In fact, promoting the highest levels of development among students seems to help them reach high academic goals (Ben-Avie et al. 2003). This is particularly the case with academic habits of mind. When faculty promote students' academic habits of mind, students gain the competency to work autonomously, handle cognitive complexity, master processes of inquiry that are common to all academic disciplines, and advocate for themselves. In short, institutions can build multidimensional predictive models combining direct and indirect evidence to strengthen student learning success.

One of the best ways to illustrate the benefits of the VALUE approach compared to other assessment methods (like standardized testing) is to examine it in practice. One MSC state—Connecticut—illustrates the robust implications of the VALUE approach as a strategy for learning assessment rather than a one-off activity. The directors of assessment from the participating Connecticut institutions initially thought that the MSC was yet another stand-alone activity, not a strategy. Stand-alone activities tend to be externally mandated, and the results—even if they are

widely shared on campus—are not effectively used to improve students' learning and developmental outcomes. The results from these stand-alone activities are rarely incorporated into longitudinal cohort datasets. However, following students from new-student orientation until they graduate or transfer to other colleges provides information that is useful when identifying important metrics for judging the effects of higher education.

For example, at Southern Connecticut State University (SCSU), the VALUE rubric scores of student work artifacts were incorporated into longitudinal cohort datasets to observe students' academic performance in relation to competencies (e.g., future orientation, interpersonal relationships, sense of belonging, self-regulation, academic habits of mind) from the developmental sciences. By incorporating scores of 444 work artifacts from SCSU students into long-term studies, VALUE and the MSC brought learning evidence into SCSU's promotion of students' success and development.

An important result from the study influenced the campus conversation about on-time graduation. Students who graduated in four years had significantly higher scores on quantitative literacy (QL) than those who graduated in five or six years. This finding attracted the interest of those involved in enrollment management and student success. Analyses determined that students' higher QL scores were predicted by the experiences they had on campus instead of their past development, demographic characteristics (e.g., ethnicity or socioeconomic status), or precollege learning (e.g., SAT scores or high school GPA). As a result, VALUE results contributed to a shift in strategy at the university to promote student success.

The research results provided evidence that practices that higher educa-

tion institutions control can promote students' competencies in written communication, quantitative literacy, and critical thinking. Thus, the MSC VALUE approach became an antidote to metrics that had been proposed for judging the effectiveness of higher education that were outside the control of the university, because the characteristic or learning either occurred prior to college or was related to postgraduation employment. Consider, for example, the *Wall Street Journal's* method of ranking colleges.

Analyses determined that students' higher QL scores were predicted by the experiences they had on campus instead of their past development, demographic characteristics (e.g., ethnicity or socioeconomic status), or precollege learning (e.g., SAT scores or high school GPA).

As Melissa Korn explains this ranking method, "Outcome scores are derived from graduation rates and academic reputation, as well as measures of loan-repayment rates and graduate salaries" (Korn 2016).

At the institutional level, VALUE can effectively leverage assessment for improvement because participation requires, for instance, assessment leaders to actively engage faculty members and other educators in the data collection process. As a result, knowledge and understanding of assessment increase among educators. In addition, educators become more committed to assessment because VALUE invites them to attend training on how to use a rubric to score students' work products. Through the use of rubrics, students' reflective thought

and insight are valued and, as a result, provide an incentive for faculty development in these areas.

At the same time, the MSC VALUE approach became an accreditation strategy when conversations about outcomes led to changes based on evidence instead of anecdotes or preconceptions. SCSU's accrediting agency, the New England Association of Schools and Colleges (2017), commented on changes that were made at the university due to participation in the MSC:

We are pleased to learn that results gleaned from analyses of student work conducted as part of Southern Connecticut State University's participation in the Multi-State Collaborative have been used to inform the restructuring of the University's access programs, developmental math curriculum, liberal education program, and writing across the curriculum program.

VALUE PLUS DEVELOPMENT EQUALS A WINNING MODEL

By students' senior year, differences in VALUE scores between students who were eligible for PELL grants and their peers were not found in written communication, quantitative literacy, and critical thinking. To improve students'
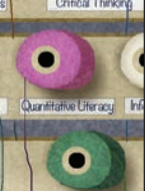
competencies in learning outcomes, faculty development events should focus on topics such as written communication, quantitative literacy, and critical thinking, because improving students' competencies also requires faculty knowledge of how to apply the developmental sciences to classroom practices without compromising the objectives of a course.

At SCSU, assessment professionals built a predictive model using specific items on the Academic Habits of Mind and College Success Inventory (AHM-CS), designed by the Office of Assessment and Planning to measure the relationship between students' learning and development. This self-assessment inventory predicted students' overall VALUE scores in written communication, quantitative literacy, and critical thinking. For example, the item from the AHM-CS that asked students to indicate whether they take the initiative to talk with their professors when issues arise predicted students' overall quantitative literacy scores. In other words, if we know the extent to which students take the initiative to talk with their professors, then we also know how they are likely to score on the quantitative literacy VALUE rubric.

On the critical thinking VALUE rubric, students are expected to demonstrate a high level of competency in evaluating issues, artifacts, and ideas before formulating a conclusion. Students are expected to analyze text and draw conclusions that either support or question the author's viewpoints. Students who "strongly agreed" on the AHM-CS that their professors were teaching them how to express their positions during classroom discussions tended to have VALUE scores that effectively met the expectations at progressively higher levels within the critical thinking rubric.

The written communication VALUE rubric expects that students demonstrate ample consideration of the audience and purpose for the writing assignment. A moderate relationship was observed between the students' "context of and purpose for writing" score on the rubric and the AHM-CS item that asked if the students were able to effectively read and comment on the work of fellow students. This indicates that the students who were required to review and critique the papers of their peers gained an awareness of how to write papers for two audiences: their professors and their peers.

The university's assessment strategies can now be used to empirically evaluate the relationship between students' learning and development. A student's ability to demonstrate learning according to the rubrics is a function of (1) academic habits of mind such as the ability to work autonomously and handle cognitive complexity, (2) content knowledge, (3) interpersonal relationships, and (4) an orientation to the future by setting goals and taking immediate actions to achieve desired futures. Mastering these habits and competencies can change students' developmental trajectory beyond what demographic characteristics and learning prior to college alone can predict. Thus, issues related to educational psychology are relevant for all students, and not only for students at risk of not thriving in college and in life. Using VALUE data, universities are able to show the impact of the education they provide as students veer from their predicted trajectories, making this "value added" one of the clearest metrics of the effectiveness of higher education.

CODA

For ten years, the VALUE rubrics have provided an approach to assessing AAC&U's LEAP Essential Learning Outcomes associated with student success in society, work, and life. Employers continue to strongly echo educators in saying that graduates need to demonstrate competence in these outcomes, and they are increasingly finding value in these cross-cutting outcomes that is equal to—if not more important than—a graduate's major field of study (Hart Research Associates 2018).

The MSC and the VALUE Institute have been able to systematically examine at scale the VALUE approach to learning assessment. The intriguing results have been gratifying for SCSU and include evidence that (1) students are learning essential outcomes; (2) engaging students in their learning (e.g., through high-impact practices) makes a positive difference in the quality of learning; (3) assessment results can engage educators and provide information useful for enhancing effectiveness in classroom practice; (4) educators and institutions can make a difference in closing equity gaps in learning quality and achievement; and (5) VALUE can lift up ways that educators can collaborate to achieve high-quality learning for all. In short, we now have evidence that higher education brings substantial value to individuals and to society. ▪

REFERENCES

Ben-Avie, Michael, Norris M. Haynes, Jayne White, Jacque Ensign, Trudy Raschkind Steinfeld, Loleta D. Sartin, and David A. Squires. 2003. "Youth Development and Student Learning in Math and Science." In *How Social and Emotional Development Add Up: Making Gains in Math and Science Education*, edited by Norris M. Haynes, Michael Ben-Avie, and Jacque Ensign, 9–35. New York: Teachers College Press.

Hart Research Associates. 2018. *Fulfilling the American Dream: Liberal Education and the Future of Work*. Washington, DC: Association of American Colleges and Universities. https://www.aacu.org/sites/default/files/files/LEAP/2018EmployerResearchReport.pdf.

Korn, Melissa. 2016. "What's the Payoff?" *Wall Street Journal*, September 29, 2016, R2.

New England Association of Schools and Colleges. 2017. Unpublished letter. February 2, 2017.

The Role of Assignments in the Multi-State Collaborative:
Lessons Learned from a Master Chef

▶ **John D. Hathcoat,** Assistant Professor of Graduate Psychology, James Madison University

Perhaps it is due to my distant British heritage, which often manifests as a cynical and dry sense of humor, but for some reason I have recently come to appreciate the work of Chef Gordon Ramsay. Chef Ramsay, who is arguably one of the best culinarians in the world, has held numerous competitions on television in which participants aim to win a prized position as head chef in one of his famous restaurants. On his shows, contestants are often asked to perform a variety of tasks over the course of several weeks. The tasks include such challenges as preparing a signature dish, blind taste tests, reproducing a meal, and turning "leftovers" into "fine-dining" cuisine. The competition can be brutal and Chef Ramsay is a difficult critic to impress. Although I find his antics generally amusing, and at times controversial, I believe we have much to learn from Chef Ramsay. There are striking parallels between the issues facing his assessment strategy and our efforts to examine student learning in higher education.

I also tend to be a critic. For example, I have been framed within the *Chronicle of Higher Education* (Berrett 2015; Berrett 2016) as a "supportive skeptic" of the Multi-State Collaborative (MSC). Like most academics, I enjoy the realm of skepticism. We like to question, debate, and critique the quality of evidence or reasoning used to support claims. In fact, I would argue that improvement is at best difficult, and at worst impossible, without some form of skeptical critique. Therefore, embracing skepticism is not our challenge in higher education; instead, we face the challenge of providing an adequate response to the skeptics. We have the task of creating an assessment system in higher education that can reasonably withstand this form of scrutiny. I believe that the MSC project is a positive step in this direction.

The MSC aims to provide a common language of student learning by assessing products that students create as a part of their curricular requirements. To date, ninety-two institutions across thirteen states have participated in the project. To partici-

> *I believe we have much to learn from Chef Ramsay. There are striking parallels between the issues facing his assessment strategy and our efforts to examine student learning in higher education.*
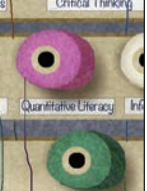
pate, institutions submitted sample student work for scoring by raters who were trained to use at least one of the VALUE rubrics. For several years, my MSC colleagues and I have conducted research related to the importance of intentional assignment design in assessment initiatives like the MSC. These themes include the problems of assignment misalignment and the tension between the competing values of generalization and directness when assessing performance. I aim to provide an intuitive overview of these issues by drawing parallels between the processes

employed by Chef Ramsay to evaluate the quality of chefs and our efforts to assess student learning in higher education.

CHEFS USE WHAT IS AVAILABLE IN THE KITCHEN

Like Chef Ramsey, who assigns tasks that allow him to make distinctions regarding the abilities of chefs, we too must make decisions about the best way to solicit performance so that students have an adequate opportunity to demonstrate what they have learned. Variations of a task sometimes referred to as a "black box challenge" occur in numerous cooking competitions. The black box challenge asks participants to cook a meal from a set of prespecified, surprise ingredients. For example, participants may be told that they have to use a specific protein, vegetable, and fruit when preparing an appetizer of their choice. Since all participants are provided with the same ingredients, we can infer that differences in the quality of appetizers are due to differences in their culinary ability. Judgments about their ability would be problematic if each chef was given a different quality of protein. This is analogous to having students complete an assignment by responding to the same, or theoretically exchangeable, prompts.

The black box challenge raises two important topics relevant to the MSC project. First, Chef Ramsay is justified when critiquing a competitor who did not use one of the specified ingredients. That mistake is often detrimental to a contestant. The contestant had an opportunity to demonstrate a skill and simply failed to do so. This situation is similar to students who do not display a skill when an assignment explicitly asked them to do so. Secondly, Ramsay would not be justified for critiquing a chef for something irrelevant, such as failing to include fish in the appetizer if the task clearly called for participants

to use chicken. In other words, there should be alignment between the task and the criteria used to distinguish quality. The problem of misalignment between assignments and assessment in higher education was reinforced by recent research that I conducted with my former graduate student, Nikole Gregg, which demonstrated that institutions participating in the MSC project need to disentangle whether a score of "zero" is a function of the student or the assignment.

The VALUE rubrics have scores that range from one to four. However, the scores given to student work products actually range from zero to four.
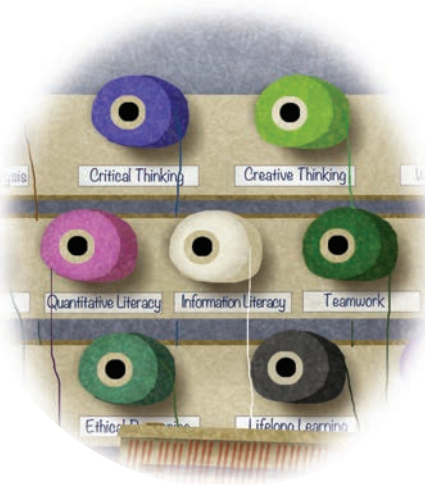
According to *On Solid Ground*, a report on the nationwide use of VALUE rubrics, "Scorers . . . assign a 'zero' score if the work product does not show evidence of any of the four levels of proficiency for the dimension in question" (McConnell and Rhodes 2017, 9). Can we attribute the zero score to a characteristic of the student, or is it instead a function of the assignment? Nikole and I examined a series of measurement models using deidentified data obtained from AAC&U in order to investigate how the raters were using the written communication, critical thinking, and quantitative literacy rubrics. The

results were amazingly clear. Simply put, the data did not "behave properly" when zeros were included in the analysis. Once we removed the zeros, the "picture" of the data aligned with what we would expect if the raters were using the rubrics properly. This story was similar for each of the three rubrics we analyzed.

To further illustrate this issue, consider an assignment in which a student was asked to write a hypothetical letter to an editor of a college newspaper about a social issue on campus. Assume that the student received a score of zero for the "sources and evidence" element of the written communication rubric. A zero may reflect that the student was asked to provide this information and failed to do so. In this case, the zero can be meaningfully applied to subsequent analyses since it reflects something about student proficiency. However, some assignments may not call for the student to evidence a specific rubric element, thus making the zero a characteristic of the assignment as opposed to the student. In this latter situation, the zeros should be deleted from subsequent analyses since they do not provide meaningful information about the student. However, in this latter case the zero score provides vital information about assignment misalignment. This information can then be used as a faculty development opportunity. For example, assignment development workshops can be held with interested faculty in which participants learn strategies for creating tasks that are capable of soliciting evidence of each rubric element (e.g., Crosson and Orcutt 2014).

In sum, there should obviously be alignment between criteria and tasks. Institutions using data from the VALUE rubrics should make sure that the zero scores are a function of the student and not the assignment. Scores that are a function of the assignment should be deleted before subsequent analyses, though this

can also serve as a potential faculty development opportunity. Lastly, just as the quality of food is influenced by the ingredients available in the kitchen, so too are student products influenced by characteristics of the task they are given. The best chef in the world cannot elicit the taste of a filet mignon from a chuck steak.

EVEN THE BEST CHEFS LOSE

Each winner of a cooking competition seems to do poorly on some of the challenges. In the ideal world, a world in which our judgments about a chef's proficiency was clear, we would have perfect consistency of performance across each of the challenges. If Chef A was the best in the first challenge then she would be consistently the best across all challenges. Unfortunately, this tidy picture is far from reality. Inconsistencies happen and are in many respects an inherent aspect of measurement. However, we hope that performances are not wildly inconsistent. Imagine the confusion of Chef Ramsay if he were faced with contestants who were perfectly inconsistent (i.e., the rank-order of chefs completely changed with each task). Perfect inconsistency would make it impossible to select the best chef. Thankfully, we do not inhabit a world of complete randomness.

The degree of inconsistency that is present is important because it reflects the amount of uncertainty that exists when generalizing from a series of limited performances to an overall metric of ability. Generalization, defined again as our capacity to draw overall conclusions about ability from a limited number of observations, tends to come at the cost of a competing value. Generalization in performance assessment is usually inversely related to directness (or what some people refer to as "authenticity"). To illustrate this point, assume we are solely interested in a person's ability to cook. We need to make decisions about the best way to solicit evidence of their ability as a cook. One option might be to administer a sixty-item test in which they were asked a series of questions about how to prepare and cook food. An alternative strategy would be to actually observe them preparing a meal, which would be then scored using some kind of rubric or checklist.

Every item on a test is treated as a mini-observation; thus, in this example, we have sixty observations with the test compared to a single observation with the alternative strategy. Generalizations are easier to make with more observations, which in this scenario favors the test. However, the test sacrifices directness since the evidence it provides is further removed from the ability or skill we are ultimately interested in measuring than the alternative strategy. Observing individuals cook a meal has the benefit of being more direct than the sixty-item test, but we end up sacrificing our ability to make generalizations since it is unlikely that I will have the time and resources to observe people cooking a variety of foods across multiple contexts.
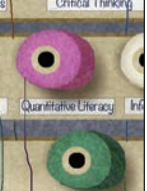
The topic of generalization has been notoriously problematic in performance assessment literature (e.g., Lane and Stone 2006). I may therefore have "good" data about the students' ability to cook a single dish, but I am unsure about how well they would cook other foods. This problem applies to outcomes that we tend to care about in higher education, such as written communication. We often observe a single performance (e.g., a paper) due to limited time and resources. However, even if we made multiple observations, there is a related problem referred to as task specificity. Our judgments about which students are doing better tends to change across multiple tasks designed to measure the same thing. I may think Students 1, 2, and 3 are my best chefs when asked to prepare a rib-eye steak, but I may come to very different conclusions if they prepare shellfish instead. Without actually making these observations, it is very difficult to determine if I would come to the same conclusions about student learning had other choices been made about what tasks to sample. In sum, alternative assessment strategies similar to those advanced by VALUE have the advantage of being more direct than many other strategies, though the directness tends to come at the expense of generalization. So how shall we proceed given these competing values?

CREATE YOUR SIGNATURE DISH

I have given much thought to these issues in the past few years. As far as I can tell, there are three possibilities for handling the problem of generalization in performance assessment. These possibilities include (1) increasing the number of observations, (2) restricting the domain of generalization, and (3) inferring what is "possible" instead of what is "typical." The first option is perhaps the most intuitive. We would never attempt to estimate a chef's overall cooking ability after they prepared a single appetizer, so why would we attempt to estimate a student's written communication skills from a single paper? If we sampled more papers, we would get a better sense of their written communication. But how many observations are necessary to obtain decent estimates? Previous research suggests that we need anywhere from ten to fifteen observations per student (e.g., Hathcoat and Penn 2012). This leads me to conclude that the first option is unrealistic.

A second possibility is to restrict the domain of generalization. Instead of asking about the number of observations that are needed for generalization, we would now also consider the type of observation that is needed. With respect to the chef example, we may decide to restrict our inference to something more

specific, such as an ability to prepare a particular cuisine like Thai, Greek, or Indian. Thus our "domain of generalization" has essentially become smaller by focusing on a specific style of food. Similarly, students may be able to write in one genre but not another (O'Neill and Murphy 2012). With this knowledge, we would be able to exert some controls during our sampling process by obtaining written work from a specific genre, thus making our inferences more limited but specific. This option is more feasible than the first, but it will require us to conduct additional research to better understand the boundaries of generalization since this will likely be different for each learning outcome.

The third option is not free of problems, but it is perhaps the most reasonable solution for the time being. When hiring a head chef, it is critical to get a sense of what is typical for them, which requires multiple observations. But do we need an estimate of what is typical with respect to our learning outcomes in higher education? Perhaps. But perhaps not. What do we learn about a chef when they are asked to create a signature dish? The signature dish does not provide information about what a chef will typically produce. Instead, the signature dish is an indication of what is possible by providing insight into what a chef is capable of creating. In higher education, senior capstone projects and specific forms of ePortfolios are analogous to a chef's signature dish since the product illustrates what a student can create. In other words, the issue of generalization virtually dissipates once we change our focus from what is typical to what is possible.

CONCLUSION

Assessment tasks in higher education are similar to the issues faced when attempting to distinguish a good chef from a not-so-good chef in a cooking competition. The product created by a chef is restricted by the tools and ingredients available in the kitchen. The products created by students are also influenced by assignment characteristics. Just as it would be unfair to critique a chef using criteria that fail to align with the task, so too should we avoid assigning numbers to students' products when they were not given an opportunity to demonstrate evidence of a particular rubric element. Lastly, when assessing student performance, there tends to be tension between the competing values of generalization and directness. There is not an easy solution to this issue, though I remain optimistic about our ability to confront this problem. If we wish to generalize, then additional research is needed to better understand the boundaries within which this is possible. However, we may also opt to investigate possibilities as opposed to generalities by sampling signature dishes. ■

REFERENCES

Berrett, Dan. 2015. "Faculty See Promise in Unified Way to Measure Student Learning." *Chronicle of Higher Education,* September 25, 2015. https://www.chronicle.com/article/Faculty-Members-See-Promise-in/233377.

Berrett, Dan. 2016. "The Next Great Hope for Measuring Student Learning." *Chronicle of Higher Education*, October 16, 2016. https://www.chronicle.com/article/The-Next-Great-Hope-for/238075.

Crosson, Pat, and Bonnie Orcutt. 2014. "A Massachusetts and Multi-State Approach to Statewide Assessment of Student Learning." *Change: The Magazine of Higher Learning* 46 (3): 24–33.

Hathcoat, John D., and Jeremy D. Penn. 2012. "Generalizability of Student Writing across Multiple Tasks: A Challenge for Authentic Assessment." *Research & Practice in Assessment* 7 (Winter 2012): 16–28.

McConnell, Kathryne Drezek, and Terrel L. Rhodes. 2017. *On Solid Ground: VALUE Report 2017.* Washington, DC: Association of American Colleges and Universities.

O'Neill, Peggy, and Sandra Murphy. 2012. "Postsecondary Writing Assessment." In *Handbook on Measurement, Assessment, and Evaluation in Higher Education*, edited by Charles Secolsky and D. Brian Denison, 405–422. New York: Routledge.

AAC&U MEETINGS

ANNUAL MEETING

Raising Our Voices:
Reclaiming the Narrative on the Value of Higher Education
January 23–26, 2019
Atlanta, Georgia

PRE-MEETING SYMPOSIUM ON THE VALUE RUBRICS

Is There a Rubric for That?
A Decade of VALUE and the Future of Higher Education
January 23, 2019

FORUM ON DIGITAL LEARNING AND EPORTFOLIOS

ePortfolios and the Value of Higher Education
January 26, 2019

NETWORK CONFERENCES

General Education and Assessment
February 14–16, 2019
San Francisco, California

Diversity, Equity, and Student Success
March 28–30, 2019
Pittsburgh, Pennsylvania

Global Engagement and Social Responsibility
October 17–19, 2019
San Antonio, Texas

Transforming STEM Higher Education
November 7–9, 2019
Chicago, Illinois

www.aacu.org/events

Making the VALUE Initiative Work for Us

▶ **Martha L. A. Stassen,** Associate Provost for Assessment and Educational Effectiveness, University of Massachusetts Amherst
Anne J. Herrington, Distinguished Professor of English Emerita and Faculty Fellow, Office of Academic Planning and Assessment, University of Massachusetts Amherst

Our participation in the Multi-State Collaborative (MSC), a collaboration led by the Association of American Colleges and Universities' VALUE (Valid Assessment of Learning in Undergraduate Education) initiative and the State Higher Education Executive Officers Association (SHEEO), was a natural progression of our efforts to enhance the collection and use of evidence to inform improvements to our undergraduate students' learning experiences. Indeed, the University of Massachusetts (UMass) Amherst's strategic plan, *Innovation and Impact*, calls for the university to promote a "culture of evidence" by demonstrating meaningful accountability, building institutional information resources, and embracing student learning outcomes assessment (University of Massachusetts Amherst 2013, 6). Given these institutional priorities, four aspects of the VALUE initiative were of interest to us: using actual student work from our courses, using rubrics developed by teams of faculty, asking faculty from our own and other institutions to score the work, and emphasizing the formative effects of assessment on learning while also working to develop a state-based and national reporting mechanism for student performance.

We joined VALUE to study the extent to which (and ways in which) the initiative could further our institutional priorities for building assessment through the increased involvement and expertise of university faculty. We pursued this goal, not only by conducting on-campus scoring of students' work in addition to submitting the work to the MSC for centralized scoring, but also through qualitative inquiry (e.g., focus groups, interviews, and surveys) with participating instructors. The focus of our inquiry was to understand the usefulness of the project to university faculty and learn how to improve our participation both in VALUE and in student learning assessment more generally on campus. Through the participation of faculty on our campus, we are advancing our campus-based assessment efforts and our overarching goal for assessment: to use valid and systematic evidence to foster reflection and inform action on student learning, pedagogy, and curriculum.

The focus of our inquiry was to understand the usefulness of the project to university faculty and learn how to improve our participation both in VALUE and in student learning assessment more generally on campus.

THE PROCESS

In preparation for participation in the MSC's 2015–16 cohort, we solicited student work through a broad call to faculty based on their perceptions as to whether they had an assignment that fit the five criteria of the Critical Thinking VALUE rubric: explanation of issues; evidence; influence of context and assumptions; the student's position (perspective, thesis/hypothesis); and conclusions and related outcomes (implications and consequences) (Association of American Colleges and Universities, n.d.). Our only other condition, based on requirements of the initiative, was that the work came from students who had completed at least 75 percent of the credits required for graduation.
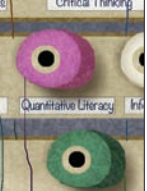
In addition to submitting work for national scoring, a team of UMass Amherst faculty scored the same work using the Critical Thinking VALUE Rubric to see how our scoring compared to that of the external national scoring and, equally important, to develop a cadre of faculty experienced with the process and able to help evaluate it. Scorers participated in a full-day norming session and then scored student work online, with additional feedback provided by the leader of the norming session in the early stages of scoring.

Once the scoring was completed and results were available, we held a follow-up meeting with all scorers. We organized this session as an informal focus group with specific prompts regarding their views on the overall process, the rubric as it defines critical thinking, the rubric's

Not only did scorers and instructors find the rubric helpful to their own thinking about what they meant by the term "critical thinking," but they could also see the value of the tool in helping to communicate their expectations to students.

fit with the student work assessed, and the potential of this approach for future campus assessments. In addition to this focused conversation with scorers, we also interviewed a set of instructors who had submitted student work for the assessment. In advance of each interview, we sent each instructor the rubric and both a high- and low-scoring paper from their course, asking them to score the work using the rubric. In the interview, we shared the external and internal scores to discuss how their own assessment fit with the external scoring and how they perceived the alignment of

the rubric with the assignment from their courses. We also discussed the critical thinking criteria more broadly and how those criteria fit with their own definition and their discipline's conception of critical thinking relevant to undergraduate student work.

Our conversations with scorers led us to examine the fit of the rubric to the student work submitted, as the scorers indicated a concern that the broad range of assignments represented might affect scoring in a manner that said less about what students can do than what the assignments asked for. Both a qualitative and quantitative analysis of the assignments and student work validated this concern. That is, we found that the assignments varied greatly in the kinds of critical thinking they called for, including some that were not well suited to the

rubric. Further, a statistical analysis showed significant correlations between average scores and both the length of the student work artifact and the number of external sources the artifact cited (University of Massachusetts Amherst 2017).

In our second year of MSC participation (2016–17), we followed many of the same steps for our formative evaluation. We also administered a survey to all scorers to collect their feedback, met with them again for an informal conversation based on the survey, and conducted interviews

with selected instructors who submitted work. The findings and observations shared here are drawn from these various sources.

THE VALUE OF PARTICIPATING IN THE PROCESS

Both years, faculty reported that they found participating in the process worthwhile, particularly for fostering reflection on their teaching and assessment more generally. Scorers commented on the value of both reading a wide range of student work and participating in discussions with colleagues about that work during the norming sessions. As one scorer commented, "I greatly enjoyed seeing work from other disciplines and hearing from faculty across the university." Others pointed to how the norming discussions and scoring prompted their thinking about what they value when they assess student work: "The experience of discussing assessment of critical thinking with faculty from a range of disciplines has been very useful. I've learned from hearing others describe what they look for in student writing and their rationale for assigning certain rubric scores."

Faculty also saw merit in a departmental approach. These faculty members said it could benefit their department to review and score student work for their own majors as a way to develop a shared understanding of both departmental expectations and their success in helping students achieve those expectations.

In fall 2017, one department used the Critical Thinking VALUE Rubric in an assessment of student work from their capstone course. Other departments are reviewing the Written Communication and Problem-Solving VALUE rubrics as possible tools for their assessment efforts; in at least one case, these efforts were catalyzed by efforts of faculty who served as scorers and submitters for the VALUE initiative.

IMPACT ON PEDAGOGY

Both the scorers and faculty who were interviewed talked about how participating in the assessment process and interviews prompted reflection on their own methods for fostering critical thinking. For example, "This experience has helped me think about what students can produce and what they need to help them produce a solid paper." Others commented that it helped sharpen their own conception of critical thinking. As one faculty member said in an interview, "We all talk about critical thinking, but neither our students nor we have a real definition of it." A scorer indicated in a survey response, "This process has given me a language for defining critical thinking and has helped me to separate the evaluation of critical thinking from the evaluation of writing."

Not only did scorers and instructors find the rubric helpful to their own thinking about what they meant by the term "critical thinking," but they could also see the value of the tool in helping to communicate their expectations to students. As a survey response, one scorer wrote, "The rubric is also a useful starting point for faculty to communicate with students about their critical thinking, particularly in gen ed classes and writing classes." In fact, following each year of participation in the MSC, a few scorers and instructors talked about using the rubric or a revised version of it in their courses, both for explaining what is entailed in critical thinking and for evaluating student work. One saw it as a tool for designing assignments, commenting, "I would use the rubric to break down components of critical thinking that could then be the focus across different assignments."

REVISIONS TO THE ASSESSMENT PROCESS AND RUBRIC

Based on what we learned after the first year of participating in the MSC, we decided to develop more precise guide-lines, in addition to those outlined by the national project, for the artifacts we would submit in the second year. To that end, when we invited faculty to participate, we specified the following criteria for student work:

1. The work is from an advanced course within a student's major.
2. Papers are at least eight pages, preferably no more than twenty pages.
3. It should be a final major paper for the course, preferably one where students have had the opportunity to revise prior to final submission.
4. The work should use primary or secondary sources.
5. It should be appropriate for assessment using the criteria identified in the rubric.

Feedback from both scorers and faculty who submitted student work pointed to the need to revise the VALUE rubric both to clarify certain aspects and to better align with our local values and student learning goals. For example, faculty were troubled by the stipulation that a high score for the "evidence" criterion required that "viewpoints of experts are questioned thoroughly," as it seems to imply such questioning is always appropriate. As one instructor commented, "I don't want [students] to challenge expertise when it's not called for. I want them to think for themselves." For this reason, we eliminated this stipulation, feeling that evaluating sources and considering others' points of view—as included elsewhere in the rubric—sufficiently covered the intention of this descriptor.

A major revision we made was to delete the criterion "conclusions and related outcomes" because scorers reported having difficulty distinguishing it from "student's position" in some student work, and because it seemed tailored 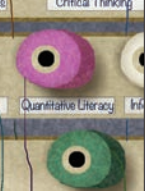more to some genres than others. On the other hand, scorers were concerned that overall logical coherence was not addressed as an important aspect of critical thinking. As one scorer said, "There is something holistic missing about the coherence of the whole piece, . . . the logical train of thought of the whole." For this reason, we added "logical coherence" as a criterion.

We used our revised rubric for our on-campus scoring in 2016–17, which means, of course, we cannot easily compare our faculty scoring with the national scores beyond the first year of our participation. However, we felt it was more important to be responsive to faculty feedback and make changes that will fit our context better than to ask faculty to use a rubric they found difficult to manage.

The scorer survey responses indicate that our revisions were well received. In response to a question about how well the revised rubric worked for scoring student work, on a five-point scale from "not well at all" to "very well," all respondents said it worked "fairly" or "very well." One added, "The rubric worked well with most papers—much better than last year!" Asked about the effectiveness of "the assessment process overall (calibration session, online system, timeline, clarity of purpose, etc.)," the scorers were even more positive, with two-thirds judging it to be "very effective" and the other third "effective." One termed the calibration sessions "extremely helpful" and another noted that they were "more helpful than last year." These responses underscored for us the value of the input we received from scorers and instructors for the revisions that we made.

LOOKING FORWARD

Our formative evaluation of our first two years of participation in the VALUE initiative demonstrated clear benefits to the individual faculty members who participated. The opportunity to review student work from across disciplines and engage in focused conversations about critical

thinking with each other has helped inform their own teaching and communication with students.

Our survey results offer further insights into the potential value of the assessment process. We asked the scorers to indicate the extent to which they felt the assessment process they participated in could be useful at the university, school/college, and department levels. They could see the value of university-level assessment, with six of the twelve scorers indicating it had "great potential" for university-wide assessment. They were even more inclined to see its value at the department level, with nine of the twelve respondents indicating that it had "great potential" for department-based assessment. One survey respondent wrote:

I learned from working with colleagues and assessment. . . . It did reinforce the belief that this is an important area to continue improving. . . . It would be outstanding to have this happen department-wide, because faculty could learn from each other and be more consistent with elements.

What is particularly promising is the extent to which the process has potential for both cross-disciplinary, university-wide assessment efforts and the more focused departmental assessment needs. The departmental efforts are supported and reinforced by the campus's enhanced program-based assessment plan (the Educational Effectiveness Plan), which streamlines and regularizes departmental planning, budgeting, and assessment into one coordinated process for improving the undergraduate experience (University of Massachusetts Amherst, 2018).

With two years under our belt, in 2017–18, we created a hybrid rubric that includes key aspects of both critical thinking and written communication and used it with good results. We also collaborated with the University Writing Program and included a selection of writing by first year students to expand our understanding

of our students' skills at two key points and to test the applicability of the rubric beyond upper level capstone work. As we enter our fourth year of participation in VALUE Institute assessments, we continue to fine-tune both the rubric and process. Still, we have increased confidence in using the VALUE approach to inform conversations with faculty about undergraduate student performance and what our results might suggest for changes at the university, department, and course levels. In addition, our campus-developed rubric is emerging as a useful tool for departments to use in their own assessment efforts. We have also been able to build a cadre of faculty with assessment experience who can work with the Office of Academic Planning and Assessment and their own departments and colleagues to build thoughtful assessment approaches that augment the evidence-based inquiry they are already conducting.

Our participation in the VALUE initiative reinforces the student learning assessment outcomes we want to communicate to faculty, students, and the administration. These coincide with the goals of the VALUE initiative: using student work, encouraging faculty participation in determining what criteria will be used for assessment and how the assessment will be conducted, and using the faculty conversations and assessment results to inform and improve student learning and the student experience on campus.

In an interview with a faculty member who submitted student work, we asked how the VALUE assessment process compared with the assessment work her department conducts for their external disciplinary accreditation. She said, "The [professional accreditation] assessment process is not very interesting. This seems more interesting."

Assessment activities that faculty view as engaging and interesting and promote self-reflection on teaching have great value to any campus. The faculty on our campus who have invested their time and effort into the process have experienced individual benefits to their teaching and see the potential for realizing the university's larger goal of enhanced student learning assessment. We look forward to what our future involvement might hold. ▪

REFERENCES

Association of American Colleges and Universities. n.d. "VALUE Rubrics." www.aacu.org/value-rubrics.

University of Massachusetts Amherst. 2013. *Innovation and Impact: Renewing the Promise of the Public Research University.* Amherst, MA: University of Massachusetts Amherst Joint Task Force on Strategic Oversight. www.umass.edu/chancellor/sites/default/files/pdf/jtfso-phase-i-report2.pdf.

University of Massachusetts Amherst. 2017. *Report on UMass Amherst's Participation in the Pilot Valid Assessment of Learning in Undergraduate Education (VALUE) Initiative, 2015–2016.* Amherst, MA: Office of Academic Planning and Assessment. http://www.umass.edu/oapa/sites/default/files/pdf/learning_outcomes/final_umass_amherst_value_initiative_participation_report_2016.pdf

University of Massachusetts Amherst. 2018. "Educational Effectiveness Plan." www.umass.edu/oapa/program-assessment/academic-department-assessment/educational-effectiveness-plan-eep.pdf.
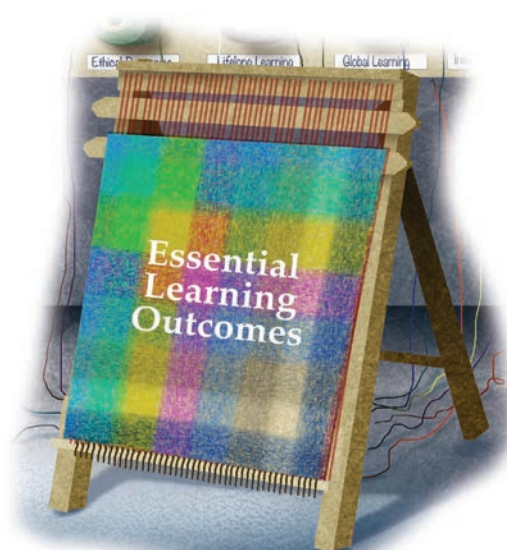
Lessons Learned from a Decade of Authentic Assessment

▶ **Kimberly Filer,** Assistant Provost for Teaching and Learning, Virginia Tech; Former Director of Institutional Effectiveness and Assessment, Roanoke College
Gail Steehler, Associate Dean for Academic Affairs and General Education, Roanoke College

In fall 2009, a reimagined faculty-developed general education curriculum was introduced at Roanoke College. The Intellectual Inquiry (INQ) curriculum was the result of five years of faculty learning, planning, and negotiating to produce a cohesive, common-core learning opportunity for students. Through the critical inquiry curriculum, students develop cross-cutting skills while being immersed in engaging themes and topics that challenge them to frame, ask, and seek answers for complex questions.

The goal of the curriculum is to produce graduates capable of addressing real world problems in diverse and collaborative settings. The capstone experience, INQ 300 Contemporary Issues, requires cross-disciplinary groups of students to problematize a contemporary issue to a local setting, apply their skills and knowledge to develop a proposal to address that problem, and then orally defend their proposal. Students synthesize skills and knowledge from their majors with those developed in earlier Intellectual Inquiry courses. All earlier INQ courses focus on skill development. The curriculum's entry point is INQ 110, a rigorous writing seminar taught by faculty across the college, in which students must apply careful reading skills and writing skills to explore topics like Formation of the Western Mind and Black Lives Matter. In the second seminar, INQ 120, students apply oral communication skills and ethical reasoning in courses such as Matters of Life and Death and Thinking Animals. A series of 200-level courses provide foundations in quantitative reasoning as well as the questions and methods of the natural sciences, social sciences, and humanities while continuing to develop students' communication and reasoning skills. Students learn to integrate knowledge and skills from many disciplines to address complex problems both in their education and in their lives after graduation.

The design, development, and implementation of a redesigned curriculum in which each course has a unique topic was a massive undertaking requiring hundreds of new courses to be

Students learn to integrate knowledge and skills from many disciplines to address complex problems both in their education and in their lives after graduation.

developed and moved through faculty governance. By the end of the three-year curriculum rollout, the number of sections designed, reviewed, and approved was equal to the number of full-time faculty. The course design and development effort continued as the curriculum was launched, while faculty made intentional efforts to maintain the quality of the previous curriculum for existing students. Faculty were busy juggling their many academic responsibilities while heavily engaged in INQ
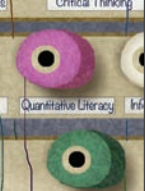
tasks. To move forward with INQ implementation while planning and piloting general education assessment, faculty adopted some guiding principles to facilitate decision making: (1) the assessment process would map onto the approved curriculum (e.g., assessment would not require curricular change); (2) faculty would take lead roles in the design, pilot, and implementation of assessment tools; (3) assessment would not reduce the curriculum to "teaching to the test"; and (4) the assessment effort would be cost-effective.

A NATURAL FIT

The Roanoke College faculty spent considerable effort specifying how cross-cutting skills would be developed in the INQ curriculum. Although not framed in the language of learning outcomes assessment, faculty spent time considering the types of assignments expected in courses and how students would engage in critical inquiry while applying their skills. The adoption of general education assessment was aided by faculty members' focus on common expectations and assignment types instead of wordsmithing learning outcomes. For example, in INQ 110, faculty committees specified how much and at what level they wanted students to write. Using this information, the assessment director offered language for a simple learning outcome to act as an organizational guidepost for the more detailed work of the faculty. Rather than focus on the language used to write about outcomes assessment, faculty focused on backwards course design, assignment development, and criteria for assessing students' writing quality.

In early conversations, faculty chose to focus assessment efforts on communication skills, reasoning skills, and the foundational methods and skills of disciplines (across the natural sciences, social sciences, and humanities) rather than more specific disciplinary knowledge connected to the course topics. For example, students were assessed on their ability to appropriately apply disciplinary theory to real-world problems, allowing flexibility in the specification of content and teaching methods. The choice to assess skills over content allowed faculty to engage in curriculum mapping and subsequent assessment mapping with the approved curriculum.
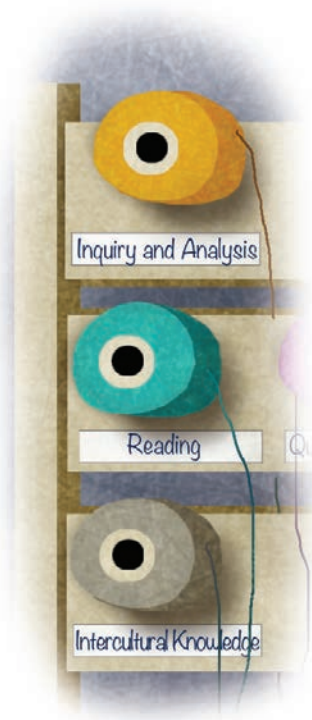
FACULTY-LED ASSESSMENT

In the first semester of curriculum implementation, a subset of INQ 110 instructors gathered in a breakfast meeting to design a rubric to pilot with their writing assignments. However, they had limited expertise with rubric development. The committee decided to begin with the Written Communication VALUE Rubric, which emerged from the Liberal Education and America's Promise (LEAP) initiative of the Association of American Colleges and Universities (AAC&U). While the VALUE rubric at that time was still in draft form, it served as a helpful template and starting point for committee deliberation. Following a suggestion from the AAC&U leadership and staff, the Roanoke faculty customized the rubric to better fit the INQ 110 design by changing developmental levels, reorganizing columns, and rewording criteria. The faculty applied the rubric to students' writing products from the last third of the course, aggregated their responses, and reconvened to discuss pilot data and suggest rubric changes. Overall, student writing data confirmed faculty members' professional opinions about student writing skills (face validity), and it seemed the data would be helpful to identify collective strengths and weaknesses in student written communication skills. However, healthy skepticism remained, and faculty questioned the likelihood of assessment data being used to inform substantive changes in instruction either at the program or course level.

AVOIDING A REDUCTIONIST APPROACH

In addition to concerns about effective data use, faculty did not want their curriculum reform efforts to be reduced to "teaching to the test." Faculty engaged in the pilot group felt the assessment exercise was useful, but what would happen when it was rolled out to all of campus? Would faculty develop "throwaway" assignments to turn in rubric scores? Would faculty be forced to use the same assignment across sections even if it wasn't relevant to the course topic? Would assessment become an exercise in collecting data to create unused reports? Would student learning be reduced to a single percentage score?

Preventing all potential problems that could stem from reformed general education assessment is impossible, but the faculty took bold steps to proactively

address concerns: (1) all faculty teaching an INQ course would apply the rubric to an existing assignment in the last third of their course; (2) for the identified assignment, work from all students would be scored using the rubric; (3) scores were aggregated at the criterion level and were not averaged across the skills; and (4) the associate dean for general education managed the data and was responsible for reporting aggregate scores to the teaching faculty.

ASSESSMENT ON A BUDGET

In the second year of INQ implementation, the written communication rubric and the oral communication rubric were used in all INQ 110 and INQ 120 classes. And while early assessment efforts in the first-year seminars moved forward relatively well, implementation decisions relating to next steps of the assessment process led to questions about unanticipated costs: How can rubric norming sessions be conducted for *all* faculty teaching INQ courses? What would it cost to pay summer faculty to be second scorers on student work? Could the college afford the costs of traditional reliability and validity efforts? The INQ curriculum was a costly endeavor for a small college and the newly formed assessment office had an annual operating budget under two thousand dollars. With these budget constraints, an unconventional solution was necessary.

THE REVIEW PROCESS

Before planning, organizing, and convening groups of faculty members to act as second evaluators to establish inter-rater reliability, the assessment office conducted a statistical analysis to see how many of our faculty scored student papers significantly higher or lower than their colleagues. Using the four-point VALUE rubric structure, the office averaged each INQ section's scores according to the five criteria (content, organization, mechanics, etc.),

resulting in five average scores per section. The director of institutional effectiveness and assessment assigned each section a random number to protect faculty identities and conducted a multivariate analysis of variance (MANOVA) to explore the magnitude of instructor effects on scores. Although far from a highly controlled research design for using a MANOVA, the results illuminated some interesting findings when considering the instructors as the independent variables and the average criteria scores as dependent variables. Overall, most faculty scored their first-year students similarly on specific criteria of the writing rubric.

For all instructors, the "use of evidence" criterion was significantly lower than the other writing criteria. Additionally, there were six instructors with average scores that were at least two standard deviations higher or lower than their colleagues. The assessment office reviewed the raw data of these six instructors to see how rubrics were used. In the cases of the four "easy graders," nearly all of their first-year students' writing projects were scored as a three or a four in every area. The two "hard graders" scored nearly all writing products as a 1 (below basic) in every area. After reflecting on the results, we determined that instructor variance in using the rubrics was not as large as we had anticipated, and plans for a large, expansive summer scoring effort were modified and a smaller group of twelve-month faculty members scored fewer samples of student work and discussed any discrepancies.

FOCUSING ON COLLECTED DATA

After reviewing the statistical analysis and the results of the summer scoring effort, the faculty committee decided to develop clear directions for rubric scoring rather than providing norming sessions for all faculty. In rubric use instruction sessions, held before the start

of classes in August, the associate dean also shared assessment data. Rather than concentrate on a perfect assessment process, the associate dean and lead faculty instructed INQ faculty on how to use data in concert with their professional judgment about student communication skills and discussed how to help students improve their performance on these important outcomes.

As INQ implementation continued, the faculty used AAC&U's VALUE rubrics as a framework to develop new rubrics to address nearly all the skills and reasoning outcomes of the curriculum. Additionally, other assessment tools were designed and tested by faculty to determine how to effectively capture information about student learning in ways that would lead to productive conversations and continuous renewal of a high-quality general education program.

LESSONS LEARNED

Roanoke College administrators and faculty learned a great deal about the assessment of student learning in the first ten years of the INQ curriculum. As we reflect on the past decade, it is evident there are opportunity costs to nearly every decision related to the assessment process. Limited by constraints on faculty time and financial resources, college faculty had to make decisions about the tasks they felt were most important for obtaining and using student learning data. Giving up some assurances of inter-rater reliability and scoring precision allowed assessment efforts to focus on faculty ownership of the process and use of the data. We chose to use assessment for discourse, affirmation, and proof of a rigorous curriculum rather than to create a perfect assessment engine. And, although there is always room for improvement, it has given continued attention to the quality of learning in the general education curriculum. ▪
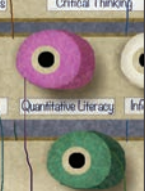
The Case for Civic Learning in the Humanities at Community Colleges

▶ **Eric Vanover,** Assistant Professor of History, Germanna Community College

The principles that led to the establishment of community colleges in the American higher education system were based on the ideologies of democracy and steeped in the early republican creed of equal opportunity. President Truman's Commission on Higher Education report in 1947 conceived the idea of community colleges as a system for providing more access to educational opportunities in local communities across the country and as a principal method for fostering a stronger nation of prosperous and engaged citizens (Truman Commission 1947). The influence of these founding principles has often resulted in their endearing designation as democracy's colleges. Given these foundational ties to democracy and citizenship, it makes sense that community colleges should serve as leaders in civic learning and assessment. Community colleges and the influence they hold in local, regional, and even national communities are in many ways best situated for the work of democracy.

CHALLENGES TO THE CIVIC ROLE OF COMMUNITY COLLEGES

The charge of serving as civic learning and assessment leaders is not without challenges. A quick look into the recent plight of community colleges will hastily reveal that they face difficult and often unique operational challenges. Examples include maintaining sustainable funding while continuing to adhere to the mission of affordability or maintaining open enrollment in an environment keen on performance-based funding. Then there is the expectation by state legislatures and public stakeholders to develop a comprehensive but flexible curriculum to meet rapidly changing local and regional workforce needs. Administrators and faculty at community colleges feel the consequences of constantly changing public demands. Efforts to address these challenges often result in the rapid initiative implementation approach to problem solving: solving a problem by doing more with less. This tactic is, in some part, a direct result of the foundational mission of the community college to be the *best* at meeting rapidly changing student, workforce, and public needs in the communities they serve. The consequence, however, is sometimes the echoing moans from faculty and administrators using the catchphrase "initiative fatigue." Prioritization is a difficult task for institutions dedicated to the democratizing notion of being all things to all people.

Just as recent national Gallup polls suggest low public confidence in democratic and political institutions like Congress and the Supreme Court, so too have the purpose and benefits of higher education come under scrutiny (Gallup 2018). From the workforce development perspective at community colleges, the trend toward a concentrated focus on the skills-to-employment mindset has in some cases meant inattention to the civic role these colleges also fulfill. If America's economic health may partly rest on how well community colleges embrace the skills-to-jobs mentality for the workforce, what responsibility for America's democratic health do community colleges still hold for the citizenry?

Perhaps no other traditional area of study has encountered a tougher existential challenge than the humanities. One does not have to browse long to find almost daily published examples of articles discussing the inadequacy of liberal arts training or the irrelevance of humanities majors like history, English literature, philosophy, or religion. Responses to these criticisms are generally defensive and often justify their existence with rhetoric expressing the intrinsic value of the humanities.

The Value of the Humanities

The humanities have traditionally held a peculiar place at the community college. Balancing the community college's dual functions of academic transfer and vocational training has led to intense discussions about what constitutes general education for all students. Courses in the humanities are most commonly visible in the general education curriculum at community colleges. One or more of these courses are found in most associate degree programs and in many career certificate programs. Courses in English, history, and other humanities disciplines are found in program requirements because they are grounded either in the college's philosophy of general education or in skill sets that employers seek, like effective communication and critical thinking.

I am a firm believer in the intrinsic value of the humanities, but I also firmly believe they serve as excellent vehicles for student skill development. Higher education and the humanities alike, especially at community colleges, now face a common challenge. As David Mathews puts it, "The most fundamental challenge that institutions of higher education face is to reestablish their public mandate" (2016, 39). The humanities have a place more relevant than ever in developing students' desirable workplace skills and civic responsibility. At Germanna Community College (GCC) in Locust Grove, Virginia, disciplines like history and other humanities are developing a new sense of importance by embracing a skills-based approach to general education core competencies like civic engagement and problem solving. The VALUE rubrics developed by the Association of American Colleges and Universities have been influential in the evolutionary process of teaching and assessing these skills with our students. While we have gradually embraced competencies like civic learning and problem solving to dif-ferent degrees over the past several years, they are becoming a central aspect of how our faculty conceptualize teaching history and other humanities courses.

DEVELOPING GENERAL EDUCATION ASSESSMENT INSTRUMENTS

In 2015, I became the cochair of a department that included the disciplines of history, humanities, religion, philosophy, and music at GCC. Our first major task that year was to develop general education assessment instruments for history and humanities courses that embraced our college's general education competencies of critical thinking and social and cultural understanding. A subcommittee of both full-time and part-time faculty from across the disciplines in the department selected the theme of citizenship for our assessments. Instructors committed to framing significant portions of the content they covered through the lens of citizenship and to continually provide students with opportunities to practice analyzing and reflecting on primary sources focused on citizenship.

In six of our history survey courses, we selected a primary source that the subcommittee felt best addressed ideas of citizenship in the context of each respective course. For example, in our US History I course, we chose an excerpt from Chief Justice Roger Taney's majority opinion concerning the citizenship status of the former slave Dred Scott. After much debate and deliberation, we agreed on three civic-minded questions, tied to corresponding general education outcomes, that could be addressed from the perspective of each document. Our emphasis was on contextualization of the sources. The assessment questions were targeted mainly at the Civic Engagement VALUE Rubric's milestone and benchmark levels and to account for the fact that our focus was on understanding the past to be better informed in the present. We learned that it was of dire importance that history instructors develop a common and clear language related to citizenship. For example, students struggled to grasp the phrase "social institution" consistently across different sections of the course, and this often led to vague responses to some aspects of the questions. We identified developing a common language among instructors as an area to focus our improvement efforts in the future.

The humanities have a place more relevant than ever in developing students' desirable workplace skills and civic responsibility.

In our Humanities 100: Survey of the Humanities course, a similarly organized subcommittee from several disciplines in the department decided to focus on political and social interpretations of art to assess the competency of social and cultural understanding. The assessment would focus on the learning unit on modernism, with *Guernica* by Pablo Picasso as the primary source and two brief summary paragraphs written by the humanities faculty to provide some historical contextualization of the artwork. As part of the final exam, students were given these materials and asked in a writing prompt to reflect on how understanding the context of the artwork affected their perceptions of its cultural
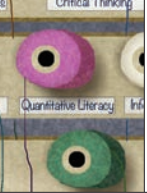
influence. They were also asked to reflect on if the political message they interpreted from the piece had any relevance to more recent social or political events. Responses to the prompt ranged from themes such as human sex trafficking to global terrorism or the conflict in Syria. Again, our assessment targeted milestones and benchmark levels for how we interpreted civic learning relevant to this course, but it was clear that our students were responding to the civic learning ideals we were trying to incorporate in these courses.

Shortly after our first assessment cycles, GCC adopted problem solving as the focus of its Quality Enhancement Plan (QEP). The QEP is an integral part of reaffirmation of accreditation for institutions in the Southern Association of Colleges-Commission on Colleges and is meant to be a project that a college or university would undertake to transform student learning or the student learning environment. After an extensive internal and external review of how we could better prepare students for the workforce, the college decided to infuse problem-based learning into our classroom instruction across all courses over the next five years. We turned to the Problem Solving VALUE Rubric for guidance in assessing student learning for problem solving.

Two survey courses in American history were selected as part of the first wave of courses to incorporate problem-based learning. As a department, we decided to merge our civic learning initiative with the new college-wide problem-solving initiative and developed a collaborative project that students would complete over the course of the semester. We presented each class with an ill-defined problem in a real-world scenario:

> The board of directors at the fictional Spotswood Museum has tasked your class with designing new gallery exhibits, selecting relevant artifacts, and organizing the museum in a way

that tells a story about America. The galleries should be reflective of five major themes, or historical questions, the class has selected. Each group will present their gallery as part of a final exam and each individual student must write a two-page reflection addressing their own learning experience in each of the Problem Solving VALUE Rubric learning outcomes.

For our history courses, the reflection prompt questions are civic in nature and are derived from both the Problem Solving and Civic Engagement VALUE rubrics. Some examples, with assignment instructions that align with specific rubric goals in parentheses, include:

- Do you think these were the best overall themes to represent American history? (Identify strategies/diversity of communities and cultures.)
- Why did you choose the objects to interpret this space the way you did? Were there alternative objects that you declined to use? (Propose multiple solutions/analysis of knowledge/civic communication.)
- Do you think the gallery you constructed represents the American experience? How do you think your gallery will be received by the public? (Evaluate the impact of the selected solution/civic action and reflection.)

Our first student responses were ready for us to score using the rubrics in fall 2018. These assignments were scored once using the Problem Solving VALUE Rubric for college-wide assessment and then again for the history content and civic learning outcomes we apply as individual history instructors. From these three projects, we have found the VALUE rubrics to be exceptional planning guides, flexible frameworks for meeting our assessment scoring needs, and compatible in merging assessments for more than one core competency in a single assignment.

WICKED PROBLEMS

Including civic learning as a strategy has often placed my students in the path of *wicked problems*. Wicked problems are complex, have numerous causes, and rarely have a single solution. As David Mathews suggests, people "don't agree on what the problems are, much less what should be done about them. . . . The disputes aren't over questions of fact but over what is the *right* thing to do" (2016, 34). As a history instructor, discussing wicked problems in today's political climate may perhaps be the most important way I can foster civic learning.

During the spring 2017 semester, a national controversy re-emerged concerning the place of Confederate monuments in the public sphere. In Virginia, the issue is a very local and personal one. At our Fredericksburg campus, we are no more than fifteen miles away from four major civil war battlefields. Controversy has erupted time and time again concerning the presence of a slave auction block monument in the center of downtown. Our proximity to Richmond and Charlottesville means that national news is really local news when it comes to controversies concerning Confederate icons in public spaces. For us, the issue has become a wicked problem.

During the 2016–17 academic year, classroom conversations in nearly all of my courses commonly returned to the debate concerning public displays of Confederate icons. By the spring, I began to notice that students were weary of the topic and more prone to irritable outbursts. I was shocked when the topic emerged in conversation again late in the spring and a student frustratingly lamented that none of their opinions mattered because the conversation belonged to the voices of social media extremes. One does not need to spend long scrolling social media feeds to realize that perhaps she

was right. If, as Theis (2016, 46) said, "democracy . . . is a mechanism for decision making among people who have a shared existence in space and time," then it was discouraging that my students felt that their opinions in *this space* and *this time* did not matter. I determined then that students learning to be engaged citizens needed a platform to have open discussions about this wicked problem. In my upcoming US History and Film course, I decided we would attempt to tackle the issue through an experiential learning process of making an amateur documentary. Students would become authoritative through their research and collaborative in their production of the film.

With the generous permission of the Fredericksburg and Spotsylvania National Military Park, a group of thirty enthusiastic students spent ten weeks in the summer researching local civil war history, monuments, and memorials and produced roughly ninety minutes of footage. Students researched and collaborated on a series of assignments like visits to the four major battlefields in the area. They did research on how Northerners and Southerners dealt with the Fourth of July in the press during the war, an assignment inspired by the Mapping the Fourth project conducted by the Virginia Center for Civil War Studies at Virginia Tech.

In small teams organized by common interests, each group wrote a short script that focused on the events and monuments that had caught their attention the most during our visits to the battlefields and other sites and that discussed how those experiences informed their opinions on the national debate about the monuments. With the instrumental help of a student intern with a passion for filmmaking, we met at various student-selected points on the battlefield sites so that each group could get the experience of filming on location and all of the challenges that came with that process. For the final exam, students were asked to reflect on and respond on camera to a series of three poignant questions about their experience and the importance of the civil war in American history. In the final week of the course, we edited the footage down to a thirty-minute film complete with each student's reflection.

We held a private showing for the students and their families and friends but saved the public viewing for GCC's History Day event that fall. At the time of the public viewing, the terrible events of the Charlottesville "Unite the Right" rally in August were still fresh in the minds of the viewing audience, so the film held even more national relevance than when we began the project. Public support for the film was overwhelming and several of the students were able to attend to discuss their experience. While some in the audience did not necessarily agree with all of the students' interpretations and opinions, there was an honest sense of mutual respect in their questions and comments. Most importantly, students were confident in their conclusions and no longer felt stifled in the public sphere. They had an authoritative and collaborative voice in *this space* and *this time*.

The Civic Engagement VALUE Rubric was instrumental in grading this experiential and reflective learning assignment. Utilizing concepts from this rubric provided better structure and more consistency in how I assessed student achievement in civic learning. The civic-learning outcomes were measured across multiple assignments that led to the final documentary content, so I was able to associate specific outcomes for certain assignments. There was a valuable lesson to be learned here as well—remember who your students are!

CONCLUSION

While I had aspirations of individual students achieving capstone-level success in these learning outcomes, it was important to remember that my first- and second-year students at the community college were enrolled mainly in introductory-level survey courses. Civic learning was a new concept for many of them, and they were still honing their other foundational skills like critical thinking and communication. As they reached benchmark and milestone levels of success, I revised my assignments to enable all my students to gain comprehensive civic learning. Understanding this greatly improved my teaching as well. In cases where students scored at the capstone level for the individual reflections, I saw this as a prime indicator for success in their future educational endeavors in the workplace and as productive citizens capable of tackling wicked problems. And what they accomplished as a group was, in my opinion, a capstone level of success for each of them.

The work we have done with civic engagement and problem solving in our department at GCC is far from finished. But the foundations we have established in embracing civic learning in our history and humanities courses and the guidance of the VALUE rubrics have helped prepare us for the greater work ahead. ▪

REFERENCES

Gallup. 2018. "Confidence in Institutions." Retrieved from https://news.gallup.com/poll/1597/confidence-institutions.aspx.

Mathews, David. 2016. "The Community College's Role in Helping to Make Democracy Work as It Should." *New Directions for Community College*s 173: 33–40.

Theis, John J. 2016. "Political Science, Civic Engagement, and the Wicked Problems of Democracy." *New Directions for Community Colleges* 173: 41–49.

Truman Commission on Higher Education. 1947. *Higher Education for Democracy: A Report of the President's Commission on Higher Education* (Vol. 1–2). New York: Harper.
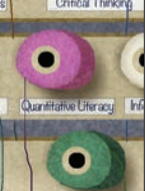
The Dependability of VALUE Scores:
Lessons Learned and Future Directions

▶ **Gary R. Pike,** Professor, Higher Education and Student Affairs, Indiana University
Kathryne Drezek McConnell, Assistant Vice President for Research and Assessment, Office of Quality, Curriculum, and Assessment, AAC&U

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.

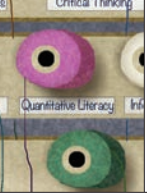Extending the landscape metaphor to the VALUE work, we are reminded

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

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.

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. ▪

REFERENCES

Brennan, Robert L. 2006. "Perspectives on the Evolution and Future of Education Measurement." In *Educational Measurement*, 4th ed., edited by Robert L. Brennan. Westport, CT: American Council on Education and Praeger.

Cronbach, Lee J., Goldine C. Gleser, Harinder Nanda, Nageswari Rajaratnam. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: Wiley.

Erwin, T. Dari. 1988. *The Analysis of Ratings Using Generalizability Theory for Student Outcomes Assessment.* Paper presented at the annual meeting of the Association for Institutional Research, Phoenix, AZ.

Finley, Ashley. 2012. "How Reliable Are the VALUE Rubrics?" *Peer Review* 13/14 (4/1): 31–34.

Haertel, Edward H. 2006. "Reliability." In *Educational Measurement,* 4th ed., edited by Robert L. Brennan, 65–110. Westport, CT: American Council on Education and Praeger.

McConnell, Kathryne Drezek, Erin M. Horan, Bethany Zimmerman, and Terrel L. Rhodes. 2019. *We Have a Rubric for That: The VALUE Approach to Assessment.* Washington, DC: Association of American Colleges and Universities.

McConnell, Kathryne Drezek, and Terrel L. Rhodes. 2017. *On Solid Ground: Value Report 2017.* Washington, DC: Association of American Colleges and Universities.

Pike, Gary R. 1995. "The Relationship between Self Reports of College Experiences and Achievement Test Scores." *Research in Higher Education* 36: 1–21.

————. 2018. "Improving the Dependability of Constructed-Response Assessment: Lessons from an Evaluation of the VALUE Rubrics." *Assessment Update* 30 (5): 8–9.

Pusecker, Kathleen Langan, Manuel Roberto Torres, Iain Crawford, Delphis Levia, Donald Lehman, and Gordana Copic. 2011. "Increasing the Validity of Outcomes Assessment." *Peer Review* 13/14 (4/1): 27–30.

Rhodes, Terrel L., and Ashley Finley. 2013. *Using the VALUE Rubrics for Improvement of Learning and Authentic Assessment.* Washington, DC: Association of American Colleges and Universities.

Rhodes, Terrel L. 2010. "Valid Assessment of Learning in Undergraduate Education." In "Rising to the Challenge: Meaningful Assessment of Student Learning." Washington, DC: Association of American Colleges and Universities, American Association of State Colleges and Universities, and Association for Public and Land-grant Universities..

————.2012a. "Emerging Evidence for Positive Change." *Peer Review* 13/14 (4/1): 4–5.

————. 2012b. "Show Me the Learning: Value, Accreditation, and the Quality of the Degree." *Planning for Higher Education* 40 (3): 36–42.

Webb, Noreen M., Glenn L. Rowley, and Richard J. Shavelson. 1988. "Using Generalizability Theory in Counseling and Development." *Measurement and Evaluation in Counseling and Development,* 21 (2): 81–90.
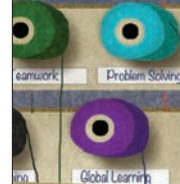
Critical Thinking VALUE Rubric

The VALUE rubrics were developed by teams of faculty experts representing colleges and universities across the United States through a process that examined many existing campus rubrics and related documents for each learning outcome and incorporated additional feedback from faculty. The rubrics articulate fundamental criteria for each learning outcome, with performance descriptors demonstrating progressively more sophisticated levels of attainment. The rubrics are intended for institutional-level use in evaluating and discussing student learning, not for grading. The core expectations articulated in all 16 of the VALUE rubrics can and should be translated into the language of individual campuses, disciplines, and even courses. The utility of the VALUE rubrics is to position learning at all undergraduate levels within a basic framework of expectations such that evidence of learning can by shared nationally through a common dialog and understanding of student success.

CAPSTONE
4
EXPLANATION OF ISSUES   *Issue/problem to be considered critically is stated clearly and described comprehensively, delivering all relevant information necessary for full understanding.*
EVIDENCE   *Selecting and using information to investigate a point of view or conclusion*   Information is taken from source(s) with enough interpretation/evaluation to develop a comprehensive analysis or synthesis.   Viewpoints of experts are questioned thoroughly.
INFLUENCE OF CONTEXT AND ASSUMPTIONS   Thoroughly (systematically and methodically) analyzes own and others' assumptions and carefully evaluates the relevance of contexts when presenting a position.
STUDENT'S POSITION (PERSPECTIVE, THESIS/HYPOTHESIS)   Specific position (perspective, thesis/hypothesis) is imaginative, taking into account the complexities of an issue.   Limits of position (perspective, thesis/hypothesis) are acknowledged.   Others' points of view are synthesized within position (perspective, thesis/hypothesis).
CONCLUSIONS AND RELATED OUTCOMES (IMPLICATIONS AND CONSEQUENCES)   Conclusions and related outcomes (consequences and implications) are logical and reflect student's informed evaluation and ability to place evidence and perspectives discussed in priority order.

DEFINITION

Critical thinking is a habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion.

Download this and other VALUE Rubrics at **www.aacu.org/value/rubrics**.

MILESTONES		BENCHMARK
3	**2**	**1**
Issue/problem to be considered critically is stated, described, and clarified so that understanding is not seriously impeded by omissions.	Issue/problem to be considered critically is stated but description leaves some terms undefined, ambiguities unexplored, boundaries undetermined, and/or backgrounds unknown.	Issue/problem to be considered critically is stated without clarification or description.
Information is taken from source(s) with enough interpretation/evaluation to develop a coherent analysis or synthesis. Viewpoints of experts are subject to questioning.	Information is taken from source(s) with some interpretation/evaluation, but not enough to develop a coherent analysis or synthesis. Viewpoints of experts are taken as mostly fact, with little questioning.	Information is taken from source(s) without any interpretation/evaluation. Viewpoints of experts are taken as fact, without question.
Identifies own and others' assumptions and several relevant contexts when presenting a position.	Questions some assumptions. Identifies several relevant contexts when presenting a position. May be more aware of others' assumptions than one's own (or vice versa).	Shows an emerging awareness of present assumptions (sometimes labels assertions as assumptions). Begins to identify some contexts when presenting a position.
Specific position (perspective, thesis/hypothesis) takes into account the complexities of an issue. Others' points of view are acknowledged within position (perspective, thesis/hypothesis).	Specific position (perspective, thesis/hypothesis) acknowledges different sides of an issue.	Specific position (perspective, thesis/hypothesis) is stated, but is simplistic and obvious.
Conclusion is logically tied to a range of information, including opposing viewpoints; related outcomes (consequences and implications) are identified clearly.	Conclusion is logically tied to information (because information is chosen to fit the desired conclusion); some related outcomes (consequences and implications) are identified clearly.	Conclusion is inconsistently tied to some of the information discussed; related outcomes (consequences and implications) are oversimplified.

VALUE@10

Valid Assessment of Learning in Undergraduate Education

First introduced in 2009 as an alternative to the predominant educational assessment paradigm—standardized tests—the VALUE rubrics have moved from the periphery of student outcomes assessment to the center of conversations about the quality of student learning within and across institutions. As of 2018, the rubrics have been downloaded more than 70,000 times by individuals representing nearly 5,900 organizations, including more than 2,188 colleges and universities. We invite you to download the rubrics—available for free at AAC&U's website—and take advantage of the following AAC&U resources.

We Have a Rubric for That: The VALUE Approach to Assessment

Kathryne Drezek McConnell, Erin M. Horan, Bethany Zimmerman, and Terrel L. Rhodes ($15 members/$25 nonmembers)
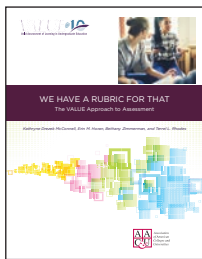
The newest publication in a series of reports on the AAC&U Valid Assessment of Learning in Undergraduate Education (VALUE) Initiative—*We Have a Rubric for That: The VALUE Approach to Assessment*—compiles ten years of evidence to provide an argument-based framework for the assessment of student learning in higher education using the VALUE rubrics. This publication presents a wide range of sources to provide timely evidence of the power of the VALUE approach to assessment. *(2019)*

Using the VALUE Rubrics for Improvement of Learning and Authentic Assessment

Terrel L. Rhodes and Ashley Finley
($15 members/$25 nonmembers)
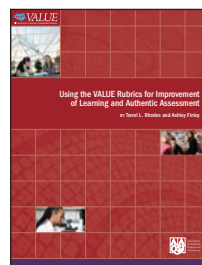
This 2013 publication addresses key elements of, and questions raised about, the development and use of the VALUE rubrics for assessment of student learning. It provides information about rubric-based assessment approaches—including validity, reliability, and rubric modification—and faculty training in the use of rubrics. Examples of how campuses are using the VALUE rubrics to improve student learning are provided. Full case studies from twelve campuses are available online at **www.aacu.org/value/casestudies.** *(2013)*

Assessing Outcomes and Improving Achievement: Tips and Tools for Using Rubrics

Edited by Terrel L. Rhodes
(eBook only (PDF); $15 members/$25 nonmembers)
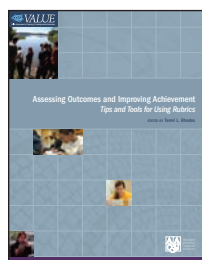
This first—and best-selling—VALUE publication provides practical advice on the development and effective use of rubrics to evaluate college student achievement at various levels. Rubrics for fifteen liberal learning outcomes are included, and can be readily adapted to reflect the missions, cultures, and practices of individual institutions and programs. *(2010)*

On Solid Ground: VALUE Report 2017

Kathryne Drezek McConnell and Terrel L. Rhodes
(Free PDF is available at www.aacu.org/OnSolidGroundVALUE)
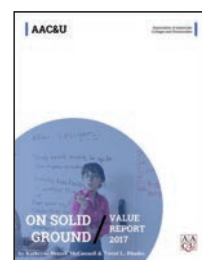
This report describes VALUE's ground-breaking approach to assessing student learning. It includes results from the first two years of data collection for the VALUE Initiative, representing the first attempt to reveal the landscape of student performance on key learning outcomes—Critical Thinking, Written Communication, and Quantitative Literacy—that educators, employers, and policy makers agree are essential for student success in the workplace and in life. *(2017)*

Association of American Colleges and Universities

For information, or to order, visit www.aacu.org/value/publications, email pub_desk@aacu.org, or call 202.387.3760.

VALUE INSTITUTE
Learning Outcomes Assessment At Its Best

VALUE INSTITUTE Registration Now Open

www.aacu.org/VALUEInstitute
Register by Friday, February 22, 2019

The VALUE Institute is an effort to create the country's most comprehensive resource for direct evidence of student learning in higher education. Thirty two-year and four-year colleges and universities participated in the VALUE Institute's first year in 2017–2018 and submitted nearly 5,000 samples of student work to be scored on five distinct learning outcomes. For 2018–2019, institutions can register to measure seven outcomes: critical thinking, written communication, quantitative literacy, global learning, intercultural knowledge and competency, civic engagement, and ethical reasoning.

Higher education institutions, departments, states, and other providers are invited to participate in the VALUE Institute by collecting samples of student work, uploading these artifacts into the digital repository, and having the work scored by certified VALUE Institute faculty scorers. Participants receive data and reports from the tested VALUE nationwide database for benchmarking student learning.

For more information about the VALUE Institute and how to participate, visit **www.aacu.org/VALUEInstitute.**

For more information about AAC&U's VALUE approach to assessment, visit **www.aacu.org/value.**

Association of American Colleges and Universities

Addressing the Assessment Paradox

▶ **David Eubanks,** Assistant Vice President, Office of Institutional Assessment and Research, Furman University

T he VALUE initiative's research can contribute to a revolution in assessment, or it can be used to justify the continuation of ineffective practices. The outcome will depend on leadership and willingness to critically analyze the assessment movement's inability to fulfill its promise to measure learning so that we might improve itself.

Assessment is a paradox: its practitioners want two things that are nearly incompatible. One is agreement among higher education stakeholders about claims like, "Our graduates can write at the college level." Such agreement—when it can be reached—legitimizes the claims but also represents a compromise between multiple

The book translated standardized test scores into a generalized conclusion, amplified across news outlets, that at the bachelor's degree level, engineers can't engineer and accountants can't count.

perspectives about what an outcome means in practice. At the same time, we want assessment measures that are grounded in empiricism, so we don't fool ourselves. Andrew Gelman, who recently (2018) proposed ethical guidelines for statistical practice and communication, raises a similar point: "Consider this paradox: statistics is the science of uncertainty and variation, but data-based claims in the scientific literature tend to be stated deterministically (e.g. 'We have discovered . . . the effect of X on Y is . . . hypothesis H is rejected')."

Problems arise when the social meaning of assessment diverges from its empirical merits, as with the publication of Richard Arum

and Josipa Roksa's *Academically Adrift: Limited Learning on College Campuses* (2011). The book translated standardized test scores into a generalized conclusion, amplified across news outlets, that at the bachelor's degree level, engineers can't engineer and accountants can't count.

On the other hand, socially constructed meaning that ignores reality is counter-productive. My favorite such story is from *Behind the Urals: An American Worker in Russia's City of Steel*, in which John Scott describes fixing tractors in the Soviet Union one weekend to help some farmers. After using parts from twelve dilapidated tractors to assemble nine that worked, the farmers were horrified: officially they were accountable for twelve tractors (whether they worked or not was immaterial), and now three of them had vanished!

Institutionalized assessment of student learning resembles Soviet tractor counting. For an assessment director facing accreditation review, it is better to have twelve reports that conform to the reviewer's (bureaucratic) expectations than to have a few good research projects (Eubanks 2017).

Assessment practice also fails at empiricism. What is typically accepted in assessment reviews has little to do with statistics and measurement. Nor could it be otherwise. The 2016 IPEDS data on four-year degrees granted shows that half of the academic programs (using CIP codes) had eight or fewer graduates that year. Such small samples are not suitable for measuring a program's quality, given the many sources of variation in student performance. By my calculations, fewer than 5 percent of four-year programs had enough graduates to make a serious measurement attempt worthwhile. It's safe to conclude that most of the 80,000+ bachelor's degree programs in the United

States are not producing trustable statistics about student learning, regardless of the nominal value of their assessment reports.

The current situation is the worst possible outcome: social acceptability of sloppy data work, creating a decades-long failure to fulfill the assessment movement's laudable empirical aims.

THE PROMISE AND PERIL OF AAC&U'S VALUE RUBRICS

The VALUE rubrics and accompanying validity studies address both the empirical and social aspects of assessment. The learning outcomes and the language used to describe them came from discussions with stakeholders in a credible attempt to define types of student achievement that are not easily measured. The ongoing research on the use of these instruments explores the validity of the rubric ratings for understanding student achievement. As such, the project is commendable in trying to resolve the paradox described above.

Assessment practice can be moved beyond its current bureaucratic impasse by projects like VALUE, but it is also possible that such efforts could worsen the situation instead. Current assessment practice relies on the language of empiricism (e.g. "measuring student learning," setting benchmarks, looking for statistical differences, etc.) for credibility, even though the conditions do not exist for actual empiricism to be broadly employed. The research done in the VALUE initiative could end up just adding a gloss of science to otherwise shoddy data work done in the name of program assessment ("My measure of program quality is valid because we used a VALUE rubric, even though it's based on a single student's performance."). To have full effect, research like VALUE requires a revolution within assessment practice.

THE FUTURE OF ASSESSMENT

Assessment as a field should take note of the multiple "replication crises" that are ongoing in other disciplines, where a significant amount of published work is being called into question. Often this is due to using too-small data sets combined with publishing conventions that relied on nominal, rather than actual, significance. Gelman's prosed ethical principles in statistical communication are directed in part at this situation. I believe that the application of his guidelines in assessment could revolutionize practice and satisfactorily resolve the paradox of social-versus-technical meaning. We can have both.

As Gelman puts it, in order to "move away from systematic overconfidence," he recommends more transparency in data work. Researchers of student achievement need access to the raw data and methods of analysis for studies. To use a cliché, the VALUE initiative can be a "guide on the side" by leading research into the psychometric properties of its rubrics while granting access to others to compare to their own work. How do our internal inter-rater agreement statistics for first-year writing compare to external ones? What is a typical growth curve for undergraduate writing over four years for low-GPA students? These sorts of questions have answers that come from collective pools of data, and VALUE is a natural hub for such a project.

Gelman also addresses the limitations of statistical ways of knowing, and he recommends a culture that embraces criticism of results and methods. The irony of the assessment movement is that it has become fixed and unresponsive to criticism, as public exchanges in the last year have shown. In contrast, in their *Change* magazine article on assignment difficulty, Daniel F. Sullivan and Kate Drezek McConnell ask the critical question of their assessment data, "Why aren't the scores of seniors much higher?" (2017).

By working through this challenge to validity, they find something interesting and useful about assignment difficulty.

Current assessment practices enforce an unworkable model of too many projects with too little data and methods that practically ensure poor results. The VALUE initiative can partly address the small-data issue through targeted research projects and in support of hermeneutic ways of knowing; see Pamela Moss's (1994) work on that subject as a guide. However, we still need larger data sets to produce generalizable hypotheses about student learning. To make progress there we need to reboot assessment's empirical expectations, eliminate the outdated rules, and seek new methods of data gathering that can address both the technical and social requirements.

Assessment in higher education can still fulfill its original promise. But we need to reflect critically on the historical ineffectiveness of the movement in comparison to the ubiquitous success of data mining in other contexts. Piecemeal solutions can't patch up the flaws; we need a complete rethink. The VALUE approach to assessment can be avant garde in this revolution. It would be a shame if instead it just becomes another checkbox on an assessment report. ▪

REFERENCES

Arum, Richard, and Josipa Roksa. 2011. *Academically Adrift: Limited Learning on College Campuses.* Chicago: University of Chicago Press.

Eubanks, David. 2017. "A Guide for the Perplexed." *Intersection of Assessment and Learning.* https://cdn.ymaws.com/www.aalhe.org/resource/resmgr/docs/Int/AAHLE_Fall_2017_Intersection.pdf.

Gelman, Andrew. 2018. "Ethics in Statistical Practice and Communications." *Significance* 15 (5): 40–43.

Scott, John. 1989. *Behind the Urals: An American Worker in Russia's City of Steel.* Bloomington, IN: Indiana University Press.

Sullivan, Daniel F., and Kate Drezek McConnell. 2017. "Big Progress in Authentic Assessment, But by Itself Not Enough." *Change: The Magazine of Higher Learning* 49 (1): 14–25.
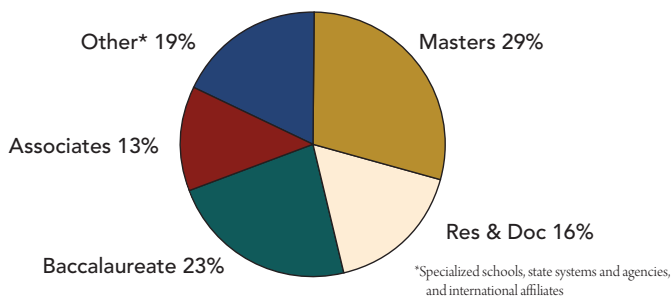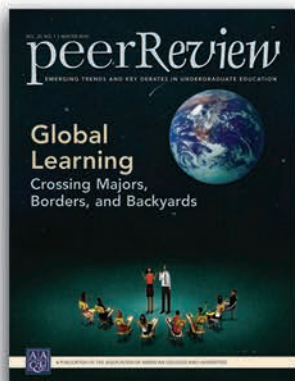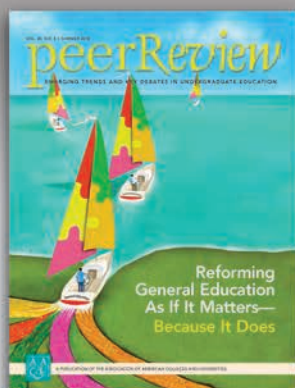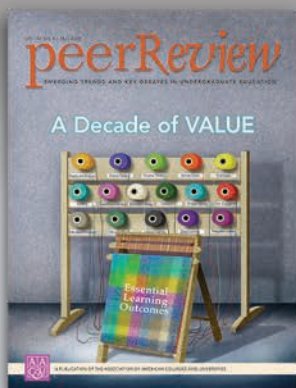
AAC&U is the leading national association dedicated to advancing the vitality and public standing of liberal education by making quality and equity the foundations for excellence in undergraduate education in service to democracy. Its members are committed to extending the advantages of a liberal education to all students, regardless of academic specialization or intended career. Founded in 1915, AAC&U now comprises 1,400 member institutions—including accredited public and private colleges, community colleges, research universities, and comprehensive universities of every type and size.

AAC&U functions as a catalyst and facilitator, forging links among presidents, administrators, faculty, and staff engaged in institutional and curricular planning. Through a broad range of activities, AAC&U reinforces the collective commitment to liberal education at the national, local, and global levels. Its high-quality programs, publications, research, meetings, institutes, public outreach efforts, and campus-based projects help individual institutions ensure that the quality of student learning is central to their work as they evolve to meet new economic and social challenges. Information about AAC&U can be found at www.aacu.org.

AAC&U Membership 2019

Other* 19%

Masters 29%

Associates 13%

Res & Doc 16%

Baccalaureate 23%

*Specialized schools, state systems and agencies, and international affiliates

peerReview

A Decade of VALUE
Fall 2018

Reforming General Education As If It Matters—Because It Does
Summer 2018

The LEAP Challenge:
Engaging in Capstones and Signature Work
Spring 2018

Global Learning:
Crossing Majors, Borders, and Backyards
Winter 2018

Single Issues: $8/$10

The first price listed is available only to individuals on AAC&U member campuses (www.aacu.org/membership/list.cfm). Discounts are available for quantities of 11 or more. Subscription information can be found online at www.aacu.org/peerreview; annual rates are $35 for individuals and $45 for libraries.

To place an order or request more information:

www.aacu.org/peerreview | 1.800.297.3775 | pub_desk@aacu.org