RESEARCH & PRACTICE IN ASSESSMENT

VOLUME EIGHTEEN | ISSUE 1 | RPAJOURNAL.COM | ISSN #2161-4120

RPA

A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

RESEARCH & PRACTICE IN ASSESSMENT

CALL FOR PAPERS

Manuscripts submitted to RPA may be related to various higher education assessment themes, and should adopt either an assessment measurement or an assessment policy/foundations framework. Contributions are accepted at any time and will receive consideration for publishing. Manuscripts must comply with the RPA Submission Guidelines and be submitted to our online manuscript submission system found at rpajournal.com/authors/.

RESEARCH & PRACTICE IN ASSESSMENT

The goal of Research & Practice in Assessment is to serve the assessment community as an online journal focusing on higher education assessment. It is dedicated to the advancement of scholarly discussion amongst researchers and practitioners in this evolving field. The journal originated from the Board of the Virginia Assessment Group, one of the oldest continuing professional higher education assessment organizations in the United States. Research & Practice in Assessment is a peer-reviewed publication that uses a double-blind review process. Approximately forty percent of submissions are accepted for issues that are published twice annually. Research & Practice in Assessment is listed in Cabell's Directory and indexed by EBSCO, ERIC, Gale, and ProQuest.

History of Research & Practice in Assessment

Research & Practice in Assessment (RPA) evolved over the course of several years. Prior to 2006, the Virginia Assessment Group produced a periodic organizational newsletter. The purpose of the newsletter was to keep the membership informed regarding events sponsored by the organization, as well as changes in state policy associated with higher education assessment. The Newsletter Editor, a position elected by the Virginia Assessment Group membership, oversaw this publication. In 2005, it was proposed by the Newsletter Editor, Robin Anderson, Psy.D. (then Director of Institutional Research and Effectiveness at Blue Ridge Community College) that it be expanded to include scholarly articles submitted by Virginia Assessment Group members. The articles would focus on both practice and research associated with the assessment of student learning. As part of the proposal, Ms. Anderson suggested that the new publication take the form of an online journal.

The Board approved the proposal and sent the motion to the full membership for a vote. The membership overwhelmingly approved the journal concept. Consequently, the Newsletter Editor position was removed from the organization's by-laws and a Journal Editor position was added in its place. Additional by-law and constitutional changes needed to support the establishment of the Journal were subsequently crafted and approved by the Virginia Assessment Group membership. As part of the 2005 Virginia Assessment Group annual meeting proceedings, the Board solicited names for the new journal publication. Ultimately, the name Research & Practice in Assessment was selected. Also as part of the 2005 annual meeting, the Virginia Assessment Group Board solicited nominations for members of the first RPA Board of Editors. From the nominees Keston H. Fulcher, Ph.D. (then Director of Assessment and Evaluation at Christopher Newport University), Dennis R. Ridley, Ph.D. (then Director of Institutional Research and Planning at Virginia Wesleyan College) and Rufus Carter (then Coordinator of Institutional Assessment at Marymount University) were selected to make up the first Board of Editors. Several members of the Board also contributed articles to the first edition, which was published in March of 2006.

After the launch of the first issue, Ms. Anderson stepped down as Journal Editor to assume other duties within the organization. Subsequently, Mr. Fulcher was nominated to serve as Journal Editor, serving from 2007-2010. With a newly configured Board of Editors, Mr. Fulcher invested considerable time in the solicitation of articles from an increasingly wider circle of authors and added the position of co-editor to the Board of Editors, filled by Allen DuPont, Ph.D. (then Director of Assessment, Division of Undergraduate Affairs at North Carolina State University). Mr. Fulcher oversaw the production and publication of the next four issues and remained Editor until he assumed the presidency of the Virginia Assessment Group in 2010. It was at this time Mr. Fulcher nominated Joshua T. Brown (Director of Research and Assessment, Student Affairs at Liberty University) to serve as the Journal's third Editor and he was elected to that position.

Under Mr. Brown's leadership Research & Practice in Assessment experienced significant developments. Specifically, the Editorial and Review Boards were expanded and the members' roles were refined; Ruminate and Book Review sections were added to each issue; RPA Archives were indexed in EBSCO, Gale, ProQuest and Google Scholar; a new RPA website was designed and launched; and RPA gained a presence on social media. Mr. Brown held the position of Editor until November 2014 when Katie Busby, Ph.D. (then Assistant Provost of Assessment and Institutional Research at Tulane University) assumed the role after having served as Associate Editor from 2010-2013 and Editor-elect from 2013-2014.

Ms. Katie Busby served as RPA Editor from November 2014-January 2019 and focused her attention on the growth and sustainability of the journal. During this time period, RPA explored and established collaborative relationships with other assessment organizations and conferences. RPA readership and the number of scholarly submissions increased and an online submission platform and management system was implemented for authors and reviewers. In November 2016, Research & Practice in Assessment celebrated its tenth anniversary with a special issue. Ms. Busby launched a national call for editors in fall 2018, and in January 2019 Nicholas Curtis (Director of Assessment, Marquette University) was nominated and elected to serve as RPA's fifth editor.

Published by:
VIRGINIA ASSESSMENT GROUP | virginiaassessment.org
Publication Design by Patrice Brown |

TABLE OF CONTENTS

FROM THE EDITOR

FROM THE PRESIDENT

ARTICLES

Assessment, Learning Outcomes, and the Digital Shift

"Assessment is not about you… it is about your students' learning. The purpose of assessment is to determine whether or not our students have learned what we want them to learn."

- John F. Kennedy

In this issue of *Research & Practice in Assessment*, we explore themes of assessment, learning outcomes, and the evolving landscape of our work in the digital age. We are grateful to our esteemed authors for their insightful contributions that shed light on these critical aspects of our discipline.

Finney, Pastor, and Silver examine the relationship between students' understanding of assessment for institutional accountability and improvement, and their test-taking effort in remote test administration. Their findings provide invaluable information for institutions seeking to optimize assessment in remote learning environments. Stafford, Cousins, Bol, and Mize delve into the reliability of assessing integrative learning using rubrics and the potential benefits of group norming. Their study offers guidance to professionals and institutions looking to enhance the accuracy and consistency of rubric-based assessments. Shapovalov and Leventhal investigate the alignment of general education and academic degree program learning outcomes, highlighting the importance of cohesion and clarity in defining educational objectives. Hutson and Hogan explore faculty engagement in student learning outcome assessment and provide insights into the essential role that professionals play in ensuring that students meet the desired learning objectives. Pusey, Valencia, Signorini, and Kranzfelder adapt the Classroom Observation Protocol for Undergraduate STEM (COPUS) for online synchronous learning, focusing on breakout rooms, polling, and chat features. Their work demonstrates the need for innovative solutions to assess and enhance learning experiences in the digital age. Lastly, Fisher, Bahl, and Mickleson present an intentional process for revising institutional learning outcomes, emphasizing the importance of regular evaluation and adaptation to ensure continuous improvement in educational institutions.

Together, these articles contribute to a deeper understanding of assessment practices, learning outcomes, and the challenges and opportunities that emerge as education continues to evolve in the digital era. We hope that our readers will find these articles both thought-provoking and informative, inspiring reflection and action within their own institutions.

Regards,

Nicholas Curtis

Editor-in-Chief,
Research & Practice in Assessment

FROM THE PRESIDENT

The *Research & Practice in Assessment* (RPA) Journal, a publication of the Virginia Assessment Group, is an online journal that focuses on the multifaceted aspects of assessment in higher education. First published in 2006, this peer-reviewed journal offers researchers an opportunity to advance the scholarly discussion amongst themselves and practitioners in the field through the continued dissemination of information about theory, scholarship, successes, and best practices in higher education assessment.

As president of the Virginia Assessment Group, I would like to extend a special thank you to each of the authors for their contribution in compiling our journal. On behalf of the Virginia Assessment Group Board members, I would also like to thank the editorial team, advisory board, and editor-in-chief, Dr. Nicholas Curtis. We are extremely proud of all of the successes of this team.

Congratulations to the authors, and thank you to our readers.

Regards,

Tia A. Minnis

President,
Virginia Assessment Group

Abstract

We examined if students' understanding about the purpose and use of institutional assessment scores was affected by moving to remote testing due to COVID-19. Moreover, we examined if students' knowledge about the purpose of outcomes assessment related to their effort on these tests. If knowledge about accountability testing and effort were positively related, we could design interventions to increase knowledge and, in turn, increase effort. We gathered data on knowledge about institutional accountability testing and test-taking effort from students differing in year in school and whether tests were completed remotely or in person. Knowledge about assessment testing was high with negligible differences in knowledge across year in school and testing context. Knowledge related positively to test-taking effort. Testing context and year in school did not moderate this relation. In sum, students who better understood that outcomes assessment was used for accountability and improvement efforts expended more effort on these assessments.

AUTHORS

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

Dena A. Pastor, Ph.D.
James Madison University

Shanti Silver, BA
Kenyon College

Students' Understanding of Assessment for Institutional Accountability and Improvement: Relation with Test-Taking Effort and Remote Test Administration

Higher education institutions engage in outcomes assessment to respond to institutional accountability mandates (U.S. Department of Education, 2006) and to inform programming changes to improve student learning and development (Fulcher & Prendergast, 2021). The student learning and development outcomes that are assessed, reported, and used for improvement are often the outcomes of multi-faceted education experiences, such as general education programming (Mathers et al., 2018; Stone & Friedman, 2002), academic degree programs (Allen, 2004), quality enhancement plans (Miller et al., 2019; Smith & Finney, 2020), and student affairs programs (e.g., Kerr et al., 2020). Outcomes assessment data for institutional accountability and improvement purposes often is not associated with an individual course, but rather tied to several academic and/or student affairs learning experiences. Thus, the assessment of these outcomes often does not inform course grades, graduation, admission into a major, or other high-stakes outcomes for students. These institutional effectiveness assessments are often low stakes for students, meaning there are no personal consequences associated with their performance.

CORRESPONDENCE
Email
finneysj@jmu.edu

Studies have shown that students perform better when assessments are perceived as high stakes versus low stakes. Wise and DeMars (2005) summarized studies that compared test performance across examinee groups who were administered the same test but under high-stakes versus low-stakes conditions. Examinees in the low-stakes condition scored .59

SD lower than those in the high-stakes condition. During institutional effectiveness testing, students may not know if the assessments they are completing are low or high stakes. Their understanding of the stakes of the assessment may influence their test performance.

Moreover, studies have shown that students expend more effort when tests are perceived as high stakes versus low stakes. Sundre and Kistansas (2004) found that self-reported effort was lower when a test was described as low versus high stakes. Unlike high-stakes tests, where students tend to expend the effort necessary to reflect their ability, low-stakes tests tend to be associated with greater variability in effort, with some students expending a high degree of effort and others not. Expended effort positively covaries with test scores, indicating that test scores reflect expended effort to some extent (Cole et al., 2008; Eklöf et al., 2014; Myers & Finney, 2021). Moreover, effort covaries with students' perceived test importance (Finney et al., 2018; Penk & Richter, 2017; Rios, 2021), test emotions (Finney, Perkins, & Satkus, 2020; Finney, Satkus, & Perkins, 2020; Penk & Schipolowski, 2015; Perkins et al., 2021; Satkus & Finney, 2021), personality (Barry & Finney, 2016; Barry et al., 2010; Freund & Holling, 2011; Kopp et al., 2011), and attitudes toward accountability testing (Zhao et al., 2020; Zilberberg et al., 2014).

Students' Knowledge of the Purpose and Use of Outcomes Assessment Data

We were interested in examining if knowledge about the purpose and use of assessment scores also related to expended effort. If knowledge about assessment for institutional effectiveness purposes and test-taking effort were positively related, we could design interventions to increase knowledge and, in turn, possibly increase effort. Interventions to influence knowledge may be easier to create than interventions to influence test importance, test emotions, or attitudes toward accountability testing, and, of course, trying to change personality is futile. With that said, we were unsure if understanding the purpose and use of institutional effectiveness test data would relate positively or negatively with effort. It may be that having an accurate understanding of the low-stakes nature of the test leads to lower expended effort.

A student's understanding of the purpose and use of test scores for institutional effectiveness may be influenced by a variety of things, including what information is shared with students, how it is shared, and how receptive students are to the information. For instance, the information itself may be high quality, but the delivery of the information may be poor. Likewise, the information and delivery may be high quality, but students may not be engaged in receiving the information (e.g., do not read or listen to information provided). When our institutional accountability testing moved online due to the COVID-19 pandemic, it provided us with an opportunity to assess students' understanding of institutional-level assessment efforts and to assess if this understanding was impacted by the modality of the testing (in-person proctored testing prior to the pandemic versus remote, unproctored testing during the pandemic).

Students' understanding of the stakes of the assessment may influence their test performance.

We were also able to examine if understanding of institutional assessment differed across student groups, specifically incoming first-year students and more advanced students (students who had completed 45-70 credit hours; typically 1.5 years at the institution). Students earlier in their college career may differ from more advanced students with respect to their understanding of institutional accountability testing and its impact on them personally. Numerous studies have found that older students exhibit lower effort on low-stakes tests than younger students (e.g., Finney et al., 2016; Rios & Guo, 2020; Thelk et al., 2009). It is unknown if the difference in effort across student age groups is due to older students understanding the low-stakes nature of the tests better than younger students. Thus, we examined if incoming first-year students differed from more advanced students in their understanding of institutional-level assessment efforts and if this difference related to differences in test-taking motivation.

Despite widespread use of institutional effectiveness testing, students' understanding of its purpose and use remains unclear, with anecdotal reports suggesting misconceptions and varied levels of motivation.

Although our institution shares a great deal of information about the purpose and use of institutional effectiveness testing data throughout a students' college career (described in the Methods section), we were unsure of the level of students' understanding. Anecdotally, we heard from some faculty that students do not understand why they are being tested, who sees the scores, or how scores are being used. Some faculty claimed this lack of knowledge resulted in low motivation to perform well on the tests. Other faculty claimed the opposite: more advanced students, unlike incoming students, do understand the purpose of testing and, in particular, understand there are no personal consequences for poor performance. These faculty would often attribute minimal increases in outcome scores to the decreased effort of more advanced students who better understood the low-stakes nature of the tests. Although we understood the logic of both claims, there were no data to support either hypothesis.

Moreover, studies examining college students' understanding of institutional accountability testing is limited. There are concerns that misunderstandings abound. "Many of the criticisms we hear about educational assessments appear to be based on misconceptions. Some of them are due to persons simply misunderstanding the meaning of test scores and their implications for instructional improvement and school accountability" (Goodman & Hambleton, 2005, p. 107). What do students understand regarding accountability testing? One study examined college students' understanding of federal K-12 accountability testing (Zilberberg et al., 2012). Performance was poor. For incoming students, item-level performance ranged from 20% of students answering correctly to 74% of students answering correctly, with less than 50% of students answering six of the nine items correctly. More advanced students had similar levels of misunderstanding, with 50% or less answering six of the nine items correctly. Paradoxically, both incoming and more advanced students indicated a moderate level of confidence in their answers. It is unknown what information was shared with these students about the purpose and use of the K-12 institutional accountability test scores. Thus, prior to conducting the current study, we did not hypothesize expected levels of understanding or how this understanding would relate to modality of testing, year in school, or expended effort on the test. Instead, this was an exploratory study to provide initial insight into students' understanding of the purpose and use of higher education institutional accountability assessment scores.

Purpose of the Current Study

Despite the widespread use of testing for institutional accountability and improvement, little is known about students' understanding of institutional accountability testing, and even less is known about how this understanding relates to students' test-taking behavior. The purpose of our study was to examine students' understanding of the purpose of institutional accountability assessment scores at our higher education institution. We examined this understanding for both incoming students at the start of the fall semester and more advanced students with over a year at the institution. Moreover, given the move to remote testing due to COVID, we examined if different modalities of testing were associated with students' understanding of testing for institutional accountability and improvement. We also examined if students' knowledge about the purpose of outcomes assessment related to their effort on these tests. We were unsure if understanding the purpose of institutional accountability test data would relate positively or negatively with effort and if this relation would be moderated by student group or testing modality. No matter the results, there would be implications for testing practices.

Methods

Information Sharing and Testing Procedures

For more than 30 years, James Madison University has used Assessment Days to collect longitudinal data on student learning outcomes. Our model ensures that all incoming students are tested twice: once in the fall semester as incoming students and again in the spring semester after accumulating 45-70 credit hours (Pastor et al., 2019). Although a student completes only four instruments each Assessment Day, 25 different assessments are typically administered, thereby allowing for examination of learning gains on a variety of outcomes.

The goal of Assessment Days is to collect data for interpretation at the program level, not the individual student level. Thus, assessment results are high stakes for the educational programs being assessed but low stakes for students. Individual student scores are not reported, nor do the scores have any individual implications. In other words, students are simply required to attend Assessment Days but are not required to receive passing scores on the assessments.

Our study used data from Fall 2019 (FA19), Spring 2020 (SP20), Fall 2020 (FA20), and Spring 2021 (SP21) Assessment Days. The four administrations differed in the year in school of students being tested, with FA19 and FA20 testing incoming first-year students and SP20 and SP21 testing more advanced students. The administrations also differed in format, with FA19 and SP20 occurring before COVID restrictions and FA20 and SP21 occurring during COVID restrictions. The differences in procedures associated with the two administrative formats are described below.

In-person Proctored Testing prior to COVID

The FA19 and SP20 administrations were typical of the Assessment Day experience at our institution. The FA19 Assessment Day took place during first-year orientation, just prior to the start of fall classes. The SP20 Assessment Day took place on a Tuesday in mid-February. As the name suggests, both Assessments Days took place on a single day. All classes were cancelled until 4:00 p.m. in SP20 to allow students, faculty, and staff to participate in assessment activities.

For both FA19 and SP20 Assessment Days, students were randomly assigned one of three two-hour sessions where they completed three to four assessments aligned with learning objectives in general education and other wide-reaching university initiatives. The majority of testing took place in classrooms where students provided their responses on Scantrons (i.e., optical answer sheets); only 20% of students were tested in computer labs where responses were collected through computer-based testing platforms. All testing sessions were facilitated by trained proctors to ensure standardized conditions.

Students were informed about Assessment Day through the undergraduate catalog, multiple emails, and alerts via university social media outlets. A link to the Assessment Day website was provided in all communications and a video about the purpose of Assessment Day was shown to students just prior to testing. University policy for nonattendance on Assessment Day is a hold placed on the student's record which blocks the student from registering for next semester's classes. Once the student completes makeup testing, the hold is removed.

Remote, Unproctored Testing during COVID

COVID necessitated changes to Assessment Day procedures in FA20 and SP21 (Pastor & Love, 2020). In both FA20 and SP21, students were asked to complete the assessments remotely, without a proctor, and during a specific testing window. In FA20, the testing window spanned from about a week before the start of classes to a little over three weeks after the start of classes. Taking advantage of classes being cancelled for Spring Assessment Day, students were asked to complete their requirement on Assessment Day or the day afterward in SP21.

Assessment Days collect program-level data, not individual student scores, making them low-stakes for students.

Thus, FA20 and SP21 Assessment Days differed from typical administrations in that all students were allowed to complete their assessments remotely, without a proctor, on a computer-based testing platform, and at any time they pleased during the testing window. Although students were told holds would be placed if they failed to complete the requirement by the deadline, to minimize disruption in academic progress, no holds were placed in FA20 and SP21.

Similar to the typical in-person administration, in FA20 and SP21 students were informed about Assessment Day through the undergraduate catalog, multiple emails, and alerts via university social media outlets. A link to the Assessment Day website was provided in all communications and a video about the purpose of Assessment Day was shown to all students just prior to the start of remote testing.

Participants

The results suggest that, despite the differences in administration format, Assessment Day scores were relatively stable over time and across cohorts.

Table 1 illustrates how the final sample for each Assessment Day administration was obtained. We begin with the number of students required to participate in each administration and end with the sample sizes used in the current study. Students were randomly assigned to assessments and only a subset of students were assigned to complete the assessments used in the present study. Because previous research indicates that students who attend make-up testing sessions differ from those who attend Assessment Day (Swerdzewski et al., 2009), we only considered students who completed the assessments on Assessment Day in FA19 and SP20 or by the deadlines in FA20 and SP21. We further limited the data to only those who provided research consent, were > 18 years of age, and provided valid, non-missing responses to all items. Because preliminary analysis indicated problems with streamlining responses (providing the same response to all items on a given scale), we also deleted students streamlining on any scale[1]. Finally, for students tested during more than one administration, we randomly selected which administration's data to retain to reduce dependencies in the data, resulting in the final sample sizes in the far right column of Table 1. A total of 7,513 students comprised the final sample, with 577 (8%), 2,660 (35%), 1,901 (25%), and 2,375 (32%) tested during the FA19, FA20, SP20, and SP21 administrations, respectively. The majority (63%) of students in the final sample self-identified as female and 77% self-identified as White with all other races/ethnicities each represented by 7% or less of the students. These demographics align with those of the institution.

Table 1
Process to Arrive at Final Sample Sizes by Assessment Day Administration

Administration	Required to participate	Assigned to complete the study assessments	Completed on Assessment Day or by deadline	≥ 18, provided research consent & valid, non-missing data	Did not provide streamlined responses	Only one record retained for those testing more than once (Final *N*)
FA19	4466	893	846	781	744	577
FA20	4462	3875	3381	2852	2700	2660
SP20	3797	2962	2600	1951	1906	1901
SP21	3524	3480	3142	2720	2578	2375
Total	16249	11210	9969	8304	7928	7513

Note. Students in a column are a subset of those students to the left of the column. For instance, of the 4,466 students in FA19 who were required to participate in Assessment Day, only 893 were assigned to complete the knowledge of institutional accountability measure (focus of this study); of those 893 students, 846 completed the measure on Assessment Day or by the deadline.

[1] The effort and importance items are responded to on a scale with values of 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Neither agree nor disagree*, 4 = *Agree*, and 5 = *Strongly agree*. With the exception of a response of 3 to all items, the same response to all items on either subscale is nonsensical given the presence of reverse-scored items. Students who provided responses of all 1s, 2s, 4s, or 5s to items on either subscale were deleted. All Assessment Day knowledge items are responded to on a scale with values of 1=*True* and 2=*False*. Because providing the same response to all items does not align with a reasonable response pattern, students doing so were considered unmotivated and deleted from the data. Interestingly, of the 376 students identified as streamliners, 87% completed the assessments under remote, unproctored conditions in either FA20 or SP21.

Measures

Knowledge about the Purpose and Use of Outcomes Assessment Data

We have never assessed the extent to which students understand the purpose and use of outcomes assessment data. Thus, we created a 12-item dichotomously-scored measure to assess their understanding. We purposefully avoided jargon related to institutional accountability and improvement, such as "value-added", "accountability", "accreditation". Instead, we constructed items to describe these purposes without using unknown terms. Because the items were not created to measure a unidimensional construct, internal consistency reliability was not computed.

Test-Taking Motivation

The Student Opinion Scale (SOS) (Thelk et al., 2009) is a 10-item measure consisting of five items reflecting students' perceived importance of the assessments they completed (e.g., "Doing well on these tests was important to me.") and five items reflecting expended effort (e.g., "I gave my best effort on these tests."). Students indicated their agreement with each statement using a 5-point scale (1 = Strongly disagree to 5 = Strongly agree). The SOS has been employed in at least nine countries, 33 universities, and 55 studies (Sessoms & Finney, 2015). A two-factor structure of scores and longitudinal invariance across one and a half years has been supported. Reliability estimates were adequate in the current study: .78 for perceived test importance and .83 for expended effort.

Results

Knowledge of Institutional Accountability Testing

Students performed incredibly well on the knowledge of institutional accountability testing measure. Across the four samples, students answered 89% to 93% of the items correctly (see Table 2) with more advanced students (SP20, SP21) performing negligibly better, on average, than incoming students. Likewise, there were trivial differences in average knowledge scores for students tested in person (FA19, SP20) versus remotely (FA20, SP21). In sum, all four samples performed well and were not practically different from one another in average scores.

Assessment data is only valuable when students understand its purpose and use.

Table 2
Average Knowledge about Institutional Accountability Testing and Test-taking Motivation

Measure	FA19 (*N*=577)		SP20 (*N*=1901)		FA20 (*N*=2660)		SP21 (*N*=2375)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Knowledge Score	10.84	1.00	11.22	0.98	10.71	1.19	11.30	1.16
Knowledge %	90.35	8.37	93.52	8.14	89.24	9.91	91.93	9.63
Expended Effort	3.77	0.65	3.83	0.64	3.74	0.67	3.61	0.66
Perceived Test Importance	3.26	0.72	2.88	0.80	3.33	0.69	3.07	0.75

Note. Knowledge Score can range from 0 to 12 with higher scores indicating more knowledge. Knowledge % is the Knowledge Score converted to a percent correct scale and can range from 0 to 100% correct. Expended Effort and Perceived Test Importance can range from 1 to 5 with higher scores indicating higher effort and importance.

Students' self-reports of expended effort on the tests was moderately high across the four samples with no noteworthy differences based on year in school, and average scores being somewhat higher for students tested in person (FA19, SP20) relative to those tested remotely (FA20, SP21). In alignment with previous research, perceived test importance was moderate and

lower than expended effort for all samples, with higher average scores for incoming students (FA19, FA20) relative to more advanced students (SP20, SP21). Perceived test importance was similar for students tested in person (FA19, SP20) and remotely (FA20, SP21).

To investigate understanding of specific aspects of institutional accountability testing, we examined performance on each of the 12 items. In particular, we were interested in differences in performance on items that did and did not reference personal consequences to the student. In Table 3, we have italicized items that reference personal consequences to students. For example, the first item assesses if students understand that scores from Assessment Day tests are not factored into grade point average. For all but one italicized item, students indicated their understanding of the low-stakes nature of these tests to them personally (i.e., no personal consequences for poor performance). Specifically, almost all students understood that their performance would not appear on their transcript, impact their grade point average, be used to determine future coursework, or affect their academic record. Of note, many incoming students (FA19, FA20) mistakenly believed that faculty could see students' individual performance on the tests. More advanced students (SP20, SP21) performed better on this item; however, this item in general was the most difficult across the four samples. Other than that item and another asking students whether the state requires all state universities to assess student learning, students performed well on the remaining 10 items and differences in performance across testing context (in person versus remote) and year in school were trivial.

Relations between Knowledge and Test-Taking Motivation

We examined the relation between knowledge and test-taking motivation via two approaches. First, we examined the bivariate linear relation between total knowledge score and both effort and perceived test importance (see Table 4). Across testing modality and student age, the relations were not practically different. Effort related positively to knowledge; students who better understood the purpose and use of institutional assessment data reported expending more effort. Knowledge was negligibly related to perceived test importance; the amount a student understood the purpose and use of institutional assessment data did not relate to their perceived value of the test. As expected based on research, effort and importance were positively related.

These results suggest that remote testing can be a viable option for institutions to continue assessment efforts during times of disruption or beyond, without sacrificing student understanding or motivation.

Next, we examined the relation between effort and knowledge by estimating the average expended effort for students who answered each item correctly versus incorrectly (see Table 5). For items that referenced personal consequences (italicized), we were interested in whether students who understood the personal low-stakes nature of the test expended less effort than those who did not. The opposite occurred — students who correctly understood the lack of personal consequences of these tests had higher average effort. In fact, for all 12 items, students who answered correctly had higher effort with effect sizes ranging from 0.04 to 1.03 SDs.

Discussion

In general, students understood the purpose and use of outcomes assessment testing, with negligible differences in knowledge across testing context and year in school. Knowledge did not relate to perceived importance of these tests but it did relate positively to test-taking effort. Testing context and year in school did not moderate these relations. In sum, students understand the low-stakes nature of outcomes assessment to them personally and increased understanding was not associated with lower expended effort. Instead, students who better understood that outcomes assessment was used for accountability and improvement efforts expended more effort on these assessments.

Implications for Remote Testing

Fortunately, our transition to remote testing was not accompanied by misunderstanding the purpose or use of institutional accountability testing. Instead, knowledge about institutional accountability testing was similarly high for students tested in-person versus remotely. Likewise, knowledge related to perceived test importance and expended test-taking effort similarly for students tested in-person versus remotely. Thus,

Table 3
Percentage of Students Answering each Knowledge Item Correctly

Item	Correct Answer	Percentage of Students Answering Item Correctly			
		FA19	SP20	FA20	SP21
Scores from the Assessment Day tests I just completed will be factored into my Grade Point Average (GPA).	False	99	100	99	99
Scores from the Assessment Day tests I completed will be used to determine which courses I enroll in next semester.	False	97	99	91	98
Scores from the Assessment Day tests I just completed will be used to evaluate the quality of James Madison University.	True	93	97	87	90
Faculty can see my individual scores on the tests I completed today.	False	44	73	54	68
James Madison University students are assessed in the Fall as entering students and again after earning 45 to 70 credits.	True	97	95	91	92
I was supposed to prepare for Assessment Day by studying.	False	99	99	95	97
Faculty use results from Assessment Day to make improvements to James Madison University programs.	True	98	98	96	96
My scores on the Assessment Day tests will appear on my transcript.	False	99	98	97	98
Students are expected to have mastered all the concepts assessed during Assessment Day.	False	93	98	88	92
Students are expected to put forth their best effort on the Assessment Day tests.	True	98	99	97	97
My performance on Assessment Day tests does not impact my academic record.	True	95	97	92	93
The state of Virginia requires all state universities to assess student learning.	True	73	71	83	81

Note. The percentage of students who answered the item correctly is often called "difficulty" by assessment experts. Items that reference personal consequences to students are italicized.

Table 4
Relations between Knowledge and Test-Taking Motivation

Variable	Expended Effort	Perceived Test Importance	Knowledge Score
In-person Proctored Testing			
Expended Effort	1.00	0.42	0.16
Perceived Test Importance	0.34	1.00	0.03
Knowledge Score	0.16	-0.01	1.00
Remote Unproctored Testing			
Expended Effort	1.00	0.48	0.15
Perceived Test Importance	0.50	1.00	0.05
Knowledge Score	0.22	0.04	1.00

Note. Correlations above the diagonal are based on the incoming student sample. Correlations below the diagonal are based on the more advanced student sample.

Encouraging results for remote testing in institutional accountability contexts, with an opportunity to increase test-taking effort through increased knowledge of testing purposes.

potential arguments to avoid remote testing due to student confusion regarding testing were not supported. These are encouraging results, particularly for assessment programs quickly transitioning to similar testing modalities.

Although students' knowledge did not differ across testing contexts, there was a small difference in expended effort, with slightly lower effort associated with the remote administration. The extent to which lower effort during remote testing affects test performance was recently considered by Alahmadi and DeMars (2022) and is worthy of continued study.

Implications for Increasing Test-Taking Effort

Across the four student samples, expended test-taking effort averaged between 3.61 and 3.83 on a 5-point scale, where higher scores indicate higher expended effort. Given these averages and variability about them, there is an opportunity to increase expended effort in institutional accountability testing contexts. Because knowledge of institutional accountability testing was positively related to expended effort, professionals may suggest increasing knowledge as a possible way to increase effort. Our results and this suggestion align with previous recommendations to explain the purpose of accountability testing, given that examinees stated they would have expended more effort if they had known this information (Zilberberg et al., 2009). Moreover, there is evidence that students do not have positive attitudes about testing, nor do these attitudes improve over time (Paris et al., 1991; Zilberberg et al., 2013; Zilberberg et al., 2014). Of importance is the negative relation between understanding the purpose of accountability tests and disillusionment toward these tests (Zilberberg et al., 2013; Zilberberg et al., 2014). If students are aware of the tests' purpose, they may be less disillusioned and might expend effort. Although we did not examine disillusionment, we did find that understanding the purpose of these tests was related to expended effort. With that said, we have two caveats regarding the suggestion of increasing knowledge of institutional accountability testing as a possible way to increase effort.

First, we cannot claim increased knowledge causes increased effort. We can simply state that those students who expended more effort during institutional accountability testing also tended to understand better the purpose of institutional accountability testing. It may be that conscientiousness influenced both variables; students higher in conscientiousness better focused on information explaining institutional accountability testing (thus, they understand it) and they responsibly put forth more effort on tests. Yet, even if increased knowledge does not directly translate into better examinee behavior, informing students as to the purpose

Table 5
Average Expended Effort for Students with Incorrect and Correct Answers on each Knowledge Item (N = 7513)

	Incorrect Answer		Correct Answer		Mean Diff	
Item	*N*	*M (SD)*	*N*	*M (SD)*		*d*
Scores from the Assessment Day tests I just completed will be factored into my Grade Point Average (GPA).	60	3.38 (0.71)	7453	3.73 (0.66)	0.35	0.50
Scores from the Assessment Day tests I completed will be used to determine which courses I enroll in next semester.	310	3.65 (0.60)	7203	3.73 (0.67)	0.07	0.12
Scores from the Assessment Day tests I just completed will be used to evaluate the quality of James Madison University.	679	3.45 (0.71)	6834	3.75 (0.65)	0.30	0.43
Faculty can see my individual scores on the tests I completed today.	2818	3.69 (0.65)	4695	3.74 (0.67)	0.05	0.07
James Madison University students are assessed in the Fall as entering students and again after earning 45 to 70 credits.	529	3.51 (0.70)	6984	3.74 (0.66)	0.22	0.33
I was supposed to prepare for Assessment Day by studying.	250	3.46 (0.68)	7263	3.73 (0.66)	0.28	0.41
Faculty use results from Assessment Day to make improvements to James Madison University programs.	257	3.27 (0.69)	7256	3.74 (0.66)	0.47	0.70
My scores on the Assessment Day tests will appear on my transcript.	158	3.43 (0.75)	7355	3.73 (0.66)	0.30	0.42
Students are expected to have mastered all the concepts assessed during Assessment Day.	584	3.54 (0.72)	6929	3.74 (0.66)	0.20	0.29
Students are expected to put forth their best effort on the Assessment Day tests.	169	3.07 (0.64)	7344	3.74 (0.66)	0.66	1.03
My performance on Assessment Day tests does not impact my academic record.	467	3.54 (0.66)	7046	3.73 (0.66)	0.19	0.29
The state of Virginia requires all state universities to assess student learning.	1600	3.70 (0.67)	5913	3.73 (0.66)	0.02	0.04

Note. Mean Diff = $M_{Correct} - M_{Incorrect}$, $d = \frac{M_{Correct} - M_{Incorrect}}{\sqrt{\frac{SD^2_{Correct} + SD^2_{Incorrect}}{2}}}$.

and use of institutional accountability testing is ethical practice. By sharing this information, we are being transparent and respectful to students who are providing the data we use to improve our educational programs and meet accountability requirements.

Increasing student knowledge about institutional accountability testing is a worthy endeavor, but it is likely to have little impact on increasing effort if knowledge is already high. Our results suggest this messaging is working.

Second, although increasing student knowledge about institutional accountability testing is a worthy endeavor, it is likely to have little impact on increasing effort if knowledge is already high. At our institution, the vast majority of students understood the purpose and use of this testing. That is, for 10 of the 12 knowledge items, over 90% of the students answered the item correctly. Thus, proposing an intervention to raise knowledge about institutional accountability assessment would have limited impact. We have great comfort knowing our students hear and understand our messaging about outcomes assessment, which includes information shared via email, university social media alerts, on the Assessment Day website, and a video about the purpose of Assessment Day featuring student actors shown just prior to testing. Our results suggest this messaging is working. With that said, we do need to explain more clearly that: 1) faculty see only aggregated results, not results from individual students, and 2) the state requires all universities to gather accountability data. For other institutions where students' understanding of institutional accountability testing is limited, increasing knowledge is not only an ethical obligation but may be an effective and cheap strategy to increase effort. However, future studies are needed to investigate the strength of the relation between knowledge and effort when knowledge is more variable than found in our context.

In closing, we return to the two hypotheses presented by faculty regarding test-taking effort and knowledge of institutional accountability testing. Recall, some faculty believed that students did not understand the purpose of the testing and this lack of knowledge resulted in low motivation to perform well on the tests. Other faculty believed that more advanced students did understand that there are no personal consequences for poor test performance and thus did not expend test-taking effort. Our study identified flaws in both arguments. Incoming and advanced students have a good understanding of institutional accountability testing and its low-stakes nature; this is not an area of concern. Moreover, understanding the low-stakes nature of these tests did not result in decreased effort during the testing process. These results provide support for our current strategies to educate students about the testing and arguments to keep the testing low-stakes in nature.

References

Alahmadi, S., & DeMars, C. E. (2022). Large-scale assessment during a pandemic: Results from James Madison University's remote assessment day. *Research & Practice in Assessment*, *17*, 5-15. https://www.rpajournal.com/dev/wp-content/uploads/2022/02/Large-Scale-Assessment-During-a-Pandemic-RPA.pdf

Allen, M. J. (2004). *Assessing academic programs in higher education*. Jossey-Bass.

Barry, C. L., & Finney, S. J. (2016). Modeling change in effort across a low stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education*, *29*, 46-64. https://doi.org/10.1080/08957347.2015.1102914

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*, 342-363. https://doi.org/10.1080/15305058.2010.508569

Cole, J., Bergin, D., & Whittaker, T. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*, 609-624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS. *Applied Measurement in Education*, *27*, 31-45. https://doi.org/10.1080/08957347.2013.853070

Finney, S. J., Myers, A. J., & Mathers, C. E. (2018). Test instructions do not moderate the indirect effect of perceived test importance on test performance in low-stakes testing contexts. *International Journal of Testing*, *18*, 297-322. https://doi.org/10.1080/15305058.2017.1396466

Finney, S. J., Perkins, B. A., & Satkus, P. (2020). Examining the simultaneous change in emotions during a test: Relations with expended effort and test performance. *International Journal of Testing*, *20*, 274-298. https://doi.org/10.1080/15305058.2020.1786834

Finney, S. J., Satkus, P., & Perkins, B. A. (2020). The effect of perceived test importance and examinee emotions on expended effort during a low-stakes test: A longitudinal panel model. *Educational Assessment*, *25*, 159-177. https://doi.org/10.1080/10627197.2020.1756254

Finney, S. J., Sundre, D. L., Swain, M. S., & Williams, L. M. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, *21*, 60-87. https://doi.org/10.1080/10627197.2015.1127753

Freund, P. A., & Holling, H. (2011). Who wants to take an intelligence test? Personality and achievement motivation in the context of ability testing. *Personality and Individual Differences*, *50*, 723-728. https://doi.org/10.1016/j.paid.2010.12.025

Fulcher, K. H., & Prendergast, C. O. (2021). *Improving student learning at scale*: *A how-to guide for higher education*. Stylus.

Goodman, D., & Hambleton, R. K. (2005). Some misconceptions about large-scale educational assessments. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 91-110). Lawrence Erlbaum Associates.

Kerr, K. G., Edwards, K. E., Tweedy, J., Lichterman, H. L., & Knerr, A. R. (2020). *The curricular approach to student affairs*: *A revolutionary shift for learning beyond the classroom*. Stylus.

Kopp, J. P., Zinn, T. E., Finney, S. J., & Jurich, D. P. (2011). The development and evaluation of the Academic Entitlement Questionnaire. Measurement and Evaluation in Counseling and Development, 44, 105-129. https://doi.org/10.1177/0748175611400292

Mathers, C. E., Finney, S. J., & Hathcoat, J. (2018). Student learning in higher education: A longitudinal analysis and faculty discussion. *Assessment and Evaluation in Higher Education*, *43*, 1211-1227. https://doi.org/10.1080/02602938.2018.1443202

Miller, A. R., Miller, K. E., Bailey, S., Fletcher, M., France-Harris, A., Klein, S., & Vickery, R. P. (2019). Partnering academics and community engagement: A Quality Enhancement Plan for a diverse and non-traditional university. *Journal of Community Engagement and Scholarship*, *12*, 1, 17-32. https://digitalcommons.northgeorgia.edu/jces/vol12/iss1/5

Myers, A. J., & Finney, S. J. (2021). Does it matter if examinee motivation is measured before or after a low-stakes test? A moderated mediation analysis. *Educational Assessment*, *26*, 1-19. https://doi.org/10.1080/10627197.2019.1645591

Paris, S. G., Turner, J. C., Lawton, T. A., & Roth, J. L. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher*, *20*(5), 12-20. https://doi.org/10.3102/0013189X020005012

Pastor, D. A., Foelber, K. J., Jacovidis, J. N., Fulcher, K. H., Sauder, D. C., & Love, P. D. (2019). University-wide Assessment Days: The James Madison University model. *The Association for Institutional Research (AIR) Professional File*, *Article 144*, 1-13. https://www.airweb.org/docs/default-source/documents-for-pages/reports-and-publications/professional-file/apf-144-2019-spring_university-wide-assessment-days-the-james-madison-university-model.pdf

Pastor, D. A., & Love, P. (2020, Fall). University-wide Assessment during Covid-19: An opportunity for innovation. *Intersection*: *A Journal at the Intersection of Assessment and Learning*, *2*(1). https://aalhe.scholasticahq.com/article/17617.pdf

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment*, *Evaluation and Accountability*, *29*, 55-79. https://doi.org/10.1007/s11092-016-9248-7

Penk, C., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences*, *42*, 27-35. https://doi.org/10.1016/j.lindif.2015.08.002

Perkins, B. A, Pastor, D. A., & Finney, S. J. (2021) Between- versus within-examinee variability in test-taking effort and test emotions during a low-stakes test. *Applied Measurement in Education*, *34*, 285-300. https://doi.org/10.1080/08957347.2021.1987905

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*, 85-106. https://doi.org/10.1080/08957347.2021.1890741

Rios, J., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, *33*, 263-279. https://doi.org/10.1080/08957347.2020.1789141

Satkus, P., & Finney, S. J. (2021). Antecedents of examinee motivation during low-stakes tests: Examining the variability in effects across different research designs. *Assessment and Evaluation in Higher Education*, *46*, 1065-1079. https://doi.org/10.1080/02602938.2020.1846680

Sessoms, J. C., & Finney, S. J. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing*, *15*, 356-388. http://dx.doi.org/10.1080/15305058.2015.1034866

Smith, K. L., & Finney, S. J. (2020). Elevating program theory and implementation fidelity in higher education: Modeling the process via an ethical reasoning curriculum. *Research & Practice in Assessment*, *15*, 5-17. https://files.eric.ed.gov/fulltext/EJ1293385.pdf

Stone, J., & Friedman, S. (2002). A case study in the integration of assessment and general education: Lessons learned from a complex process. *Assessment & Evaluation in Higher Education*, *27*:2, 199-211. https://doi.org/10.1080/02602930220128760

Sundre, D. L., & Kistantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*, 6-26. https://doi.org/10.1016/S0361-476X(02)00063-2

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2009). Skipping the test: Using empirical evidence to inform policy related to students who avoid taking low-stakes assessments in college. *Journal of General Education*, *58*, 167-195. https://doi.org/10.1353/jge.0.0043

Thelk, A., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale (SOS) to make valid inferences about student performance. *Journal of General Education*, *58*, 129-15. doi:10.1353/jge.0.0047.

U.S. Department of Education. (2006). *A test of leadership*: *Charting the future of American higher education* (Report of the commission appointed by Secretary of Education Margaret Spellings). Washington, DC: Author.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1-17. http://dx.doi.org/10.1207/s15326977ea1001_1

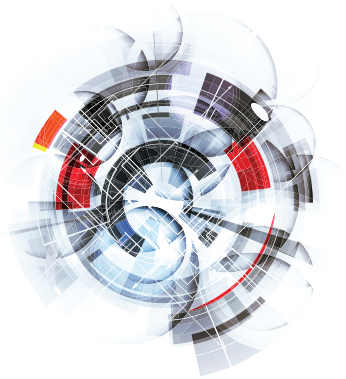
Zhao, A., Brown, G., & Meissel, K. (2020). Manipulating the consequences of tests: How Shanghai teens react to different consequences. *Educational Research and Evaluation*, *26*, 221-251. https://doi.org/10.1080/13803611.2021.1963938

Zilberberg, A., Anderson, R. A., Finney, S. J., & Marsh, K. R. (2013). American college students' attitudes toward institutional accountability testing: Developing measures. *Educational Assessment*, *18*, 208-234. https://doi.org/10.1080/10627197.2013.817153

Zilberberg, A., Anderson, R. A., Swerdzewski, P. J., Finney, S. J., & Marsh, K. R. (2012). Growing up with No Child Left Behind: An initial assessment of the understanding of college students' knowledge of accountability testing. *Research & Practice in Assessment*, *7*, 12-25. http://www.rpajournal.com/dev/wp-content/uploads/2012/07/A1.pdf

Zilberberg, A., Brown, A. R., Harmes, J. C., & Anderson, R. D. (2009). How can we increase student motivation during low-stakes testing? Understanding the student perspective. In D. M. McInerney, G. Brown, & G. Liem (Eds.), *Student perspectives on assessment*: *What students can tell us about assessment for learning* (pp. 255-277). Information Age.

Zilberberg, A., Finney, S. J., Marsh, K. R., & Anderson, R. A. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. *International Journal of Testing*, *14*, 360-384. https://doi.org/10.1080/15305058.2014.928301



Abstract

Integrative learning is an important outcome for graduates of higher education. Therefore, it should be well-defined and assessed reliably. The American Association of Colleges & Universities has developed a rubric to define and assess integrative learning, but it has low reliability. This pilot study examines whether this rubric's reliability can be improved by training users on how to use the rubric in a group setting rather than individually. Twelve faculty were trained to score undergraduate ePortfolios using the Integrative and Applied Learning VALUE Rubric. Half of the faculty were trained in an individual setting and half in a group setting using a popular norming protocol. Results indicate that group training does not improve interrater reliability, though it does improve rater confidence in their rubric scores. Implications include the need for more research comparing individual and group training as well as investigating the efficacy of current training protocols.

AUTHORS

Lanah Stafford, M.A.
Doctoral Student,
Educational Foundations
and Leadership
Old Dominion University

Erin Cousins, MS.Ed.
Doctoral Student,
Educational Foundations
and Leadership
Old Dominion University

Linda Bol, Ph.D.
Professor, Educational
Foundations and Leadership
Old Dominion University

Megan Mize, Ph.D.
Director, ePortfolios
& Digital Initiatives
Old Dominion University

Improving Reliability in Assessing Integrative Learning Using Rubrics: Does Group Norming Help?

CORRESPONDENCE

Email
lstaffor@odu.edu.

Higher education is increasingly focused on ensuring that students are thinking critically, reflecting, synthesizing, applying learning, and developing clear writing skills to succeed in school and the workplace (Demeter et al., 2019; Ferren et al., 2014). Fostering such skills – broadly termed integrative learning – will serve students well as professionals, community members, and lifelong learners (AAC&U, 2009; D'Amico, 2020). One component of effectively fostering skills and improving student learning is the ability to identify clearly its presence or absence (Fulcher et al., 2014). In this case, properly assessing integrative learning requires clear definitions, standards, and processes that improve the reliability of raters to score examples of its demonstration (McClellan, 2010). The American Association of Colleges and Universities (AAC&U) developed a rubric – called the Integrative and Applied Learning Valid Assessment of Learning in Undergraduate Education (VALUE) Rubric – to support universities in these efforts (AAC&U, 2009). However, research on the psychometric properties of this rubric produced low reliability coefficients (i.e., kappa scores), signaling opportunities to improve its reliability (Finley, 2011).

Much of the literature utilizing VALUE rubrics invokes rater calibration (or group norming) as a best practice to improve the reliability of results without empirical evidence to support such a claim (Gray et al., 2017). This pilot study examines whether the Integrative and Applied Learning VALUE Rubric's reliability can be improved by using such a rater calibration process. The following research questions guide our study:

1. To what extent does training among raters in a group versus an individual setting impact the reliability of the Integrative and Applied Learning VALUE Rubric when used to score undergraduate ePortfolios?
2. How confident are raters before and after training in the validity and reliability of their rubric scores using the Integrative and Applied Learning VALUE Rubric to score undergraduate ePortfolios, and does this confidence differ by condition?

Literature Review

Broadly, integrative learning focuses on finding connections between one's gained knowledge and experiences (Reynolds et al., 2014; Gallagher 2019) and using those connections in some manner (Huber & Hutchings, 2004; Reynolds et al., 2014). Connections might be made between "seemingly disparate information" (AAC&U, 2002, p. 21) or among "skills and knowledge from multiple sources and experiences" (Huber & Hutchings, 2004, p. 15). Integrative learning might require "challenging and complex settings" (Green & Hutchings, 2018, p. 42) or "interdisciplinary understanding" (Lardner & Malnarich, 2009, p.32), and be used to "make decisions" (AAC&U, 2002, p. 21) or solve problems (Gallagher, 2019). Or, holistically, integrative learning might simply be considered the "ability to learn across context and over time" (Reynolds et al., 2014, p. 26). The Integrative and Applied Learning VALUE Rubric defines integrative learning as, "an understanding and a disposition that a student builds across the curriculum and co-curriculum, from making simple connections among ideas and experiences to synthesizing and transferring learning to new, complex situations within and beyond the campus" (AAC&U, 2009, p. 1).

ePortfolios are one tool for facilitating and documenting students' developing integrative learning skills (Buyarski & Landis, 2014; Cheng et al., 2015; Yastibas & Yastibas, 2015). ePortfolios are multi-modal collections of electronic evidence that can showcase students' integrative learning, critical thinking, and written communication (Benander et al., 2016; Buyarski & Landis, 2014; Douglas et al., 2019). When designed appropriately, ePortfolios can promote self-directed learning (Beckers et al., 2016) and encourage both student reflection (Dalal et al., 2012; Jenson, 2011) and metacognitive awareness (Kohler & Van Zile-Tamsen, 2020). They have been identified as a high impact practice due to their relation to positive academic outcomes such as improved grades, retention, and graduation (Watson et al., 2016).

At the same time, assessing meaningful integration is a complex endeavor (Huber & Hutchings, 2004), and attempts to assess it soundly vary greatly among institutions (Dawson, 2017; Demeter et al., 2019). Rubrics are one tool for accomplishing this task. They involve specific, defined criteria for evaluation (Dawson, 2017) and their use can increase the transparency of assessment while supporting student self-regulation, self-assessment, and revision through clear standards and formative feedback, often leading to improved achievement and learning (Hattie & Timperley, 2007; Jonsson, 2014; Panadero & Jonsson, 2013; Reddy & Andrade, 2010). With heightened focus on interrater reliability, calibration, and norming in higher education (Reddy & Andrade, 2010; Schoepp et al., 2018), rubrics that are standardized and applied consistently across raters can provide scores that are reliable and trustworthy to stakeholders (Fulcher & Orem, 2010; McCellan, 2010; Schoepp et al., 2018).

At the same time, assessing meaningful integration is a complex endeavor, and attempts to assess it soundly vary greatly among institutions.

Not only does reliability contribute to providing trustworthy scores to stakeholders, the perceived reliability of evidence can also influence one's confidence in making decisions (Boldt et al., 2017). In addition, previous experiences can shape one's confidence; this can, in turn, prepare one for making future decisions (Boldt et al., 2019). This confidence, in turn, can play an active role in both learning and performance by influencing one's motivation and subsequent behaviors (Hainguerlot et al., 2018; Rouault et al., 2019). As it relates to this study, confidence to assign rubric scores is important not only for faculty raters, but also for students to trust that their scores were confidently assigned (O'Connell et al., 2016). Thus, providing reliable evidence might improve the confidence in raters to assign scores to student artifacts, the confidence in students to respect the validity of these scores, and the confidence in institutional personnel to hold university-wide discussions about the state of student learning as identified through these scoring efforts. This, then, might support adaptive behaviors at the student, faculty, and institutional level to improve student learning.

Training to score with rubrics can improve raters' ability to interpret scoring items reliably and improve interrater reliability beyond practice or previous experience with the rubric.

One opportunity to improve a rubric's reliability is through training. Training to score with rubrics can improve raters' ability to interpret scoring items reliably (Stuhlmann et al., 1999) and improve interrater reliability beyond practice or previous experience with the rubric (Attali, 2016). Some scholars propose that interactive or collaborative group training can improve interrater reliability of rubric scoring by allowing raters to develop a shared understanding of rubric dimensions and performance criteria through group discussion and peer feedback (Cole et al., 2012; Finley, 2011; Stuhlmann, et al., 1999; Weigle, 1999). Cole et al. (2012) employed collaborative group training under the assumption that "group discussion and problem solving" fostered shared understanding of rubric criteria (p. 4). The Educational Testing Service considers group norming a best practice in training raters to score constructed-response items (McClellan, 2010). These assertions are supported by encouraging results (Cole et al., 2012; Marshall et al., 2017). Existing studies have demonstrated improved reliability as a result of collaborative group training (Cole et al., 2012; O'Connell et al., 2016; Marshall et al., 2017), as well as improved individual rater confidence (Marshall et al., 2017; O'Connell et al., 2016). Marshall et al.'s (2017) results demonstrated that collaborative group training increased faculty confidence in assessing ePortfolios using an institutionally developed rubric and O'Connell et al. (2016) reported that raters' confidence increased following a collaborative group workshop.

Method

This study employed a true experimental design. Half of the participants were randomly assigned to the group training while the other half of the participants were randomly assigned to the individual training condition. Participant names were entered into an Excel spreadsheet, assigned a random number, and then sorted by that random number. The first six names were assigned to the group training and the last six names were assigned to the individual training. The study was reviewed and approved by the university's institutional review board committee.

Participants

A convenience sample of 12 participants was recruited from a larger pool of faculty who had been trained to teach integrative learning, demonstrated in an ePortfolio. Specifically, these faculty were trained to teach integrative learning as defined by the Integrative and Applied Learning VALUE Rubric. Participants were recruited by email announcement from one of the authors who leads these training efforts at their institution. A $250 stipend was offered as compensation for completing the study. Institutional data were used to identify important characteristics of these participants who varied in rank (tenured, tenure track, and non-tenured instructors) and discipline. A table illustrating the number of faculty from various departments across each condition is included in Appendix A. Although the goal was to represent proportionally the total population of trained faculty, the convenience sample included an overrepresentation of faculty from the English department.

Other faculty from this same pool of trained instructors were recruited to submit their undergraduate students' ePortfolios for use in this study. All faculty trained in integrative learning were required to implement an ePortfolio in at least one of their courses following training and submit these assignments to an ePortfolio repository. Notification letters were distributed to all students enrolled in the courses taught by these faculty volunteers with an option to opt out of the study. Of the resulting pool of ePortfolios, 30 were randomly selected for inclusion. They represented multiple disciplines such as Biology, Communications, and Mechanical Engineering Technology at the 200-, 300-, and 400-levels. Content included semester-long projects, individual assignments, and reflective prompts. All were created in WordPress or Wix. Twenty student ePortfolios were assigned to the experimental and comparison groups, respectively, with 10 that overlapped across groups. No rater reviewed work produced by a student in his/her course.

Procedure

The experimental group followed a procedure outlined by many popular group training protocols (Rhode Island Department of Education, n.d.; Stanford Center for

Assessment, Learning, & Equity, 2017; Virginia Department of Education, 2019). Participants in the experimental group engaged in a three-hour group discussion facilitated by the lead author. First, the participants jointly reviewed the rubric, defining and discussing the criteria and corresponding levels of performance. Then, raters independently scored three practice ePortfolios, describing their ratings and reasoning/evidence to support these ratings with the group between each round.

Participants in the individual condition followed the procedure outlined by Finley (2011). Raters reviewed the rubric in a one-on-one session with the lead author, defining and discussing the criteria and corresponding levels of performance. After reviewing the rubric, raters scored three practice ePortfolios, asking follow-up questions about the rubric or its application between rounds. Each session was allotted three hours, though actual duration varied from one to two-and-a-half hours.

After training, raters in both groups received their assignment of 20 ePortfolios and rated them independently over two weeks. Scores were submitted electronically with identification numbers assigned to both raters and ePortfolios. The lead author verified that all ePortfolios received scores from their assigned raters.

Measures

Rubric Scores

The Integrative and Applied Learning VALUE Rubric (AAC&U, 2009) provides a definition of integrative learning, additional context about integrative learning and higher education, a glossary of key terms, and the dimensions, performance levels, and descriptors for each performance level. This rubric categorizes integrative learning into five dimensions: (1) connections to discipline, (2) connections to experience, (3) transfer, (4) integrated communication, and (5) reflection and self-assessment. There are four progressive levels of performance per dimension: 1-Benchmark (lowest performance level), 2- and 3-Milestones, and 4-Capstone (highest level of performance). The rubric additionally encourages raters to assign a score of 0 to any dimension in which the student artifact does not reach the level of the 1-Benchmark criteria. In this study, the score of 0 was also used if the rater determined that the ePortfolio was missing the evidence needed to make a scoring decision.

AAC&U has determined that the rubric has face and content validity due to its development by national teams of interdisciplinary faculty experts. Reliability indices that include the percent agreement and kappa scores are also included in Appendix B (Finley, 2011).

The Integrative and Applied Learning VALUE Rubric provides a definition of integrative learning, additional context about integrative learning and higher education, a glossary of key terms, and the dimensions, performance levels, and descriptors for each performance level.

Confidence

Confidence was determined by having participants predict and postdict their rating accuracy and alignment with peers. Following training but prior to receiving their assignments, participants responded to two prediction questions: 1) How confident are you that you will give valid ratings on these ePortfolios?, 2) How confident are you that your scores will align with other raters? Response options were: 1-Not at all confident, 2-Slightly confident, 3-Moderately confident, and 4-Very confident. After completing their assignments, participants were asked the same two questions using the same scale.

Analyses

Analyses were conducted using 10 ePortfolios which were scored by six raters who had been trained in an individual setting and six raters who had been trained in a group setting. In alignment with Finley (2011), each analysis was run using the original five-point scoring scale of 0-4, a collapsed four-point scale, and a further collapsed three-point scale. To collapse from five to four points, the mean, median, and mode scores were calculated within each rubric category. These calculated values were used to determine which rating scores would be combined. In instances in which all three values were the same, rating frequencies were used to make consolidation decisions. This process was replicated to collapse from four to three points for analyses, again in alignment with Finley (2011). Interrater reliability was

determined for both groups by calculating percentage agreement and Randolph's (2005) free-marginal multi-rater kappa using the 10 ePortfolios which were scored by all raters. This was calculated for overall scores as well as for individual student learning outcomes (SLOs). Both percentage agreement and multi-rater kappa scores were reported in Finley (2011), allowing for direct comparisons.

Findings for percent agreement and multi-rater kappa for individual trained raters were remarkably similar to those reported by Finley.

Percentage agreement represents the percentage of cases that raters agreed upon determined by dividing the number of agreed upon cases by the total number of cases (Allen, 2017). This statistic is simple to interpret but does not address the probability of raters agreeing by chance and, therefore, is not a comprehensive representation of reliability (Fleiss, 1981). Randolph's (2005) multi-rater kappa takes into consideration the likelihood that raters agreed by chance, making it more comprehensive than percent agreement. Randolph's (2005) multi-rater kappa was selected for this study, given that "raters' distributions of cases into categories are not restricted" and because the raters were non-unique; the same 12 raters graded each of the ePortfolios (Randolph, 2005, p. 2). In line with other reliability coefficients, this multi-rater kappa can range in value from -1 to 1, with values of 0 representing agreement which is equal to chance and values of 1 representing perfect agreement beyond chance (Randolph, 2005).

Due to the small sample sizes, raters' confidence in the validity and reliability of their rubric scores before and after training was analyzed for each condition using the Mann-Whitney U test and Wilcoxon Signed-Rank test. The Mann-Whitney U test is a non-parametric alternative to the t test of independent samples (Salkind & Frey, 2020) and was used to compare the pre- and post-confidence of raters. The Wilcoxon Signed-Rank test is a non-parametric alternative to the t test of dependent samples and was used to compare the changes between pre- and post-training confidence of raters for both training groups (Salkind & Frey, 2020).

Results

Interrater Reliability

Results for interrater reliability analyses can be found in Table 1. Expectedly, inter-rater reliability improved as rubric scores were collapsed into fewer categories. Findings for percent agreement and multi-rater kappa for individual trained raters were remarkably similar to those reported by Finley (2011). Individually trained raters achieved a percent agreement value of 73.47% and a kappa score of .60 for the collapsed 3-point score category. In contrast to our hypothesis, raters trained in a group setting had lower interrater reliability across all measures compared to raters trained individually. Group trained raters achieved a percent agreement value of 67.33% and a kappa score of .51 for analyses of 3-point scoring categories.

Table 1
Reliability Results - Comparing Reliability for 10 Overlapping Rubrics Scored by Both Individually and Group Trained Raters

	Perfect Agreement (Original 5 categories)	Approximate Agreement (Using 4 categories)	Approximate Agreement (Using 3 categories)
Percentage of agreement - individually trained raters / group trained raters	29.60 / 23.73	51.33 / 44.67	73.47 / 67.33
Randolph's multi-rater kappa* score - individual trained raters / group trained raters	.12 / .05	.35 / .26	.60 / .51

*Interpreted like other reliability coefficients with 0 indicating no agreement and 1 indicating perfect agreement

Percent agreement and multi-rater kappa were also calculated for each dimension of integrative learning as defined by the rubric; these results are available in Table 2. Again, in contrast to our hypothesis, individually-trained raters had greater agreement than group-trained raters on all dimensions of the rubric for nearly all scoring schemes (5-point, 4-point, and 3-point). The few exceptions were: Integrated Communication when collapsed to a 4-point scoring scale and Transfer when collapsed to 4-point and 3-point scoring scales.

Table 2
Reliability Results - Comparing Reliability for 10 Overlapping Rubrics Scored by Both Individually and Group Trained Raters

		% Agreement	Randolph's multi-rater kappa* score
		Individually Trained / Group Trained	Individually Trained / Group Trained
Perfect Agreement (Original 5 categories)	Connections to Experience	30.00 / 20.67	.12 / .01
	Connections to Discipline	29.33 / 24.00	.12 / .05
	Transfer	30.00 / 26.00	.12 / .07
	Integrated Communication	28.00 / 24.67	.10 / .06
	Reflection/Self-Assessment	30.67 / 23.33	.13 / .04
Approximate Agreement (Using 4 categories)	Connections to Experience	50.67 / 42.67	.34 / .24
	Connections to Discipline	58.67 / 44.67	.45 / .26
	Transfer	58.00 / 40.67	.44 / .21
	Integrated Communication	49.33 / 54.00	.32 / .39
	Reflection/Self-Assessment	40.00 / 41.33	.20 / .22
Approximate Agreement (Using 3 categories)	Connections to Experience	75.33 / 68.00	.63 / .52
	Connections to Discipline	82.00 / 71.33	.73 / .57
	Transfer	76.00 / 60.00	.64 / .40
	Integrated Communication	79.33 / 76.67	.69 / .65
	Reflection/Self-Assessment	54.67 / 60.67	.32 / .41

Confidence

Mann-Whitney U test results showed that confidence about the predicted validity of raters' scores was greater for individually-trained raters (M=3.33) than for group-trained raters, but that this difference did not reach statistical significance (M=3.20, U=13.00, p=.792). Mann-Whitney U results also showed that confidence about the predicted alignment of raters' scores with one another was greater for individually-trained raters (M=3.17) than for group-trained raters, but that this difference did not reach statistical significance (M=2.80, U=16.00, p = .818). However, individually-trained raters reported lower post-scoring confidence in the validity (Mdn=3.00) of their rubric scores than group-trained raters (Mdn=3.00, U=16.00, p=.818) and equal post-scoring confidence in the alignment of rubric scores (Mdn=3.00, Mdn=3.00, U=18.00, p=1.00), though neither of these differences reached levels of statistical significance.

In contrast to our hypothesis, individually-trained raters had greater agreement than group trained raters on all dimensions of the rubric for nearly all scoring schemes.

Raters' confidence scores were compared before and after rating ePortfolios. The Wilcoxon Signed-Rank test was used to analyze the data. There were no significant differences from pre- to post-confidence scores within either group. The descriptive results are presented in Table 3. Though these differences did not reach statistical significance, they do reflect an interaction effect as shown in Figures 1 and 2.

Table 3
Comparison of Individual and Group Training Means on Confidence to Provide Valid Rubric Scores

	Individually-Trained Raters	Group-Trained Raters
Pre-Validity	Mean: 3.33 Std. Dev.: .52	Mean: 3.20 Std. Dev.: .45
Pre-Alignment	Mean: 3.17 Std. Dev.: .75	Mean: 2.80 Std. Dev.: .45
Post-Validity	Mean: 3.00 Std. Dev.: .89	Mean: 3.20 Std. Dev.: .45
Post-Alignment	Mean: 3.00 Std. Dev.: .89	Mean: 3.00 Std. Dev.: .00

Figure 1
Changes in Validity Confidence Pre- and Post-Rubric Scoring by Treatment

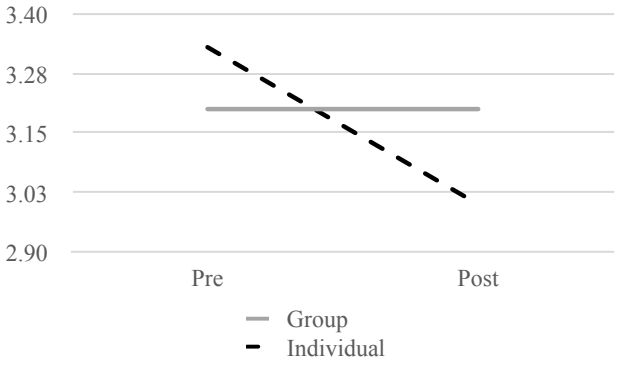
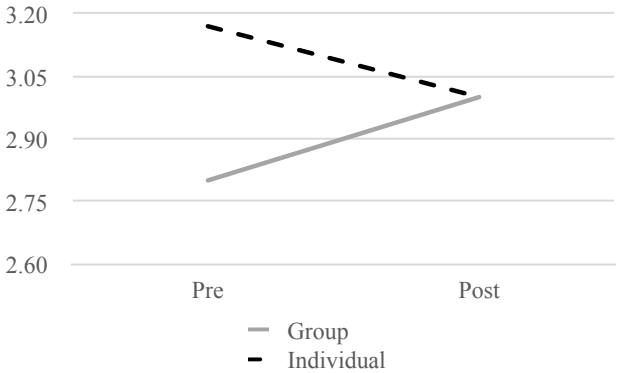


Figure 2
Changes in Reliability Confidence Pre- and Post-Rubric Scoring by Treatment



Discussion

It was expected that collaborative group training would improve interrater reliability beyond the levels produced in Finley (2011), as well as those produced by the raters in the individually-trained condition in this study. However, few results from this study indicated

improved interrater reliability for group-trained raters. Rather, interrater reliability for those trained in a group setting was slightly lower across nearly all analyses compared to individually-trained raters. The only instances in which group-trained raters were more reliable than those trained individually were when reliability was examined for specific dimensions of integrative learning. At the more focused level, group-trained raters had greater reliability in scoring for two individual dimensions, but only at certain levels of collapsed scoring. Group-trained raters never had stronger reliability for a dimension of integrative learning when scores were left at the original five scoring categories.

There are a few plausible explanations for this finding. One explanation may be that the group training provided was insufficient in some regard. Perhaps there is a minimal threshold for effective group training and a single session with three practice samples is not enough practice to truly norm the rubric. When collapsing scales, the two most frequent scores were combined into one point. As Finley (2011) explained: "in practicality, when working with faculty on campuses it is often not assumed that 'perfect agreement' is necessary. It is assumed, rather, that close scores also count as agreement" (para. 8). While the scores combined for the individually-trained raters were largely the 2-Milestone and 3-Milestone scores, the scores combined for the group training were entirely 1-Benchmark and 2-Milestone. In this way, the most frequent disagreement among raters in the individual training condition was on *which* Milestone level of performance was achieved, while group-trained raters could not agree on whether a student ePortfolio even reached the Milestone level of performance.

The only instances in which group-trained raters were more reliable than those trained individually were when reliability was examined for specific dimensions of integrative learning.

Another explanation may be that there were significant differences between the two groups prior to training. This explanation is supported by the greater disciplinary diversity among faculty members in the group training condition compared to those in the individual training condition. Because five of the six raters trained in the individual condition came from the English department, it is possible that these raters possessed more consistent disciplinary training to make scoring decisions and therefore were more reliable in their scoring.

Finally, it is possible that the results may stem from limitations with the Integrative and Applied Learning VALUE Rubric itself. The faculty recruited to serve as raters were familiar with the concept of integrative learning, the demonstration of integrative learning via student ePortfolio, and the Integrative and Applied Learning VALUE Rubric. Yet comments made during both the group and individual training sessions highlighted problematic statements within the rubric, such as key words which served to delineate performance levels (ex. "in a basic way") and how that was operationalized in practice. Such descriptors leave room for interpretation and, therefore, contribute to more error and lower reliability results.

Other authors have reported similar results regarding a lack of improved interrater reliability with collaborative group training. Knoch et al. (2007) reported mixed results when comparing self-paced and collaborative group training methods for scoring a direct writing assessment with a rubric. Raczynski et al. (2015) reported that the reliability of raters trained in a collaborative group setting was not significantly different from those trained individually to score essays using a rubric. While each of these studies involved scoring essays, ePortfolios involve a high proportion of written material and are considered a type of digital composition (Cicchino et al, 2019; Clark, 2010; Yancey, 2009) and are therefore an appropriate comparison.

We found the confidence levels for raters did not significantly differ between groups before or after training. However, the results showed an interaction pattern. Raters in the group-trained condition began their scoring with less confidence than raters in the individually-trained condition. It may be that the normative group activities led to lower initial confidence due to public disagreement among raters' scores. When trained individually, the lead author did not compare the raters' training scores against any kind of anchor score. Discussions about the scores given were couched solely within the context of the rubric language and its specific application to the ePortfolio being scored. This is juxtaposed with the public score comparisons made in the group-trained condition. Although the discussion in the group-trained condition was likewise couched within the context of the rubric and its application, the experience of producing differing scores introduced unreliable evidence about raters' ability to score this work; evidence that was not present in the individually-trained condition. Per Boldt et al. (2017; 2019), this could contribute to lower predicted confidence in undertaking the scoring task.

Participants may also have questioned their ability to provide ratings that aligned with their peers because there were social comparisons (Hacker & Bol, 2004). Though not measured as part of this study, it is possible that raters in the group-trained condition might have lost some motivation to persist with the scoring task (Rouault et al., 2019).

After scoring all ePortfolios, the confidence of raters in the group-trained condition increased notably. This may be attributed to the similarities among the sample ePortfolios selected for calibration and those scored for record. When limited to their experiences with the group training, the raters only experienced evidence of unreliability due to the differing scores each rater assigned to the respective ePortfolios. However, scoring 20 ePortfolios that were similar to their training experiences introduced evidence of reliability of their skill. This might have introduced new evidence of the reliability of the rubric and their training, thus contributing to this improved confidence (Boldt et al., 2017). This would explain why confidence increased for the group-trained raters but remained constant for the individually-trained raters. Finally, because scoring for record was an individual experience, there could have been a diminishing effect of social comparisons between the start and the end of the scoring process.

The finding that increased confidence in the reliability of instructors' scoring did not directly align with improved inter-rater reliability is somewhat counterintuitive. One would predict a positive correlation between these variables. However, a durable phenomenon in the literature is the negative relationship between overconfidence in performance and performance itself. That is, the lowest performing individuals tend to be overconfident and the highest performing students are much more accurate in their predictions of performance (Hacker et al., 2000). The relationship is diminished as individuals become more competent at a task (Hacker & Bol, 2019). It seems plausible that as raters become more reliable with extended training, their confidence would more precisely reflect the accuracy of their judgments.

Implications and Future Directions

The explanations outlined above align with facets highlighted in generalizability studies conducted on other VALUE rubrics (Pike, 2018; Pike & McConnell, 2018). As Pike (2018) reported, variation across raters and assignments were the largest two sources of error. Future generalizability studies on the Integrative & Applied Learning VALUE Rubric might identify if other facets contribute to the variance in results and, if so, to what extent. At present, recommended actions to improve the dependability of other VALUE rubrics include enhancing rater training, aligning assignments, and modifying the rubrics themselves (Pike & McConnell, 2018). This study contributes additional empirical evidence in support of these actions and extends them to the assessment of integrative learning via the Integrative & Applied Learning VALUE Rubric.

The finding that increased confidence in the reliability of instructors' scoring did not directly align with improved inter-rater reliability is somewhat counterintuitive.

Because some of the differences observed between group- and individually-trained raters may have been influenced by the instructors' discipline, training within disciplines may improve its effectiveness. For example, it may not make good sense for an English scholar to review and score an engineering portfolio. Discipline-based training may increase both the reliability and validity of rubric scores. This strategy would afford comparisons both within and between disciplines to potentially uncover an interaction between group versus individual training contexts and subject areas. That is, group training may be more effective in some disciplines compared to others.

The process followed in this study aligns with the process and duration of popular group training protocols (Rhode Island Department of Education, n.d.; Stanford Center for Assessment, Learning, & Equity, 2017; Virginia Department of Education, 2019), yet the reliability of raters trained in this manner was worse than those trained individually. The present reliability results call for cautiousness in espousing the benefits of these protocols. Additional research is needed to investigate the reliability of such group training protocols – particularly for applications of VALUE rubrics to student work (Gray, et al., 2017).

As Pike (2018) states, "there is no substitute for well-trained raters" (p. 9). It is plausible that collaborative group training would be more successful when increased in duration and activities, such as the two-day training institute employed by Marshall et al. (2017). If choosing

to move forward with using multiple raters to review student work, future practitioners might consider extending the duration of training and/or introducing anchor papers (Pike, 2018) to ground rater scores. In practicality, however, these results suggest that institutional trainers may be able to leverage individual rater training sessions to increase its available rater pool beyond those who have availability to attend synchronous group training sessions.

While rubrics may be beneficial by providing students with clear performance expectations and potentially support self-regulation, students should also be able to trust in the reliability of the scores given to them. Given the present results, it bears repeating AAC&U's directive that the Integrative and Applied Learning VALUE Rubric is not appropriate for grading individual student's assignments (AAC&U, 2009). However, the limitations of the reliability of this instrument are also relevant to institutional-level uses. Even at the institutional level, the use of this rubric could be high-stakes for students who will be the recipients of any pedagogical or programmatic alterations that may occur as a result of the data produced from such work. This reinforces the need for reliability in order to avoid unsupported decisions which could ultimately have a negative impact on students.

Even at the institutional level, the use of this rubric could be high-stakes for students who will be the recipients of any pedagogical or programmatic alterations that may occur as a result of the data produced from such work.

Furthermore, institutions and/or individual faculty may choose to ignore AAC&U's directives. The use of rubrics in higher education, particularly with their integration into learning management systems, can vary widely and lead to some institutions mandating their use (Dawson, 2017). The proliferation of interest in VALUE rubrics across individuals, organizations, and colleges and universities (Pike & McConnell, 2018) offers insight into the potential for both use and misuse. Although that is not the fault (nor intention) of this rubric as it is designed, it may be an outcome. Therefore, it remains important that its reliability be improved as much as possible.

Per Finley (2011), the standard interpretation of a high or acceptable kappa score is 0.70. Only one finding – the reliability of the individually trained raters on the Connections to Discipline dimension at the 3-point score level – achieved this threshold. This is especially problematic when considering that the collapsed 3-point scale was: 0-Missing/Unable to determine, 1/2/3, 4-Capstone. Combining the scales in this way may have improved the quantitative reliability to a high level, but qualitatively, the scores are meaningless. Pike (2018) and Pike & McConnell's (2018) potential solution for improving the reliability of the VALUE rubrics by utilizing better assignment design to align more explicitly to the criteria of the rubric proved unsuccessful in this study, as the ePortfolios rated were designed to align with the Integrative and Applied Learning VALUE Rubric. Changes to the Integrative and Applied Learning VALUE Rubric are needed and are already underway (Pike & McConnell, 2018).

Perfectly reliable assessment tools are not sufficient alone to instigate widespread improvements to learning (Eubanks et al., 2021). Yet methodologically sound assessment designs remain an integral piece of the puzzle. Integrative learning will continue to serve as a driving goal of a college education in at least the near future. The faculty empowerment and professional development needed to spur larger gains in integrative learning must rest soundly on a foundation of reliable assessments of its demonstration. This requires a rubric with clearly and appropriately defined criteria which can be applied reliably across raters and student work. The present study investigated whether the reliability of AAC&U's Integrative and Applied Learning VALUE Rubric could be improved when human raters were trained in collaborative group settings. Although the findings did not support our hypotheses, they contribute empirical evidence to the literature on group training, interrater reliability, and the application of nationally-normed rubrics to locally-designed ePortfolios.

References

Allen, M. (Ed.) (2017). *The sage encyclopedia of communication research methods* (Vols. 1-4). SAGE Publications, Inc. http://dx.doi.org/10.4135/9781483381411

American Association of Colleges & Universities (AAC&U). (2002). Greater expectations: A new vision for learning as a nation goes to college. Association of American Colleges and Universities. https://files.eric.ed.gov/fulltext/ED468787.pdf

American Association of Colleges and Universities (AAC&U). (2009). *Integrative and applied learning VALUE rubric*. https://www.aacu.org/initiatives/value-initiative/value-rubrics/value-rubrics-integrative-and-applied-learning

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99-115. https://doi.org/10.1177/0265532215582283

Beckers, J., Dolmans, D., & van Merrienboer, J. (2016). e-Portfolios enhancing students' self-directed learning: A systematic review of influencing factors. *Australasian Journal of Educational Technology*, *32*(2), 32-46. https://doi.org/10.14742/ajet.2528

Benander, R., Robles, R., Brawn, D., & Refaei, B. (2016). Assessment without standardization: Can general education competencies be assessed from ePortfolios across the university? *The Journal for Research and Practice in College Teaching 1*(1), 1-10. https://journals.uc.edu/index.php/jrpct/article/view/618

Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(8). 1520-1531. http://dx.doi.org/10.1037/xhp0000404

Boldt, A., Schiffer, A., Waszak, F., & Yeung, N. (2019). Confidence predictions affect performance confidence and neural preparation in perceptual decision making. *Scientific Reports*, *9*(1). 4031. https://doi.org/10.1038/s41598-019-40681-9

Buyarski, C. A., & Landis, C. M. (2014). Using an ePortfolio to assess the outcomes of a first-year seminar: Student narrative and authentic assessment. *International Journal of ePortfolio*, *4*(1). 49-60. http://www.theijep.com/pdf/ijep133.pdf

Cheng, S.-I., Chen, S.-C., & Yen, D. C. (2015). Continuance intention of E-portfolio system: A confirmatory and multigroup invariance analysis of technology acceptance model. *Computer Standards & Interfaces*, *42*, 17-23. http://dx.doi.org/10.1016/j.csi.2015.03.002

Cicchino, A., Efstathion, R., & Giarrusso, C. (2019). Revisualizing the composition process. In K. Yancey (Ed), *ePortfolio as Curriculum: Models and practices for developing students' ePortfolio literacy* (pp. 13-29). Stylus Publishing. https://ebookcentral.proquest.com/lib/odu/detail.action?docID=5747189

Clark, J. E. (2010). The digital imperative: Making the case for a 21st century pedagogy. *Computers and Composition*, *27*, 27-35. https://files.eric.ed.gov/fulltext/EJ1120704.pdf

Cole, T. L., Cochran, L., & Troboy, K. (2012). Efficiency in assessment: Can trained student interns rate essays as well as faculty members? *International Journal for the Scholarship of Teaching and Learning*, *6*(2), 1-11. https://doi.org/10.20429/ijsotl.2012.060206

Dalal, D. K., Haekl, M. D., Sliter, M. T., & Kirkendall, S. R. (2012). Analysis of a rubric for assessing depth of classroom reflections. *International Journal of ePortfolio*, *2*(1), 75-85. http://www.theijep.com/pdf/IJEP11.pdf

D'Amico, C. (2020, February 23). How to increase consumer confidence in higher education. *Forbes*. https://www.forbes.com/sites/stradaeducationnetwork/2020/02/23/how-to-increase-consumer-confidence-in-higher-education/?sh=ff7769d36d0c

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment and Evaluation in Higher Education*, *42*(3), 347-360. https://doi.org/10.1080/02602938.2015.1111294

Demeter, E., Robinson, C., & Frederick, J. G. (2019). Holistically assessing critical thinking and written communication learning outcomes with direct and indirect measures. *Research & Practice in Assessment*, *14*(1), 41-51. https://www.rpajournal.com/dev/wp-content/uploads/2019/07/A3.pdf

Douglas, M. E., Peecksen, S., Rogers, J., & Simmons, M. (2019). College students' motivation and confidence for ePortfolio use. *International Journal of ePortfolio*, *9*(1), 1-16. https://www.theijep.com/pdf/IJEP316.pdf

Eubanks, D., Fulcher, K., & Good, M. (2021). The next ten years: The future of assessment practice? *Research & Practice in Assessment*, *16*(1), 1-6. https://www.rpajournal.com/dev/wp-content/uploads/2021/03/The-Future-of-Assessment-Practice.pdf

Ferren, A., Anderson, C., & Hovland, K. (2014). Interrogating integrative learning. *Peer Review: Emerging Trends and Key Debates in Undergraduate Education*, *16*(4), 4. https://www.aacu.org/peerreview/2014-2015/fall-winter/ferren

Finley, A. (2011). How reliable are the VALUE rubrics? *Peer Review*, *13*(4). 31-33. https://www.aacu.org/publications-research/periodicals/how-reliable-are-value-rubrics

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. Wiley. https://doi.org/10.1002/0471445428

Fulcher, K. H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014, December). A simple model for learning improvement: Weigh pig, feed pig, weigh pig. (Occasional Paper No. 23). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://files.eric.ed.gov/fulltext/ED555526.pdf

Fulcher, K. H., & Orem, C. D. (2010). Evolving from quantity to quality: A new yardstick for assessment. *Research & Practice in Assessment*, *5*, 13-17. http://www.rpajournal.com/dev/wp-content/uploads/2012/05/A25.pdf

Gallagher, C. W. (2019). College made whole: Integrative learning for a divided world. ProQuest Ebook Central. https://ebookcentral.proquest.com/lib/odu/detail.action?docID=5880856.

Gray, J. S., Brown, M. A., & Connolly, J. P. (2017). Examining construct validity of the quantitative literacy VALUE Rubric in college-level STEM assignments. *Research & Practice in Assessment*, *12*, 20-31. http://www.rpajournal.com/dev/wp-content/uploads/2017/07/A2.pdf

Green, K., & Hutchings, P. (2018). Faculty engagement with integrative assignment design: Connecting teaching and assessment. *New Directions for Teaching and Learning*, *2018*(155). 39-46. https://doi.org/10.1002/tl.20301

Hacker, D. J., & Bol, L. (2004). Metacognitive theory: Considering the social influences (pp. 275-297). In S. Van Etten & D. McInerny (Eds.), *Research on sociocultural influences on motivation and learning*. *Volume 4, Big Theories Revisited*. Information Age Press. ISBN: 1593110537

Hacker, D. J., & Bol, L. (2019). Calibration and self-regulated learning: Making the connections (pp. 647-677). In J. Dunlosky & K. Rawson (Eds.), *Cambridge handbook on cognition and education*. Cambridge University Press. ISBN: 1108416012

Hacker, D. J., Bol, L., Horgan, D., & Rakow, E. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*, 160-170. https://doi.org/10.1037/0022-0663.92.1.160

Hainguerlot, M., Vergnaud, J., & deGardelle, V. (2018). Metacognitive ability predicts learning cue-stimulus associations in the absence of external feedback. *Scientific Reports*, *8*(1). 5602-5608. https://doi.org/10.1038/s41598-018-23936-9

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112. http://www.jstor.org/stable/4624888

Huber, M. T., & Hutchings, P. (2004). *Integrative learning: Mapping the terrain*. Association of American Colleges and Universities. https://files.eric.ed.gov/fulltext/ED486247.pdf

Jenson, J. D. (2011). Promoting self-regulation and critical reflection through writing students' use of electronic portfolio. *International Journal of ePortfolio*, *1*(1), 49-60. https://eric.ed.gov/?id=EJ1107586

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment and Evaluation in Higher Education*, *39*(7), 840-852. https://doi.org/10.1080/02602938.2013.875117

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26-43. https://doi.org/10.1016/j.asw.2007.04.001

Kohler, J. J., & Van Zile-Tamsen, C. (2020). Metacognitive matters: Assessing the high-impact practice of a general education capstone ePortfolio. *International Journal of ePortfolio*, *10*(1), 33-43. https://www.theijep.com/pdf/IJEP331.pdf

Lardner, E., & Malnarich, G. (2009). When faculty assess integrative learning: Faculty inquiry to improve learning community practice. *Change (New Rochelle, N.Y.)*, *41*(5), 28-35. https://www.jstor.org/stable/20696178

Marshall, M. J., Duffy, A. M., Powell, S., & Bartlett, L. E. (2017). ePortfolio assessment as faculty development: Gathering reliable data and increasing faculty confidence. *International Journal of ePortfolio*, *7*(2), 187-215. https://www.theijep.com/pdf/IJEP267.pdf

McClellan, C. A. (2010). Constructed-response scoring - Doing it right. *R&D Connections*, *13*, 1-7. http://www.ets.org/Media/Research/pdf/RD_Connections13.pdf

O'Connell, B., De Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B., & Watty, K. (2016). Does calibration reduce variability in the assessment of accounting learning outcomes? *Assessment and Evaluation in Higher Education*, *41*(3), 331-349. https://doi.org/10.1080/02602938.2015.1008398

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129-144. https://doi.org/10.1016/j.edurev.2013.01.002

Pike, G. (2018). Improving the dependability of constructed-response assessment: Lessons from an evaluation of the VALUE rubrics. *Assessment Update*, *30*(5), 8-9. https://doi.org/10.1002/au.30147

Pike, G., & McConnell, K. (2018). The dependability of VALUE Scores: Lessons learned and future directions. *Peer Review: Emerging Trends and Key Debates in Undergraduate Education*, *20*(4), 22-25. https://d38xzozy36dxrv.cloudfront.net/qa/content/magazines/PR_FA18_Vol20No4.pdf

Raczynski, K. R., Cohen, A. S., Engelhard Jr., G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, *52*(3), 301-318. https://doi.org/10.1111/jedm.12079

Randolph, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss´ fixed-marginal multirater kappa. *Paper presented at the Joensuu University Learning and Instruction Symposium*. https://eric.ed.gov/?id=ED490661

Randolph, J. J. (2008). *Online kappa calculator*. http://justus.randolph.name/kappa

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4). https://doi.org./10.1080/02602930902862859

Reynolds, C., Patton, J., & Rhodes, T. (2014). *Leveraging the ePortfolio for integrative learning [e-book]: A faculty guide to classroom practices for transforming student learning*. Stylus Publishing. https://ebookcentral.proquest.com/lib/odu/detail.action?docID=3037636

Rhode Island Department of Education. (n.d.) *Calibration Protocol for Scoring Student Work*. https://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Online-Modules/Calibration_Protocol_for_Scoring_Student_Work.pdf

Rouault, M., Dayabn, P., & Fleming, S. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, *10*(1), 1141. https://doi.org/10.1038/s41467-019-09075-3

Salkind, N. J., & Frey, B. B. (2020). *Statistics for people who (think they) hate statistics (7th Edition)*. SAGE Publications, Inc. ISBN-13: 978-1544381855, ISBN-10: 1544381859

Schoepp, K., Danaher, M., & Kranov, A. A. (2018). An effective rubric norming process. *Practical Assessment, Research, and Evaluation*, *23*(11), 1-12. https://doi.org/10.7275/z3gm-fp34

Stanford Center for Assessment, Learning, & Equity (SCALE). (2017). *Semi-structured calibration activity protocol*. https://www.performanceassessmentresourcebank.org/sites/default/files/addendum/spring2017/SCALE_Semi_Structured_Calibration_Protocol.pdf

Stuhlmann, J., Daniel, C., Dellinger, A., Denny, R. K., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Reading Psychology*, *20*(2), 107-127. https://doi.org/10.1080/027027199278439

Virginia Department of Education (2019). *Calibration Protocol*. https://www.doe.virginia.gov/instruction/mathematics/professional_development/institutes/2019/k-2/6a-calibration-protocol.pdf

Watson, C. E., Kuh, G. D., Rhodes, T., Light, T. P., & Chen, H. L. (2016). Editorial: ePortfolios – The eleventh high impact practice. *International Journal of ePortfolio*, *6*(2), 65-69. http://www.theijep.com/pdf/IJEP254.pdf

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2). 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

Yancey, K. (2009). *Writing in the 21st century*. https://cdn.ncte.org/nctefiles/press/yancey_final.pdf

Yastibas, A. E., & Yastibas, G. C. (2015). The use of e-portfolio-based assessment to develop students' self-regulated learning in English language teaching. *Procedia - Social and Behavioral Sciences*, *176*, 3-13. http://dx.doi.org/10.1016/j.sbspro.2015.01.437

Appendix A

Departments Represented by Faculty amongst Conditions

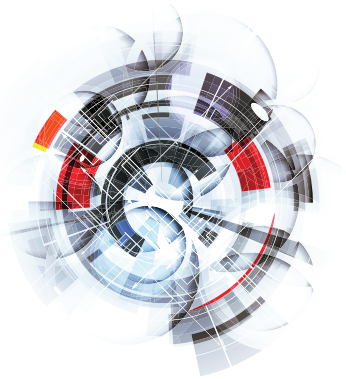
Department	Number of Faculty Raters Assigned to Individually-Trained Condition	Number of Faculty Raters Assigned to Group-Trained Condition
English	5	1
Electrical and Computer Engineering	0	1
STEM Education and Professional Studies	0	1
Teacher Education	1	0
Communication & Theatre Arts	0	1
Psychology	0	1
Political Science	0	1

Appendix B

Reliability of Integrative and Applied Learning VALUE Rubric as Reported in Finley (2011)

	Perfect Agreement (Original 5 categories)	Approximate Agreement (Using 4 categories)	Approximate Agreement (Using 3 categories)
Percentage of Agreement	28% (3%)	49% (8%)	72% (8%)
Kappa Score	0.11 (0.04)	0.31 (0.11)	0.58 (0.11)

Note. Standard deviations are provided in parentheses for each score.



Abstract

General education is an essential feature of American universities and colleges. While many educators value general education, many students do not. Students view general education requirements as interfering with their major requirements, perpetuating a negative sentiment about a disconnect between general education and academic degree programs. We aim to investigate the validity of this story using empirical data collected from robust assessment practice. We examined the extensive assessment records at a liberal arts university to evaluate overlap between the learning outcomes of the general education program and academic degree programs. Findings suggest that the outcomes-oriented general education program was well integrated — nearly half of the academic degree learning outcomes were linked to a general education outcome. Further research should explore if the sentiment of disconnect stems from a lack of implementation fidelity and how findings regarding general education interrelatedness can contribute to the larger conversations and concerns surrounding higher education.

AUTHORS

Yelisey A. Shapovalov, M.A.
James Madison University

Brian C. Leventhal. Ph.D.
James Madison University

CORRESPONDENCE

Email
shapovyx@dukes.jmu.edu

Investigation Of The Alignment Of General Education And Academic Degree Program Learning Outcomes

General education became a mainstay in American institutions of higher education in the early 20th century (Crooks, 1979; O'Banion, 2016). Since then, general education has endured waves of reforms, criticisms, and periods of popularity as well as times of disillusionment and fragmentation. Nonetheless, postsecondary institutions continue to feature general education programs and some continue to experiment with emerging ideas and revisions (O'Banion, 2016).

Defining General Education

General education falls under the larger ideal of a liberal education, "a philosophy of education that empowers individuals, liberates the mind from ignorance, and cultivates social responsibility" as defined by the American Association of Colleges and Universities (AAC&U, 2002). As such, general education has been said to serve multiple purposes, from imparting students with broader knowledge for life (Bastedo, 2002) to developing transdisciplinary skills and values that foster success in academic pursuits and beyond (Glynn et al., 2005; Scott, 2014). General education programs strive to empower students to consider diverse cultures, lifestyles, and backgrounds from well-reasoned and informed perspectives (Glynn et al., 2005). Moreover, these programs aim to educate students on how to grow into responsible, caring members of society (Benander et al., 2000; Melville et al., 2013). AAC&U synthesized the broad goals of a liberal education into a practical definition of general education as the "part of a liberal education curriculum shared by all students. It

provides broad exposure to multiple disciplines and forms the basis for developing important intellectual and civic capacities" (AAC&U, 2002).

The disconnect between educator and student perspectives highlights the need for reform and innovation in general education programs to better serve the needs and expectations of all stakeholders.

To meet their broad goals, universities and colleges can design general education programs in various ways, such as following a more traditional distribution requirement model, adopting a learning outcomes model, or using a hybrid approach. Currently, most general education programs follow the traditional model characterized by requiring students to complete introductory level courses across a range of disciplines from a list of pre-selected options (Bourke et al., 2009). Gump (2007) identified eleven characteristics shared by courses that fit the requirements of general education programs across universities. Several of these characteristics reinforce the introductory nature of general education courses as many are also the initial courses of a major in a discipline. Specifically, these courses are aimed at nonspecialized audiences, assume no background knowledge or skills in the subject area, and thus carry no prerequisites. Moreover, these courses, typically of large class size, emphasize breadth instead of depth and are designed to help students acquire a knowledge base through lectures or practice exercises. Unless integrated in a common core curriculum, such courses carry no inherent expectations for connections with material covered in other courses and tend to stand alone, with learning outcomes that are linked only or primarily to activities carried out as part of the course requirements. Consequently, students can complete these courses at any point in their academic careers, resulting in students considering these courses to be distractions from their interests or academic majors (Gump, 2007).

The traditional model has also been described as a "cafeteria" model and criticized as "a smorgasbord of courses loosely connected to core disciplines from which students must make choices of two or three helpings from a buffet of sometimes a hundred or more offerings" (O'Banion, 2016, p. 332). An alternative, among many suggested general education reform proposals, is adopting a learning outcomes model in which general education programs consist of courses aligned to common program-level learning outcomes. A key feature of this model is the intentional alignment of general education, institutional, and academic program learning outcomes to form a well-integrated liberal education experience (Galle & Galle, 2010). However, an outcomes-based curricular design and distribution model for general education are not mutually exclusive. That is, higher education institutions, particularly larger universities, may design academic program curriculum based on outcomes but rely on a distribution model for their general education program. Regardless of program design, by the time students graduate, they must have satisfied general education requirements by completing courses in writing, mathematics, foreign languages, social science, natural science, and the humanities in addition to their major-related coursework (Stevens, 2001). Although students receive a variety of experiences, educator and student perspectives on the state of general education have not been overwhelmingly positive.

Educator and Student Perspectives

General education programs have been critiqued for being too heavily "supply side" oriented, as in, focused on what the educators consider valuable, rather than attending to the concerns, attitudes, and opinions of the "demand side" or students (Johnston et al, 1991). Many educators value having a general education program (Paulson, 2012) and have a sense of ownership or responsibility for designing and delivering a general education (Beld & Booth, 2010). However, many faculty are also not convinced that the traditional model is effective in achieving general education goals (Paulson, 2012). Mintz (2020) criticizes this dominant form of general education programs for being juvenile and not challenging students as a college program should, thereby influencing student opinion of general education programs as being obsolete.

Indeed, few students find favor in the current state of general education programs and tend to value general education outcomes less than other college experiences (Humphreys & Davenport, 2005). A common theme throughout studies is that students primarily attend university for vocational purposes, meanwhile general education programs are not specifically designed toward a vocation (Abowitz, 2006; Burns, 2020; Astin et al., 1989; Boyer, 1987; Krukowski, 1985; Moffatt, 1989). Some educators are concerned about the increased focus on vocationally oriented education as more higher education institutions are losing a

traditional classification of liberal arts criteria; arguing that "American higher education will be diminished if the number of liberal arts colleges continues to decline" (Baker et al., 2012). Others advocate for the integration of vocational education into higher education, suggesting that general education programming can be complementary to vocationally oriented studies (Abowitz, 2006; Burns, 2020). However, students experience a tension between major requirements and general education requirements — viewing general education requirements as detracting from their major (Humphreys & Davenport, 2005).

Students feel that their general education program was completely disconnected from their major program, which students tend to view as more related to their vocational goals (Humphreys & Davenport, 2005). They report that general education courses taught them nothing that they had not already learned in high school (Peruski, 2005). As a result, more students are opting to take courses to meet general education requirements in the summer or at a community college to save time and money (Thompson et al., 2015). Increasingly, others are completing general education requirements as quickly and cost-efficiently as possible through Advanced Placement (AP) and dual-credit courses available during their secondary education (Felder, 2018). Essentially, students view general education requirements as a waste of time, money, and energy that interferes with their major- or career-oriented goals; students would not take general education courses if they were not required (Humphreys & Davenport, 2005; Mintz, 2020; Peruski, 2005; Thompson et al., 2015). A recent report on graduates between 2007 and 2018 prepared for the State Council of Higher Education for Virginia, found that, at least a quarter of two- and four-year undergraduate students across Virginia were "not at all appreciative" or only "slightly appreciative" of their general education experience, primarily citing four reasons — corroborating what has been commonly found in the literature: (1) not receiving meaningful value, knowledge, or skills from general education courses; (2) not being relevant to vocational goals nor useful for their career; (3) taking time away or detracting from major, concentration, or field of study; and (4) not being worth the cost or time (Survey and Evaluation Research Laboratory, 2021). However, students holding these perspectives may not be well informed.

The same students that hold negative views of general education and prefer general education courses that align with their major are also fairly unfamiliar with the purposes and requirements of their general education program (Thompson et al., 2015). Only a small portion of students report ever receiving a clear explanation of the purpose of general education from an instructor or an advisor (Thompson et al., 2015). Students report simply selecting general education courses that fit their schedule in a preferable manner (Peruski, 2005). Fortunately, faculty are picking up on student perceptions of a disconnect between academic programs and general education programing. Accordingly, most faculty seek to foster the integration of general education programing with departmental majors and make this connection more evident for students (Paulson, 2012).

It is increasingly common for students to view general education requirements as a waste of time, money, and energy that interferes with their major-or career-oriented goals.

While the rhetoric about a disconnect between general education and academic programs has become pervasive, there is a lack of formal inquiries into its basis. Perhaps students are operating under a misconception that general education requirements are just a checklist to get through with no connection to their major. Some researchers have looked at student outcome results regarding their perception of the relevance of their general education coursework (Walters, 2018), and other studies have examined student performance on learning outcomes with measures of interest, such as student engagement (Carini et al., 2006). In a different approach, as Johnston et al. (1991) proposed for research, in this study we examine the learning outcomes of the general education program at a liberal arts university and cross-examine them with the learning outcomes and related assessment data of all undergraduate programs. Our methodology is more aligned to the research conducted by Thomas et al. (2019) where researchers used qualitative coding. However, whereas Thomas et al. (2019) created inductive codes from student open-ended response, we treated the general education learning outcomes as our predetermined set of codes and holistically evaluated each academic learning outcome to determine if it matched the criteria to fit a general education outcome. Evaluating qualitative data according to a priori codes is a common methodology in qualitative research called deductive coding (Saldana, 2021).

Although assessment data commonly refers to scores from cognitive and noncognitive instruments, we define assessment data to include information at all stages of the assessment cycle. This includes the stated learning outcomes, their links to planned programming and the assessment instruments, as well as implementation fidelity data. In this sense, assessment data consists of all the information and rationale employed in the assessment process, not just assessment results, because assessment professionals and program stakeholders rely on this information to contextualize result interpretation and better inform their assessment related decision-making (Smith et al., 2017; 2019).

Recognizing that information at all stages of the assessment process is data addresses a major problem in educational assessment. Namely, widespread outcomes assessment is being conducted on campuses — often only to document student learning for accreditation — without using the assessment results to make a meaningful change to program outcomes (Banta & Blaich, 2011). Researchers and practitioners have responded to this gap in the assessment literature by providing models for using assessment data to evidence improvement in student learning at the program level (Fulcher, Good, et al., 2014; Fulcher, Smith, et al., 2017). Our work also addresses the gap in use of assessment results by demonstrating how rigorous assessment practice can be leveraged into research to answer relevant questions and form more accurate, empirical claims. Specifically, in this study we innovatively address the perceptions of a general education program using extensive assessment data that was already available at our institution.

Assessment data is more than just scores from tests-it includes all information and rationale employed in the assessment process.

The purpose of our study is to address one of the key cited reasons for disgruntlement with contemporary general education programs: the claim that they are disconnected and unrelated to students' major(s). To do this, we investigate assessment data to evaluate the degree to which general education and academic program learning outcomes are related. Using standardized program assessment records, we aim to collect evidence of whether there exists a connection between general education and academic programs.

Methods

We investigated a general education program at a large, mid-Atlantic, public university – hereafter referred to as the University. The University's general education program is not a traditional distribution requirement model. Rather, the University has invested significant resources in building a learning outcomes backbone to support the general education program. General education courses are built around domains with common learning outcomes where the curriculum is flexible, but the students are expected to attain the common learning outcomes. Specifically, the general education curriculum is built to align with five sets of related domains totaling 61 student learning outcomes. Although each course in the general education program is aligned to common domain outcomes, these courses are offered in multiple departments. For example, common critical thinking outcomes are found on the syllabi of history, business, and philosophy department courses, among others. Even within courses of the same title, students will have unique experiences as faculty are given flexibility to design the courses themselves, as long as they align with the general education outcomes. This structure provides students with multiple opportunities to achieve the same outcomes. Consequently, students are able to curate their education to their own interest or find new interests, which is favorable, as general education courses typically comprise one-third of students' total coursework. While many general education programs follow a traditional distribution requirement model, the current study investigates the degree of outcome interrelatedness in the local general education program based on a learning outcomes model. We do not pose a hypothesis comparing the models because data was not collected on both types of general education programs.

Our investigation focused on the alignment between the general education outcomes and the learning outcomes of the University's academic programs. Due to the University's investment in an outcomes assessment framework and to meet accreditation requirements, all programs submit similarly structured yearly program-level assessment reports to the assessment center on campus. These organized assessment reports encompass each academic program's assessment data: (1) a specification of student learning outcomes, (2) outcome alignment to curricular design, (3) the instruments used for assessment, (4) documenting

assessment methodology, (5) an analysis of results, and (6) how the program uses empirical results for program-related decisions to improve student learning. Each of these components are developed by faculty to benefit student learning. For the general education program, outcomes must go through a rigorous development and approval process by faculty representing multiple disciplines with stakes in the program. Assessment instruments are selected, modified, and/or developed by combining the expertise of general education faculty and assessment professionals at the University. Program outcomes, assessment instruments, and mapped programming are developed and reviewed by faculty within each academic program with consultation, if requested, from professionals in the on-campus assessment center. Assessment coordinators explicitly describe and show evidence of these processes in the annual assessment reports. With the goal of improving assessment practice, the assessment reports from undergraduate and graduate programs are reviewed by trained assessment professionals and faculty during a summer lockdown session.

To investigate the alignment of program-level outcomes with general education learning outcomes, two members of the research team independently reviewed all assessment reports, identifying program-level outcomes of academic programs (hereafter referred to as program outcomes) that matched general education learning outcomes. Evaluators treated the general education outcomes as predetermined coding categories using a deductive coding methodology. Then evaluators judged whether the general education outcomes overlapped or scaffolded up to the program outcomes in terms of content knowledge, skills, or abilities. Although this may seem straight forward, the specificity or generality of learning outcomes caused ambiguity. Consequently, evaluators investigated the assessment instruments used to assess each program outcome in addition to the mapped curriculum. Researchers referenced descriptive information about the instruments provided in the annual assessment reports, specifically the rationale linking instruments to learning outcomes, which helped develop a deeper understanding of the operationalization of the outcomes for each evaluator to independently determine alignment.

Through an adjudication process across the two independent evaluators, we identified two sets of learning outcomes: (a) those in which there was agreement in alignment and (b) those in which there was disagreement in alignment. Agreement meant that both researchers coded the program outcome with the same general education outcome, having agreed that the program outcomes and general education outcomes were similar in terms of skills or content. Disagreement meant that only one evaluator provided a general education outcome code for a program outcome. Even with agreement, evaluators still compared rationales for alignment or lack of alignment. It became evident that in nearly all instances where evaluators agreed, it was because the learning outcome text, along with instrument and content coverage, provided sufficient detail for judgment. For cases in which evaluators disagreed, they relied on subjective judgements as to whether there was alignment because of limited information in the assessment reports. After discussion, if disagreement persisted, a third researcher independently evaluated to adjudicate. Fortunately, this situation was quite rare. Interrater reliability was bolstered by the adjudication process between the two independent evaluators and the input of the third researcher for special cases of disagreement such that all outcome alignment was based on the judgement of at least two evaluators.

Assessment reports provide evidence-based insights for improving student learning.

To elucidate the judgement of the evaluators, we provide two examples: first, an example where alignment between outcomes was evident by evaluating the outcome statements alone; second, an example where alignment was determined after examining supplemental information in the assessment reports to determine the appropriate general education outcome code. As evaluators read program outcomes, they considered the content and skills students would need to know, think, or do to fulfill outcome requirements. Then evaluators holistically evaluated whether the content and/or skill(s) of the program outcome overlapped with the content and/or skill(s) specified by any of the general education outcomes, the a priori coding categories. For example, a Computer Science outcome states that students will be able to "express themselves clearly on technical matters both orally and in writing; communicate effectively with individuals who have a technical background and with individuals who do not." Evaluators would match this outcome to a general education outcome from the Human Communication domain: "construct messages consistent with

the diversity of communication purpose, audience, context, and ethics" because alignment between skills is clear from the outcome text; specifically, both outcomes delineate elements of rhetorical awareness as a key skill for effective communication. In most cases, evaluators agreed between outcome alignment following this independent reasoning process.

In other cases, the degree of alignment was not evident from the outcome text alone. For example, the Dietetics program features a learning outcome stating that "upon completion of the program, graduates are able to apply critical thinking skills." While critical thinking skills should align to the critical thinking general education domain, evaluators were not able to determine which code is appropriate without examining additional assessment data, such as how the Dietetics program assesses critical thinking skills. Referring to their annual report, the evaluators found that the Dietetics program used a course embedded assessment as part of a capstone project where students were required to find, analyze, and use information as the basis for evidence-informed practices. Specifically, the critical thinking skills outcome was evaluated on students' ability to apply research to practice with coherent and valid rationale and evidence. To that end, evaluators determined that students must be able to evaluate research *claims and sources* for *relevance* to their practice, as well as *credibility* and *accuracy* for their practice recommendations, which aligns to the general education outcome code from the Critical Thinking domain: "evaluate claims and sources for clarity, credibility, reliability, accuracy and relevance." These examples illustrate the two-step judgement strategy the evaluators used. First, evaluators holistically evaluated whether students would engage in overlapping or scaffolded skills to accomplish outcome statements as written. If the outcomes did not provide sufficient information to determine the degree of alignment, then evaluators referred to data in annual assessment reports to supplement their judgement. Results based on this evaluation strategy are presented next.

Results show that writing and presentation skills are common targets of program outcomes, and Human Communication, Writing, Quantitative Reasoning, and Information Literacy are the most linked general education domains.

Results

In the review year, the 2018-2019 academic year, 51 undergraduate academic degree programs with 633 learning outcomes submitted assessment reports. It is the case that in some years, academic programs are given exemptions to have more time to focus on long-term assessment projects, such as learning improvement. Each program, on average, had 12 learning outcomes with a standard deviation of 11. No academic degree program had less than one learning outcome and the maximum was one program with 64 learning outcomes. Within the submitted reports, 293 (46%) program-level outcomes were found to be related to general education learning outcomes.

The number of academic programs with linked outcomes to general education outcomes are shown in Table 1. The Human Communication (30) and Writing (22) general education domains linked to the greatest number of academic programs, while Wellness (2) linked to the least. Program outcomes targeting writing and presentation skills accounted for most of these mappings. Even for general education domains with few links to academic programs, within the linked programs there was still a considerable amount of alignment between outcomes. For example, the American Experience as well as the Visual and Performing Arts domains each only linked to four academic programs, but American Experience was aligned with 20 program outcomes and the Visual and Performing Arts domain was aligned with 12 program outcomes.

The four general education domains with the highest number of academic programs linked were Human Communication, Writing, Quantitative Reasoning, and Information Literacy. The Human Communication domain was comprised of four learning outcomes; each aligned with multiple program outcomes. Figure 1 displays the thirty academic degree programs with outcomes that mapped to the four Human Communication outcomes. These programs were quite variable in their disciplines. For example, Hospitality and Tourism Management had six (out of 18) outcomes aligned with Human Communication, while Integrated Science and Technology had four (out of 64) of their program outcomes aligned to the Human Communication domain. The "message construction" outcome accounted for the most links.

Table 1
Distribution of academic degree programs with linked outcomes by general education domain

General Education Domain	Number of Academic Programs with Linked Outcomes (% out of all 51 potential programs)	Number of Linked Outcomes across Academic Programs
American Experience	4 (7.8)	20
Critical Thinking	9 (17.6)	14
Global Experience	9 (17.6)	22
Human Communication	30 (58.8)	57
Human Questions and Context	7 (13.7)	10
Information Literacy	16 (31.4)	26
Lab Experience	3 (5.9)	3
Literature	2 (3.9)	4
Natural Principles	6 (11.8)	15
Physical Principles	9 (17.6)	17
Quantitative Reasoning	20 (39.2)	40
Sociocultural Domain	7 (13.7)	8
Visual and Performing Arts	4 (7.8)	12
Wellness	2 (3.9)	3
Writing	22 (43.1)	42
Total	150	293

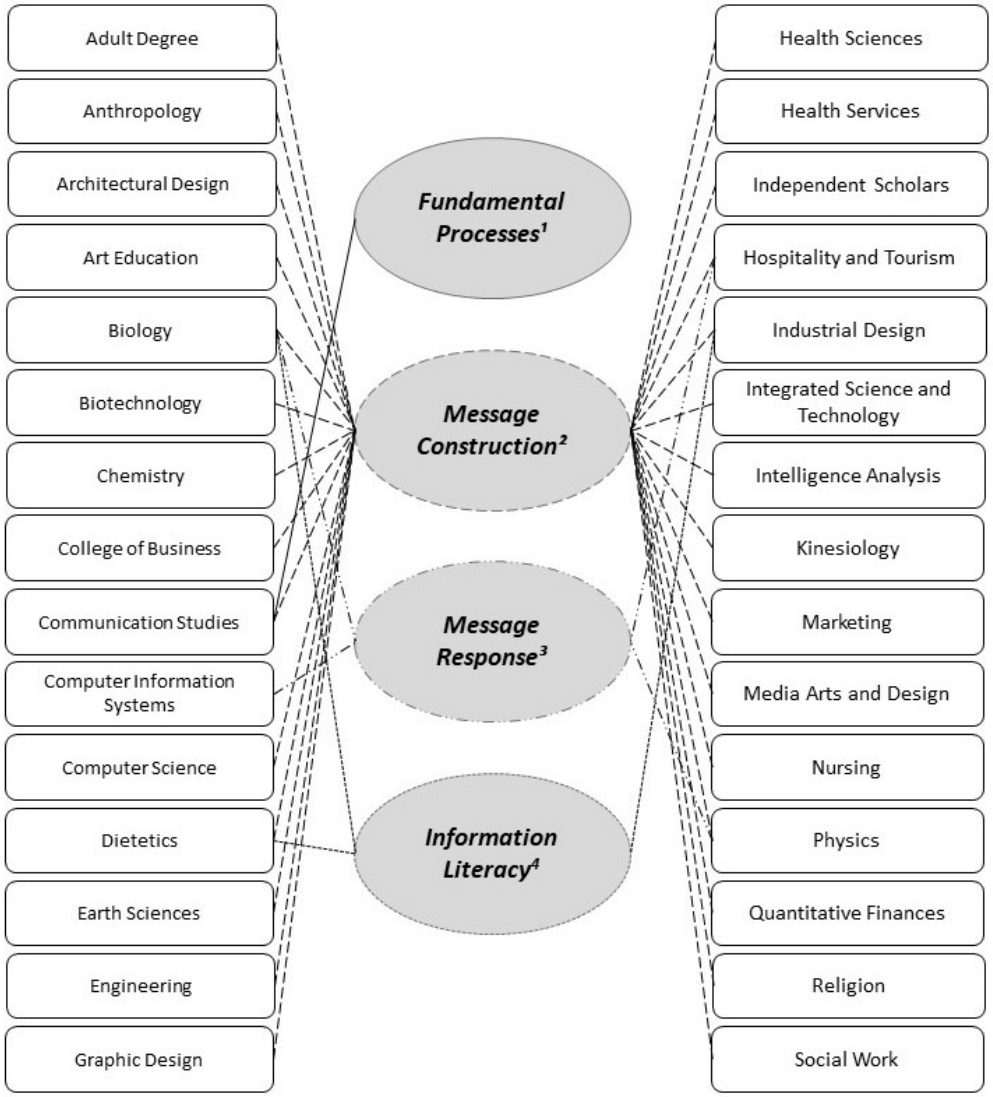
Writing, Quantitative Reasoning, and Information Literacy general education domains show strong alignment with program outcomes across academic degree programs.

Courses in the general education Writing domain were developed for students to achieve five learning outcomes. Figure 2 displays the twenty-two academic degree programs with outcomes that mapped to the five Writing outcomes. Although twenty-two academic programs had aligned outcomes, the "writing in multiple environments" general education outcome was not linked to any program outcomes. Unsurprisingly, the English program had seven (out of nine) outcomes linked to Writing and the Writing, Rhetoric and Technical Communication program had six (out of six) of their program outcomes linked to this domain. The "rhetorical awareness" and "writing process" general education outcomes accounted for the most links.

The Quantitative Reasoning general education domain consisted of three learning outcomes, each aligned with numerous program outcomes. Figure 3 displays the twenty academic programs with outcomes that mapped to the three Quantitative Reasoning outcomes. The Integrated Science and Technology program had five (out of 64) outcomes linked to Quantitative Reasoning whereas several academic programs had three of their program outcomes aligned, including Anthropology (out of 24), Biology (out of 23), Engineering (out of 40), International Affairs (out of 15), and Political Science (out of 17). The "methods of inquiry" general education outcome accounted for the most links.

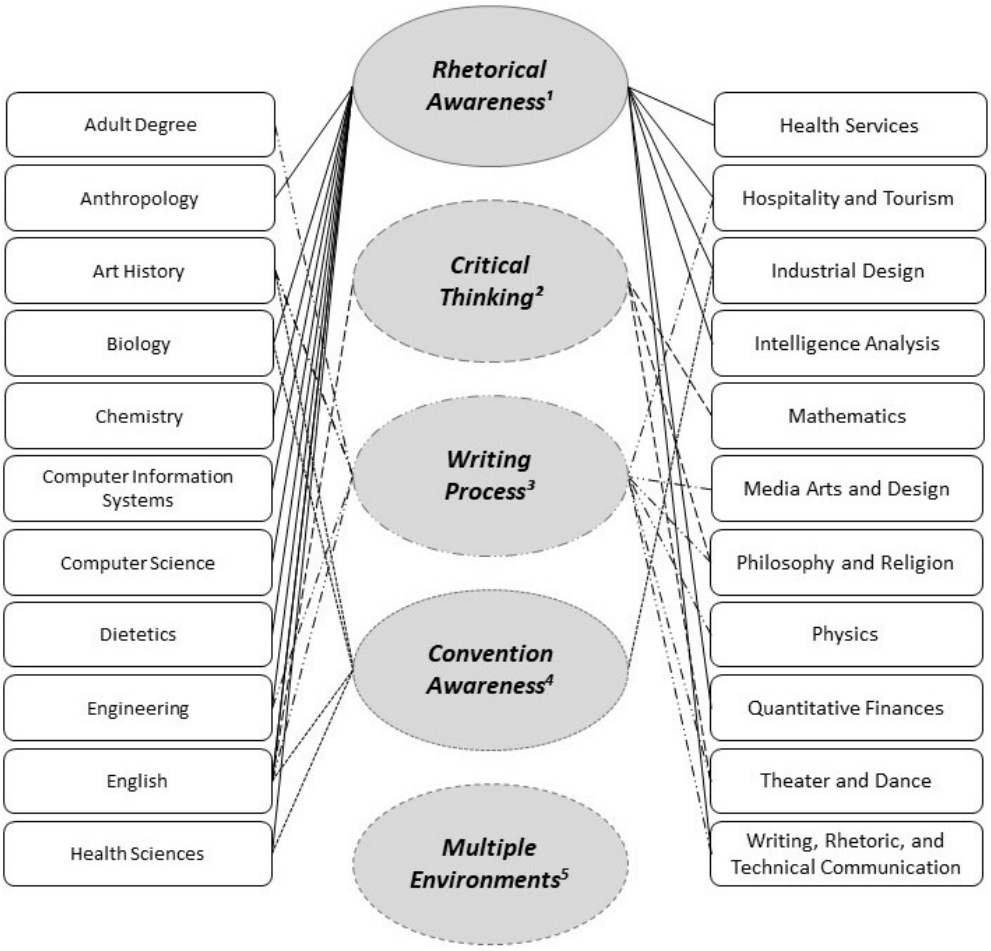
Programming in the Information Literacy general education domain was designed to cover six learning outcomes. Figure 4 displays the sixteen academic degree programs with outcomes mapped to the six information literacy outcomes. Each learning outcome in the Information Literacy domain was linked to at least one of sixteen academic programs. Both the Engineering (out of 40) and the Media Arts and Designs (out of 12) programs had three outcomes linked to Information Literacy. The "persistency" and "information quality" general education outcomes accounted for the most links to program outcomes.

Figure 1
Academic programs (in rectangles) with outcomes mapped to the Human Communication domain outcomes (in ovals)



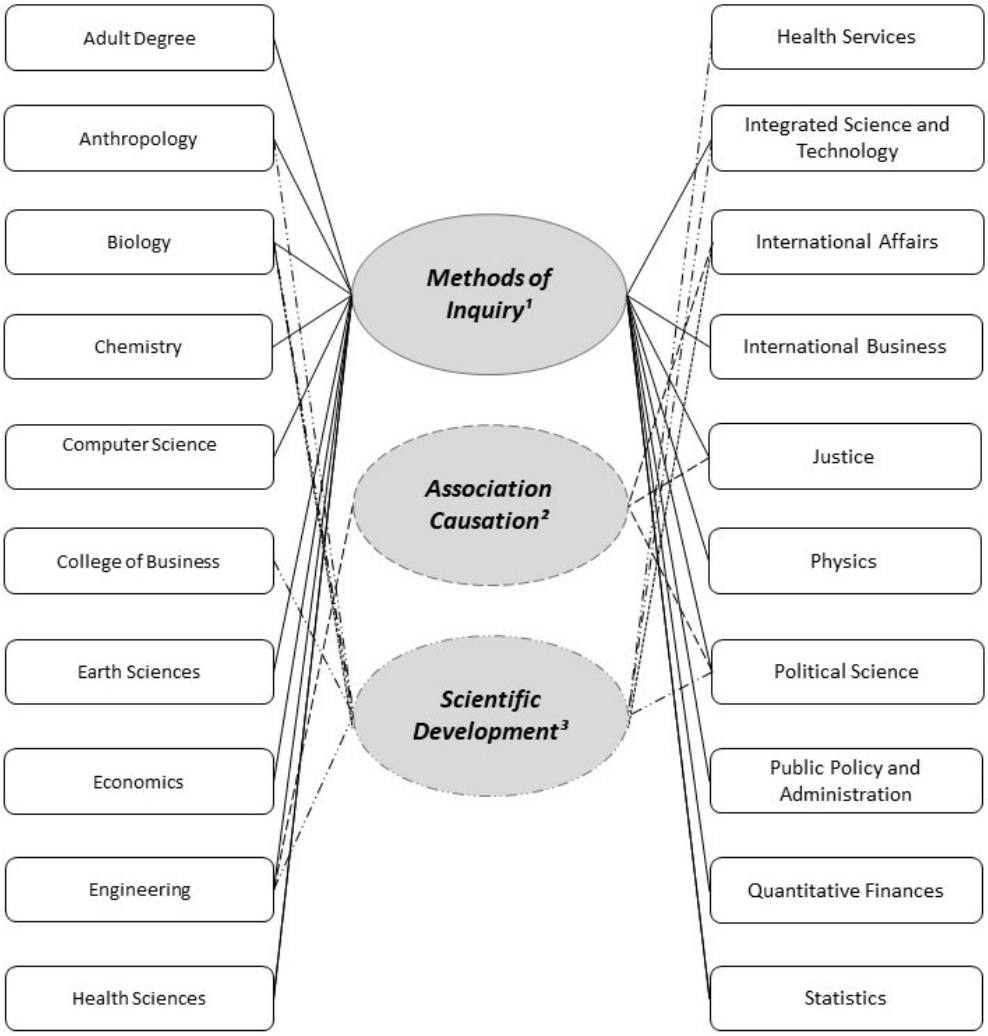
[1] Explain the fundamental processes that significantly influence communication.
[2] Construct messages consistent with the diversity of communication purpose, audience, context, and ethics.
[3] Respond to messages consistent with the diversity of communication purpose, audience, context, and ethics.
[4] Utilize information literacy skills expected of ethical communicators.

Figure 2
Academic programs (in rectangles) with outcomes mapped to the Writing domain outcomes (in ovals)



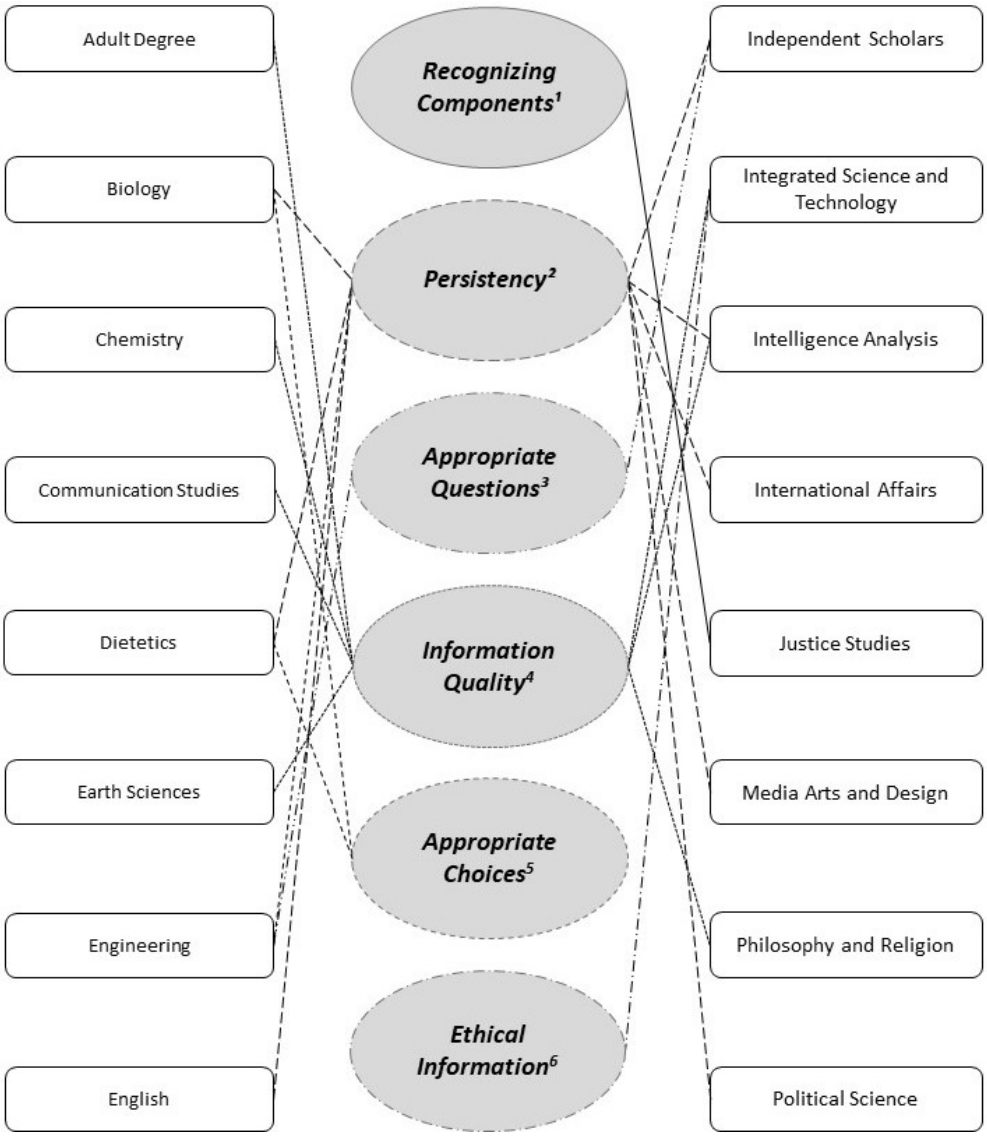
[1] Demonstrate an awareness of rhetorical knowledge, which may include the ability to analyze and act on understandings of audiences, purposes and contexts in creating and comprehending texts.
[2] Employ critical thinking, which includes the ability through reading, research and writing, to analyze a situation or text and make thoughtful decisions based on that analysis.
[3] Employ writing processes.
[4] Demonstrate an awareness of conventions, the formal and informal guidelines that define what is considered to be correct and appropriate in a variety of texts.
[5] Compose in multiple environments using traditional and digital communication tools.

Figure 3
Academic degree programs (in rectangles) with outcomes mapped to the Quantitative Reasoning domain outcomes (in ovals)



[1] Describe the methods of inquiry that lead to mathematical truth and scientific knowledge and be able to distinguish science from pseudoscience.
[2] Discriminate between association and causation, and identify the types of evidence used to establish causation.
[3] Evaluate the credibility, use and misuse of scientific and mathematical information in scientific developments and public-policy issues.

Figure 4
Academic degree programs (in rectangles) with outcomes mapped to the Information Literacy domain outcomes (in ovals)



[1] Recognize the components of scholarly work and that scholarship can take many forms.
[2] Demonstrate persistence and employ multiple strategies in research and discovery processes.
[3] Identify gaps in their own knowledge and formulate appropriate questions for investigations in academic settings.
[4] Evaluate the quality of information and acknowledge expertise.
[5] Use information effectively in their own work and make contextually appropriate choices for sharing their scholarship.
[6] Use information ethically and legally.

The remaining general education domains had fewer outcomes aligned to academic program outcomes. Table 2 presents academic program information for the remaining general education domains that had at least 10 links. The number of outcomes a general education domain had did not relate to the number of links aligned to academic program outcomes, $r(13) = .15$, $p = .59$. However, academic programs with many outcomes were linked with more general education outcomes than programs with fewer outcomes, $r(49) = .61$, $p < .01$. Moreover, the academic programs with the most linked outcomes varied by general education domain.

Table 2
Academic program information for the general education domains that had at least 10 links

General Education Domain (number of learning outcomes)	Total Links	Academic programs with most linked outcomes	Number of Linked Outcomes within Academic Program (Number of total outcomes)	General Education Outcome(s) accounting for the most links
Global Experience (5)	22	International Affairs	6 (15)	Global Systems
		Political Science	4 (17)	
American Experience (6)	20	Political Science	9 (17)	Democratic Institutions
		Public Policy and Administration	8 (21)	
Physical Principles (2)	17	Biology	5 (23)	Formulate Hypothesis
		Engineering	3 (40)	
		Physics	3 (14)	
Natural Principles (2)	15	Biology	6 (23)	Numerical Methods
		Physics	3 (14)	Theories Models
Critical Thinking (4)	14	Biology	3 (23)	Argument Evaluation
		Dietetics	3 (9)	Argument Components
Visual and Performing Arts (6)	12	Art History	4 (6)	Disciplinary Literacy
		Theater and Dance	4 (28)	Art and Works
		Studio Art	3 (5)	
Human Questions and Context (5)	10	Philosophy and Religion	3 (8)	Appropriate Concepts
		Integrated Science and Technology	2 (64)	Understanding Context

Our findings provide evidence that general education programs have connections to academic degree programs...the knowledge, skill, and abilities taught in the general education program align to the knowledge, skill, and abilities students develop in their academic programs.

Discussion

General education programs are a staple of higher education institutions in America (Crooks, 1979; O'Banion, 2016). However, these general education programs have come under considerable criticism from students and some educators. Many students question the utility of these programs and consider them an interference to their academic degree pursuit or a drain of time and resources (Humphreys & Davenport, 2005; Mintz, 2020; Peruski, 2005; Survey and Evaluation Research Laboratory, 2021; Thompson et al., 2015). Many educators tend to see the value in a core liberal education experience for all students and become invested in providing a general education program — some recognize that there is room for improvement; others see a disconnect between academic degree programs and their local general education program (Beld & Booth, 2010; Paulson, 2012). Yet many of these sentiments are based on anecdotal evidence and self-reports. The present investigation of program-level student learning outcomes provides empirical clarity to this dilemma.

Our findings provide evidence that general education programs have connections to academic degree programs. Nearly every learning outcome of the general education program at the University aligned with at least one learning outcome of an academic degree program,

many of which had several linked learning outcomes. There also existed general education outcomes in which we were unable to conclude direct alignment, but that were still likely related to program outcomes. For example, consider the general education Writing outcome "writing in multiple environments." Although we found zero links with academic program learning outcomes, achievement of this outcome by students is likely necessary for success in multiple academic programs, each requiring writing in their own unique environment.

Overall, results showed considerable overlap between general education and academic program learning outcomes, suggesting that the knowledge, skill, and abilities taught in the general education program align to the knowledge, skill, and abilities students develop in their academic programs. General education programs aim to set students up for success in a variety of academic degree programs (Glynn et al., 2005; Scott, 2014). Our results show that the academic outcomes linked tended to vary by general education domain. In other words, the domains of the general education program aligned differently to a variety of academic degree programs. For instance, the academic outcomes linked to the Global Experience domain were from different degree programs than the outcomes linked to the Physical Principles domain or the Visual and Performing Arts domain. Nonetheless, the learning outcomes of four general education domains each linked to at least a third of the academic programs. These domains focus on the higher order skills students need to be successful in most, if not all, academic degree programs: Human Communication, Writing, Quantitative Reasoning, and Information Literacy. In these outcomes, we can see a scaffolding of skills from general education into students' major, reflecting important learning steps that all students are intended to experience and develop from. Skills targeted in these learning outcomes are critical to multiple disciplines across higher education. These findings oppose the reported perception students have that general education programs are unrelated to academic degree programs and interfere with major requirements (Humphreys & Davenport, 2005). However, our results do not provide evidence for inferences about students' actual experience in the learning environments that may lead to a perception of disconnect nor do our results address the feeling of redundancy, specifically that students report having learned general education outcomes in high school (Peruski, 2005). Thus, a well-integrated and well-planned general education program may still not alleviate the negative sentiments associated with the sense of redundancy.

In addition to results on assessment instruments, assessment data consists of the stated learning outcomes, mapped curriculum, instruments used, and implementation fidelity, the last of which was not considered in this study. Implementation fidelity refers to the extent to which learning experiences were delivered as designed (Gerstner & Finney, 2013; Smith et al., 2017; 2019). In other words, implementation fidelity data would allow us to consider the student experience in the classroom. Although we evidence that the academic degree programs and general education program may be well integrated by design, students' perceptions are informed by their lived experiences, and are a reflection of whether the related programming is implemented as planned. No matter how much alignment we see on paper, a lack of adherence to the programing as designed will influence the sentiment of disconnect experienced by students. In the present study we investigated the planned programming rather than the programing students actually received, due to limited available implementation fidelity data; thus, we were not able to assess how well the intended curriculum was adhered to.

Our results show that the academic outcomes linked tended to vary by general education domain. In other words, the domains of the general education program aligned differently to a variety of academic degree programs.

Misconceptions about the interrelatedness of academic programs and general education may also be confounded by a messaging problem. Paulson (2012) suggests that educators are attempting to make the connection between general education and academic programs more evident; however, they lack the empirical data demonstrating this connection. By examining outcomes-based assessment records, we were equipped with the necessary documentation to provide evidence of a connection, which can be used to strengthen the messaging efforts of educators. Having a rigorous and thorough assessment practice in an outcomes-based framework adopted throughout the University was integral for producing this evidence.

In addition to directly and clearly educating incoming students on the purposes and requirements of their general education program, as well as the philosophy behind a well-round liberal education, educators should consider how their assessment practice can strengthen this messaging effort by providing empirical evidence of the connection between general education and academic programs. Clarifying the misconceptions early on can prevent students from

developing them throughout their educational career and letting this misconception take root. Moreover, faculty may consider integrating general education programing into academic program curriculum, explicitly linking the skills students are using in the academic degree program to the skills the general education program develops; thus, making this connection more evident for students. wFinally, while the study presented a story from the general education perspective, it may be worthwhile to see how an academic degree program maps to the existing general education programing to provide suggestions and recommendations on how to fill gaps in the general education program or how to reframe general education content to better fit and better scaffold toward academic degree programs.

Conclusion

By leveraging valuable assessment data, our study has taken a significant empirical step toward clarifying the connection between a general education and academic programs. Moreover, assessment data can be used to specifically form partnerships across general education academic program stakeholders, including instructors, advisors, department chairs, and admissions staff. As seen in the learning improvement literature, collaboration within and across disciplines is a key component for making meaningful change in higher education institutions (Fulcher, Good, et al., 2014; Fulcher, Smith, et al., 2017). Using assessment data to show the connection between program outcomes can warrant the basis for faculty partnerships and foster collaboration.

The effort to explicate the connection between general education and academic degree program learning outcomes can be addressed by higher education professionals at various levels and in various roles across higher education. Stakeholders can reference our results to strengthen their messaging regarding the alignment of general education with academic degree programs to represent higher education as an integrated learning experience, rather than separate tracks.

Assessment professionals can use our methodology and results to describe and visualize the alignment among academic program outcomes and general education outcomes. Furthermore, the outcome mapping could be useful for pulling together outcome-based accreditation reports. The outcome mapping can be used to organize reports by outcome alignment or through inclusion of information and visuals that demonstrate outcome integration across programs. Instructors will also benefit from using our methodology and findings to determine alignment between their academic program or course outcomes and general education outcomes, which can inform how they plan and execute learning activities to foster a more integrated education for students.

Collaboration across positions to facilitate educational connectedness is an important pursuit to reduce the perceived tension between students' career-oriented goals and general education roles (Humphreys & Davenport, 2005; Survey and Evaluation Research Laboratory, 2021; Thompson et al., 2015). Student critique of general education experience builds toward a greater discontent with contemporary conditions of higher education, particularly the devaluation of a liberal arts education (Gerber, 2012; Siegel, 2013), the rising cost of a post-secondary education (Hill & Pisacreta, 2019; Schneider & Seligman, 2018), and the challenges of securing a comparable position in the workforce (NCES, 2021; Richard et al., 2013). Irrespective of the accuracy or extent to which these arguments are well founded, and the efforts made for counter-persuasion (AAC&U, 2002; Abel & Deitz, 2014), the negative narratives building against higher education are mounting. It is our hope that continuing the research on the interrelatedness of general education with academic programming at large, drawing on the strength provided through rigorous and extensive assessment practice, will help bring a positive focus that can aid in turning the tide of the conversation around higher education.

References

Abowitz, K. K. (2006). The interdependency of vocational and liberal aims in higher education. *About Campus*, *11*(2), 16-22. https://doi.org/10.1002/abc.162

Association of American Colleges and Universities (AAC&U). 2002. *Greater expectations: A new vision for learning as a nation goes to college*. Washington, DC: Association of American Colleges and Universities.

Astin, A. W., Korn, W. S., & Berz, E. R. (1989). *The American freshman: National norms for fall 1989*. Los Angeles: Higher Education Research Institute, University of California, Los Angeles.

Baker, L. V., Baldwin, R. G., & Makker, S. (2012). Where are they now: Revisiting Breneman's study of liberal arts colleges. *Liberal Education*, *98*(3).

Banta, T. W., & Blaich, C. (2011). Closing the assessment loop. *Change: The Magazine of Higher Learning 43*(1), 22-27. https://doi.org/10.1080/00091383.2011.538642

Bastedo, M. (2002). General education. In J. J. F. Forest & K. Kinser (Eds.), *Higher education in the United States: An encyclopedia* (Vol. 1, pp. 273-276). Santa Barbara: ABC-CLIO.

Beld, J., & Booth, D. (2010). *Fostering faculty ownership of general education assessment*. Paper presented at the Association of American Colleges and Universities Network for Academic Renewal Conference "General Education and Assessment: Maintaining Momentum, Achieving New Priorities," Seattle.

Benander, R., Denton, J., Page, D., & Skinner, C. (2000). Primary trait analysis: Anchoring assessment in the classroom. *Journal of General Education*, *49*(4), 279-302. https://doi.org/10.1353/jge.2000.0025

Bourke, B., Bray, N. J., & Horton, C. H. (2009). Approaches to the core curriculum: An exploratory analysis of top liberal arts and doctoral-granting institutions. *Journal of General Education*, *58*(4), 219-240. https://doi.org/10.1353/jge.0.0049

Boyer, E. L. (1987). *College: The undergraduate experience in America*. New York: Harper and Row.

Burns, C. J., & Natale, S. M. (2020). Liberal and vocational education: The Gordian encounter. *Education & Training*, *62*(9), 1087-1099. http://dx.doi.org/10.1108/ET-03-2020-0064

Carini, R.M., Kuh, G.D., & Klein, S.P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education*, *47*(1), 1-32. https://doi.org/10.1007/s11162-005-8150-9

Crooks, J. B. (1979). History's role in general education. *Journal of General Education*, *31*(2), 109-121. http://www.jstor.com/stable/27796755

Felder, B. (2018, November 30). *How colleges are adapting to the decline in liberal arts majors*. PBS News Hour. https://www.pbs.org/newshour/education/how-colleges-areadapting-to-the-decline-in-liberal-arts-majors

Fulcher, K. H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014). A simple model for learning improvement: Weigh pig, feed pig, weigh pig. (NILOA occasional paper no. 23). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Fulcher, K. H., Smith, K. L., Sanchez, E. H., Ames, A. J., & Meixner, C. (2017). Return of the pig: Standards for learning improvement. *Research and Practice in Assessment*, *11*, 10–27.

Galle, J. K., & Galle, J. (2010). Building an integrated student learning outcomes assessment for general education: Three case studies. *New Directions for Teaching and Learning*, *121*, 79-87. https://doi.org/10.1002/tl.390

Gerber, S. (2012, September 24). How liberal arts colleges are failing America. *The Atlantic*. https://www.theatlantic.com/business/archive/2012/09/how-liberal-arts-colleges-are-failing-america/262711/

Gerstner, J. J., & Finney, S. J. (2013). Measuring the implementation fidelity of student affairs programs: A critical component of the outcomes assessment cycle. *Research and Practice in Assessment*, *8*, 15–28.

Glynn, S. M., Aultman, L. P., & Owens, A. M. (2005). Motivation to learn in general education programs. *Journal of General Education*, *54*(2), 150-170. https://doi.org/10.1353/jge.2005.0021

Gump, S. E. (2007). Classroom research in a general education course: Exploring implications through an investigation of the sophomore slump. *Journal of General Education*, *56*(2), 105-125. https://doi.org/10.1353/jge.2007.0020

Hill, C. B. & Pisacreta, E. D. (2019). The economic benefits and costs of a liberal arts education. *The Andrew W. Mellon Foundation*. https://mellon.org/news-blog/articles/economic-benefits-and-costs-liberal-arts-education/

Humphreys, D. & Davenport, A. (2005). What really matters in college: How students view and value liberal education. *Liberal Education*, *91*(3), 36-43.

Johnston, J. S., Reardon, R. C., Kramer, G. L., Lenz, J. G., Maduros, A. S., & Sampson, J. P. (1991). The demand side of general education: Attending to student attitudes and understandings. *Journal of General Education*, *40*, 180-200. https://www.jstor.org/stable/27797136

Krukowski, J. (1985). What do students want? Status. *Change: The Magazine of Higher Learning*, *17*(3), 21-28. https://doi.org/10.1080/00091383.1985.9939796

Melville, K., Dedrick, J., & Gish, E. (2013). Preparing student for democratic life: The rediscovery of education's civic purpose. *Journal of General Education*, *62*(4), 258-276. https://doi.org/10.1353/jge.2013.0026

Mintz, S. (2020, June 23). The general education curriculum we need. *Inside Higher Ed*. https://www.insidehighered.com/blogs/higher-ed-gamma/general-education-curriculum-we-need

Moffatt, M. (1989). *Coming of age in New Jersey: College and American culture*. Rutgers University Press.

National Center for Education Statistics (NCES). (2021). *Employment outcomes of bachelor's degree holders*. *Condition of Education*, U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/programs/coe/indicator/sbc

O'Banion, T. (2016). A brief history of general education. *Community College Journal of Research and Practice*, *40*, 327-334. https://doi.org/10.1080/10668926.2015.1117996

Paulson, K. (2012). Faculty perceptions of general education and the use of high-impact practices. *Peer Review*, *14*(3), 25-28. https://www.proquest.com/scholarly-journals/faculty-perceptions-general-education-use-high/docview/1243370541/se-2?accountid=11667

Peruski, A. M. (2005). *Student opinion survey on the general education program at CMU*. Central Michigan University. https://silo.tips/download/student-opinion-survey-on-the-general-education-program-at-cmu-may-2005

Saldana, J. M. (2021). *The coding manual for qualitative researchers* (4th ed.). SAGE Publications.

Schneider, M., & Seligman, M. (2018), *Saving the liberal arts: Making the bachelor's degree a better path to labor market success*. American Enterprise Institute. https://www.burning-glass.com/wp-content/uploads/Saving-the-Liberal-Arts.pdf

Scott, R. A. (2014). The meaning of liberal education. *On The Horizon*, *22*(1), 23-34. https://doi.org/10.1108/OTH-09-2013-0036

Siegel, L. (2013, July 12). Who ruined the humanities? *The Wall Street Journal*. https://www.wsj.com/articles/SB10001424127887323823004578595803296798048

Smith, K. L., Finney, S. J., & Fulcher, K. H. (2017). Actionable steps for engaging assessment practitioners and faculty in implementation fidelity research. *Research and Practice in Assessment*, *12*, 71-86.

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

Stevens, A. H. (2001). The philosophy of general education and its contradictions: The influence of Hutchins. *Journal of General Education*, *50*(3), 165-191.

Survey and Evaluation Research Laboratory. (2021). *Virginia educated: A post-college outcomes study of Virginia public college and university graduates from 2007 to 2018*. Virginia Commonwealth University, L. Douglas Wilder School of Government and Public Affairs. https://schev.edu/docs/default-source/reports-and-studies/2021-reports/virginia-educated-survey-2021-full-report-no-appendices.pdf

Thompson, C. A., Eodice, M., & Tran, P. (2015). Student perceptions of general education requirements at a large public university: No surprises? *Journal of General Education*, *64*(4) 278-293. https://doi.org/10.1353/jge.2015.0025

Vedder, R., Denhart, C., & Robe, J. (2013). *Why are recent college graduates underemployed: University enrollments and labor-market realities*. Center for College Affordability and Productivity. https://files.eric.ed.gov/fulltext/ED539373.pdf

Walters, H. D., & Bockorny, K. M. (2018). Relevance of general education: An assessment of undergraduate business students. *e-Journal of Business Education and Scholarship of Teaching*, *12*(3), 34-43.

Abstract

Faculty engagement in assessment processes is critical but remains limited, especially in public doctoral research universities. In this article, we propose an engaged assessment model that emerged from our work at a leading doctoral university. Through the engaged assessment process, faculty members are involved throughout, centering on student learning. Using the assessment process of the institution's quality enhancement plan as an example, we detail how the engaged assessment model can be implemented through faculty learning communities. We also elaborate on core activities where faculty members explored assessment design, examined assessment data, and celebrated assessment as scholarship.

AUTHORS

Bryant L. Hutson, Ph.D
University of North Carolina at Chapel Hill

Kelly A. Hogan, Ph.D
University of North Carolina at Chapel Hill

Faculty Engagement in Student Learning Outcome Assessment

The importance of faculty engagement in the assessment of student learning has been widely discussed (Ewell, 2009; Hutchings, 2010; 2016; Jankowski & Marshall, 2017; Middaugh, 2010; Reder & Crimmins, 2018). Kinzie et al. (2019) described the potential of greater integration of assessment and faculty development efforts in promoting a shared institutional commitment to student learning. There is also evidence that higher education institutional assessment is moving beyond the compliance-oriented approach to a more classroom-centered embedded approach (Hutchings, 2010; Kinzie et al., 2019). Instead of centering assessment on institutional compliance of external accountability measures, the classroom-centered embedded assessment approach offers the potential of faculty members being more engaged in institutional assessment design and implementation, as well as involved in using assessment data for curriculum improvement to support student learning.

However, even while there are more efforts toward classroom-centered assessment, faculty engagement in assessment processes is still limited, especially in public doctoral research universities (Grunwald & Peterson, 2003; Jankowski et al., 2018; Kuh & Ikenberry, 2009; Kuh et al., 2014). Based on a survey involving provosts and chief academic officers at U.S. higher education institutions about assessment activities and how assessment results are used at their institutions, Kuh and Ikenberry (2009) found that four-fifths of provosts of doctoral research universities report greater faculty engagement with assessment of student learning as their leading challenge. Similarly, Kuh et al. (2014) surveyed provosts or their designates at 1,202 institutions regarding assessment activities and how their institutions

CORRESPONDENCE

Email
bhutson@email.unc.edu

use assessment results. Their findings reiterated the importance of faculty's role in assessment and reported public institutions' expressed needs to have more faculty involvement in assessment, increase the use of assessment results, and have more professional development for faculty and staff. A 2018 National Institute for Learning Outcomes Assessment (NILOA) survey also confirmed that provosts of doctoral institutions were more likely to express a desire for increased faculty engagement in assessment than their peers at other types of institutions (Jankowski et al., 2018).

Engaged Assessment challenges traditionally-defined boundaries between instruction and assessment, emphasizing the importance of faculty involvement and design thinking in collecting and analyzing assessment data.

In this article, we propose an engaged assessment model that emerged from our work at a leading doctoral university. We describe the components of the model and provide an example of how this model can be implemented in the higher education context.

Engaged Assessment

Even though faculty members design and implement the curriculum to support students in achieving specified learning outcomes, not all of them perceive engagement in the institutional assessment process as an integral aspect of their primary responsibilities (Banta & Blaich, 2011). If institutions adopt a traditional assessment model that centers on assessment reporting rather than learning improvement (Hundley & Kahn, 2019), faculty members and assessment professionals may work in isolation and the engagement of faculty in the assessment process may be peripheral and limited.

The ideal assessment process requires expanded faculty engagement and ownership. Inasmuch as assessment informs teaching and curriculum (Angelo & Cross, 1993; Maki, 2010; Suskie, 2014), faculty members have an incentive to be engaged in discussions about assessment. Their experience with evaluating student mastery of content is invaluable in determining methods and criteria for measuring student learning, while their research backgrounds and expertise can be leveraged to promote the scholarship of assessment (AAHE, 1992; Arum & Roksa, 2011; Metzler & Kurz, 2018).

An engaged assessment model challenges the traditionally-defined boundaries between instruction and assessment. Assessment professionals with expertise in assessment design and measurement work collaboratively with faculty members with expertise in the content area and pedagogy throughout the assessment process. Together they engage in ongoing dialogs to negotiate the assessment designs, instruments, and protocols to collect and analyze assessment data. Applying design thinking in the process (Benson & Dresdow, 2014; Brown, 2008), they engage in iterative assessment cycles to not only modify instruction and curriculum based on assessment data but also make assessment adaptations. Through the engaged assessment model, faculty members and assessment professionals develop deeper understandings of the assessment process and outcomes and continue to increase the institutional capacity for assessment and instruction. Table 1 provides a summary that highlights the features of the engaged assessment process.

Table 1
Features of Engaged Assessment

Structure - Boundary	Boundary crossing is encouraged based on individuals' backgrounds, experiences, and interests
Process - Design	Dialog space is created to negotiate meanings among team members
	Assessment designs, instruments, and protocols are emerging and adaptable
Product - Data Use	Iterative assessment circles - Immediate interpretation and use of data directly for program/curriculum and assessment adaptations
	Summative reporting reflects more nuanced contextualized interpretation for program improvement
Sustainability - Mutual Learning	Boundary crossing dialogs provide learning opportunities and capacity building

In this section, we provide the program context and elaborate on the intentionality, process, product, and sustainability of the engaged assessment process at University of North Carolina at Chapel Hill (UNC). We describe core activities where faculty members explored the assessment design, examined assessment data, and celebrated assessment as scholarship through the assessment efforts for the implementation of course-based undergraduate research experiences or "CUREs" as part of the quality enhancement plan (QEP) at the university.

QEP Context

A key component of the Southern Association of Colleges and Schools Commission on Colleges (SACSCOC) reaccreditation requirements, the QEP is a plan of action that addresses an "issue that the institution considers important to improving student learning outcomes and / or student success" (SACSCOC, 2020, p. 1). As a leading research university, UNC identified providing meaningful research experiences to undergraduates as a student learning priority (Sathy et al., 2020; 2021).

The QEP and its assessment have been central to transforming undergraduate education at UNC to the point that the effort has been sustained by incorporating many of the findings into the new general education curriculum. UNC's QEP comprises four programs: integrated curricula (co-taught first-year seminars that integrate the arts and humanities with the sciences and support other interdisciplinary connections across campus), Makerspace (engaging students in the design and creation of physical objects to promote creativity, problem-solving, research, and entrepreneurship), research exposure opportunities (building support and infrastructure and learning opportunities to ensure research experiences for all students) and course-based undergraduate experiences or "CUREs" (an introduction to research that engages an entire class in a semester-long, hypothesis-driven research problem). We focus our discussion on the assessment of CUREs in this article.

Structure – Professional Learning Communities

One of the key structures that needs to be in place to support the engaged assessment process is a space where faculty from a variety of disciplinary traditions and assessment professionals with expertise in documenting and reporting student learning can share and collaborate in meaningful ways. Instead of working in separate communities first and then sharing research and assessment outcomes at the end of the process, it is critical that the engagement of all partners is integrated throughout program and assessment discussions.

Faculty members from diverse disciplines collaborate in a faculty professional learning community to enhance teaching and learning through the implementation and assessment of course-based undergraduate research experiences (CUREs).

Building upon the principles of communities of practice (Wenger et al., 2002), a faculty professional learning community (FLC) was formed at UNC focusing on the implementation and assessment of CUREs. An FLC is a variety of community of practice; that is, a group of individuals who share a concern or a passion for an area of practice and learn how to do it better through regular interaction (Cox, 2003; Wenger, 1998). Through FLCs, a group of faculty members engage in a collaborative and sustained program of exploration that enhances the quality of teaching and learning (Cox, 2003; Cox & Richlin, 2004; Huber, 2008; Hutson & Downs, 2015). Participation in FLCs has been associated with faculty becoming more aware and respectful of others' viewpoints, cultures, and learning preferences, as well as increased research and publication in the scholarship of teaching and learning (Cox, 2003).

The FLC for CUREs at UNC included faculty members with differing content and professional expertise. Reflecting the CURE model's origin in the sciences (Corwin et al., 2015), several faculty members from various science disciplines joined the FLC. For example, FLC participants included biology faculty members who designed a seafood forensic lab course to support students' inquiries focusing on food mislabeling (Korzik et al., 2020; Spencer et al., 2020) and chemistry faculty who taught a large-enrollment introductory organic chemistry CURE course (Cruz et al., 2020). In addition, faculty members from the humanities also developed CURE courses and participated in the FLC. For example, a digital humanities CURE course was developed to engage students in converting life histories from the Federal Writers' Project into data to make text within the documents machine-readable and therefore easily searchable for future research (Rivard, 2019; Rivard et al., 2019). A law and literature CURE course supported students to read and analyze landmark court decisions alongside plays that were written in response to those cases (Larson & Rivard, 2018). Through the FLC,

faculty members from different disciplinary traditions had the opportunity to share their expertise and ways they adapt assessment measures and interpret assessment outcomes to enhance student learning in their respective courses.

Different from sporadic professional development sessions, an FLC has a coherent focus and typically leads to a shared product or outcome. At UNC, faculty members would like to increase the number of CURE courses to maximize student participation in these experiences and support students in developing research identities and become more empowered and capable of conducting research. Even though the content of the CURE courses may vary, the shared goals among faculty members include enhancing retention and graduation rates, increasing the inclusion and diversity of the research community, and expanding the number of research experiences in undergraduate education (University of North Carolina at Chapel Hill, 2017). The FLC discussions regarding program innovations and the assessment process centered on these shared goals and assessment resources were shared to support all faculty members for their CURE implementation.

Design thinking offers a solution to the assessment challenges of CUREs by integrating inspiration, ideation, and implementation in the assessment process, allowing for the measurement of the complex and evolving teaching and learning process.

Process – Design Thinking

CUREs offer a scalable and accessible research experience occurring within the context of a credit-bearing course, providing undergraduate students, regardless of past research experience, an opportunity to participate in an authentic research project (compared to simply enrolling in a regular course or lab). These courses offer students opportunities to develop and test their own hypotheses, collect their own data, experience iteration and failure, and potentially achieve discoveries that are new to the field (Sathy et al., 2020; 2021). This offers unique challenges for assessment since opportunities for applied research skill development may take priority over reviewing content and the outcomes of student work are not known ahead of time. The innovative and emergent nature of CUREs necessitates the integration of design thinking not only in curriculum and pedagogy (Koh et al., 2015; Wrigley & Straker, 2017), but also in the assessment process (Benson & Dresdow, 2014). As Brown (2008) stated, the design process is "a system of spaces rather than a predefined series of orderly steps" (p. 4). Integrating design thinking in the assessment process makes it possible to measure the complex and evolving process in teaching and learning and to inform decision-making in the process (Benson & Dresdow, 2014; Wehlburg, 2008).

To explore the assessment design for CUREs, FLCs engage in core design thinking activities including inspiration (surfacing problems and innovative solutions through interdisciplinary dialogs and collaborations), ideation (engaging in iterations of assessment process generation, prototyping, and experimentation), and implementation (conducting the assessment plan and sharing learning to further enhance instruction and the assessment process). One of the key challenges to the engaged assessment process is to integrate the assessment process in the least obstructive ways through instruction. With faculty members engaged in CUREs from different disciplinary areas, divergent ideas and perspectives emerged when discussing assessment design. These ideas contributed to the emergent design of the QEP assessment process and were implemented with the support of multiple stakeholders. Table 2 includes sample questions we used in FLC discussions to explore outcomes and measures, data collection and interpretation, and data use for program improvement.

As a result of these discussions, one of the CURE assessment measures, the Laboratory Course Assessment Survey (LCAS, Corwin et al., 2015) was adapted and used across CURE courses. The LCAS is a 17-item instrument used to differentiate CUREs from other courses by measuring student perceptions of the level of collaboration, discovery and relevance, and iteration that occurred within a given class. The instrument is comprised of three subscales: 1) collaboration; 2) discovery and relevance; and 3) iteration. Table 3 details each subscale and response options.

The instrument was originally piloted in biology labs with small enrollments (Corwin et al., 2015). Validation of the instrument followed Benson's (1998) three-stage process to specify the dimensionality, reliability, and validity. The three subscales on the 17-item instrument were established and confirmed through multiple iterations of exploratory factor analysis. The reliability of the instrument was measured using Cronbach's alpha. For traditional student

Table 2
Sample Assessment Discussion Questions

	Questions
Outcomes and Measures	How are course student learning outcomes aligned with program goals and the institutional mission? What are the direct and indirect measures embedded in the course design? What are the ideal measures that can be integrated in a meaningful way to measure student learning?
Data Collection and Interpretation	How can assessment data collection be least obstructive and most supportive of course delivery and program implementation? What assumptions and contexts do we need to consider when analyzing the data and interpreting the results?
Data Use for Program Improvement	How can data be used immediately for course or program improvement? How can data be used for long-term program improvement and development?

Table 3
Laboratory Course Assessment Survey (LCAS, Corwin et al., 2015)

Subscale	Focus	Number of Items	Response Options
Collaboration	student perceptions of how frequently they were encouraged to work together and share feedback, as well as their sense of developing metacognition toward research	6 items	1 = never, 2 = one or two times, 3 = monthly, or 4 = weekly
Discovery & Relevance	the degree to which students perceive themselves as having opportunities to create novel knowledge in the discipline and provide support for their findings	5 items	1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = somewhat agree, 5 = agree, and 6 = strongly agree
Iteration	students' perceptions of having opportunities to revise or repeat their work as problems or questions that arise in their research	6 items	1 = strongly disagree, 2 =disagree, 3 = somewhat disagree, 4 = somewhat agree, 5 = agree, and 6 = strongly agree

groups, the Cronbach's alpha was 0.83 for collaboration, 0.84 for discovery/relevance and 0.90 for iteration (Corwin et al., 2015).

As CURE courses expanded across the sciences, social sciences, and arts and humanities, faculty members from different disciplinary areas contributed to the adaptation of the instrument so that we can use the instrument to explore how CURE features vary across disciplines and course sizes. In fact, much of what we have learned about how well the instrument performs in these new contexts has originated in engaged collaboration among faculty, assessment staff, and students (Sathy et al., 2020; 2021).

Product – Data Use

The integration of design thinking allowed us to adapt the assessment measures for local use and provided more meaningful data discussion. The use of the LCAS as a common measure made it possible for FLC members to explore the relationship between student characteristics and LCAS data to draw implications for CURE instruction at the university level.

In addition to collecting student LCAS data, faculty also completed a somewhat modified version of the LCAS instrument that asked respondents to what degree they perceived their students as having experienced collaboration, discovery and relevance, and iteration in their classes. Asking faculty to answer questions aligned to the student LCAS questions provides faculty data about whether students experience the course the way they planned (Hogan et al., 2019). Discrepancies between student and faculty perceptions help faculty fine tune design, instructor talk, and teaching assistant training. Differences in faculty and student responses of sample courses are reflected in Table 4. Negative scores represent when the instructor's LCAS score was greater than the student's score, whereas positive scores represent when the student's LCAS score was greater than the instructor's LCAS score.

Table 4
Differences between faculty and student responses in sample courses

	Enrolled Students	Survey Respondents	Collaboration	Discovery & Relevance	Iteration
Polymer Chemistry	6	6	+0.33	-1.17	-3.50
Astronomy	12	8	-1.12	-2.5	-.87
Analytical Chemistry	16	10	+1.20	-6.3	-2.50
Statistics in Psychology	220	186	+3.69	-7.24	+5.49
Organic Chemistry	411	313	-5.52	-4.47	-6.97

In reviewing findings, faculty reacted to those comparisons. When examining the difference between faculty and student responses on iteration, a genetics instructor noted that the time constraints in the course led to limitations in the ability for students to repeat experiments:

> I would have liked to allow students more control in choosing experimental methods to use. However, students are presented with a novel problem in genetics. This means that it is not easy to determine how to solve that problem. It would take more than one semester to teach them all the potential methods before they have to choose some to use to solve the research problem.

The use of design thinking and LCAS as a common measure allowed for meaningful data discussion and exploring the relationship between student characteristics and CURE instruction.

In Statistics in Psychology–a large enrollment course–the instructor reflected on the difference in student and faculty scores on discovery and relevance, saying "I suspect the larger class size limited the perception of discovery in some aspects of formulating hypotheses because of the consensus approach we took with hypothesis generation." What this instructor learned led them to reframe how the discussion around hypothesis formulation occurred.

In the Organic Chemistry class, senior graduate teaching assistants, known as Graduate Research Consultants (GRCs), designed procedures and materials for the lab course and, in partnership with the lead instructor, transitioned the course from a traditionally structured lab to a CURE. When reviewing and discussing the LCAS assessment data, the lead instructor noted "I'd be curious to see how my TAs answered these questions. I'd like to think more about TA training in the future and how the TAs communicate with students about the process of

science." As answering one question often leads to many more new questions, the instructor engaged the GRCs in the assessment process to develop measures of student learning for the new CURE course and collected and analyzed student response data on the LCAS as well as measures of student project ownership and self-efficacy. The GRCs evidentially published a Scholarship of Teaching and Learning (SoTL) article describing how they transformed the course and delineating their assessment efforts (Cruz et al., 2020). The Organic Chemistry course exemplifies how the engaged assessment model may broaden opportunities to include other stakeholders, inform instructional decisions, and expand program-specific assessment efforts.

Sustainability – Assessment as Scholarship

Engaged assessment reframes assessment as scholarship, expanding opportunities for discipline experts to contribute to SoTL through collaborative dialogues and enhanced assessment capacity.

The engaged assessment effort offered an opportunity to reframe assessment as scholarship by supporting faculty members and GRCs in reframing the common measures used to better reflect their discipline. In addition to using assessment reports and outcomes to inform institution-specific program innovations, the enhanced assessment capacity and the collaborative dialogues also made it possible for discipline experts to contribute to SoTL through the engaged assessment process.

As Hutchings et al. (2011) note, assessment resembles the scholarship of teaching and learning in that they share "a focus on student learning, a more systematic evidence-based approach to educational quality, and a commitment to being more public about what and how well students are learning in college and university classrooms" (p. 6). However, SoTL inquiries generally originate in faculty interest in the impact of classroom practices while assessment has been associated with external and internal concerns about institutional effectiveness. That is, the scholarship of teaching and learning has tended to be a more decentralized, grassroots, classroom-centered effort by faculty while assessment has been seen as a centralized, directed initiative originating with the administration (Hutchings et al., 2011). Hutchings et al. (2011) noted that in the past, "assessment and the scholarship of teaching and learning have proceeded on more or less separate tracks—with their different histories, methods, and champions—each somewhat wary of the other." However, there is evidence that this perception is changing with the connection between the assessment of student learning and the scholarship of teaching and learning becoming increasingly apparent (Beach et al., 2016; Hutchings et al., 2011; Jankowski et al., 2018).

Through the engaged assessment process at UNC, content area experts expanded their collaborations among themselves and with the assessment team to share their insights through SoTL. For example, two faculty members from the English department considered the question of how CUREs–an instructional approach originating in the sciences–would differ when implemented in the humanities (Larson & Rivard, 2018). In addition to creating direct measures to reflect the CURE approach implemented in their classes, these faculty reviewed and modified the LCAS to reflect the nomenclature and practices found in a humanities classroom, collected data across their classes, and worked with the QEP assessment team to analyze data and review their findings. Two SoTL articles (Rivard, 2019; Rivard et al., 2019) have already been published, and other work is pending. Several other SoTL publications using the assessment data have been published, including two studies related to adapting the CURE model to psychology (Sathy et al., 2020; 2021) and a study led by a group of graduate students examining scaling up CUREs to high enrollment chemistry courses (Cruz et al., 2020).

Discussion and Implications

As we illustrated through the assessment process regarding the implementation the CUREs at UNC, the engaged assessment process differs from traditional institutional assessment models in terms of the structure, process, product, and sustainability (see Table 5). The monthly professional learning community meetings offered a platform to engage faculty from different disciplinary backgrounds in dialogues with the assessment team. With a focus on shared student learning outcomes at the institutional level, individuals crossed traditional disciplinary boundaries to engage in discussions regarding program innovations and the assessment process. Following the design-thinking principles, the engaged assessment process emphasized the iterative nature of assessment generation, adaptation, and validation.

The collaboration also made it possible to embed the assessment process in program implementation, instead of having it as a separate, add-on component. Ongoing dialogues centering on the interpretation and use of the data provided immediate input to inform program improvement and assessment enhancement. These discussions also contributed to assessment capacity building at the institution. With the development of assessment capacity and growing ownership of the assessment data, it was very encouraging for us to report the contribution to SoTL building upon the assessment data in this engaged process.

Table 5
Engaged Assessment Activities

	Activities	Outcomes
Structure -Boundary	Monthly professional learning community meetings Participation of faculty across disciplinary areas and assessment professionals	Program innovation and assessment focusing on shared student learning outcomes
Process - Design	Inspiration - surfacing problems and innovative solutions through interdisciplinary dialogs and collaborations Ideation - engaging in iterations of assessment process generation, prototyping, and experimentation Implementation - conducting the assessment plan and sharing learning to further enhance instruction and the assessment process	Identification, adaptation, and validation of common assessment measures Collaborations on data collection processes
Product - Data Use	Dialogues centering on interpretation and use of data to inform program innovations and assessment adaptations	Program improvement Instructor development Assessment enhancement
Sustainability - Mutual Learning	Contribution to the Scholarship of Teaching and Learning (SoTL)	Enhanced collaborative assessment capacity Generative impact through SoTL

The engaged assessment process offers implications for institutional leaders, assessment professionals, and faculty and teaching assistants interested in contributing to SoTL.

What we learned from our experience with the engaged assessment process may also offer further implications for institutional leaders, assessment professionals, and faculty and teaching assistants who are interested in contributing to SoTL through the engaged assessment process.

At the institution level, the creation of professional learning communities, or FLCs on our campus, offers the platform for cross-disciplinary collaborations centering on shared goals. At UNC, faculty members were incentivized to participate in these learning opportunities through FLCs, develop and implement assessment measures in their courses, share the results, and make their assessment tools available for other faculty members to adapt. In addition, a research summit offering a forum for faculty to share their research findings based on their teaching and to co-present with their students was held every fall. These engagement opportunities expand the dialogs campus-wide, augmenting faculty discussions about improving student learning, and supporting the assessment capacity building across units on campus.

Further, assessment professionals have a key role to play in the engaged assessment process. In addition to assisting faculty members with adapting common measures or identifying new ones, our assessment team engages in individual and group discussions with faculty members, tracks changes to the instruments over time, maintains the data collected across all courses, and provides the overall analysis of the outcomes and impact of CURE courses. Through this collaboration, the assessment team also developed its capacity to offer

more contextualized data collection, analysis, and reporting support to extend the use of assessment data beyond summative reporting that captures program effectiveness and impact. Instead of serving as a silent observer walking alongside faculty members implementing the curriculum innovations, the assessment team and faculty members negotiated the directions and methods throughout the program implementation journey and had more opportunities to offer just-in-time assessment support to inform program decision making.

Dialogs and collaborations among faculty members across disciplinary areas enriched the assessment discussions and augmented the use of assessment for program improvement.

Finally, our engaged assessment experiences through CUREs also led to faculty members and TAs' scholarly development, especially in terms of the collaborative contribution to the SoTL research regarding CUREs. The use of the common measure across disciplinary areas allowed the team to contribute to the larger research agenda regarding the expansion of CUREs in higher education settings in a more systematic manner. This type of interactions and scholarly engagement can be especially beneficial for future faculty who are developing their professional network and connections in an interdisciplinary manner.

Conclusion

Faculty engagement is critical in student learning outcome assessment in higher education settings. The engaged assessment model at UNC exemplified the potential of faculty engagement at a leading doctoral university. FLCs offered space for the exploration of assessment designs, the examination of assessment data, and the celebration of assessment as scholarship. Dialogs and collaborations among faculty members across disciplinary areas enriched the assessment discussions and augmented the use of assessment for program improvement. As CUREs were integrated into the new general education requirements, the assessment activities were also carried over. To sustain and further expand the engaged assessment model at the institution, the assessment leadership team plans to continue these FLC discussions. This offers continued opportunities for faculty members to have a central role in assessment of general education, collaborating with the assessment team to identify appropriate measures and designs to illustrate the impact of the new general education curriculum on student learning. In addition to meeting accountability requirements, the design thinking activities of inspiration, ideation, and implementation are integrated into instructional practices to normalize the engaged assessment model to support curriculum innovation and program improvement.

Future collaborations with other stakeholders, including students and community partners, in the engaged assessment process could further strengthen and realize the implementation of transformative assessment to enhance teaching and learning. We have already observed the positive impact of students becoming involved in assessment activities, including serving on general education committees and working groups, and participating in internships with the assessment team in which they took responsibility for identifying assessment tools and collecting data. Student involvement in these efforts have deepened our understanding of student perceptions of how they are assessed, and how they use the information they receive through assessment processes. Extending the engaged assessment model to include students in collaboration and dialog with faculty members and assessment staff may prove to be the next stage in promoting a more holistic and empowering approach to assessing student learning at our university.

References

American Association for Higher Education (AAHE). (1992). *Principles of good practice for assessing student learning*. American Association for Higher Education.

Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers* (2nd ed.). Jossey-Bass.

Arum, R., & Roksa, J. (2011). Limited learning on college campuses. *Society*, *48*(3), 203-207.

Banta, T. W., & Blaich, C. (2011). Closing the assessment loop. *Change: The Magazine of Higher Learning*, *43*(1), 22-27.

Beach, A. L., Sorcinelli, M. D., Austin, A. E., & Rivard, J. K. (2016). *Faculty development in the age of evidence: Current practices, future imperatives*. Stylus.

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

Benson, J., & Dresdow, S. (2014). Design thinking: A fresh approach for transformative assessment practice. *Journal of Management Education*, *38*(3), 436-461. https://doi.org/10.1177/1052562913507571

Brown, T. (2008). Design thinking. *Harvard Business Review*, *86*(6), 84-92.

Corwin, L. A., Runyon, C., Robinson, A., & Dolan, E. L. (2015). The Laboratory Course Assessment Survey: A tool to measure three dimensions of research-course design. *CBE Life Sciences Education*, *14*(4), ar37. https://doi.org/10.1187/cbe.15-03-0073

Cox, M. D. (2003). Fostering the scholarship of teaching through faculty learning communities. *Journal on Excellence in College Teaching*, *14*(2/3), 161-198.

Cox, M. D., & Richlin, L. (Eds.). (2004). *Building faculty learning communities*. Jossey-Bass.

Cruz, C. L., Holmberg-Douglas, N., Onuska, N. P. R., McManus, J.B., MacKenzie, I.A., Hutson, B.L., Eskew, N. A., & Nicewicz, D.A. (2020). Development of a large-enrollment course-based research experience in an undergraduate organic chemistry laboratory: Structure–function relationships in pyrylium photoredox catalysts. *Journal of Chemical Education*, *97* (6), 1572-1578. https://doi.org/10.1021/acs.jchemed.9b00786

Ewell, P. T. (2009, November). *Assessment, accountability, and improvement: Revisiting the tension* (NILOA Occasional Paper No. 1). Urbana, IL: University of Illinois and Indiana University, National Institute of Learning Outcomes Assessment.

Grunwald, H., & Peterson, M. W. (2003) Factors that promote faculty involvement in and satisfaction with institutional and classroom student assessment. *Research in Higher Education*, *44*, 173-204. https://doi.org/10.1023/A:1022051728874

Hogan, K. A., Bruno, J. F., Steinwand, B. J., Sathy, V., Robertson, S., Nasiri, M., Strauss, C., and Hutson, B. L. (2019, July). *Do faculty in a college-wide CURE program achieve the design goals they planned via a year-long faculty learning community?* Poster presented at The Society for the Advancement of Biology Education Research (SABER), Twin Cities, MN.

Huber, M. T. (2008). *The promise of faculty inquiry for teaching and learning basic skills*. The Carnegie Foundation for the Advancement of Teaching.

Hutchings, P. (2010). *Opening doors to faculty involvement in assessment*. (NILOA Occasional Paper No. 4). University of Illinois and Indiana University, National Institute of Learning Outcomes Assessment.

Hutchings, P. (2016, January). Aligning educational outcomes and practices. (Occasional Paper No. 26). University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Hutchings, P., Huber, M. T., & Ciccone, A. (2011). Getting there: An integrative vision of the scholarship of teaching and learning. *International Journal of the Scholarship of Teaching and Learning*, *5*(1). https://doi.org/10.20429/ijsotl.2011.050131.

Hutson, B. L., & Downs, H. (2015). The College STAR faculty learning community: Promoting learning for all students through faculty collaboration. *Journal of Faculty Development*, *29*(1), 25-32.

Hundley, S. P., & Kahn, S. (2019). *Trends in assessment: Ideas, opportunities, and issues for higher education*. Stylus Publishing.

Jankowski, N. A., & Marshall, D. W. (2017). *Degrees that matter: Moving higher education to a learning systems paradigm*. Stylus Publishing.

Jankowski, N. A., Timmer, J. D., Kinzie, J., & Kuh, G. D. (2018). *Assessment that matters: Trending toward practices that document authentic student learning*. University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Kinzie, J., Landy, K., Sorcinelli, M. D., & Hutchings, P. (2019). Better together: How faculty development and assessment can join forces to improve student learning. *Change: The Magazine of Higher Learning*, *51*, 46-54. https://doi.org/10.1080/00091383.2019.1652076

Koh, J. H. L., Chai, C. S., Wong, B., & Hong, H.-Y. (2015). *Design thinking for education: Conceptions and applications in teaching and learning*. Springer. https://doi.org/10.1007/978-981-287-444-3

Korzik, M. L., Austin, H. M., Cooper, B., Japerse, C., Tan, G., Richards, E., Spencer, E. T., Steinwand, B., Fodrie, J., & Bruno, J. F. (2020). Marketplace shrimp mislabeling in North Carolina. *PLoS ONE 15*(3): e0229512. https://doi.org/10.1371/journal.pone.0229512

Kuh, G. D., & Ikenberry, S. O. (2009, October). *More than you think, less than we need: Learning outcomes assessment in American higher education*. University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. (2014, January). *Knowing what students know and can do: The current state of learning outcomes assessment at U.S. colleges and universities*. University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Larson, J., & Rivard, C. (2018, March). Creating course based undergraduate research experiences (CUREs) in the humanities. UNC System-Wide Undergraduate Research Development Summit, Greensboro, NC, United States.

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

Metzler, E. T., & Kurz, L. (2018). Assessment 2.0: An organic supplement to standard assessment procedure. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Middaugh, M. F. (2010). *Planning and assessment in higher education: Demonstrating institutional effectiveness*. Jossey-Bass.

Reder, M., & Crimmins, C. (2018). Why assessment and faculty development need each other: Notes on using evidence to improve student learning. *Research & Practice in Assessment*, *13*, 15-19.

Rivard, C. (2019) Turning archives into data: Archival rhetorics and digital literacy in the composition classroom. *College Composition and Communication*, *70*(4), 527-559.

Rivard, C., Arnold, T., & Tilton, L. (2019). Building pedagogy into project development: Making data construction visible in digital projects. *Digital Humanities Quarterly*, *13*(2). http://www.digitalhumanities.org/dhq/vol/13/2/000419/000419.html

Sathy, V., Nasiri, M., Strauss, C., & Hutson. B. (2020). The CURE for broadening participation in undergraduate teaching. In T. M. Ober, E. Che, J. E. Brodsky, C. Raffaele, & P. J. Brooks (Eds.). *How we teach now: The GSTA guide to transformative teaching* (pp. 429-444). Society for the Teaching of Psychology. http://teachpsych.org/ebooks/howweteachnow-transformative

Sathy, V., Strauss, C. L., Nasiri, M., Panter, A. T., Hogan, K. A., & Hutson, B. L. (2021). Cultivating inclusive research experiences through course-based curriculum. *Scholarship of Teaching and Learning in Psychology*, *7*(4), 312-322. https://doi.org/10.1037/stl0000215

Southern Association of Colleges and Schools, Commission on Colleges (SACSCOC). (2020). *The Quality Enhancement Plan*. Available at: https://sacscoc.org/app/uploads/2020/01/Quality-Enhancement-Plan-1.pdf

Spencer, E. T., Richards, E., Steinwand, B., Clemons, J., Dahringer, J., Desai, P., Fisher, M., Fussell, S., Gorman, O., Jones, D., Le, A., Long, K., McMahan, C., Moscarito, C., Pelay, C., Price, E., Smith, A., VanSant, A., & Bruno, J. F. (2020) A high proportion of red snapper sold in North Carolina is mislabeled. *PeerJ 8*:e9218 https://doi.org/10.7717/peerj.9218

Suskie, L. (2014). *Five dimensions of quality: A common sense guide to accreditation and accountability*. Jossey-Bass/Wiley.

University of North Carolina at Chapel Hill (2017). *Quality Enhancement Plan: Creating Scientists – learning by connecting, doing, and making*. University of North Carolina.

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

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press. https://doi.org/10.1017/CBO9780511803932

Wenger, E., McDermott, R., & Snyder, W. M. (2002). Seven principles for cultivating communities of practice. In E. Wenger, R. McDermott, & W. M. Snyder (Eds.), *Cultivating communities of practice: A guide to managing knowledge* (pp. 49-64). Harvard Business School Press.

Wrigley, C., & Straker, K. (2017). Design thinking pedagogy: The educational design ladder. *Innovations in Education and Teaching International*, *54*(4), 374-385. https://doi.org/10.1080/14703297.2015.1108214

Abstract

As a result of the COVID-19 pandemic, many instructors were forced to adjust from in-person to emergency remote teaching; however, classroom observation protocols, like the Classroom Observation Protocol for Undergraduate STEM (COPUS), have only been developed and validated for in-person instruction. Therefore, we developed and validated an adapted version of COPUS, called Online COPUS (O-COPUS), to measure teaching and learning practices during online synchronous instruction. We collected COPUS and O-COPUS data from 35 STEM instructors teaching via in-person and online synchronous instruction at a research-intensive, minority-serving institution (MSI). By identifying emergent codes from live observations and using an exploratory coding process, we adjusted six instructor and six student COPUS code descriptions for O-COPUS. As we prepare for teaching in the future, it is important to have formative assessment tools, like classroom observation protocols, designed for all course formats to be able to measure and improve pedagogical practices in college STEM classrooms.

AUTHORS

Téa S. Pusey
University of California, Merced

Andrea Presas Valencia
University of California, Merced

Adriana Signorini, Ed.D.
University of California, Merced

Petra Kranzfelder, Ph.D.
University of California, Merced

Breakout Rooms, Polling, and Chat, Oh COPUS! The Adaptation of COPUS for Online Synchronous Learning

Instructors and the teaching practices they employ play a critical role in improving student learning in science, technology, engineering, and mathematics (STEM) courses (Smith et al., 2013b). Active learning is an evidence-based teaching practice which requires students to engage cognitively and meaningfully with the course materials (Bonwell & Eison, 1991; Chi & Wylie, 2014). There are many benefits associated with the implementation of active learning (Chickering & Gamson, 1987; Crouch & Mazur, 2001; Freeman et al., 2014), such as improved student attitudes (Armbruster et al., 2009) and retention of course material (Pérez-Sabater et al., 2011; Schwartz et al., 2011; Vanags et al., 2013) as they are practices that improve learning for all students, particularly persons excluded because of their ethnicity or race (Asai, 2020). Recently, Denaro et al. (2021) noted a national focus on implementing active learning to improve the quality of STEM education promoted by, among others, the National Research Council (2012), Olson and Riordan (2012), and Laursen (2019). Therefore, shifting large numbers of STEM faculty to include even small amounts of active learning in their teaching may effectively educate far more students and raise retention of undergraduate STEM students (Owens et al., 2017).

Ongoing formative assessment is a key instructional practice in student-centered learning environments (MacIsaac, 2019; Offerdahl et al., 2018; Rosenberg et al., 2018). Historically, undergraduate teaching has been predominantly transmissionist in nature with the goal of conveying information to students (Barr & Tagg, 1995). However, with the benefits associated with active learning, it is imperative that STEM instructors consider

CORRESPONDENCE

Email
pkranzfelder@ucmerced.edu

how they can effectively implement student-centered practices in all formats of their classrooms. Therefore, an array of tools have been developed over the past two decades to measure enacted teaching practices in the in-person context, especially in STEM courses (Eddy et al., 2015; Owens et al., 2017; Sawada et al., 2002; Smith et al., 2013b). To name a few, the Practical Observation Rubric To Assess Active Learning (PORTAAL) (Eddy et al., 2015), the Decibel Analysis for Research in Teaching (DART) (Owens et al., 2017), the Reformed Teaching Observation Protocol (RTOP) (Sawada et al., 2002), and the Classroom Observation Protocol for Undergraduate STEM (COPUS) (Smith et al., 2013b) provide a way of collecting unbiased data by a trained third-party or an application like the Generalized Observation and Reflection Platform (GORP) (Tomkin et al., 2019; University of California Davis, 2018; van der Lans, 2018). Out of all of these classroom observation protocols, COPUS has commonly been used to examine instructor and student behaviors using 25 codes occurring during in-person teaching in STEM classrooms (Smith et al., 2013b). Table 1 offers a description of the COPUS codes: 12 instructor behaviors, such as *lecturing* and *moving and guiding*, and 13 student behaviors, such as *listening* and *asking questions*.

For faculty professional development and education research efforts, COPUS codes have been characterized as traditional lecturing or active learning through various descriptive methods or the presence of various COPUS codes among different population of instructors (Akiha et al., 2018; Kranzfelder, Lo, et al., 2019; Lewin et al., 2016; Reisner et al., 2020; Smith et al., 2013b). For example, Akiha et al. (2018) calculated the frequency of various COPUS codes to analyze the differences in how instructors implemented traditional lecturing and active learning in middle school, high school, first-year university, and advanced university STEM classes. Results showed that students in middle school and high school participated in more active learning activities, such as group work, than first-year or advanced university STEM classes. Furthermore, middle school and high school instructors were also more active in the classroom, such as moving and guiding their students in active learning tasks, than first-year or university classes (Akiha et al., 2018). Also, analyses have been conducted through statistical methods to describe how instructors' teaching practices are related across COPUS codes (Commeford et al., 2022; Denaro et al., 2021; Stains et al., 2018; Tomkin et al., 2019). Tomkin et al. (2019) used regression models to assess the differences in teaching practices between instructors who were involved in a community of practice (CoP) (Wenger, 1996) versus those that were not. They found that instructors who were a part of a CoP implemented more active learning practices compared to non-CoP instructors (Tomkin et al., 2019). Additionally, COPUS data have been used in combination with other tools, such as the Classroom Discourse Observation Protocol (CDOP) (Alkhouri et al., 2021; Kranzfelder, Bankers-Fulbright, et al., 2019) and Instructor Talk (Lane et al., 2021; Seidel et al., 2015). Lane et al. (2021) examined what instructors teaching STEM courses do on their first day of class. They paired COPUS and Instructor Talk data to suggest that negatively phrased Instructor Talk was less common among instructors who used student-centered teaching practices (Lane et al., 2021). As a result, COPUS data have been analyzed and applied in a variety of ways to characterize STEM classroom behaviors.

COPUS codes have been used to analyze differences in traditional lecturing and active learning in STEM classrooms, and have been applied in various ways to characterize STEM classroom behaviors, including during the sudden shift to emergency remote teaching during the COVID-19 pandemic.

As a result of the COVID-19 pandemic, many instructors were forced to transition their course modality rapidly from in-person to online instruction. The sudden shift was constituted as emergency remote teaching (ERT) (Hodges et al., 2020). In contrast to traditional online teaching and learning, which has been studied and implemented for decades (e.g., Ally, 2004; Darby & Lang, 2019; Means et al., 2014; Nilson & Goodson, 2021), ERT is a temporary shift from in-person, blended, or hybrid courses to fully online teaching due to crisis circumstances. It provided access to online instruction and instructional supports in a manner that was quick to set-up, reliable (Hodges et al., 2020), and required instructors who were mostly inexperienced with online teaching to become familiar with new teaching tools and techniques including asynchronous and synchronous formats (Giesbers et al., 2014; Nilson & Goodson, 2021; Skylar, 2009). Some instructors choose to implement an asynchronous format (e.g., recorded lectures, discussion boards, and at-home assignments) if they were concerned about their own or their students' abilities to attend and participate in live online lectures (Lemke, 2022; Van Heuvelen et al., 2020). In contrast, instructors more often adopted a synchronous format (e.g., videoconference call or live online lectures) as this can encourage student-instructor interactions and group work (Heiss & Oxley, 2021;

Table 1
COPUS Coding Scheme

	COPUS Codes	COPUS Code Descriptions
Instructor Doing	Lecturing (Lec)	Lecturing (presenting content, deriving mathematical results, presenting a problem solution, etc.)
	Real-time Writing (RtW)	Realtime writing on board, doc. projector, etc. (often checked off with Lec)
	Demo or Video (D/V)	Showing or conducting a demo, experiment, simulation, video, or animation
	Follow-up (Fup)	Follow-up/feedback on clicker question or activity to entire class
	Posing a question (PQ)	Posing non-clicker question to students (nonrhetorical)
	Clicker question (CQ)	Asking a clicker question (mark the entire time the instructor is using a clicker question, not just when first asked
	Answering questions (AnQ)	Listening to and answering student questions with the entire class listening
	Moving and guiding (MG)	Moving through class guiding ongoing student work during active learning tasks
	One on one (1o1)	One on one extended discussion with one or a few individuals, not paying attention to the rest of the class
	Administration (Adm)	Administration (assign homework, return tests, etc.)
Students Doing	Group Clicker Question (CG)	Discuss clicker question in groups of 2 or more students
	Group Worksheet (WG)	Working in groups on worksheet activity
	Other Group Work (OG)	Other assigned group activity, such as responding to instructor question
	Answering questions (AnQ)	Student answering a question posed by the instructor with rest of class listening
	Student Question (SQ)	Student asks question
	Whole class discussion (WC)	Engaged in whole class discussion by offering explanations, opinion, judgment, etc. to whole class, often facilitated by instructor
	Prediction (Prd)	Making a prediction about the outcome of demo or experiment
	Student Presentation (SP)	Presentation by student(s)
	Test or Quiz (TQ)	Test or quiz
	Waiting (W)	Waiting (instructor late, working on fixing AV problems, instructor otherwise occupied, etc.)
	Other (O)	Other – explain in comments

Note: COPUS codes and code descriptions from Smith et al. (2013a).

Lang, 2020; Van Heuvelen et al., 2020). In the systematic review of research on ERT in higher education before the pandemic, text-based tools used for asynchronous instruction (e.g., discussion forums) were most often used by instructors (Bond et al., 2020). During the pandemic, they found that there was a much higher use of synchronous collaboration tools, especially video conferencing platforms like Zoom, Teams, and Google Meet (Bond et al., 2021). Nevertheless, teaching in a synchronous format does not guarantee student participation; for example, Reinholz et al. (2020) found an overall decrease in student participation in biology classrooms as the class transitioned from in-person to synchronous instruction during ERT. However, the transition from an in-person to online synchronous instruction mid-course presented many challenges, including maintaining active student engagement (Giordano & Christopher, 2020).

Shift to online synchronous instruction during ERT led to innovative classroom pedagogical practices, including the use of breakout rooms, virtual whiteboards, and live polling tools to engage students actively, but there is a need for reliable assessment tools to measure these practices.

During the COVID-19 pandemic, there have been a few studies that have documented teacher and student behaviors because of the shift to online synchronous instruction during ERT. Some instructors were not able to implement the best active learning strategies for online learning (Youmans, 2020), but others approached this challenge with creativity leading to opportunities for classroom pedagogical innovations including adaptations of student-centered activities. To recreate activities that were administered before ERT, Tan et al. (2020) utilized the Zoom breakout room function which creates small videocall rooms within the main virtual meeting, as well as Padlet, a virtual whiteboard. In pre-assigned breakout rooms, students would have discussions facilitated by an instructor or teaching assistant who were present in each breakout room. By asking students to turn their microphone functions on, groups were highly engaged in discussions. Singhal (2020) also utilized breakout rooms when assigning group active learning activities and moved between groups as they worked collaboratively. Tan et al. (2020) also utilized Poll Everywhere, an online tool for live polling to engage students actively during online synchronous instruction. Similarly, Christianson (2020) utilized Socrative, another online tool for live polling, to assign their students group quizzes at the beginning of class. During the administration of the quiz, students used Microsoft Teams to engage in group discussion on the quiz questions. Tan et al. (2020) found that the Zoom chat function, a messaging system within the video call room that allows participants to send messages to the group or direct messages to each other, was valuable to maintain interactions among faculty, teaching assistants, and undergraduates in the course. Researchers also found that students in the course seemed to respond to more questions and participate more in the chat compared to in-person discussions. In a large-enrollment biochemistry course, Dingwall (2020) designed templates for metabolic pathways that students could actively fill out during lecture. Students agreed that these templates were useful in online synchronous instruction because it allowed them to engage in lecture material rather than passively listen to their instructor. Therefore, while there have been studies documenting what classroom technology tools were successful in implementing active learning strategies during ERT (Christianson, 2020; Dingwall, 2020; Singhal, 2020; Tan et al., 2020), studies have not been conclusive about how to document the specific teaching and learning practices that have implemented in online synchronous instruction. More specifically, formative assessment tools are needed to measure these practices in a reliable and valid manner.

During ERT, instructors, institutional assessment programs, and biology education researchers faced a problem of not having reliable, validated classroom observation tools that could be easily implemented online by trained observers to measure teaching and learning behaviors. As described above, Smith et al. (2013b) developed and validated COPUS for in-person instruction, so we responded to the need to adjust some COPUS code descriptions to document teaching and learning practices more efficiently for online synchronous instruction. Although adjusting the original COPUS code descriptions to fit online synchronous instruction may seem like a seamless transition, many universities stopped conducting COPUS observations during ERT due to its complexities. For instance, UC Irvine put COPUS observations on hold because they "were lacking the resources to validate a novel observation protocol in the face of the numerous other COVID-19-related challenges" (personal communication with Brian Sato, 07/20/2021). Additionally, UC San Diego commented: "We had trained undergrads to do the observations and didn't think we could ask them on the fly to adjust things" (personal communication with Melinda Owens,

07/20/2021). Some institutions utilized COPUS as an assessment tool to support instructors while transitioning to online synchronous instruction (Clark et al., 2020); however, they did not validate the tool for this new study context (i.e., ERT).

Study goal and objectives

Therefore, out of necessity, the goal of this study was to adjust the original COPUS code descriptions to document online synchronous teaching and learning practices as well as online functions that instructors may incorporate into their future teaching practices. This adaptation was not intended to capture asynchronous teaching and learning practices as the original COPUS was developed around live in-person class sessions. By adapting COPUS for online synchronous instruction, instructors will have the ability to make comparisons between past, present, and future teaching and learning practices as they move to other instructional modalities, such as in-person, hybrid, or hybrid-flexible (HyFlex) (Beatty, 2019). The objectives for this case study were to:

In adapting COPUS for online synchronous instruction, instructors will have the ability to make comparisons between past, present, and future teaching and learning practices as they move to other instructional modalities, such as in-person, hybrid, or hybrid-flexible.

1. Adapt and validate COPUS for online synchronous instruction (O-COPUS) and effectively train observers to collect reliable COPUS and O-COPUS data.
2. Create an O-COPUS codebook that captures commonly observed online teaching and learning practices.
3. Showcase sample data that instructors or researchers might obtain from original COPUS compared to the adapted O-COPUS protocol.

Case Study

This study was approved by UC Merced's Institutional Review Board, and all participating instructors provided informed consent to anonymously participate in the study (Protocol ID UCM2020-3).

Participants and instructional context

For this study, we drew on previously collected classroom data from 40 undergraduate and graduate STEM courses taught by 35 instructors at the University of California, Merced (UC Merced), a research-intensive, minority-serving institution (MSI) in the western United States. This larger ongoing research project was funded by two research grants: the Howard Hughes Medical Institute Inclusive Excellence (HHMI IE) awarded to the biology program and the National Science Foundation Hispanic Serving Institution (NSF I) awarded to the chemistry and biochemistry department with the goal of understanding, documenting, planning, and enacting meaningful initiatives to improve teaching and student learning at UC Merced. As part of that project, we collected data from STEM instructors each semester before the transition to ERT (fall 2018 through spring 2020), during the transition to ERT (spring 2020), and/or during the continuation of ERT (fall 2020 and spring 2021) (Table 2). We chose instructors for that larger project who had: 1) taught either a lower or upper division undergraduate or graduate STEM course; 2) taught a lecture course (excluding laboratory, discussion, or seminar courses); and 3) either taught in-person or via online synchronous instruction (excluding asynchronous instruction). Lund et al. (2015) found that at least three successive classroom observations are necessary to characterize adequately an instructor's teaching practices; therefore, we conducted at least three classroom observations per instructor.

Descriptive information about the instructors and courses included in this study can be found in Table 2. Instructors taught mainly lower division undergraduate courses from a variety of STEM disciplines with the majority being in biology or chemistry. All three instructor types from our institution (tenure-track research faculty, tenure-track teaching faculty, and non-tenure track contingent faculty, i.e., lecturers) were observed with the majority being tenure-track research faculty. Course class sizes ranged from four to 292 students with a mean of 110 students. One of the authors was also one of the participating instructors; we did not collect or analyze the data.

Table 2
Instructor and course demographics

	In-person		Online	
Characteristics	Fall 2018 – Spring 2020	Spring 2020	Fall 2020	Spring 2021
Discipline				
Biology	16	9	6	7
Chemistry	9	1	5	5
Mathematics	4	4	0	0
Physics	4	2	0	0
Engineering	2	0	0	0
Instructor type				
Research faculty	14	8	5	5
Teaching faculty	8	4	2	3
Lecturers	13	4	4	4
Course Size				
Small (≤60 students)	12	6	1	1
Medium (61-100 students)	3	3	0	0
Large (>101 students)	20	7	10	11
Class level				
Lower division	25	9	10	10
Upper division	7	5	1	1
Graduate	3	2	0	0
Total	**35**	**16**	**11**	**12**

Note: Classroom observation data were collected from 40 STEM courses taught by 35 different instructors. During the transition and continuation of emergency remote teaching, some instructors from our original study did not continue their participation.

Methods

Students Assessing Teaching and Learning (SATAL)

Undergraduate interns from UC Merced's SATAL program play a vital role in implementing a wide range of assessment tools, including COPUS, to improve classroom practices and gather student perspectives.

Since 2009, the undergraduate interns from UC Merced's Students Assessing Teaching and Learning (SATAL) collaborate with faculty who are focused on pedagogical and curricular exploration and have the desire to understand their students' experiences and perspectives in order to inform classroom practices (Signorini & Pohan, 2019). To accomplish this, SATAL implements a wide range of assessment tools for gathering student perspectives, including classroom interviews, surveys, and COPUS. SATAL interns work with faculty to provide assessment results and feedback to improve students' experience in their class. Since 2018, the undergraduate interns from the SATAL program have partnered with faculty to conduct multiple COPUS research projects for different purposes. Therefore, the adaption of COPUS during ERT came naturally to the SATAL program as they continued to collaborate with faculty during ERT on COPUS research projects.

COPUS Data Collection

During in-person COPUS observations, SATAL interns followed the COPUS codebook in Smith et al. (2013b) to document instructor and student behaviors in 2-minute intervals throughout the duration of the class session. We created a COPUS Frequently Asked Questions (FAQs) to describe in further detail whom we code (leading instructor, co-instructor, or teaching assistant), what behaviors we categorize under different COPUS codes, and what codes we pair together when a particular behavior occurs (File S1, S2).

We refined the codes descriptions using focused coding and then we reached coder agreement through group consensus.

COPUS Reliability

We trained 15 SATAL interns for four hours between two training sessions for these in-person COPUS observations. Each of the two sessions consisted of pre- and post-activities as well as a 45-minute coding activity which utilized in-person lecture recordings by instructors of our home institution. These training sessions followed an adapted and extended version of the COPUS training in Smith et al. (2013b) (File S3). To quantify the degree of agreement between observers, we calculated inter-rater reliability (IRR) using Fleiss' Kappa. Landis and Koch (1977) suggested the following interpretations of Fleiss' Kappa (κ): 0.0-0.20 poor to slight agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and 0.81-1.0 almost perfect agreement. Before conducting in-person observations, we trained observers until we reached moderate agreement ($\kappa = 0.56$,7 95% CI: 0.55-0.56) (File S4).

Transitioning to O-COPUS

As we transitioned to online synchronous instruction during ERT, observers continued to code observations using the original COPUS codebook (Smith et al., 2013b). Observers took detailed notes of instructor behaviors and their uses of online functions, such as the messaging function, as well as student behaviors. When observers attended online classroom observations, instructors would make them co-host of the online meeting, allowing them to see reactions better, such as raised hands and thumbs up, as well as permission to move throughout breakout rooms. Observers coded classroom observations synchronously so private features in the messaging function, such as direct messaging, could not be captured. During observations, observers took detailed notes of how instructors and students utilized online functions, such as those mentioned above, which would be the basis for identifying new codes.

Qualitative Coding Process to Adapt COPUS Codes for O-COPUS

We began developing O-COPUS code descriptions using an exploratory coding process with data-driven (inductive) choices (Saldaña, 2015). To do this, during online synchronous observations, we identified the new online teaching and learning behaviors in the COPUS notes section (Smith et al., 2013b). Then, we used a deductive approach to assign the observed online teaching and learning behaviors to a pre-existing COPUS code that related to the students and instructor's behaviors. During the last cycle of analysis, we refined the codes descriptions using focused coding and then we reached coder agreement through group consensus. In this last cycle, our research team determined the inclusion (and exclusion) of O-COPUS codes and code descriptions (File S5, S6).

O-COPUS Validity

Originally, COPUS was developed and validated to measure student and instructor behaviors inside the in-person classroom (Smith et al. 2013). To ensure trustworthiness of the COPUS results, authors of the original COPUS protocol obtained feedback from science education specialists and K-12 instructors to ensure the code and code descriptions were valid in capturing student and instructor behaviors (Smith et al. 2013). This is an example of validation through peer review or debriefing (Creswell & Miller, 2000). To ensure trustworthiness of O-COPUS results, multiple validity approaches were implemented including using prolonged engagement in the field, peer review, and rich, thick descriptions (Creswell & Miller, 2000). First, as described in detail in the sections above, our team has

been collecting COPUS data since 2018 indicating our prolonged engagement in the field and understanding of the tool. Second, like Smith et al. (2013), we used peer review or debriefing by getting feedback from a group of STEM educators and discipline-based education researchers (DBER) at a research-intensive university unrelated to the institution in this study (n = 11). The expert feedback panel activities were organized into two parts. In the first part, authors TP and APV collectively presented a subset of the instructor and student codes for the panelists to evaluate the O-COPUS code descriptions. To do this, we showed panelists the original COPUS code descriptions compared to our O-COPUS code descriptions and asked panelists if: 1) these behaviors were applicable to them in online synchronous instruction; 2) there were any additional behaviors that they had observed which we should be included; and 3) any of our code descriptions needed further clarification. Authors AS and PK were present to take notes on the feedback. Originally, our O-COPUS codebook contained the original COPUS descriptions (Smith et al., 2013b), our research team's clarifications of the COPUS code descriptions, as well as our O-COPUS code descriptions.

However, the expert panel suggested that we focus our code descriptions on online behaviors, not clarifications to in-person COPUS codes. For instance, in our original codebook, we clarified that our research team included the instructor going over learning outcomes as a *lecturing* behavior instead of *administration*. Our expert panel pointed out that this was not a change that emerged because of online synchronous instruction, rather our team's interpretation of the code. Hence, we created our COPUS FAQ (Files S1, S2) files to demonstrate our interpretations of the in-person COPUS codes and not distract from our O-COPUS code descriptions. Additionally, based on the feedback received from the expert panel, we realized that some of our original O-COPUS code descriptions were not specific enough. For example, our original O-COPUS code description for *moving and guiding* was: "Moving throughout breakout rooms guiding students or guiding students while they are working on a problem or clicker question (hints/working through a problem) using the microphone or chat function during active learning tasks." However, our expert panel determined that "guiding students while they are working on a problem or clicker question" was not specific enough to translate to a *moving and guiding* behavior. Therefore, we adjusted our O-COPUS code description for *moving and guiding* to include specific instructor behaviors: "Moving through breakout rooms guiding ongoing student work during active learning task or guiding students while they are working on an active learning task by providing hints or working through a problem using the microphone or messaging function." Lastly, the expert panel pointed out that the original vocabulary we were using in our code descriptions, such as "chat", "microphone", and "annotation tool," were specific to the platform Zoom; however, not all institutions utilized Zoom as their online meeting platform for synchronous online instruction. Therefore, we adjusted the terminology used in the code descriptions to fit multiple meeting applications, such as Skype or Google Meet, so other institutions could adapt our codebook even if they used a different application (Tables 2-3). Our original O-COPUS codebook with the feedback from the expert panel can be found in supplemental materials (File S7) as well as our development flowchart of O-COPUS (File S8). Third, we used rich, thick descriptions of the codes and code descriptions to convey our O-COPUS findings in the results that follow.

Our expert panel suggested that we focus our code descriptions on online behaviors, not clarifications to in-person COPUS codes.

O-COPUS Reliability

Once we developed and validated our O-COPUS codebook, we trained 23 SATAL interns following the same training as mentioned above, but we utilized both in-person and online synchronous lecture recordings. To quantify the degree of agreement between observers, we calculated IRR using Fleiss' Kappa. Before conducting online observations, we trained observers until we reached substantial agreement (κ = 0.67, 95% CI: 0.66-0.67) (File S9). In addition to training for both in-person and online observations, observers met for up to 30 minutes after each observation to discuss codes and resolve any inconsistencies until reaching 100% agreement.

Results

Instructor O-COPUS Code Descriptions

We adjusted six of the 12 instructor COPUS code descriptions to document the teaching behaviors we observed during online synchronous instruction as illustrated in Table 3. We did not change any of the original COPUS codes, but rather adjusted their code descriptions to fit the new usage of online tools seen in online synchronous instruction such as the messaging function. Most adjustments to the COPUS code descriptions were the result of new online functions, such as breakout rooms, polling, and the messaging function, rather than pedagogical changes. Therefore, we did not add any new codes or remove any of the original COPUS codes from our final codebook. Tables 3 and 4 present O-COPUS codes and code descriptions for instructor and student behaviors, respectively. To make our O-COPUS code descriptions more inclusive for other meeting software programs, we used the terms "messaging function" and "chat function" interchangeably. See Supplementary Materials for full descriptions of the codebook which includes inclusion and exclusion criteria for instructor and student behaviors (File S6, S7).

We adjusted COPUS code descriptions to fit online synchronous instruction, incorporating new online functions such as breakout rooms and polling without adding or removing codes.

Breakout Rooms

Moving and Guiding (MG)

We changed the code description for *moving and guiding* by adding newly observed behaviors as well as excluded behaviors that no longer fit online synchronous instruction. During online synchronous instruction, instructors could no longer physically move around the classroom, so we utilized this code when the instructor moved in and out of breakout rooms and guided students in their active learning activity. We also found instructors engaged in *moving and guiding* behaviors without having to move throughout breakout rooms. For example, we also coded *moving and guiding* when an instructor assigned an active learning activity and provided students hints to answer a problem or showed students how to solve the problem as they were working on it. We agreed that this was also a moving and guiding behavior even though the instructor did not create breakout rooms because they were still guiding students in an active learning activity (Table 3).

One-on-One (1o1)

We changed the code description for *one-on-one* to better describe online synchronous instruction. Specifically, it occurred when the instructor was moving between breakout rooms and staying with one group for an extended period of time. This behavior would be similar to the instructor walking around the classroom and spending extended time with student groups during group work.

Administration (Adm)

We adjusted the description of *administration* to include scenarios that we frequently encountered during online synchronous instruction, like assigning breakout rooms or assigning an *individual thinking* question that was not a *clicker question* (e. g. think-pair-share). While these behaviors could be interpreted in the "etc." of the original description, we included them to ensure consensus of coding these behaviors during observations.

Polling

Clicker Question (CQ)

Next, for the *clicker question* code description, we added online functions that appeared during online synchronous instruction. The most prominent online activity we observed were online polls such as those used on Zoom or third-party sites (e. g. Poll Everywhere, Socrative, or Mentimeter) which we coded as a *clicker question*. While not identical, online polls allowed students to think individually and submit their answer to a multiple-choice question as well as see the distribution of student responses like a *clicker question*.

Table 3
O-COPUS Instructor Coding Scheme

	Individual COPUS Code	In-person COPUS Code Description	Online COPUS Code Description
Instructor is Doing	Moving and guiding (MG)	Moving through class guiding ongoing student work during active learning task	Moving through **breakout rooms** guiding ongoing student work during active learning task **or guiding students while they are working on an active learning task by providing hints or working through a problem using the microphone or messaging**
	One-on-one (1o1)	One on one extended discussion with one or a few individuals, giving undivided attention to one or a group of students	One on one extended discussion with one or a few individuals, giving undivided attention to one or a group of students **in a breakout room**
	Posing a question (PQ)	Posing non-clicker question to students (non-rhetorical) and waiting for students to respond	Posing non-clicker question to students (non-rhetorical) **using the microphone or messaging function and** waiting for students to respond
	Answering questions (AnQ)	Listening to and answering student questions with the entire class listening	Listening to and answering student questions **using the microphone or messaging function** with the entire class listening
	Clicker question (CQ)	Asking a clicker question (mark the entire time the instructor is using a clicker question, not just when first asked)	Asking a clicker question **or online poll** (mark the entire time the instructor is using a clicker question, not just when first asked)
	Administration (Adm)	Administration (assign homework, return tests, etc.)	Assigning homework, returning tests, **class announcements/agenda, assign to breakout rooms, etc.), when the instructor is waiting for students to answer a non-clicker question (i.e., think-pair-share), or administering a test or quiz**

Note: Descriptions of the in-person COPUS code descriptions adapted from Smith et al. (2013a). Modifications to online COPUS code descriptions are noted in bold.

Chat

Posing a Question (PQ) and Instructor Answering Questions (AnQ)

Lastly, for the codes *posing a question* and *answering questions*, we found that the chat function allowed instructors to ask and answer questions in two ways - verbally or written through the chat. Therefore, we slightly changed these code descriptions to include both modalities.

Student O-COPUS Code Descriptions

We adjusted six of the 13 student COPUS code descriptions to document the teaching behaviors we observed during online synchronous instruction as illustrated in Table 4. While most student codes were easily adaptable to online synchronous instruction, some codes required more adjustments. Most of these code descriptions were adapted to include the online functions used during online synchronous instruction, such as the chat, as well as any new behaviors that emerged because of the implementation of these functions.

Breakout Rooms

Group Clicker Question, Group Worksheet, and Other Group Work (CG, WG, and OG)

During online synchronous instruction, we found group work could be seen in two ways: 1) when the instructor assigned students to work on an active learning activity in breakout rooms; or 2) when students engaged in group work by discussing an active learning activity in the chat without instructor facilitation. For example, in one observation, a group of five students used the chat to work on a clicker question together without any instructor intervention. Since this discussion was not facilitated by the instructor, we concluded it was not a *whole class discussion* but a *group clicker question* instead.

Chat

Student Answering Questions (AnQ)

The code description for *answering questions* includes all the ways that students could answer a question during online synchronous instruction. The first and most direct way a student could answer a question posed by the instructor was by responding verbally using the microphone function while the rest of the class was listening. The second way a student could answer a question posed by the instructor was by using the chat function available for everyone in the class to read. We determined this to be interchangeable with "the rest of the class listening" as the original code description for *answering questions* stated. However, in some observations, we noticed that some students' responses in the chat were unnoticed by the instructor. Furthermore, while it was possible for students to answer an instructor's question through private messaging, observers were unable to see these responses during online synchronous observations. Therefore, the description to the code *answering questions* was adjusted to explicitly state "student answering a question posed by the instructor using the microphone or chat function *and* the instructor acknowledges the answer with the rest of the class listening." Additionally, we noticed that throughout the class session some students would ask and respond to each other's questions in the chat, sometimes without the instructor's intervention. This is a behavior that went unseen during live COPUS observations. To acknowledge that these students received an answer to their questions, we deemed it appropriate to code *answering questions* and added "or student answering another students' question using the chat function" to the O-COPUS code description (Table 4).

Online synchronous instruction requires adaptations to traditional COPUS coding, including accounting for chat functions and group work in breakout rooms.

Whole Class Discussion (WC)

Online synchronous instruction allowed students to be involved in a *whole class discussion* utilizing different functions, including the chat, writing, or drawing function. If multiple students answered an instructor's question using the chat, writing, or drawing function, then we coded *whole class discussion*. For example, if the instructor asked the class to use the drawing function to draw a cell structure on a slide and multiple students participated, then observers coded it as *whole class discussion*. Another example of a *whole class discussion* would be if the instructor posed a question to the class and multiple students responded in the chat.

Student Question (SQ)

The description for the code *student question* was slightly altered to account for the modalities that a student could ask a question during online synchronous instruction -

Table 4
COPUS Coding Scheme

	Individual COPUS Code	In-person COPUS Code Description	Online COPUS Code Description
Student is Doing	Answering questions (AnQ)	Student answering a question posed by the instructor with rest of class listening	Student answering a question posed by the instructor **using the microphone function, reaction function, annotating function, or messaging function and the instructor acknowledges the answer with the rest of the class listening or student answering other students' questions using the messaging function**
	Whole class discussion (WC)	Engaged in whole class discussion by offering explanations, opinion, judgment, etc. to whole class, often facilitated by instructor	Instructor poses a question or facilitate a whole class discussion in which 2 or more students answer **verbally, using messaging function, or drawing function while the rest of the class is listening**
	Group Clicker Question (CG)	Discuss clicker question in groups of 2 or more students	Discussing clicker question in groups of 2 or more students **in breakout rooms or messaging function**
	Group Worksheet (WG)	Working in groups on worksheet activity	Working in groups of 2 or more students on worksheet activity **in breakout rooms or messaging function**
	Other Group Work (OG)	Other assigned group activity, such as responding to instructor question	Working in groups of 2 or more students on other assigned group activity, such as responding to instructor question **or collaborating on a shared document/ website, in breakout rooms in breakout rooms or messaging function**
	Student Question (SQ)	Student asks question	Student asks question **using the microphone or messaging function**

Note: Descriptions of the in-person COPUS code descriptions adapted from Smith et al. (2013a). Modifications to online COPUS code descriptions are noted in bold.

verbally or through the chat function. Similar to *answering questions*, students could ask the instructor questions privately through the chat function; however, observers are unable to see these behaviors during online synchronous instruction unless the instructor explicitly acknowledges they received a private message with a question. Additionally, the original code description for *student question* did not specify that the whole class must be listening, unlike

the original code description for *answering questions*. Therefore, regardless of whether the instructor acknowledged a students' question, we used *student question* to code this behavior.

Analyzing Sample O-COPUS Data

To understand how online synchronous instruction impacted instructor and student behaviors, we compared one instructor's in-person and online instructor and student behaviors (Figure 1) using the finalized O-COPUS coding scheme (Tables 2-3). This instructor was observed three times in fall 2019 (in-person) and in three times in spring 2021 (online synchronous instruction during ERT). To compare this instructor's COPUS and O-COPUS codes, we took the number of two-minute time intervals marked for each code and divided it by the total of two-minute time intervals for the class session (Kranzfelder et al., 2020; Lewin et al., 2016; Lund et al., 2015). We visualized the changes between in-person and online synchronous teaching and learning behaviors by using pie charts. This is an example of how instructors who teach in-person, synchronously online, or both can utilize O-COPUS to understand better how their teaching practices may differ between the two learning environments (Figure 1).

O-COPUS can be a valuable tool for instructors teaching in-person, synchronously online, or both, to gain insights into their teaching practices and adapt accordingly.

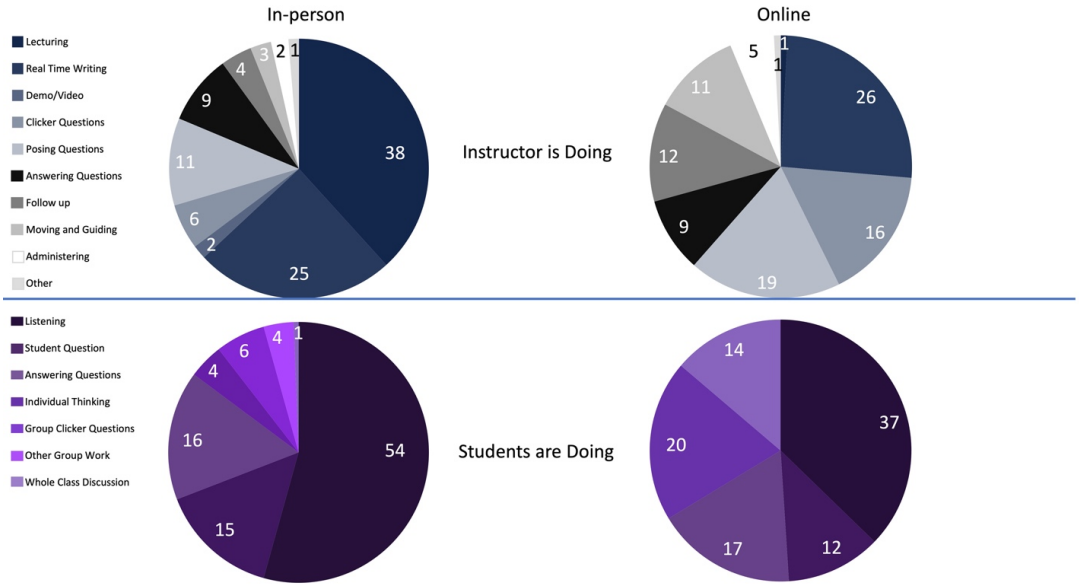


Figure 1
A comparison of the average percentage of two-minute time intervals spent on individual instructor and student COPUS/O-COPUS codes from one STEM instructor averaged across three different class sessions during in-person and online synchronous instruction.

Discussion

We developed and validated O-COPUS to measure teaching practices in the synchronous online format. More specifically, we adapted six instructor COPUS code descriptions to better represent the observed online teaching practices: *moving and guiding*, *one-on-one*, *administering*, *clicker questions*, *posing questions*, and *instructor answering questions*. Moreover, we adapted six student COPUS code descriptions that better represent the observed online synchronous learning behaviors: student answering questions, *whole class discussion*, *group clicker question*, *group worksheet*, *other group work*, and *student question*. The changes in our O-COPUS code descriptions demonstrated that the overall teaching and learning practices in the classrooms we observed did not change; however, the utilization of online functions caused these practices to look different during online synchronous instruction. From the development of O-COPUS, we saw that in-person active learning practices could be translated to online synchronous instruction using online functions, such as breakout rooms, chat, and polling.

Applications of O-COPUS

The innovative teaching practices and online functions instructors adopted to engage their students during online synchronous instruction can be used as instructors teach in different learning environments. O-COPUS results can benefit instructors, institutional assessment programs, and biology education researchers in many ways. Overall, O-COPUS can be applied to: 1) understand the teaching and learning behaviors instructors adapted during online synchronous instruction during ERT; 2) how these behaviors changed between in-person and online synchronous instruction; and 3) make decisions on what teaching practices to implement when instructors return to in-person/hybrid instruction, or if instructors return to online synchronous instruction in the future.

First, O-COPUS allows instructors to continue utilizing formative assessments to explore the teaching practices that could be implemented in online synchronous instruction. O-COPUS data informs instructors if their perceptions of their teaching and learning practices are aligned with observed teaching and learning practices. Additionally, O-COPUS data provide insight to instructors about how they are utilizing their class time to help them realize if their teaching and learning practices might be promoting or inhibiting student cognitive engagement (Chi & Wylie, 2014; Fredricks et al., 2004). For instance, an instructor using a clicker question via a polling function during online synchronous instruction (an example of a constructive mode of engagement) is more likely to promote student cognitive engagement than lecturing and writing notes on an electronic whiteboard (an example of a passive mode of engagement) (Chi & Wylie, 2014). If observers were not capturing the chat function, COPUS data would not reflect the whole class discussions and questions being asked and answered by the instructors and students in the chat. Furthermore, if observers did not capture breakout rooms, online group work would go unnoticed in the COPUS data. Therefore, this adaptation of COPUS for online synchronous instruction was essential to capture the active learning strategies that could not be captured with the original COPUS code descriptions and provide instructors with formative assessment data for their own teaching professional development purposes.

Second, if instructors have previously received in-person COPUS data, then they can compare their two sets of data to see if and how their teaching practices have changed, or not, between in-person and online synchronous instruction. As we examined in our sample data, visualizations such as these can be helpful to show the instructor what teaching and learning practices were used during online synchronous instruction to inform future iterations of courses. More specifically, instructors can use their O-COPUS data to explore what online functions promoted the most student cognitive engagement during in-person and online active learning strategies (Chi & Wylie, 2014). For example, if instructors notice an increase of student cognitive engagement in online synchronous instruction due to the use of the chat function, they may consider how they can continue to use these online functions during in-person, hybrid, or hybrid-flexible instruction (Keiper et al., 2021; Kohnke & Moorhouse, 2021; Miller et al., 2021).

O-COPUS can be used by institutional assessment programs to assess and improve the quality of online synchronous instruction and to ensure that instructors are meeting the learning objectives and goals of their courses.

Third, as some universities begin to return to in-person, hybrid, or hyflex instruction, O-COPUS and COPUS can be used to record both the online and in-person teaching and learning practices. By having a standard protocol for both the in-person and online synchronous instruction, it will allow for consistent classroom observations between the two class formats. If instructors decide to incorporate online functions into the in-person learning environment, COPUS and O-COPUS can be implemented together to document instructor and student behaviors. We hope this tool will support instructors in understanding and improving their own teaching practices as well as provide researchers with a tool that can be used consistently during online synchronous instruction. Furthermore, if universities continue or revert to online instruction in the future, then instructors can refer to their O-COPUS data to determine what practices were effective for them in the past as well as where they can improve.

Limitations and Future Directions

We acknowledge there are several factors that limit our study, providing opportunities for future studies. First, we conducted a convenience sample at one MSI, UC Merced; therefore, our results have limited generalizability. Our selected participants were teaching introductory

chemistry and biology courses based on the focus of the larger grant-funded studies, so we did not employ a systematic approach to ensure even distribution of faculty and students across STEM disciplines at our institution or other institutions. In the future, it would be interesting to collect O-COPUS data across several universities, especially other MSIs, to determine if the teaching and learning practices at our institution are similar to others.

O-COPUS offers a consistent protocol for documenting teaching practices in both online and in-person synchronous instruction, allowing instructors to improve their teaching and researchers to study online instruction.

In addition, we developed O-COPUS while observing instructors using Zoom during synchronous online instruction; therefore, we did not examine if there were differences in teaching and learning practices across different meeting software programs, such as Skype or Google Meet. We recommend future studies utilize O-COPUS to document online behaviors with other software programs to see if new teaching or learning behaviors emerge and if our current code descriptions are applicable outside of the Zoom meeting software program. Furthermore, we hope that future studies will utilize O-COPUS to document how instructors incorporate newfound online functions, such as the chat, during in-person instruction.

As we continue to assess online synchronous instruction, O-COPUS could be complemented by pairing it with other tools to study other variables that influence online instruction. Teacher discourse moves, or the conversational strategies used by instructors to encourage student engagement in science content (Kranzfelder et al., 2020; Warfa et al., 2014), has not yet been studied during online synchronous instruction. Observing instructor discourse alongside instructor behaviors can reveal the quality of active learning strategies used by instructors in online synchronous instruction. For example, instructors may be using student-centered, guiding teaching practices, but taking a teacher-centered, authoritative discourse approach with their students (Kranzfelder et al., 2020). By pairing O-COPUS with discourse protocols, such as the Classroom Discourse Observation Protocol (CDOP) (Kranzfelder, Bankers-Fulbright, et al., 2019), instructors can assess if their teaching practices align with how they are talking to their students.

Finally, we focused on the teaching and learning practices at an MSI, but we did not study how the different student demographics were impacted by changes in the teaching practices because of the transition to online synchronous instruction. In the future, it would be relevant to examine aspects of equity and inclusion as well as power dynamics during online synchronous instruction by taking a closer look at student behaviors. Based on recent studies, Barber et al. (2021) found that first-generation and underrepresented minority students were more likely to have limited access to the internet and computers compared to their white counterparts, suggesting that flexibility on policies and assignments would create more equitable online synchronous instruction. Also, Lee and Mccabe (2021) found that male students dominated in-person discussions in science courses compared to their female counterparts. Furthermore, they found that male students frequently spoke without raising their hands and used assertive language when speaking (Lee & Mccabe, 2021). Online synchronous instruction is unique in that it allows students to participate using both the messaging function and verbally, which may lead to female students participating as frequently as their male counterparts. In the future, we recommend documenting who is talking and students' modes of communication during online, hybrid, and/or in-person instruction.

AUTHOR NOTE

Petra Kranzfelder, ORD ID: https://orcid.org/0000-0003-4146-7929

This study was funded by the Howard Hughes Medical Institute (HHMI) Award #GT11066, National Science Foundation Hispanic-Serving Institution (NSF HSI) Award #1832538, and start-up funding from the Department of Molecular & Cellular Biology at UC Merced for PK.

We have no known conflict of interest to disclose. We would like to thank all the instructors who welcomed us into their classrooms. We would also like to thank all the SATAL interns for their contributions to COPUS data collection and brainstorming O-COPUS code descriptions: Andrea Alvarez Gallardo, Ashley Argueta, Ayooluwa Babalola, Gurpinder Bahia, Avreen Bal, Guadalupe Covarrubias-Oregel, Sandy Dorantes, Dafne Garcia, Jesus Lopez Lopez, Matias Lopez, Stephanie Medina, Andrew Olmeda Juarez, Sara Patino, Monica Ramos, Simrandip Sandhu, Kim Ta, Christian Urbina, Shaira Vargas, Abrian Villalobos, Riley Whitmer, and Isabella Woodruff. Special thanks to members of the Schuchardt and Warfa research teams at the University of Minnesota for their valuable input as our expert panel feedback and manuscript revisions.

References

Akiha, K., Brigham, E., Couch, B. A., Lewin, J., Stains, M., Stetzer, M. R., Vinson, E. L., & Smith, M. K. (2018, January 22). What types of instructional shifts do students experience? Investigating active learning in science, technology, engineering, and math classes across key transition points from middle school to the university level. *Frontiers in Education*, *2*(68). https://doi.org/10.3389/feduc.2017.00068

Alkhouri, J. S., Donham, C., Pusey, T. S., Signorini, A., Stivers, A. H., & Kranzfelder, P. (2021). Look who's talking: Teaching and discourse practices across discipline, position, experience, and class size in STEM college classrooms. *BioScience*, *71*(10), 1063-1078. https://doi.org/https://doi.org/10.1093/biosci/biab077

Ally, M. (2004). Foundations of educational theory for online learning. *Theory and Practice of Online Learning*, *2*, 15-44.

Armbruster, P., Patel, M., Johnson, E., & Weiss, M. (2009). Active learning and student-centered pedagogy improve student attitudes and performance in introductory biology. *CBE-Life Sciences Education*, 8(3), 203-213.

Asai, D. J. (2020). Race matters. *Cell*, *181*(4), 754-757.

Barber, P. H., Shapiro, C., Jacobs, M. S., Avilez, L., Brenner, K. I., Cabral, C., Cebreros, M., Cosentino, E., Cross, C., & Gonzalez, M. L. (2021). Disparities in remote learning faced by first-generation and underrepresented minority students during COVID-19: Insights and opportunities from a remote research experience. *Journal of Microbiology & Biology Education*, *22*(1), ev22i21. 2457.

Barr, R. B., & Tagg, J. (1995). From teaching to learning-A new paradigm for undergraduate education. *Change*: *The Magazine of Higher Learning*, *27*(6), 12-26.

Beatty, B. J. (2019). Hybrid-flexible course design. *Implementing student*.

Bond, M., Bedenlier, S., Marín, V. I., & Händel, M. (2021, August 30). Emergency remote teaching in higher education: Mapping the first global online semester. *International Journal of Educational Technology in Higher Education*, *18*(1), 50. https://doi.org/10.1186/s41239-021-00282-x

Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., & Kerres, M. (2020, January 22). Mapping research in student engagement and educational technology in higher education: A systematic evidence map. *International Journal of Educational Technology in Higher Education*, *17*(1), 2. https://doi.org/10.1186/s41239-019-0176-8

Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom*. 1991 ASHE-ERIC higher education reports. ERIC.

Chi, M. T. H., & Wylie, R. (2014, October 02). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Chickering, A. W., & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *AAHE bulletin*, *3*, 7.

Christianson, A. M. (2020). Using socrative online polls for active learning in the remote classroom. *Journal of Chemical Education*, *97*(9), 2701-2705.

Clark, R. M., Besterfield-Sacre, M., & Dukes, A. (2020). Supportive classroom assessment for remote instruction. *Advances in Engineering Education*, *8*(4), n4.

Commeford, K., Brewe, E., & Traxler, A. (2022). Characterizing active learning environments in physics using latent profile analysis. *Physical Review Physics Education Research*, *18*(1), 010113.

Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory Into Practice*, *39*(3), 124-130.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, *69*(9), 970-977.

Darby, F., & Lang, J. M. (2019). *Small teaching online*: *Applying learning science in online classes*. John Wiley & Sons.

Denaro, K., Sato, B., Harlow, A., Aebersold, A., & Verma, M. (2021). Comparison of cluster analysis methodologies for characterization of classroom observation protocol for undergraduate STEM (COPUS) data. *CBE-Life Sciences Education*, *20*(1), ar3. https://doi.org/10.1187/cbe.20-04-0077

Dingwall, S. (2020). Lessons learned from active engagement in a large-enrollment introductory biochemistry course during a remote quarter. *Journal of Chemical Education*, *97*(9), 2749-2753.

Eddy, S. L., Converse, M., & Wenderoth, M. P. (2015, September 21). PORTAAL: A classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. *CBE-Life Sciences Education*, *14*(2), ar23. https://doi.org/10.1187/cbe.14-06-0095

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, *74*(1), 59-109.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014, June 10). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410-8415. https://doi.org/10.1073/pnas.1319030111

Giesbers, B., Rienties, B., Tempelaar, D., & Gijselaers, W. (2014). A dynamic analysis of the interplay between asynchronous and synchronous communication in online learning: The impact of motivation. *Journal of Computer Assisted Learning*, *30*(1), 30-50. https://doi.org/https://doi.org/10.1111/jcal.12020

Giordano, A. N., & Christopher, C. R. (2020). Repurposing best teaching practices for remote learning environments: Chemistry in the news and oral examinations during COVID-19. *Journal of Chemical Education*, *97*(9), 2815-2818.

Heiss, E. M., & Oxley, S. P. (2021, February 01). Implementing a flipped classroom approach in remote instruction. *Analytical and Bioanalytical Chemistry*, *413*(5), 1245-1250. https://doi.org/10.1007/s00216-020-03147-w

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). *The difference between emergency remote teaching and online learning*. Educause. Retrieved March 27, 2020 from https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning

Keiper, M. C., White, A., Carlson, C. D., & Lupinek, J. M. (2021). Student perceptions on the benefits of Flipgrid in a HyFlex learning environment. *Journal of education for business*, *96*(6), 343-351.

Kohnke, L., & Moorhouse, B. L. (2021). Adopting HyFlex in higher education in response to COVID-19: Students' perspectives. *Open Learning: The Journal of Open, Distance and e-Learning*, *36*(3), 231-244.

Kranzfelder, P., Bankers-Fulbright, J. L., García-Ojeda, M. E., Melloy, M., Mohammed, S., & Warfa, A.-R. M. (2019). The Classroom Discourse Observation Protocol (CDOP): A quantitative method for characterizing teacher discourse moves in undergraduate STEM learning environments. *PLoS One*, *14*(7), e0219019. https://doi.org/https://doi.org/10.1371/journal.pone.0219019

Kranzfelder, P., Bankers-Fulbright, J. L., García-Ojeda, M. E., Melloy, M., Mohammed, S., & Warfa, A.-R. M. (2020). Undergraduate biology instructors still use mostly teacher-centered discourse even when teaching with active learning strategies. *BioScience*, *70*(10), 901-913. https://doi.org/https://doi.org/10.1093/biosci/biaa077

Kranzfelder, P., Lo, A. T., Melloy, M. P., Walker, L. E., & Warfa, A.-R. M. (2019). Instructional practices in reformed undergraduate STEM learning environments: A study of instructor and student behaviors in biology courses. *International Journal of Science Education*, *41*(14), 1944-1961. https://doi.org/https://doi.org/10.1080/09500693.2019.1649503

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Lane, A. K., Meaders, C. L., Shuman, J. K., Stetzer, M. R., Vinson, E. L., Couch, B. A., Smith, M. K., & Stains, M. (2021). Making a first impression: Exploring what instructors do and say on the first day of introductory STEM courses. *CBE-Life Sciences Education*, *20*(1), ar7. https://doi.org/10.1187/cbe.20-05-0098

Lang, J. M. (2020). On now drawing conclusions about online teaching now-or next fall. *The Chronicle of Higher Education*, *18*.

Laursen, S. (2019). Levers for change: An assessment of progress on changing STEM instruction: Executive Summary.

Lee, J. J., & Mccabe, J. M. (2021). Who speaks and who listens: Revisiting the chilly climate in college classrooms. *Gender & Society*, *35*(1), 32-60.

Lemke, T. (2022). How much Zoom is too much? Making asynchronous learning work. In A. A. Szarejko (Ed.), *Pandemic pedagogy: Teaching international relations amid COVID-19* (pp. 73-96). Springer International Publishing. https://doi.org/10.1007/978-3-030-83557-6_5

Lewin, J. D., Vinson, E. L., Stetzer, M. R., & Smith, M. K. (2016, March 20). A campus-wide investigation of clicker implementation: The status of peer discussion in STEM classes. *CBE-Life Sciences Education*, *15*(1). https://doi.org/10.1187/cbe.15-10-0224

Lund, Pilarz, M., Velasco, J. B., Chakraverty, D., Rosploch, K., Undersander, M., & Stains, M. (2015, June 1). The best of both worlds: Building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice. *CBE-Life Sciences Education*, *14*(2), 1-12. https://doi.org/10.1187/cbe.14-10-0168

MacIsaac, D. (2019). US government releases Charting a Course for Success: America's strategy for STEM education, report guiding federal agencies that offer STEM funding opportunities. *The Physics Teacher*, *57*(2), 126-126.

Means, B., Bakia, M., & Murphy, R. (2014). *Learning online: What research tells us about whether, when and how*. Routledge.

Miller, A. N., Sellnow, D. D., & Strawser, M. G. (2021). Pandemic pedagogy challenges and opportunities: Instruction communication in remote, HyFlex, and BlendFlex courses. *Communication Education*, *70*(2), 202-204.

National Research Council. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. The National Academies Press. https://doi.org/doi:10.17226/13362

Nilson, L. B., & Goodson, L. A. (2021). *Online teaching at its best: Merging instructional design with teaching and learning research*. John Wiley & Sons.

Offerdahl, E. G., McConnell, M., & Boyer, J. (2018). Can I have your recipe? Using a fidelity of implementation (FOI) framework to identify the key ingredients of formative assessment for learning. *CBE-Life Sciences Education*, *17*(4), es16.

Olson, S., & Riordan, D. G. (2012). Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. Report to the President. *Executive Office of the President*.

Owens, M. T., Seidel, S. B., Wong, M., Bejines, T. E., Lietz, S., Perez, J. R., Sit, S., Subedar, Z.-S., Acker, G. N., & Akana, S. F. (2017). Classroom sound can be used to classify teaching practices in college science courses. *Proceedings of the National Academy of Sciences*, *114*(12), 3085-3090.

Pérez-Sabater, C., Montero-Fleta, B., Pérez-Sabater, M., Rising, B., & De Valencia, U. (2011). Active learning to improve long-term knowledge retention. Proceedings of the XII Simposio Internacional de Comunicación Social.

Reinholz, D. L., Stone-Johnstone, A., White, I., Sianez, L. M., Jr., & Shah, N. (2020). A pandemic crash course: Learning to teach equitably in synchronous online classes. *CBE-Life Sciences Education*, *19*(4), ar60. https://doi.org/https://doi.org/10.1187/cbe.20-06-0126

Reisner, B. A., Pate, C. L., Kinkaid, M. M., Paunovic, D. M., Pratt, J. M., Stewart, J. L., Raker, J. R., Bentley, A. K., Lin, S., & Smith, S. R. (2020). I've been given COPUS (Classroom Observation Protocol for Undergraduate STEM) data on my chemistry class... Now what? *Journal of Chemical Education*, *97*(4), 1181-1189.

Rosenberg, M., Hilton, M. L., & Dibner, K. A. (2018). *Indicators for monitoring undergraduate STEM education*. National Academies Press Washington, DC.

Saldaña, J. (2015). *The Coding Manual for Qualitative Researchers* (3rd Edition ed.). SAGE Publications. https://books.google.com/books?id=jh1iCgAAQBAJ

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics*, *102*(6), 245-253. https://doi.org/10.1111/j.1949-8594.2002.tb17883.x

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, *103*(4), 759-775. https://doi.org/10.1037/a0025140

Seidel, S. B., Reggi, A. L., Schinske, J. N., Burrus, L. W., Tanner, K. D., & Tomanek, D. (2015). Beyond the biology: A systematic investigation of noncontent instructor talk in an introductory biology course. *CBE-Life Sciences Education*, *14*(4), ar43. https://doi.org/10.1187/cbe.15-03-0049

Signorini, A., & Pohan, C. (2019). Exploring the impact of the students assessing teaching and learning program. *International Journal for Students as Partners*, *3*(2), 139-148. https://doi.org/10.15173/ijsap.v3i2.3683

Singhal, M. K. (2020). Facilitating virtual medicinal chemistry active learning assignments using advanced Zoom features during COVID-19 campus closure. *Journal of Chemical Education*, *97*(9), 2711-2714.

Skylar, A. A. (2009). A comparison of asynchronous online text-based lectures and synchronous interactive web conferencing lectures. *Issues in Teacher Education*, *18*(2), 69-84.

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013a). The Classroom Observation Protocol for Undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE-Life Sciences Education*, *12*(4), 618-627.

Smith, M. K., Jones, F. H. M., Gilbert, S. L., & Wieman, C. E. (2013b, December 21). The Classroom Observation Protocol For Undergraduate Stem (COPUS): A new instrument to characterize university STEM classroom practices. *CBE-Life Sciences Education*, *12*(4), 618-627. https://doi.org/10.1187/cbe.13-08-0154

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan, M. K., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M., McDonnell, L. M., McKay, T. A., Michelotti, N., Musgrove, A., Palmer, M. S., Plank, K. M.,…Young, A. M. (2018). Anatomy of STEM teaching in North American universities [10.1126/science.aap8892]. Science, 359(6383), 1468-1470. http://science.sciencemag.org/content/359/6383/1468.abstract

Tan, H. R., Chng, W. H., Chonardo, C., Ng, M. T. T., & Fung, F. M. (2020). How chemists achieve active learning online during the COVID-19 pandemic: Using the Community of Inquiry (CoI) framework to support remote teaching. *Journal of Chemical Education*, *97*(9), 2512-2518.

Tomkin, J. H., Beilstein, S. O., Morphew, J. W., & Herman, G. L. (2019, January 14). Evidence that communities of practice are associated with active learning in large STEM lectures. *International journal of STEM education*, *6*(1), 1. https://doi.org/10.1186/s40594-018-0154-z

University of California Davis. (2018). *Generalized Observation and Reflection Platform* (GORP). Retrieved July 5, 2021 from https://cee.ucdavis.edu/GORP

Van der Lans, R. M. (2018). On the "association between two things": The case of student surveys and classroom observations of teaching quality. *Educational Assessment, Evaluation and Accountability*, *30*(4), 347-366.

Van Heuvelen, K. M., Daub, G. W., & Ryswyk, H. V. (2020). Emergency remote instruction during the COVID-19 pandemic reshapes collaborative learning in general chemistry. *Journal of Chemical Education*, *97*(9), 2884-2888.

Vanags, T., Pammer, K., & Brinker, J. (2013). Process-oriented guided-inquiry learning improves long-term retention of information. *Advances in Physiology Education*, *37*(3), 233-241.

Warfa, A.-R. M., Roehrig, G. H., Schneider, J. L., & Nyachwaya, J. (2014, June 10). Role of teacher-initiated discourses in students' development of representational fluency in chemistry: A case study. *Journal of Chemical Education*, *91*(6), 784-792. https://doi.org/10.1021/ed4005547

Wenger, E. (1996). How we learn. Communities of practice. The social fabric of a learning organization. The Healthcare Forum Journal.

Youmans, M. K. (2020). Going remote: How teaching during a crisis is unique to other distance learning experiences. *Journal of Chemical Education*, *97*(9), 3374-3380.

Abstract

This article describes a faculty-led project to assess and revise institutional student learning outcomes at a small urban community college. The revision process involved four stages: (1) exploring stakeholders' explicit and implicit understandings through an experimental assessment; (2) using statistical tools to identify redundancies and opportunities for regrouping and revising the learning outcomes; (3) triangulating findings through focus group discussions and test assessments; and (4) drafting and refining the revised learning outcomes. By grounding revisions in stakeholders' explicit and implicit understandings of the existing outcomes, the school was able to streamline and significantly improve institutional student learning outcomes without starting completely from scratch.

AUTHORS

Forest Fisher, Ph.D.
City University
of New York

Tara Bahl, Ph.D.
City University
of New York

Nate Mickleson, Ph.D.
New York University

An Intentional Process for Revising Institutional Learning Outcomes

The AAC&U's (2009) VALUE Rubrics and the Lumina Foundation's (2014) Degree Qualifications Profile offer crucial frameworks for defining learning outcomes and using them at the assignment, course, program, and institutional levels. There is growing literature on the validity, affordances, and limitations of these frameworks (e.g., Colson et al., 2018; Stevenson et al., 2016). Despite increasing adoption of these sophisticated tools, it remains a challenge to ensure that assessment work is meaningful to faculty, staff, and students. Indeed, faculty and other stakeholders often experience learning outcomes assessment as an exercise in institutional box-checking that is irrelevant or even detrimental to their work with students (Stanny, 2018). As Schoepp & Tezcan-Unal (2017) have found, misperceptions about the purposes and uses of outcomes assessment can inhibit participation and limit an institution's ability to use assessment work to improve student outcomes. As Colson et al. (2018) have noted, faculty are also less likely to embrace and use learning outcomes frameworks they perceive to be unnecessarily complex.

This article describes a faculty-led project to assess and revise institutional student learning outcomes (ISLOs) at a small urban community college. We launched the project in spring 2019 and completed it in spring 2020 when the revised ISLOs were approved through the college governance. As we describe below, it proceeded in three general stages: gathering data and perspectives about the existing ISLOs from the college community; analyzing data, triangulating findings, and drafting revised ISLOs; and refining the revised ISLOs through consultation with the college community. The project began as a result of limitations we discovered through several cycles of assessment. These limitations motivated us to undertake the project we describe here to assess and revise our ISLOs. The

CORRESPONDENCE

Email
forest.fisher@guttman.cuny.edu

growing literature on outcomes assessment suggests several principles for ensuring the work is meaningful and effective:

- Develop clear outcomes that faculty, students, and other stakeholders perceive as directly aligned with curricula;
- Design collaborative processes that build from the bottom up, or from the classroom and co-curricular experience to the program and institutional levels;
- and Use evidence gathered through assessment to continually revise and refine outcomes and processes.

Effective assessment is not about checking boxes but about gathering evidence and using it to continually improve student learning outcomes.

The literature shows that distributions of labor and accountability within assessment processes play a key role in their effectiveness (Kinzie & Jankowski, 2015). Matuga & Turos (2018), for example, have found that misalignments can feed faculty disengagement and distrust. Engaging faculty insights and leadership in designing and enacting learning outcomes assessment is a key lever for transforming a "culture of assessment," with its bureaucratic overtones, into a "culture of improvement" that enables faculty, staff, administrators, and students alike to gather, reflect, and act on evidence of learning-in-progress (Stanny, 2018, p. 114; see also Roscoe, 2017).

We add to this literature in three ways. First, we explain how we used evidence from prior assessments as a foundation for improving our learning outcomes rather than starting from scratch. Second, we describe several methods we used to engage faculty and other stakeholders to build from the bottom up rather than from the top down. Third, we highlight the importance of taking time to examine and reflect on explicit and implicit understandings of existing ISLOs in order to ensure they are clear, concise, and aligned with curricula.

Context/Background

Outcomes

Our original ISLOs identified 24 skills in five categories: Broad Integrative Knowledge; Applied Learning; Specialized Knowledge; Intellectual Skills for Lifelong Learning; and Civic Engagement. The ISLOs' initial purpose was to provide a framework for us to follow in developing the college's curriculum. The Broad Integrative Knowledge and Intellectual Skills for Lifelong Learning outcomes informed the first-year experience and general education requirements and the Specialized Knowledge outcomes defined fundamental skills for the college's degree programs. The Applied Learning and Civic Engagement outcomes were intended to infuse the whole curriculum by orienting the work students might do toward engagement with the surrounding communities and preparation for emerging careers. The ISLOs were developed by a team of administrators and faculty through an iterative design process that drew from two national models: AAC&U's Essential Learning Outcomes and associated VALUE rubrics and The Lumina's Foundation's Degree Qualifications Profile (DQP). Each ISLO category has a corresponding rubric. Figure 1 below shows an excerpt from the Broad Integrative Knowledge rubric, an example we will revisit throughout this article.

Our framework shows the contrasting influences of these models. Some of the ISLOs correspond to stages in a student's progress toward their degree, similar to the schema outlined in the DQP. Other ISLOs describe broader areas of learning which mirror the approach embedded in AAC&U's Essential Learning Outcomes and VALUE rubrics. The project we describe in this article helped us disentangle these elements and develop revised ISLOs that are clear, concise, and aligned with our evolving curriculum.

It is quite common for schools to do what we did, adapting language from the DQP and the AAC&U's VALUE rubrics without raising questions about their construct validity. As Knekta et al. (2019) warns, "validity must be considered each time an instrument is used" since it may be valid for one population and purpose but not another (pp. 2). The VALUE rubrics were designed with a general population in mind and could not anticipate the many different, specific student populations with which they would be used. Even though we adapted these

Figure 1
Excerpt from Original Broad Integrative Knowledge ISLO

Broad, Integrative Knowledge: General Education *The outcomes in this category demonstrate that students can integrate learning from broad fields of general study and connect different academic disciplines and multiple perspectives.*				
Criteria or Domain	**Capstone 4**	**Milestones 3**	**Milestones 2**	**Benchmark 1**
a. Engages with issues that have contemporary, historical, scientific, economic, technological, or artistic significance	Applies new knowledge on an issue to academic and/or experiential contexts. Independently evaluates information from multiple sources. Can articulate multiple perspectives on an issue to others.	Situates an issue in a broader context to provide in-depth explanation. Independently gathers information from multiple sources. Can articulate own position on an issue.	Explores an issue with some depth by applying skills or presenting evidence provided in classes. Provides occasional insight and/or connection to self.	Explores issues at surface level, providing little insight and/or information beyond the basic facts. Can state ideas from other sources.
b. Exhibits an understanding of how different disciplines create knowledge and approach questions.	Synthesizes knowledge and approaches from at least two disciplines in planning and conducting research geared toward answering questions.	Considers that different disciplines ask and answer questions in different ways. Presents a rationale for following one disciplinary approach over another in specific cases.	Recognizes knowledge in a specific discipline. Asks and answers questions using general assumptions and approaches of one's own discipline.	Lists academic disciplines and expresses interest in one or more subject areas.
c. Evaluates multiple perspectives on key issues connected to societal concerns.	Synthesizes multiple perspectives through comprehensive evidence-based analysis of positions.	Analyzes multiple perspectives on a key issue connected to societal concerns. Provides some evidence to support an argument.	Acknowledges two sides of a key issue connected to societal concerns. Describes both perspectives by clarifying each position.	States a single perspective on a key issue connected to societal concerns with basic description.

rubrics to our local context in designing our initial ISLOs, our assessments showed that some outcomes were less relevant than others to our students' experiences and the student work we assessed. As a result, some faculty members' "mental models" of assessment defined our institution-level work as being separate and distinct from assessments they use in their classrooms (Heinrich, 2017).

The development of clear and concise ISLOs is crucial for any institution's curriculum development.

Structures

Our institution includes two structures designed to make the assessment process more collaborative: (1) a faculty-led Academic Assessment & Learning Committee charged with assessing student learning and recommending improvements and (2) dedicated Assessment Days at the beginning, middle, and end of the semester when no classes are held. Assessment Days are organized by the Assessment & Learning Committee and funded by the Office of Academic Affairs. These days provide time and space (and lunch!) for faculty, staff, and administrators to collaboratively assess student work, discuss and reflect on emerging evidence, and engage in curricular and professional development activities. Participation rates have been consistently high: a majority of the full-time faculty participate regularly along with a considerable number of staff from the Office of Student Engagement and other units. The Assessment & Learning Committee works in collaboration with two deans: one located in the Office of Academic Affairs and reporting to the Provost and the other located in the Office of Institutional Effectiveness and Strategic Planning and reporting to the President.

Assessment Plan

The college's assessment plan charges working groups co-chaired by two elected members of the Assessment & Learning Committee to assess each of the five ISLOs on staggered two-year cycles. Participants at the Assessment Days rate student work from first-

semester, first-year, and capstone courses for evidence of learning particular to each ISLO. The working groups then analyze, reflect, and report on the evidence and offer recommendations. Even with the working group structure and staggered assessment plan, assessing 24 skills across five ISLOs has proven difficult. When we looked across working group reports, a number of limitations arose repeatedly:

1. **A lack of consistency between levels in the rubrics.** For example, the benchmark (lowest level) for skill B on the Broad Integrative Knowledge rubric (see figure 1 above) states that a student "lists academic disciplines and expresses interest in one more subject area" while milestone 2 requires that a student "ask and answer questions using the general assumptions and approaches of one's own discipline." If a student asks and answers questions in their own discipline but does not list any other disciplines, should they receive a 1 or a 2 for this skill?
2. **Many outcomes appeared to measure similar skills.** For instance, on the Broad Integrative Knowledge rubric, the capstone (highest level) for both skills A and C is about synthesizing or engaging multiple perspectives.
3. **Some skills did not reflect the type of learning that was intended for the classroom.** For example, skill B on the Broad Integrative Knowledge rubric focuses on integrating different academic disciplines. Faculty wondered if this discipline-heavy language reflected the kind of integrated learning we envisioned for first-semester college students and instead suggested that we focus on integrating "perspectives" or "methods of inquiry."
4. **Difficulty finding examples of student work that were appropriate to assess with these rubrics.** One of the skills addressed by the ISLOs was "collaboration", and faculty found it difficult to assess collaboration through the end product of group work. "No Evidence" ratings ranged from 10% to 80% in our working groups' assessments.

As a result of these challenges, the working groups consistently observed **low inter-rater reliability** in their assessment. For example, when evaluating skill B on the Broad Integrative Knowledge rubric, only 43% of participants agreed in their ratings. We measure agreement when raters' scores are within 1 or when they agree that the work offers no assessable evidence of a given skill.

The Assessment & Learning Committee considered rewriting the outcomes from scratch to resolve these limitations. However, since our assessment work had already taught us quite a bit about the ISLOs, we determined that starting from scratch would mean introducing a whole new set of unknown issues. Our goal was to fix known issues, not introduce new unknown issues. Therefore, the committee decided to take the approach of identifying which skills from the original ISLOs might be eliminated, which might be maintained, and which might be combined. We designed an assessment project with four stages: (1) exploring stakeholders' explicit/implicit understandings through an experimental assessment; (2) using statistical tools to identify redundancies and opportunities for regrouping and revising the ISLOs; (3) triangulating findings through focus group discussions and test assessments; and (4) drafting and refining the revised ISLOs.

Even with the working group structure and staggered assessment plan, assessing 24 skills across five ISLOs has proven difficult.

Explicit/Implicit Understandings of ISLOs (Spring 2019 - Summer 2019)

In this section, we describe the process we used to revise the ISLOs and explain how different steps in that process explored stakeholder's implicit and explicit understandings of the ISLOs. As a first step, we collected two different types of data about stakeholder understandings of the ISLOs. The first type of data measured faculty and staff's *explicit understandings* of the learning outcomes in relation to one another. We revised the language of the learning outcomes and their rubrics following the recommendations of previous working groups. We then printed each outcome on a separate piece of paper and gave copies of these outcomes to groups of faculty and staff at one of the Assessment Days. We asked each group to discuss the outcomes and reorganize them in the way that they felt made the most sense.

This process produced nine different potential ways of reorganizing the learning outcomes. For instance, group 1 organized the outcomes into four categories: Communication, Cultural Background and Identity, Problem Posing, and Knowledge of the Field or Program of Study while group 5 used five: Critical Thinking & Practice or Applied Learning, Research Process, Disciplinary Fluency, Self-Aware Learning, Civic and Community Engagement. We looked for commonalities across groupings that would allow us to better understand how faculty and staff explicitly envisioned the learning outcomes in relation to one another. For instance, both groups 1 and 5 put skills 3B ("Connections to Experience") and 4E ("Cultural Background and Identity") together, group 1 in the category Cultural Background and Identity, and group 5 in the category Critical Thinking & Practice or Applied Learning. This suggested an overlap in what our assessment of these skills might measure.

We also surveyed a small group of faculty to understand better their experiences using the rubrics. The survey asked participants three questions: "What was clear, effective, or useful about using the rubric?", "What was confusing, ineffective, or difficult about using the rubric?", and "In what ways would you recommend revising this rubric?" Some survey results corroborated findings from other types of data. For instance, one survey respondent observed that skill "1C [on collaboration] was difficult [to assess] because the assignments that I read did not specify group or individual work." Likewise, another respondent observed that skill 4B ("Synthesize Multiple Perspectives") was "confusing because of the term 'discipline-specific issues', which only shows up in Level 2." Overall, respondents noted "the progression [of skills] didn't seem logical."

We found a conflict between faculty and staff's explicit and implicit understandings of ISLOs, highlighting potential inaccuracies in assessment practices.

The results also suggested that faculty members' *explicit* descriptions of their understandings of ISLOs may differ from how they use them in practice. For example, a faculty member may say that two outcomes are related, but in practice, they may actually score these two outcomes very differently. This suggests a conflict between explicit and implicit understandings of the outcomes. To identify these potential inaccuracies, we decided to collect a second type of data that would help us assess *implicit understandings* of the learning outcomes. We asked faculty and staff to assess student work using the rubrics and then used a statistical technique called Exploratory Factor Analysis to identify groups of outcomes that tended to receive similar ratings.

To do this, we selected 80 samples of student work from recent course and program-level assessments: 40 from courses in the first-year core curriculum and 40 from courses in the programs of study. Half of the samples in each selection pool had received lower ratings in previous assessments (an average rating of 2 or below on a scale of 1-4) and half had received higher ratings (an average above 2 on a scale of 1-4). Figure 2 provides more detailed information about the sample.

Figure 2
Sample for Re-Assessment

Course	Total sample	Student work with average ratings 1-2	Student work with average ratings above 2
First-year Social Science Course	20	10	10
Interdisciplinary Freshman Seminar	20	10	10
Business Administration Capstone	10	5	5
Human Services Capstone	10	5	5
Liberal Arts & Sciences Capstone	10	5	5
Urban Studies Capstone	10	5	5
Total	**80**	**40**	**40**

To re-assess these pieces of student work, we designed test rubrics that divided the 24 skills listed in our ISLOs across four rubrics containing six skills each. The groupings did not align with our existing categories. In fact, we made an explicit effort to place similar or related

skills on different test rubrics because we worried that faculty might give side-by-side skills similar ratings just because they appeared superficially similar. For example, we placed the skills "Quantitative Data Analysis" and "Quantitative Problem Solving" on different test rubrics because we worried faculty might rate them similarly just because they both have the word "quantitative" in their titles.

Eight faculty were paid a small stipend through an internal grant to use the test rubrics to assess student work over two days during the summer. On the first day, each participant assessed roughly 20 pieces of student work using one of the four test rubrics. On the second day, they used a different test rubric to assess roughly 20 more pieces of student work. In this way, each individual participant was responsible for assessing student work for 12 of the 24 ISLO skills. In total, we assessed 72 pieces of student work for all 24 skills. While we could have generated more robust data by asking each participant to assess each piece of student work for all 24 skills, we determined this would have been too cognitively taxing.

[The results suggest] a conflict between explicit and implicit understandings of the outcomes.

Exploratory Factor Analysis (Fall 2019)

Next, we used Exploratory Factor Analysis (EFA) to analyze this data. EFA is a statistical technique used to extract underlying latent variables that might characterize a data set (Tabachnick et al., 2007). Social scientists frequently use it to measure the construct validity of different statistical instruments like a survey (Knetka et al., 2019). The core idea is to look for correlations in the responses to different survey questions. A researcher might find that respondents tend to answer three questions similarly suggesting that there is one underlying factor explaining the answers to all three of these questions. In this way, EFA allows researchers to identify a smaller number of factors that might explain their survey responses and then determine whether or not these factors align with constructs they are studying. As explained below, we decided that descriptive statistics were more appropriate for our small data set, but we have included a short discussion of EFA here because it offers an innovative approach that other institutions with larger, more robust data sets might want to consider.

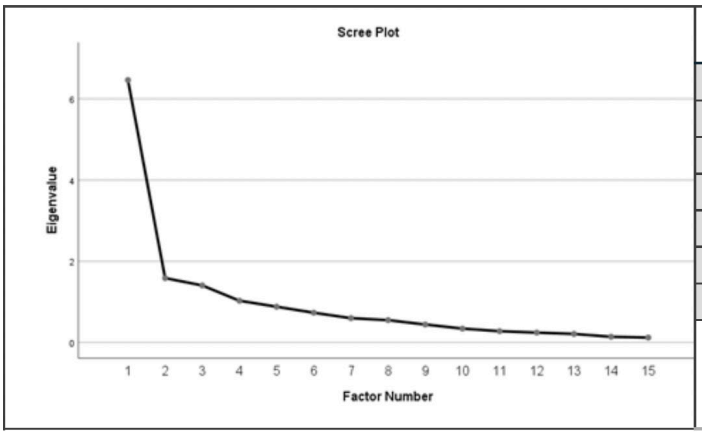
The EFA results are shown in figures 3 and 4 below. Figure 3 shows the Scree Plot and eigenvalues of the factor analysis. These measures are used to determine the appropriate number of factors to extract. A best practice is to consider only factors with an eigenvalue of 1 or higher (4 factors in our case); but since our data are messy and our sample size small, we opted to consider six factors. The Scree Plot provides a way of visualizing this information by plotting the eigenvalues on the vertical axis and the factors on the horizontal axis. It is also common to include only factors that appear before the "knee" of the Scree Plot where the graph levels out.

Figure 4 shows the "rotated factor matrix" for our EFA. The columns along the top list the factors indicated by the Scree Plot and the rows list the skills that these factors "load onto." The values inside the matrix range from 0 to 1, indicating the extent to which each factor loads onto the corresponding skill. The closer this number is to 1, the more that this skill is a determining component of the factor. For example, the first cell in the matrix has a value of 0.888, indicating that skill 3D ("Analysis of Ideas") is a deciding piece of the first factor. As Knetka et al. (2019) observe, there is "no clear rule for when an item has a factor loading that is too low to be included" (pp. 11). However, it is common to omit values less than 0.3 and we follow this convention in figure 4.

The rotated factor matrix in figure 4 was created using an orthogonal rotation algorithm (Varimax). Orthogonal rotations produce uncorrelated factors whereas oblique rotations (like Promax or Direct Oblimin) allow the factors to correlate. We found similar results using Oblimin. We chose to include the Varimax rotation here because orthogonal rotations tend to produce fewer factors and uncorrelated factors tend to be easier to interpret. This six factor Varimax model explained 70.5% of the variance across the 15 variables.

EFA is used to *explore* a data set and look for patterns. There is a related technique called Confirmatory Factor Analysis (CFA) which is used to confirm that data support a previously hypothesized model. Best practice is to conduct an EFA with one sample and then confirm these results with a different sample using CFA. As Knekta et al. (2019) explain, "This confirmation should never be conducted on the same sample as the initial EFA. Doing so does not provide generalizable information, as the CFA will be (essentially) repeating many of the relationships

Figure 3
Scree Plot from Exploratory Factor Analysis



Factor	Eigenvalue
1	6.464
2	1.585
3	1.404
4	1.026
5	.878
6	.731
7	.598

Figure 4
Rotated Factor Matrix

	Factor 1	2	3	4	5	6
skill_3D	.888					
skill_3E	.817					
skill_3C	.811					
skill_3F	.623					
skill_4F		.787				
skill_4A		.646				
skill_4D	.317	.644		.373		
skill_1F			.805			
skill_1A			.661		.353	
skill_1B			.637	.378		
skill_2B				.575		
skill_2E				.538		
skill_4B		.443			.730	
skill_1E				.314	.449	
skill_2F				.357		.853

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 7 iterations.

that were established through the EFA. Additionally, there could be something nuanced about the way the particular sample responds to items that might not be found in a second sample" (pp. 8). We did not confirm our results with CFA because we did not have resources to conduct a second test assessment. Without CFA confirmation, we worried about the reliability of our EFA results.

…we made an explicit effort to place similar or related skills on different test rubrics because we worried that faculty might give side-by-side skills similar ratings just because they appeared superficially similar.

Our small sample presented several additional challenges. Wolf et al. (2013) observe that in some situations a sample as small as 30 observations will suffice for EFA, but in general, researchers recommend 150-300 observations and a minimum of 5-10 observations per variable (MacCallum et al., 1999). With limited resources, we assessed only 72 pieces of student work for 24 skills (or variables) which amounts to only three observations per variable. However, nine of the 24 skills were not included in the EFA because of missing values. (i.e., a very small portion of the student work in our sample proved appropriate for assessing these skills). Fifteen variables put our data right on the margins of an appropriate sample size. However, researchers (Tabachnick et al., 2007) also typically recommend no more than one factor for every three variables, meaning our data should only produce about five factors. It was not our intention to collapse the 24 skills to only five skills. With these considerations in mind, we ultimately favored descriptive statistical results (the correlation matrix in Appendix 1) over the results of the EFA. Nonetheless, the EFA still proved instructive.

For example, enterprising readers will notice that the first four factors align closely with the skills grouped together in the test rubrics. Participants tended to rate items on the

same rubric similarly even though the test rubrics were designed to combine skills that appeared to have nothing in common. This pattern also appears in the correlation matrix described below. Additionally, the values reported for the fifth factor suggest there is a latent variable that partially explains both skill 4B ("Synthesize Multiple Perspectives") and skill 1E ("Interdisciplinary Knowledge"). In surveys and focus groups, faculty repeatedly told us that they felt these two skills were redundant; factor five presents one place where our implicit and explicit understandings of the learning outcomes seem to align.

To make better use of our data, we decided to focus our analysis on the correlation matrix shown in Appendix 1. The values in this matrix are the Pearson correlation coefficients which range from -1 to 1. Values closer to 0 indicate that the two skills are mostly unrelated. Values closer to 1 indicate that these skills tend to be rated in a similar manner, and values closer to -1 indicate that these skills tend to be rated in an opposite manner. We have omitted values less than 0.4 for ease of readability. Since the correlation matrix is a descriptive tool rather than an inferential tool like the EFA, we were able to include data for 22 of the 24 skills (instead of 15). Two of the skills, 1C ("Collaboration") and 4C, ("Connections to the First Year Experience") were still omitted because fewer than 10 of the 72 examples of student work provided evidence of these skills. We decided that if these skills produced so little evidence from student work, they could not be assessed using the test rubrics and should be eliminated or addressed elsewhere in the college's assessment plan.

Triangulating Findings - A first draft of new ISLOs (Fall 2019)

At this point, we had accumulated multiple types of data about our ISLOs. Some of the data (the sorting activity and the survey results) gave us insights into how faculty and staff explicitly understood our learning outcomes while other data (the correlation matrix and EFA results) examined implicit understandings of these outcomes. To make sense of all this data, the Assessment and Learning Committee formed subgroups to review each piece of data in turn. The subgroups identified learning outcomes that should be eliminated or combined based on the data they were reviewing. Recommendations were collected on a white board and discussed with the larger group.

We looked for similar findings across the different types of data and revised the learning outcomes accordingly. For example, we noted skills 3A ("Reflection on Learning") and 3B ("Connections to Experience") had a high correlation (0.79) in the correlation matrix. Likewise, in the sorting activity, many groups combined skill 3B ("Connection to Experience") with skill 4E ("Cultural Background and Identity"). We thus decided to collapse all three of these skills into one learning outcome with a rubric based on skill 3B since participants felt this rubric was the clearest. Similarly, we decided to eliminate skill 1C ("Collaboration"). We found very few items in the student work that could be used to assess this learning outcome; and during the sorting activity, many groups suggested eliminating it. By cross-referencing different types of data, we were able to make informed decisions about how to reduce the number of learning outcomes.

Finalizing the revised ISLOs (Fall 2019 - Spring 2020)

We wanted to maintain the collaborative, participatory process that guides our assessment work . . .

After analyzing explicit and implicit understandings of the ISLOs and using statistical tools to analyze redundancies and opportunities for revision, we presented a draft of the revised ISLOs at an Assessment Day in December 2019. We wanted to maintain the collaborative, participatory process that guides our assessment work, so we presented data from the EFA and correlation matrix alongside the draft revised ISLOs and asked faculty and staff to use the revised rubrics to assess examples of student work. Our goal was not to collect assessment data but rather to gain a better understanding of the process and experience using the new draft ISLOs. Roughly 35 faculty and staff participated. After engaging with the draft rubrics, participants completed an interactive survey that asked them to reflect on their experience using the rubric to evaluate a piece of student work.

This activity helped to corroborate evidence from prior assessments, the results of our statistical analysis, and insights from a broad range of stakeholders about the draft revised ISLOs. In addition, it helped us determine that some of the revised ISLOs required additional attention and a more serious overhaul. For example, we determined that the outcomes related

to technology needed further development. The framing of our original ISLOs in this area no longer reflected the kinds of work the college was asking students to do. Similarly, the outcomes related to civic engagement and global learning required a deeper dive. To revise these ISLOs, we met with faculty and staff with expertise and leadership roles in related areas - technology, science, and civic engagement/global learning. During these meetings, we reviewed the second draft of the revised ISLOs and discussed how they might better reflect student work in these areas.

We developed a second, "semi-final" draft of the revised ISLOs by meticulously incorporating data and insights collected from the activities and conversations described above. We presented this draft during a college Assessment Day in June 2020, six months after we presented the first revised draft. As part of the presentation, we again asked participants to assess student work using the draft rubrics and collected survey data about participants' experiences using the rubrics. This semi-final draft did not provide names for any of the ISLO skills. We additionally asked survey participants to suggest names for these skills to get a sense of whether the correct ideas were coming across in the rubrics. Thirty-two faculty and staff participated in this activity. We collected assessment data during these activities in addition to survey feedback. The data and survey feedback confirmed that our semi-final draft ISLOs were aligned with the explicit and implicit understandings of faculty and other stakeholders and with the work students were currently doing in our classes. We presented the revised ISLOs for ratification through the college governance process in fall 2020.

The subgroups identified learning outcomes that should be eliminated or combined based on the data they were reviewing.

The New ISLO Framework

Figure 5 shows an example rubric from the revised ISLOs we developed. The new framework addresses several of the challenges described above:

1. **Consistency between levels in the rubrics.** Since the new ISLOs were adapted from the original set (as opposed to starting from scratch), we were able to respond to stakeholder feedback about inconsistencies in the rubrics. The new ISLOs use more consistent language. For example, the lowest level of skill B on the original Broad Integrative Knowledge rubric (see figure 1) asked students to "list academic disciplines" in "one or more subject areas" while subsequent levels only required them to ask/answer questions based on a single discipline. Stakeholders told us that compiling a list of disciplines was not a compelling way for students to demonstrate integrative learning, and it seemed misguided to give students credit for integrative learning if they only asked or answered questions using a single discipline. We eliminated this language. The lowest level on the new rubric (see "Synthesizing Methodologies" in figure 5) requires students to attempt to "ask and answer questions using the general assumptions and approaches of two or more disciplines / methodologies."
2. **Redundant outcomes.** The new framework reduces the number of skills from 24 to 15 by consolidating redundancies identified in the process described above. For instance, skills B and C from the original Broad Integrative Knowledge rubric were both consolidated into the "Synthesizing Methodologies" skill on the new rubric since multiple forms of data suggested these two skills were measuring similar things.
3. **Some skills did not reflect the type of learning that was happening in the classroom.** For example, skill B from the original Broad Integrative Knowledge rubric focused on integrating academic "disciplines." One of our first-year courses asked students to conduct research using both qualitative and quantitative data. Faculty felt that this class required students to synthesize different approaches, but those different techniques were not necessarily representative of specific academic disciplines. We thus refined the rubric to refer to "methodologies" rather than disciplines.

1. **Difficulty finding examples of student work that were appropriate to assess with these rubrics.** The "Collaboration" skill was removed from our ISLOs. Multiple pieces of data suggested it was difficult to assess collaboration using the type of student work our assessment system provided. We still think collaboration is an important skill for students to learn, but we believe its assessment belongs elsewhere.

Figure 5
New Integrative Knowledge Rubric

Integrative Knowledge

*Integrative learning is the process of making connections between ideas and experiences from different contexts in order to leverage knowledge in new and more meaningful ways. This rubric, especially skill D1, is informed by Veronica Boix-Mansilla's notion of "integrative leverage" which suggests that quality work integrates different disciplines/methodologies "to generate a new and preferred understanding." Expert practitioners of these skills will integrate knowledge and modes of thinking from multiple disciplines or perspectives. They will situate issues in broader contexts and relate them to their own lived experiences. In particular, integrative knowledge is not exclusive to curricular experiences; it also applies to co-curricular experiences like student leadership, peer mentoring, tutoring, etc. In this rubric, we use the word **perspectives** to refer to perspectives of specific cultures or stakeholders as opposed to disciplinary perspectives. We use the word **methodologies** to refer to the approaches that different fields use to ask or answer questions.*

Skill	Level 1	Level 2	Level 3	Level 4
Synthesize Methodologies	Attempts to ask and answer questions using the general assumptions and approaches of two or more disciplines / methodologies, but does so ineffectively.	Effectively asks and answers questions using the general assumptions and approaches of two or more disciplines / methodologies, but does not integrate these approaches.	Integrates knowledge and approaches from at least two different disciplines / methodologies in planning and conducting research.	Integrates knowledge and approaches from at least two different disciplines / methodologies in planning and conducting research, and critically compares these different approaches.
Connections to Personal Experience	Identifies connections between one's own life experiences and/or prior knowledge to academic texts/ ideas.	Explains connections between one's own life experiences and/or prior knowledge to academic texts/ideas using basic examples, facts, or theories.	Explains connections between one's own life experiences and/or prior knowledge to academic texts/ideas using multiple, rich examples, facts, or theories.	Connects examples of one's own life experiences and/or prior knowledge to academic texts/ideas to illustrate concepts from multiple perspectives.
Contextualize an Issue	Explores an issue at the surface level, providing little insight and/or information beyond the basic facts.	Moves beyond basic facts to demonstrate an awareness of multiple perspectives on an issue.	Provides some historical/social context around an issue to explain how different perspectives relate to one another.	Situates an issue in a broader historical/ social context to demonstrate an understanding of the issue from multiple perspectives.

The new framework reduces the number of skills from 24 to 15 by consolidating redundancies identified in the process described above.

The new framework organizes the skills in ISLOs categories that allow students (and faculty and staff) to see at a glance the range of work they will be asked to do across their degrees. The skill "Synthesize Methodologies," for example, is listed under the Integrative Knowledge ISLO alongside two others: "Connections to Personal Experience" and "Contextualize an Issue." These three skills describe three distinct ways we expect students to integrate knowledge and methodologies they are studying in their classes. Taken together, the three skills in the Integrative Knowledge ISLO signal to students that we conceive of the "knowledge" we expect them to construct across their careers holistically, as combining lived experience, academic disciplines, and social contexts. The predecessor skills were listed in a more amorphous and curriculum-focused category, "Broad, Integrative Knowledge: General Education," as if they were relevant primarily to the student's general education classes rather than to both those classes and their program of study. This framing was helpful to us as we designed the curriculum in the college's early years; but, as our assessment results and this research project shows, it has hindered our ability to define and communicate our overall expectations for student learning.

Discussion

Looking back on the years-long project we undertook to assess and revise our ISLOs, three critical features emerged that might transfer to other institutional settings. These features were central in making the process unfold in an effective and inclusive manner and ensuring there were consistent spaces for reflection. We offer them up as "best practices."

Start from where you are, rather than starting over

When the revision process began, there was a strong push from many faculty, staff, and administration to start from scratch. The limitations associated with our original ISLOs had caused significant frustration; as a result, many stakeholders wanted to develop completely new ISLOs. However, the college had collected several cycles of assessment data and the working groups had issued detailed reports about how individual outcomes could be revised to reflect the kinds of teaching, learning, and values that underpin the college. These reports provided a trove of information and insights about our current ISLOs and allowed us to use what we already knew to begin the process of revising them.

The new framework organizes the skills in ISLOs categories that allow students (and faculty and staff) to see at a glance the range of work they will be asked to do across their degrees.

In other words, we approached the revision process with the goal of addressing challenges and issues we already knew existed. Rather than introduce new, unknown challenges by completely rewriting the ISLOs, we used a large body of existing knowledge to rebuild them. This approach also helped to continually propel the work forward, rather than getting caught up in new, unfamiliar challenges and issues.

Ensure faculty and staff remain at the center throughout the process

The process outlined above prioritized faculty and staff agency. Faculty and staff lead the revision process and constantly came back to the larger college community not only to reflect on the data but also to help generate new data that guided the process. This iterative, inclusive process maximized participation and increased buy-in from faculty, staff, and administration as we shared the draft revised ISLOs. The bottom-up approach we used is markedly different from an approach spearheaded by administrators or a small group of faculty/staff.

One example of how this approach worked in practice is the way we used Assessment Days throughout the year to keep stakeholders engaged. We used activities during these days to do much of the revising work. For instance, we invited participants to regroup the ISLOs and evaluate our test rubrics rather than simply presenting our findings. During several Assessment Days, we asked faculty and staff to use draft ISLO rubrics to assess student work and then provide feedback on their experiences. This iterative and inclusive process helped capitalize on the expertise of the practitioners who will use these rubrics.

A holistic approach to gain insight from stakeholders

Earlier, we outlined the process of accessing faculty and staff members' *explicit*, as well as *implicit*, understandings of our learning outcomes. For instance, we facilitated an activity during an Assessment Day that invited members of the college community to regroup existing ISLOs in order to understand how they think about them. Additionally, faculty, staff, and administrators engaged in a structured discussion about larger institutional values that we identify as critical to our college. Activities along this vein provided us with data and insights about how community members explicitly think and feel about ISLOs. We also accessed their implicit understandings by way of analyzing assessment data of student work with Exploratory Factor Analysis. These activities provided us with data and insight about how community members use the ISLOs and rubrics in practice. We were then able to compare these different types of understandings to identify redundancies in our ISLOs.

Rather than pursuing one over the other, bringing together these different strands of assessment data and analysis provided us with a more comprehensive snapshot of our ISLOs. Comparing these two data also produced evidence of potential inconsistencies between how community members explicitly think and talk about ISLOs and how they make use of them to assess student work. For example, faculty and staff explicitly identified "Quantitative Data Analysis" and "Quantitative Problem Solving" as referring to the same skill. However, our

EFA indicated that these skills were not assessed similarly and did not overlap. This holistic approach provided us with more nuanced, inclusive perspectives on current ISLOs and enabled us to make targeted revisions.

Conclusion

Our college started using the revised ISLOs in the 2020-2021 academic year. Preliminary feedback has generally been positive. This article provides a deep reflection on our revision process, lifting up key themes and considerations we identify as recommendations to any institution grappling with developing or revising ISLOs. As we note above, three principles emerged through our work that seem particularly salient for colleges embarking on similar outcomes revision projects: starting our revision process from where we were rather than replacing our existing ISLOs wholesale; striving to ensure faculty and staff stakeholders played leading roles throughout the process; and using multiple approaches, including inferential and descriptive statistics, pilot assessments, surveys, and small and large group discussions to develop a holistic understanding of our existing ISLOs and possible revisions.

We… found that these consultations increased buy-in and ensured that each subsequent draft of the rubrics better reflected our values as a community.

These principles helped us mitigate challenges identified by other practitioners related to stakeholder misperceptions of purposes and uses of ISLOs (see Colson et al., 2018; Schoepp & Tezcan-Unal, 2017). They also helped us negotiate the limitations of our own work, including the restricted sample we used for our test assessment and the well-intentioned complexities of our existing ISLOs while we worked with our colleagues to construct what Stanny (2018) has described as a "culture of improvement" (p. 114). As a concluding point, we note that we plan to continue consulting with the college community as we roll out the revised ISLOs. We did this during each step in the revision process and found that these consultations increased buy-in and ensured that each subsequent draft of the rubrics better reflected our values as a community.

References

Association of American Colleges and Universities. (2009). *Valid assessment of learning in undergraduate education (VALUE)*. https://www.aacu.org/initiatives/value

Colson, T., Berg, B., Hunt, T., & Mitchell, Z. (2018). Simple, transparent, and less burdensome: Re-envisioning core assessment at a regional public university. *Journal of Assessment and Institutional Effectiveness*, *7*(1-2), 92-114. https://doi.org/10.5325/jasseinsteffe.7.1-2.0092

Heinrich, W.F. (2017). Toward ideal enacted mental models of learning outcomes assessment in higher education. *Journal of Applied Research in Higher Education*, *9*(4), 490-508. https://doi.org/10.1108/JARHE-10-2016-0064

Kinzie, J., and Jankowski, N. A. (2015). Making assessment consequential: Organizing to yield results. In G. D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. Hutchings & J. Kinzie (Eds.), *Using Evidence of Student Learning to Improve Higher Education* (pp. 74-93). Jossey-Bass.

Knekta, E., Runyon, C., & Eddy, S. (2019). One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE—Life Sciences Education*, *18*(1), 1-17. https://doi.org/10.1187/cbe.18-04-0064

Lumina Foundation. (2014). *The degree qualifications profile*. https://www.luminafoundation.org/resource/dqp/

Matuga, J. M., & Turos, J. M. (2018). Infrastructure support for using assessment data for continuous improvement. *New Directions for Teaching & Learning*, *2018*(155), 81-88. https://doi.org/10.1002/tl.20306

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological methods*, *4*(1), 84. https://doi.org/10.1037/1082-989X.4.1.84

Roscoe, D. D. (2017). Toward an improvement paradigm for academic quality. *Liberal Education*, *103*(1), 14-21.

Schoepp, K., & Tezcan-Unal, B. (2017). Examining the effectiveness of a learning outcomes assessment program: A four frames perspective. *Innovative Higher Education*, *42*(4), 305-319. https://doi.org/10.1007/s10755-016-9384-5

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

Stevenson, J. F., Hicks, S. J., & Hubbard, A. (2016). Evaluating a general education program in transition. *New Directions for Evaluation*, *2016*(151), 37. https://doi.org/10.1002/ev.20197

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5, pp. 481-498). Pearson.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and psychological measurement*, *73*(6), 913-934. https://doi.org/10.1177/0013164413495237

Appendix 1 – Correlation Matrix

Note: Positive correlation coefficients with values less than 0.4 have been omitted for ease of reading. Higher values have been highlighted in progressively darker shades of gray to emphasize where the strongest correlations exist. There were very few negative correlations, but these have been italicized to help distinguish them from the positive values.

	1A	1B	1D	1E	1F	2A	2B	2D	2E	2F	3A	3B	3C	3D	3E	3F	4A	4B	4D	4E	4F
1A	1																				
1B	0.54	1																			
1D	0.44		1																		
1E			0.47	1																	
1F	0.68	0.63			1																
2A	0.40				0.44	1															
2B			0.40		0.40	0.49	1														
2D								1													
2E						0.77	0.44		1												
2F					0.47	0.54	0.48	*-0.22*	0.41	1											
3A				*-0.15*						0.44	1										
3B										0.50	0.79	1									
3C		0.42			0.52						0.77	0.79	1								
3D					0.45						0.73	0.72	0.83	1							
3E					0.43						0.77	0.69	0.73	0.79	1						
3F		0.43				0.41					0.52	0.66	0.61	0.62	0.61	1					
4A		0.44											0.50	0.47			1				
4B	0.50		0.51	0.46										0.41			0.56	1			
4D	0.45	0.49					0.46			0.49			0.51	0.50	0.40	0.45	0.69	0.57	1		
4E			*-0.20*														0.40			1	
4F																	0.55	0.46	0.54		1