RESEARCH & PRACTICE IN ASSESSMENT

VOLUME SEVENTEEN | 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 | Copyright © 2022

TABLE OF CONTENTS

FROM THE EDITOR

ARTICLES

Planting the Seeds of Effective Assessment

Don't judge each day by the harvest you reap but by the seeds that you plant.
-Robert Louis Stevenson

"In this issue of Research & Practice in Assessment, we share six 'seeds' of assessment research that we hope take root and grow within your own assessment practices.

Volume 17, Issue 1 of RPA includes six articles with a wide range of foci. First, Alahmadi and DeMars share their findings from analyzing a shift in assessment practices during the COVID-19 pandemic. Merrell, et.al., then share their work developing a course-based undergraduate research experience. Drewes, Scantlebury, and Soslau evaluated coteaching using a mixed methods modality to develop the evaluation. Dorimé-Williams, Cogswell, and Baker share their work to connect assessment research and student learning. Chang and Menzies share lessons learned from their experiences with data-informed decision making.

First, Cook-Sather and Woodworth provide a compelling piece exploring the intersection of the impacts of COVID and on-going inequities in US higher education. Horst, et.al., then discuss the varying credibility of program effectiveness studies focusing specifically on student affairs journals. Stanny and Bryan provide another excellent example of the effectiveness of meta-assessment Finally, Finney, Gilmore, and Alahmadi provide a guide to finding existing measures to assist in the outcomes assessment process. Schafer, Pastor, and Harmon share lessons from a content analysis following a shift to unproctored internet testing.

We hope that these articles provide fertile ground for your own growth in your assessment work!

Regards,

Nicholas Curtis

Editor-in-Chief,
Research & Practice in Assessment

Abstract

The COVID-19 pandemic posed many disruptions to higher education assessment in 2020. At James Madison University (JMU), ensuing modifications to long-standing, university-wide assessment necessitated unproctored remote testing instead of the typically proctored, onsite assessment. Applying such modifications to low-stakes educational assessment raises validity concerns. JMU's assessment model allowed us to explore the effect of the different test administrations, taking into account pre-existing trends in cohorts' performance. We compared assessment results on three tests (history, global issues, and scientific reasoning) between the 2020 entering class (tested remotely) and the previous four cohorts (tested in-person). Our results revealed lower test performance and a bimodal distribution of effort scores in students tested remotely in 2020, but only on the more cognitively demanding scientific reasoning test, compared to the less arduous tests, history and global issues. Implications and limitations are discussed.

AUTHORS

Sarah Alahmadi, M.S.
James Madison University

Christine E. DeMars, Ph.D.
James Madison University

Large-Scale Assessment During a Pandemic: Results from James Madison University's Remote Assessment Day

Assessment efforts in higher education were among the many domains and practices that COVID-19 has disrupted in 2020. A report published by the National Institute for Learning Outcomes Assessment (NILOA) revealed that 97% of 813 higher education professionals who held assessment-related roles indicated that changes to their assessment were necessitated in response to COVID-19, especially with regards to modifying assignments or assessments (Jankowski, 2020). At James Madison University (JMU), assessment modifications were required not only at the course- and program-level, but also at the university-level. For more than 30 years, JMU has been collecting longitudinal data assessing learning outcomes for every cohort. Students are assessed twice, first as incoming first-year students (i.e., before completing any classes) and again after completing 45-70 credit hours. Such a model allows for a longitudinal assessment of learning growth. Additionally, having assessed students for the last 30 years allows us to observe larger trends in learning improvement across cohorts.

JMU's Assessment Day model and its logistics were comprehensively described by Pastor et al. (2019). The Assessment Days typically involve around 4,000 students tested in one of three proctored, 2-hour sessions. Different groups of incoming students are randomly assigned different configurations of Assessment Day instruments. Some assessments are completed using paper-and-pencil while others are computer-based. Proctors play an

CORRESPONDENCE

Email
alahmasi@jmu.edu

important role on Assessment Days as they ensure that tests are completed properly, noises are minimized, and students are motivated and aware of the importance of the Assessment Day. However, changes were necessary for the 2020-2021 academic year: Assessment was conducted remotely due to the COVID-19 pandemic, whereas all assessment was conducted in person in previous years.

Conducting a remote Assessment Day constituted many modifications to the abovementioned procedures (Pastor & Love, 2020). Instead of being tested on a specific day, students were allowed a three-week[1] window to complete the assessments via the links they received by email. The format of testing changed from paper-and-pencil to computer-based. Participation rates were somewhat lower. These changes raised several questions: Do students tested remotely in Fall 2020 score comparably to students tested in-person in the previous cohorts? If there are differences, are the differences similar across different tests? Also, do students tested remotely in Fall 2020 report test-taking effort similar to the effort reported by students tested in-person in the previous cohorts?

These findings indicate that the results of low-stakes testing may not precisely reflect individual differences in proficiency; rather the results are confounded by other factors, such as motivation or effort, rendering the validity of such results questionable.

Assessment Day testing is considered low-stakes testing, because students' performance bears no direct personal consequences. Thus, students could vary in the amount of effort they expend on assessment tests. Low effort has been found to affect performance by producing scores that underestimate ability (Wise & DeMars, 2005). In their review of examinee effort in low-stakes testing, Wise and DeMars computed differences between groups tested under motivating and less motivating conditions. Across 12 studies, they found that, on average, students tested under more motivating conditions performed more than one-half standard deviation higher. These findings indicate that the results of low-stakes testing may not precisely reflect individual differences in proficiency; rather the results are confounded by other factors, such as motivation or effort, rendering the validity of such results questionable. There are several strategies that could be employed to improve students' motivation, such as increasing the stakes of testing and selecting less cognitively taxing test designs. JMU utilizes both strategies by (1) making Assessment Days *semi*-consequential by not allowing students to register for future semesters if they did not attend Assessment Day, and (2) devising tests that contain mainly multiple-choice questions as opposed to essay questions; a strategy that has been shown to be less cognitively-overwhelming, maintaining higher levels of effort from students (DeMars, 2000). Also, students are made aware of the importance and value of Assessment Day before they complete their tests. In a typical year, proctors would ensure that students completed all the tests within the allotted time and that no students left the testing room early.

Moving Assessment Day online in Fall of 2020 raised several validity concerns that often accompany low-stakes, unproctored internet test (UIT) administrations. In general, implementing a UIT design entails unstandardized testing conditions among examinees regarding, to name a few, the amount of time spent completing the tests, environmental noise, and technological issues. While—specific to our interest—the fact that the test is *low-stakes* alleviates the usual UIT concerns around examinee cheating, it brings about questions related to examinee motivation and effort. Empirical evidence is mixed with regards to whether low-stakes UIT produces differences across test scores by introducing construct-irrelevant variance. One study that compared examinee performance in proctored versus unproctored online settings found no significant differences (Hollister & Berenson, 2009). Another study that examined the effect of *web-based* tests in several conditions—including proctored, in-person and unproctored, remote—reported no differences (Templar & Lange, 2008). Conversely, there is some evidence suggesting higher performance in web-based, remote unproctored cognitive tests (Karim, Kaminsky, & Behrend, 2014).

These findings collectively provide some evidence that differences in performance may occur. However, one study that looked specifically at performance differences between low-stakes online-proctored tests and online-unproctored tests found some reassuring results (Rios & Liu, 2017). The study examined differential performance and test-taking behavior based on whether online tests were proctored. Test-taking behavior was examined

[1] The window was later extended due to disruptions in on-campus courses early in the semester.

via keystroke data (the frequency of item views, items omitted, and items not-reached), and response time data (total testing time and rapid-guessing time). The results showed negligible and insignificant differences in terms of test-taking behavior as well as test scores between those whose online test was proctored and those whose online test was unproctored. These findings suggest that in low-stakes online testing, there are no meaningful implications for the absence of proctoring.

The question remains whether administering low-stakes tests remotely versus in-person would have differential implications for assessments. There is not yet any individual study that compares performance differences among college students on cognitive *low-stakes* tests in an in-person proctored, paper-and-pencil administration versus an online unproctored administration. By sharing the results of our remote Assessment Day, we hope to shed some light on this unexplored area. In this paper, we compare the scores from several tests delivered remotely in 2020 to the scores from the same tests administered in person in previous years, to see if there are performance differences and if those differences vary by test. We then examine differences in self-reported effort and in time spent testing as possible explanations of differences in test performance.

Method

Participants

Participants were first-year students entering the university in 2016-2020. All students were required to participate in Assessment Day, but different students were randomly assigned to each assessment instrument. For this study, data were used from all students who consented to having their results used for research and completed one of the three selected instruments, described below. Demographic information about the participants is shown in Table 1. In 2016-2019, students who did not complete their assessments were prevented from registering for the next semester until they participated in a make-up session. In 2020, there were no consequences for not participating. As described earlier, the assessments in 2016-2019 were completed at an assigned time, on paper, in a group setting, supervised by a proctor, whereas the 2020 assessments were completed anytime within a 3-week window, on computer, in a setting of the student's choice (generally home or dorm room), and unproctored.

There is not yet any individual study that compares performance differences among college students on cognitive lowstakes tests in an in-person proctored, paper-and-pencil administration versus an online unproctored administration. By sharing the results of our remote Assessment Day, we hope to shed some light on this unexplored area.

Assessment Instruments

Three of the General Education assessments were chosen for this study because they have been administered for at least five years and thus have a history from which to judge whether scores in 2020 were within the range of year to year fluctuation or represented a departure from past trends. These assessments span different subject areas and test lengths. The selected instruments were developed by faculty to target students' knowledge in history, global issues, and scientific reasoning. We also administered an assessment of test-taking motivation and effort, the Student Opinion Survey (Sundre & Moore, 2002).

Knowledge of history and political science is assessed using a 40-item test, with a possible number correct score range of 0 to 40. Knowledge of global issues is assessed by 31 items, with a possible number correct score range of 0 to 31. Scientific reasoning is assessed by 66 items, with a possible range for number correct score between 0 and 66. Lastly, effort and motivation are measured by a 5-item survey. Students indicate their agreement level with statements regarding how much effort they expended on a 5-point Likert scale ranging from 1 = *Strongly Disagree* to 5 = *Strongly Agree*. The possible total score range is 1 to 5 after taking the average over the five items.

Results

Test Scores

To make comparisons of cohort performance across the differently scaled assessments, we standardized the scores. The standardization was based on students with no course credit tested in 2016-2019; for these students, the mean was set to zero and

Table 1
Participants

Year	Test	*N*	% Female	% In-State Residents	% non-Hispanic White
2016	History	1041	60%	75%	76%
	Global Issues	821	61%	73%	77%
	Scientific Reasoning	817	61%	73%	78%
2017	History	996	59%	73%	79%
	Global Issues	1148	60%	74%	76%
	Scientific Reasoning	734	58%	73%	75%
2018	History	1027	58%	73%	77%
	Global Issues	767	60%	73%	75%
	Scientific Reasoning	745	62%	69%	78%
2019	History	1178	58%	77%	77%
	Global Issues	1031	60%	75%	76%
	Scientific Reasoning	458	60%	74%	77%
2020	History	841	62%	75%	76%
	Global Issues	840	63%	77%	77%
	Scientific Reasoning	457	67%	76%	77%

We observe a pattern of decreasing scores over the years on history and global issues, but fluctuating scores on the scientific reasoning test, with a large drop in 2020. In the scientific reasoning assessments, however, there was not a clear trend prior to 2020, and the 2020 group exhibited a more drastic decrease.

the within-group pooled-standard deviation was set to one. See Figure 1 for standardized mean comparisons across the last five cohorts for only those with no credits. Because the within-group standard deviation was set to one, differences in Figure 1 can be interpreted similarly to Cohen's *d*. We observe a pattern of decreasing scores over the years on history and global issues, but fluctuating scores on the scientific reasoning test, with a large drop in 2020. It seems that students in 2020 conformed to the general pattern of slightly decreasing scores year by year on the history and global issues assessments. The linear trend was statistically significant (history: $F_{1, 5078} = 21.35$, $p < .001$; global issues: $F_{1, 4602} = 47.70$, $p < .001$), but there were no significant differences among the cohorts beyond the linear trend (history: $F_{3, 5078} = 2.91$, $p = .4677$; global issues: $F_{3, 4602} = 188$, $p = .1303$).[2] In the scientific reasoning assessments, however, there was not a clear trend prior to 2020, and the 2020 group exhibited a more drastic decrease. A contrast between 2020 and the mean of the previous years showed that 2020 scores were significantly different ($F_{1, 3206} = 180.63$, $p < .001$). The 2020 mean was 0.75 standard deviations below the mean for the previous years.

Additional information about student test performance can be gained by examining the distribution of scores. In Figure 2, the score distribution for scientific reasoning did not just shift lower—the shape of the distribution changed. The mode of the distribution in 2020 was located just below the mode of previous cohorts, but there was a secondary mode of lower scores. A substantial portion of the students scored much lower than previous cohorts.

[2] There were five groups, so the omnibus F-test was partitioned into a 1-*df* linear trend a 3-*df* test of the remaining variance. The latter test was of interest in this study, and answered the question: Beyond the linear trend, were there any significant differences in the group means?

Figure 1
Mean Standardized Scores across Cohorts

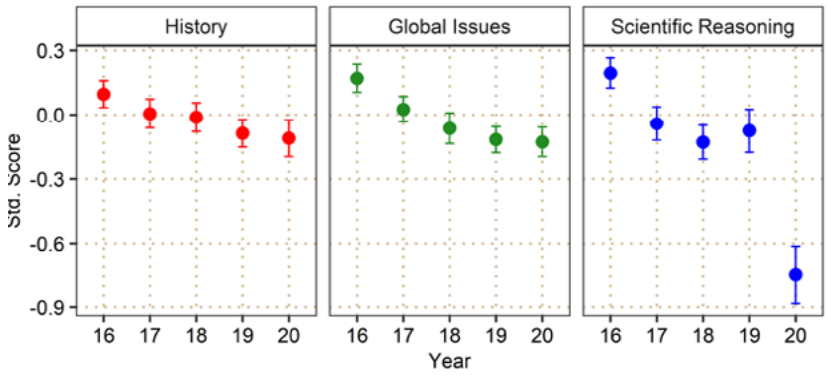
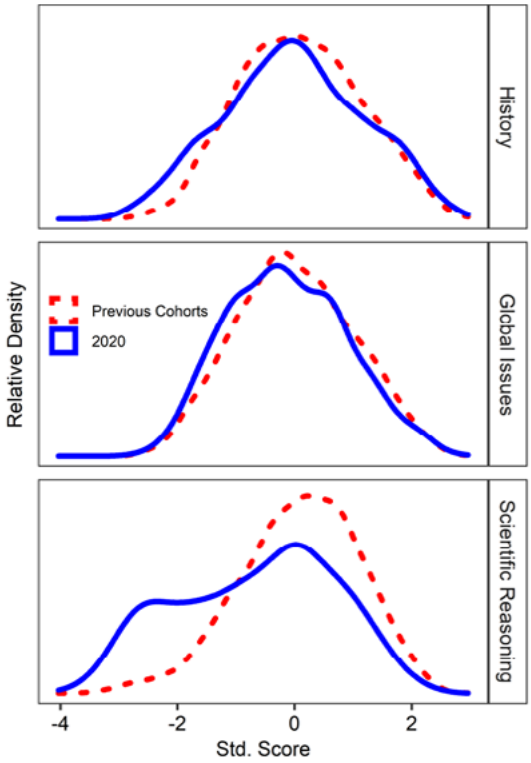


Figure 2
Distribution of Test Scores



Students enter the university with varying levels of course credit, such as transfer or AP credit. The comparison in Figure 1 used data only from students with no course credit, to avoid the possibility of confounding administration conditions with differences in the proportion of students with course credit. However, the performance of students with course credit may also be of interest. Table 2 presents raw mean scores on the three tests assessing students in history, global issues, and scientific reasoning, overall and broken down by course credit. For simplicity, we report mean scores for this year's cohort, 2020, and the previous four cohorts (from 2016 to 2019) combined. Students in the "No credit" column and the 2016-2019 row were used for setting the standardized metric in Figure 1. Incoming students in 2020 scored slightly lower on the history test; as discussed earlier, this was due to a decreasing linear trend, not to an unexpected drop in 2020. The 2020 students had considerably larger variability among their scores compared to students from the previous years, except in global issues. Typically, students with AP credit in US history or political

science scored the highest on the history test, followed by those with transfer credit, and then those with no credit. A similar pattern is observed for scientific reasoning. Overall, this pattern holds for 2020. For the global issues assessment, very few students had AP or transfer credits so we did not separate the students into subgroups. As in Figure 1, the largest differences in Table 2 between 2020 and previous years are found on the scientific reasoning assessment. Students in 2020—regardless of whether they had previous credit or not—scored distinctly lower than those in previous years with much higher variability among the scores, particularly on scientific reasoning. Could the interaction between cohort and assessment subject be due to differences in effort? We turn to answering this question next.

Comparable effort was reported by all five cohorts on all assessments, except on the scientific reasoning assessment. Slightly lower levels of effort were reported in 2020 across all assessments; however, they did not seem to deviate much from previous cohorts.

Table 2
Performance across Cohorts

Test	Cohort	Raw Score Mean (*SD*) *N*			
		All	AP	Transfer	No credit
History					
	2020	21.77 (7.26) 841	28.84 (5.81) 57	21.65 (6.50) 80	21.21 (7.16) 704
	2016-19	22.48 (6.39) 4242	30.69 (4.77) 280	22.03 (5.74) 414	21.89 (6.12) 3548
Global Issues					
	2020	16.76 (5.05) 840			
	2016-19	17.40 (4.99) 3767			
Scientific Reasoning					
	2020	38.66 (10.29) 457	45.89 (9.37) 35	38.43 (9.69) 44	38.01 (10.21) 378
	2016-19	44.35 (7.88) 2754	51.65 (6.40) 201	44.54 (7.95) 174	43.72 (7.68) 2379

Note. Subgroup scores are not reported for global issues, because few students had AP or transfer credits in this domain ($N = 22$ in 2020, $N = 64$ in 2016-2019). Students were removed if they omitted more than 25% of the items.

Effort

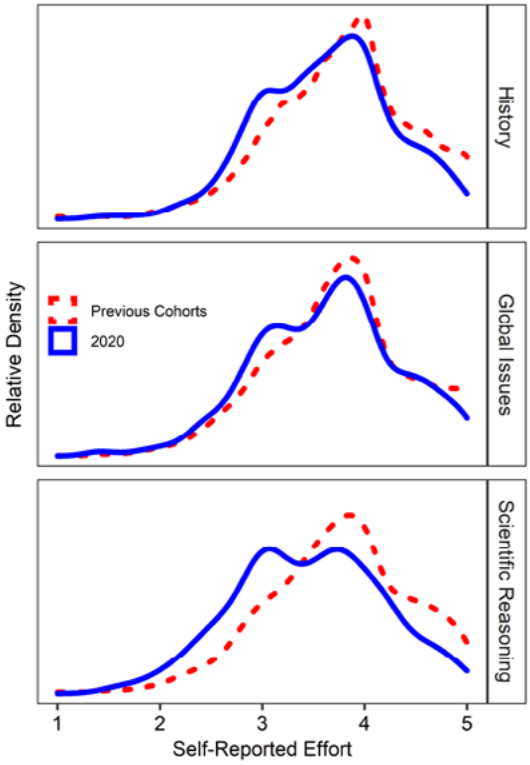
Comparable effort was reported by all five cohorts on all assessments, except on the scientific reasoning assessment. Slightly lower levels of effort were reported in 2020 across all assessments; however, they did not seem to deviate much from previous cohorts (see Table 3). For example, in history, the mean dropped 0.10 (on the 5-point scale) from 2019 to 2020, similar to the 0.13 drop from 2016 to 2017. In global issues, the 2020 mean was only 0.01 below the 2018 mean. Similarly, in scientific reasoning, the 2020 mean was 0.04 below the 2018 mean. These differences are not far from the normal year-to-year fluctuations.

The means and standard deviations, however, do not give a full comparison of effort across cohorts. Refer to Figure 3 for density plots of effort comparing 2020 cohort and previous cohorts combined. The 2020 effort appears bimodal, especially in scientific reasoning. There was a larger bump in students responding *neutral* (3) in 2020. This bump may be related to the greater density of very low scores seen earlier in Figure 2.

Table 3
Means and Standard Deviations of Self-Reported Effort across Cohorts and Assessments

Test	*M* (*SD*) *N*				
	2016	2017	2018	2019	2020
History	3.87 (0.67)	3.74 (0.68)	3.74 (0.67)	3.73 (0.70)	3.63 (0.65)
	1,030	986	1,009	1,164	828
Global Issues	3.69 (0.68)	3.82 (0.67)	3.65 (0.68)	3.73 (0.68)	3.64 (0.70)
	819	1,140	763	1,027	825
Scientific Reasoning	3.86 (0.68)	3.87 (0.70)	3.50 (0.65)	3.71 (0.73)	3.46 (0.72)
	772	667	745	441	456

Figure 3
Density of Self-Reported Effort Scores



To examine the possible relationship between effort and test performance, we have computed the squared correlations between effort and test scores. The squared correlation measures the amount of variation in test scores that can be attributed to exerted effort as reported by the students (see Table 4). Generally, effort seems to be most strongly associated with the scientific reasoning test across cohorts. We also observe an increase in the amount of variation in test scores that is explained by effort in 2020. It appears, overall, that higher levels of effort were associated with higher test performance, especially on the scientific reasoning test.

Time Spent Testing

Another measure of effort is the time students spend taking the test. For each test in 2020, the total time the student spent viewing the test, including short videos at the beginning with information about the test, was recorded. In Figure 4, the standardized score is plotted as a function of the total testing time. Students with transfer or AP credit are not shown.

Table 4
Squared Correlation between Test Score and Self-Reported Effort

Test	Cohort				
	2016	2017	2018	2019	2020
History	.05	.07	.07	.07	.12
Global Issues	.02	.06	.10	.07	.09
Scientific Reasoning	.11	.10	.17	.17	.21

Do students tested remotely in 2020 show less or greater likelihood of correctly answering specific items on the tests than students tested in person, after controlling for ability?

The relationship between time and score appears to be non-linear, especially in science. For students who spent little time on the test, scores increased as time increased. For students who spent at least moderate amounts of time testing, there was little relationship between time and score. Only the first 30 minutes are shown in Figure 4; after that point, the lack of relationship between time and score continued. A regression line was fit to the relationship between the natural log of time and scores. The analysis for fitting the regression line included students not shown in the graph, beyond the 30-minute point. However, students were omitted from the analyses if their time was more than three times the median testing time; it did not seem plausible these students were spending that much time actually focused on the test. This impacted 4.1%, 5.7%, and 3.7% of the students on the history, global issues, and scientific reasoning tests, respectively. The regression accounted for 13% of the variance in history, 10% in global issues, and 27% in scientific reasoning. Testing each pair of correlations at $\alpha = .017$ for a Bonferroni-corrected familywise $\alpha = .05$, the history and global issues correlations were each significantly different from the scientific reasoning correlation, but not significantly different from each other. Time spent on the test was a better predictor of performance for the scientific reasoning test than for the other two tests.

In the history and global issues tests, the time spent per item was also recorded. From this, an adjusted time was calculated. First, a median time was calculated for each item. When a student spent more than three times the median time on an item, the student's time for that item was replaced with an imputed time[3] and the total testing time was recalculated (here labelled the adjusted time). The log of the adjusted time accounted for 21% of the variance in test scores for both history and global issues. The scientific reasoning test might have shown a comparable increase in the correlation, but item-level response times were not available for this adjustment.

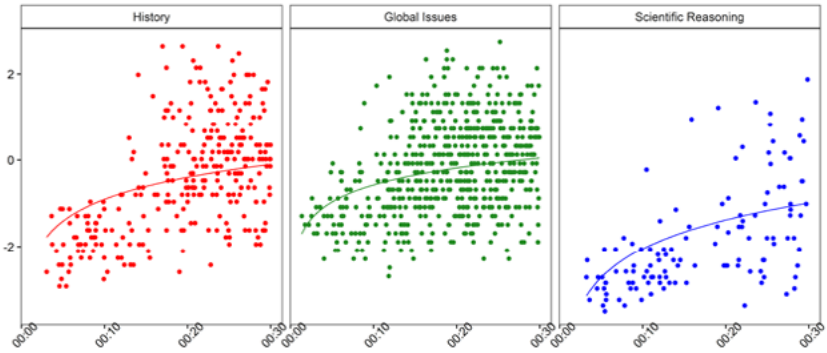
Differential Item Functioning

Remote testing appears to impact students' average performance specifically on the scientific reasoning test. This raises the question of whether remote testing could produce differences not just at the test level, but also at the item level. Do students tested remotely in 2020 show less or greater likelihood of correctly answering specific items on the tests than students tested in person, after controlling for ability? We conducted a differential item functioning (DIF) analysis to examine whether individual items exhibit differential performance between the past four cohorts (2016-2019 combined) and the 2020 cohort after controlling for ability or knowledge level.

We utilized the Mantel-Haenszel method (Holland & Thayer, 1988) to calculate α_{MH}, which is a ratio of the odds of answering an item correctly for the reference group (i.e., past cohorts) over the odds of answering an items correctly for the focal group (i.e., 2020 cohort).

[3] The student's median response time was first estimated as the median across items, excluding any item on which the student took more than 3 times the group's median time for that item. Then the student's ratio was calculated as the ratio of the student median to the group median (overall, across items). Finally, for the excluded items, the response time was imputed as the student's ratio times the group's item-specific median for that item. For example, student Q's median response time across items was 10% more than the group median. On item W, student Q took a break and spent 600 seconds on the item, when the group median time was 22 seconds. Student Q's time for item W would be adjusted to 1.1*22 = 26.4 seconds.

Figure 4
Correlation between Testing Time and Test Score



Note. The fitted line shows the regression of test score on the natural log of time spent. The points cluster closer to the line for scientific reasoning than for the other tests.

The Mantel Haenszel procedure statistically tests the null hypothesis that $\alpha_{MH} = 1$, indicating that the odds for the reference group and focal group are the same. We controlled for false positive rate using the Benjamini-Hochberg procedure (1995). To estimate the effect size of the DIF, we employed ETS classification (Zwick, 2012), which uses the index Δ_{MH} : $\Delta_{MH} = 2.35 \ln(\alpha_{MH})$. According to the ETS classification, an item is classified level A DIF if the absolute value of Δ_{MH} is less than 1 **or** if Δ_{MH} is not statistically significantly different from 0. To be classified as level C DIF, an item has to show an absolute value of Δ_{MH} that is equal to or greater than 1.5 with a Δ_{MH} that is statistically significantly different from 1. Level B classification includes items that do not meet level A or C requirements.

JMU's remote Assessment Day was an exceptional opportunity to study performance differences attributable to testing settings (in-person versus remote) in low-stakes, student learning assessment.

For the history test, item 8 and item 13 showed level C DIF. For the global issues test, none of the items showed DIF with a large effect size (i.e., $\Delta_{MH} \geq 1.5$). For the scientific reasoning test, only item 33 was identified as exhibiting level C DIF. All three items favored the reference group (i.e., previous cohorts) over the focal group (i.e., 2020 cohort). That is, after matching the 2020 examinees with examinees from the previous cohorts with the same total scores, the previous cohorts scored higher on these three items. Inspecting the content of said items, we could not find any plausible explanation as to why these items functioned differently. The lower performance on the scientific reasoning test in 2020 seems to be a pervasive effect, not limited to specific items.

Conclusion

JMU's remote Assessment Day was an exceptional opportunity to study performance differences attributable to testing settings (in-person versus remote) in low-stakes, student learning assessment. The results from the remote Assessment Day were contrasted with results from the previous four cohorts tested in person to control for any pre-existing trends. In terms of mean performance, students tested remotely in 2020 followed the preceding trend of decreasing scores on the history and global issues tests. However, the 2020 cohort exhibited significantly lower scores on the scientific reasoning test than their counterparts in previous years. Those students also showed a different distribution of effort on the scientific reasoning test than students in previous cohorts due to lower effort levels produced by a subgroup of the 2020 students, producing a bimodal distribution. Test performance on scientific reasoning also exhibited this shift in distribution. The scientific reasoning test was longer than the other two tests (66 items vs. 40 and 31), and science may be perceived as more difficult by students. Thus, the different patterns of effort and performance may be attributable to the higher cognitive demand of the scientific reasoning test. Effort was also measured in the 2020 cohort as the time spent taking the test, which predicted performance better for the scientific reasoning test than for the other tests. In future work, as suggested by an anonymous reviewer, we plan to look further at the group of students who gave reasonable effort to assess how their test performance compares to previous cohorts.

We also assessed whether the observed score differences were consistent across items or if instead there was differential item functioning (DIF). Only three items showed large and significant DIF effects between the 2020 cohort and previous cohorts. Evaluating the content of those items judiciously did not yield a reasonable explanation for the DIF.

Overall, the results from JMU's remote Assessment Day suggest that the differences in performance in low-stakes educational assessments observed in students who tested remotely in 2020 can be mainly ascribed to differences in test types. The more arduous scientific reasoning test was the only test showing a significant drop in scores compared to the history and global issues tests which may have required less exertion of cognitive resources. Our findings also highlight the promising potential of remote, large-scale assessment. While a main disadvantage of conducting assessment remotely seems to manifest in the differential performance and effort based on test type, some advantages include less demand for resources (e.g., hiring proctors, reserving rooms, etc.) and the opportunity to collect item-level data on effort. Collecting item-level data allows us to better assess how much effort a student put forth on a test as evidenced by time spent on each individual item rather than the test as a whole. We plan to apply the same remote administration procedures of the 2020 Assessment Day to at least one more assessment day at JMU to further examine the effect of test type and effort levels using data collected at the item level.

We recognize a few limitations of the current study. Effects of remote testing in 2020 may have been impacted by anxiety or other construct-irrelevant factors besides effort due to the pandemic. Lower scores exhibited by the students may have also been affected by events in the semester previous to their enrollment at JMU, when secondary school classes were abruptly moved online. Nonetheless, the current results provide insight into some factors that may impact remote testing. We will continue to study those factors as we assess the same 2020 cohort after completing 45-70 credit hours.

AUTHORS NOTE:
Sarah Alahmadi, https://orcid.org/0000-0002-9985-6807

Christine E. DeMars, https://orcid.org/0000-0003-0050-3655

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* (Series B), 57, 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*, 55–77. https://doi.org/10.1207/s15324818ame1301_3

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel Procedure. In Wainer H & Braun HI (Eds.), Test validity (pp. 129–145). Hillsdale, NJ, US.

Hollister, K. K., & Berenson, M. L. (2009). Proctored versus unproctored online exams: Studying the impact of exam environment on student performance. *Decision Sciences Journal of Innovative Education*, *7*(1), 271-294. https://doi.org/10.1111/j.1540-4609.2008.00220.x

Huff, K., Cline, M., & Guynes, C. S. (2012). Web-based testing: Exploring the relationship between hardware usability and test performance. *American Journal of Business Education (AJBE)*, *5*(2), 179-186. https://doi.org/10.19030/ajbe.v5i2.6820

Jankowski, N. A. (2020, August). *Assessment during a crisis: Responding to a global pandemic*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://www.learningoutcomesassessment.org/wp-content/uploads/2020/08/2020-COVID-Survey.pdf

Karim, M. N., Kaminsky, S. E., & Behrend, T. S. (2014). Cheating, reactions, and performance in remotely proctored testing: An exploratory experimental study. *Journal of Business and Psychology*, 29(4), 555-572. https://doi.org/10.1007/s10869-014-9343-z

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*, *144*, 1-13.

Pastor, D., & Love, P. (2020). University-wide assessment during Covid-19: An opportunity for innovation. *Intersection*: *A Journal at the Intersection of Assessment and Learning*, *2*(1), 17617.

Rios, J. A., & Liu, O. L. (2017). Online proctored versus unproctored low-stakes internet test administration: Is there differential test-taking behavior and performance? *American Journal of Distance Education*, *31*(4), 226-241. https://doi.org/10.1080/08923647.2017.1258628

Sundre, D. L. & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update*, *14*, 8-9.

Templer, K. J., & Lange, S. R. (2008). Internet testing: Equivalence between proctored lab and unproctored field conditions. *Computers in Human Behavior*, *24*(3), 1216-1228. https://doi.org/10.1016/j.chb.2007.04.006

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

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1), i-30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

Abstract

Course-based undergraduate research experiences (CURE) can improve student skills, views toward research, and identity as a scientist. Many barriers exist for implementing program-wide CUREs, including assessment of these programs. This paper addresses the direct assessment of a required senior CURE in one high-volume (400+ students per year) academic program. Research groups (45-50 groups per semester, four-six students each) design, implement, and analyze data in a research study culminating in a poster symposium and paper write-up. This paper discusses the iterative process of developing the assessment procedures tied to program-level student learning outcomes, including suggestions for implementation at other institutions. Programs that wish to create an assessment of CURE should include the collaboration of key stakeholders in developing processes and tools to ensure findings guide course content and teaching strategies.

AUTHORS

Laura K. Merrell, Ph.D., C.P.H.
James Madison University

Dayna S. Henry, Ph.D., MCHES
James Madison University

Stephanie L. Baller Ph.D.
James Madison University

Audrey J. Burnett, Ph.D.
James Madison University

Andrew A. Peachy, Dr.PH.
James Madison University

Yu Bao, Ph.D.
James Madison University

Developing an Assessment of a Course-Based Undergraduate Research Experience (CURE)

Introduction

Student research, particularly at the undergraduate level, is considered a High Impact Practice (HIP) (National Survey of Student Engagement, 2007; Kuh & Association of American Colleges & Universities [AAC&U], 2008). According to the Association of American Colleges and Universities (AAC&U), the goal of undergraduate research is to "involve students with actively contested questions, empirical observation, cutting-edge technologies, and the sense of excitement that comes from working to answer important questions" (Kuh & AAC&U, 2008, para 7). It can broadly be defined as scholarship, creative activities, or scientific inquiry that leads to the production of original work (Kinkead, 2003). The benefits of participating in research as an undergraduate student are numerous and include increased interest in pursuing graduate education, viewing themselves as scientists, improved writing skills, ethical conduct, understanding others' research, inquiry and analysis, independence, communication, and teamwork (Hunter et al., 2007; Lopatto, 2010; Russell et al., 2007). In particular, faculty-mentored research experiences have the potential to increase students' identities as scientists in their field (Auchincloss et al., 2014). Typical mentoring activities involve honors projects and independent studies with faculty mentoring one or a few students. These types of experiences require students to apply, and oftentimes high-performing students or those with a greater understanding of the university system, self-select into these opportunities. This can exacerbate inequities in access to mentored research (Bangera & Brownell, 2014).

CORRESPONDENCE

Email
merrellk@jmu.edu

Whereas HIPs have the potential to positively impact all students in terms of learning outcomes, retention, and graduation, research has indicated they are particularly impactful for historically underrepresented students (see for example Collins et al., 2017). Unfortunately, studies show these groups are less likely to participate in HIPs, such as student research, for a variety of reasons (Kinzie, 2012). According to the 2019 annual results of the National Survey of Student Engagement, research with faculty is among the least common HIPs among most Carnegie Classifications. Additionally, inclusivity in such activities differs by student characteristics, including race/ethnicity and non-traditional, first-generation, or transfer student status. Barriers to engaging in traditional individual faculty-mentored research experiences for underrepresented students include lack of awareness of research opportunities and their benefits, perceived barriers of interaction with faculty, and personal and financial barriers (see Bangera & Brownell, 2014 for a review). However, given that the positive outcomes of engaging in research are numerous, it is important to consider how to remove barriers and increase access to research opportunities. One example is course-based undergraduate research experiences (CURE) which provide the opportunity for many students to access mentored research with faculty while gaining course credit, rather than needing to apply or spend time outside of class (Auchincloss et al., 2014).

Given that the positive outcomes of engaging in research are numerous, it is important to consider how to remove barriers and increase access to research opportunities.

Given research demonstrating the benefits of CUREs and increased accessibility to students, it is important to consider how they may be assessed at the course or program level (Auchincloss et al., 2014). Most assessments of undergraduate research rely on self-report data that measure advances in skills, such as collaboration, written and oral presentations, and conducting research studies (Corwin et al., 2015; Weston & Laursen, 2015), as well as others related to their attitudes toward science (Hanauer et al., 2016). In a review of over 60 articles published on the impact of undergraduate research, fewer than 10% had direct measures of student learning despite calls for better assessments (Linn et al., 2015). Shortlidge and Brownell (2016) suggest that direct assessments of CURE should align with course learning outcomes and could potentially use existing 'off-the-shelf' assessments of skills, such as data analysis, experimental design, scientific reasoning, and scientific literacy.

Given the need to better assess student research experiences, and the benefits of using CURE to reduce barriers to access, this case study presents the recurring, iterative systematic assessment of student research outcomes in a faculty-mentored and course-based undergraduate research experience at a large public university located in the central Atlantic region of the United States. This paper will first describe the course, followed by the assessment process at the institution and the specific development of the assessment process for the CURE. The paper ends with a discussion of challenges and suggestions for implementation.

Description of CURE

First, it is important to understand the context in which the CURE described in this case study exists. The institution is a large public 4-year master's-granting university located in the central Atlantic. The vision of the institution-James Madison University-is to be "the national model of the engaged university," including engaged learning, civic engagement, and community engagement. Engaged learning is defined as "developing deep, purposeful and reflective learning, through classroom, campus, and community experiences in the pursuit, creation, application and dissemination of knowledge" (James Madison University, 2021).

The department in which the CURE is required, the Department of Health Sciences, offers a Bachelor of Science degree that prepares students to pursue entry-level, non-clinical health careers, or to apply to graduate programs in a variety of health fields, including but not limited to, athletic training, dentistry, medicine, occupational therapy, physical therapy, physician assistant studies, and public health. The anticipated growth in these career fields, and the flexibility of the curriculum within the degree, helped to make the program the largest producer of graduates at the university. There are approximately 1,600 students majoring in this program with 450 students graduating each year (of which 20% are minority race/ethnicity identified). Given the size and nature of the program, in 2016, the faculty re-envisioned the curriculum and subsequently aligned the assessment of the program objectives when the curriculum was implemented in 2018.

Through collaborative processes guided by the departmental curriculum and instruction and assessment committees, with the support of the department head and the university assessment center, the faculty affirmed that the inclusion of HIPs, particularly undergraduate research, was critical to achieving student learning outcomes of the program and supporting the vision of the university. Inclusion of HIPs was added as a source of evidence for excellence in teaching within the annual evaluation and tenure and promotion guidelines of the department to acknowledge the contributions of faculty teaching the course. During program modification, the faculty generated 10 program objectives and mapped the curriculum to these program objectives. Two of the program objectives related to research and communication were mapped to a senior-level research methods course required of all majors in the program (see description below). As noted in the introduction, close mentoring from a faculty member is a critical component of a CURE. Therefore, the prioritization of the HIP within the curriculum required that the class size be limited relative to most other courses offered within the department to ensure high-quality mentoring that meets the diverse learning needs of students. Teaching two sections of the research methods course would comprise half of a faculty member's teaching load per semester.

The course serves more than 400 students per year in faculty-guided Institutional Research Board (IRB)-approved research projects. Early iterations of this project were described previously in Peachey & Baller (2015). Enrollment for the course is typically capped at 25 students per section and student research teams are comprised of four-six students (see Table 1). Each student completes human subjects' ethics training (CITI training). Teams then select a topic, develop research questions and instrumentation for data collection, submit a proposal to the IRB for approval, collect and analyze data, present findings at a bi-annual research symposium, and prepare a final written report of the research project. The research poster symposium is a shared experience across all sections of the course where students can see the breadth and scope of peer accomplishments, as well as present their own group work in a quasi-professional setting. Departmental faculty and other university administrators regularly attend the symposium to discuss the research projects with student groups. In some semesters, judges provided feedback to top-performing teams, leading to award recognition for students. During the COVID-19 pandemic, adjustments were made to the course and project, including exclusion of the poster symposium. The symposium is scheduled to resume fall of 2021.

The prioritization of the HIP within the curriculum required that the class size be limited relative to most other courses offered within the department to ensure high-quality mentoring that meets the diverse learning needs of students.

Table 2
Student enrollment and faculty teaching committed to undergraduate research course.

Semester	Students Enrolled in HTH 408	Course Sections (n)	Faculty	FTE	Average Class Size	Student Projects (n)
Fall 2017	186	8	4	1	23.3	38
Spring 2018	253	10	5	1.25	25.3	50
Fall 2018	181	8	4	1	22.6	39
Spring 2019	258	11	5	1.25	23.5	55
Fall 2019	203	10	5	1.25	20.3	44
Spring 2020	254	11	5	1.25	23.1	n/a*
Fall 2020	203	10	5	1.25	20.3	52
Spring 2021	243	10	4	1	24.3	43
TOTAL	1,781	78	-	-	-	321
Average	223	10	5	1	22.8	46

**Student projects were not assessed the first semester of the COVID-19 pandemic*
FTE = Full-Time Equivalent for tenure line faculty teaching a 4/4 course load

Because each student in the program completes the research project as a member of a research team, it can be used to assess two of the program-level objectives. The poster presentation serves as a health-specific communication tool to convey the methods, analysis, and results of public health research questions.

Assessment of CURE

A four-person committee, three of whom do not teach the research methods course, in consultation with the department head and the university assessment center, conducts the assessment activities for the program. All committee members are tenure-track faculty who regularly conduct and publish research requiring statistical analysis and interpretation. The assessment committee works closely with university assessment center staff in developing, analyzing, and interpreting assessment data. The assessment committee chair also attends extensive university-required assessment and measurement training. The committee is charged with conducting the required annual assessment of the program's student-centered learning outcomes (SLOs). Assessment activities across the institution contain a variety of indirect and direct measures of student learning. For example, in the Department of Health Sciences, the majority of SLOs in the department are assessed with a pre- and post-test of majors in their first major course and last semester within the program, respectively. Given that all students complete the research methods course, it provides the opportunity to assess two program-level, higher-order SLOs beyond the level of knowledge recognition and comprehension as indicated by Bloom's taxonomy in cognitive domain of educational goals (Bloom et al., 1956; Huitt, 2011). As stated above, these SLOs broadly cover a variety of topics related to the research methods course:

Given that all students complete the research methods course, it provides the opportunity to assess two program-level, higher-order SLOs beyond the level of knowledge recognition and comprehension.

As a result of participating in the Department of Health Sciences curriculum, graduating students will be able to:

> **SLO1.** Utilize the basic concepts, methods, and tools of public health science data collection, analysis (statistics), and evaluation,
>
> **SLO2.** Utilize basic concepts of public health-specific communication, including technical and professional writing, and the use of mass media and electronic technology.

These two SLOs are stated with clarity and specificity by including rich descriptions of the content and skills that are required in the CURE. Clear and specific SLOs can aid the design of instructional courses, clarifying what students should comprehend and teachers should evaluate (Bloom et al., 1956). In addition, SLOs can promote the development of assessment tools by providing guidelines about the student population to be assessed, the type of assessment to be used, and the type of inferences to be made from results (Kuh & Ewell, 2010). For example, the two SLOs mapped to the CURE indicate that a performance assessment that evaluates students' skills and behaviors, as evidenced through certain products or performances (e.g., research posters), is the most appropriate approach. Therefore, a rubric that articulates the criteria to address the expectations of the performance tasks, as well as the specification of different levels of success for each criterion, is often selected as the instrument to evaluate students' mastery levels of the desired knowledge and skills (Andrade, 2000; Arter & Chappuis, 2007; Moskal, 2002; Stiggins, 2001). In assessment, the rubric is also considered a direct measure, since students must explicitly demonstrate their ability to conduct important research-related processes and communicate practical findings to a lay audience (Allen, 2003, p.88; Suskie, 2018). The poster evaluation rubric was developed by the committee in conjunction with the university assessment center using the poster instructions and rubrics used for grading by the instructors of this course (see Figure 1, for example). The next section describes the iterative process of rubric development.

Development of Poster Assessment

Based on initial discussions, the committee developed a rubric that assessed 15 criteria covering research elements (e.g., quality of research questions, appropriate statistical analysis) and writing and style elements (e.g., grammar, writing quality, and layout) of the

Figure 1
Example of faculty-provided course poster rubric used in poster assessment rubric development.

	Excellent	Good	Satisfactory	Unsatisfactory
Information Presented /8	All required information is included. No unneeded information is included. 8	Most of the required information is included. Minor edits (included more or remove unneeded) only. 6.5	Much of the required information is included but could need more than minor edits (include more or removed unneeded). 5.5	Missing a lot of required information or includes a lot of unneeded information. 0
Layout/ Presentation /7	Poster is visually appealing, organized well, made good use of graphics/tables/figures. Font, headings, and spacing is consistent throughout. 7	Minor errors in poster formatting. Could use some minor editing or re-formatting of headings, fonts, spacing, and graphics/tables/figures. 5.5	Major errors in poster formatting. Could use some major editing or re-formatting of headings, fonts, spacing, and graphics/tables/figures. 5	Poster is disorganized and inconsistently formatted. 0
Writing Style /5	Easy to read. Used 3rd person and past tense. Minimal spelling/grammatical errors. 5	Minor errors with flow, spelling, grammar or writing style. Used mostly 3rd person and past tense. Should use shorter sentences in some places. 4	Major errors with flow, spelling, grammar, or writing style. Used incorrect tense or 1st person. Needed much shorter sentences. 2	Poster was very difficult to understand. Use of incorrect tense throughout. 0
TOTAL				/20

poster (see Figure 2). Initially, rubric elements were scored on a three-point scale from 'Poor,' which earned zero points, to 'Excellent,' which earned two points, for a maximum score of 30 points. This rubric was piloted using 20 posters from spring 2018. All posters were independently scored by assessment committee members involved in the development of the rubric and who did not teach the research methods course. Initial inter-rater reliability analysis showed inconsistencies in how the three raters assessed each poster.

Figure 2
First iteration poster assessment rubric.

Rater: _____

Poster IRB # _____	**Excellent (2)**	**Good (1)**	**Poor (0)**
Title Represented major purpose of study including variables in RQ			
Purpose of the Study Provided a clear and concise rationale for the study based on previous literature and/or theory			
RQ/Hypotheses RQ are stated and measurable			
Procedures Described sampling and data collection procedures			
Instruments/Measures Identified instruments that measure the variables			
Analysis Selected appropriate information for each analysis			
Results Presented appropriate information for each analysis			
Discussion: Limitations Identified obvious limitations regarding procedures			
Layout: Graphics Used appropriate graphs/tables; well organized			
Layout: Text Used appropriate amount of text			
Layout: Organization Well organized			
Writing Quality: Formality Used formal, concise language			
Writing Quality: Grammar/Spelling/Tense Was free of any grammar or spelling errors			
Writing Quality: Consistency Used a consistent style			
Overall Quality			
Total			/30
Notes.			

As a result of this analysis, the rubric elements were refined to address issues identified in the rating *process* (see Figure 3). The committee identified that ratings were potentially subjective. What one rated as 'excellent' may have been viewed as 'good' by another rater. Therefore, ratings were shifted to reflect the perceived understanding of the process of research, rather than the subjective rating of the research project itself. Elements were rated as 'Absent' if they did not meet the description of the element or if important information was missing, 'Not Clear' if the poster met the description of the element but required improvement, and 'Present' if the poster met the description of the element. For example, a poster title that represented the overall purpose of the study, including major independent and dependent variables, would be scored as 'Present.' A poster that lacked a title would be scored as 'Absent.'

Figure 3
Second iteration of the poster assessment rubric.

Rater: _____

Poster IRB# _____	Present (2)	Not Clear (1)	Absent (0)
Title Represented major purpose of study including variables in RQ			
Purpose of the Study Provided a clear and concise rationale for the study based on previous literature and/or theory			
RQ/Hypotheses RQs are stated and measureable			
Procedures Described sampling and data collection procedures			
Instruments/Measures Identified instruments that measure the variables			
Analysis Selected appropriate information for each analysis			
Results Presented appropriate information for each analysis			
Discussion Results tied back to literature and identified obvious limitations regarding procedures			
	Excellent (2)	**Good (1)**	**Poor (0)**
Layout: Graphics, Text, Organization Used appropriate graphs, tables, appropriate amount of text. Well organized			
Writing Quality: Formality, Grammar/Spelling/Tense, Consistency Used formal, concise language. Was free of any grammar or spelling errors, used a consistent style.			
Total			/20
Overall Quality:			
Notes:			

Additionally, presentation elements related to layout and writing initially carried equal weight as research elements. Therefore, a poster that was well presented, but had major methodological issues, could obtain the same score as a well-done research study with sub-standard layout and writing. The six elements for layout and writing were combined into two criteria and continued to be rated on the 'Poor' to 'Excellent' scale, reflecting the subjective nature of those elements. While the committee felt it was important to include an overall rating of the poster, they recognized it should not be scored as other elements. Rather a rating of the overall quality without points should provide a way to check if the subjective perception of the quality of a poster aligned with the score it received.

Each time the poster assessment rubric is revised, it is shared with the instructors of the research methods courses.

Despite the change to fewer evaluative ratings, issues with inter-rater reliability persisted when the next set of posters was rated during the fall 2018 semester. The committee identified the need to provide in-depth descriptions of poster criteria and each of the ratings with examples that raters may refer to when assessing each poster. The rubric currently utilized to assess the posters (see Figure 4) includes 10 elements (i.e., title, purpose of the study, research questions/hypotheses, procedures, instruments/measures, analysis, results, discussion, layout, and writing quality). Posters are scored on a scale from 'Present' to 'Absent' for all research elements and from 'Good' to 'Poor' for layout and writing elements with two points possible for each element. Each time the poster assessment rubric is revised, it is shared with the instructors of the research methods courses.

Figure 4
Current iteration of poster assessment rubric.

Poster Letter# _____	Present (2)	Not Clear (1)	Absent (0)
Title Represented major purpose of study including variables in RQ	Contains all major variables, can determine general purpose of the study from the title.	Contains some but not all major variables OR hard to understand purpose of the study from the title/doesn't match major variables used in the study.	Title is absent OR not connected to the major purpose of the study.
Purpose of the Study Provided a clear and concise rationale for the study based on previous literature and/or theory	Presents a concise review of previous literature that is connected to the purpose and RQ for the study.	Presents a summary of previous literature; the purpose of the study is not completely connected to the previous literature in terms of importance and need OR summary of previous literature is not complete in terms of making a case for the study.	There is no connection between the study background and purpose of the study (i.e., variables in background do not match variables in the study).
RQ/Hypotheses RQ are stated and measurable	Clear which variables are being examined and the direction of predicted association (if applicable).	Direction of predicted association not clear (if applicable) OR variables are unclear OR wording is unclear.	No measurable RQ OR hypotheses stand (i.e., RQ implies bivariate relationship but only one variable is included).
Procedures Described sampling and data collection procedures	Describes major characteristics of sample (e.g., JMU students, 18+), method of data collection, and sampling strategy.	Missing one or two requirements from "present".	Missing all three requirements from "present".
Instruments/Measures Identified instruments that measure the variables	Described how they measured all the major variables in the study (e.g., instrument names, cut-offs).	Not clear how each variable was measured OR names of measures not clearly identified OR identified for some but not all variables.	No instruments identified OR measurement not described.
Analysis Selected appropriate analysis for RQ	Analysis is clearly identified either in methods, or in the results by proper reporting procedures AND is correct based on their data. Variables are operationalized clearly enough so proper analysis can be determined.	Analysis is mentioned, but it is not clear if it is correct because variable(s) operationalization is unclear.	No statistics are mentioned OR they are incorrect given the operationalization of the data.
Results Presented appropriate information for each analysis	Presented ALL necessary information including test statistics, degrees of freedom, and p value. Results answer RQ.	Some of the necessary information was presented, but not all (e.g., report F statistic without degrees of freedom). Some results answer RQ but not all are presented.	Results not presented OR do not answer RQ.
Discussion Results tied back to literature and identified obvious limitations regarding procedures	Discusses most salient findings based on results; relate findings back to previous literature; includes most salient limitations.	Missing one or two requirements from "present".	Missing all three requirements from "present".
	Good (2)	**Fair (1)**	**Poor (0)**
Layout: Graphs, Text, Organization Used appropriate graphs/tables, appropriate amount of text, well organized			
Writing Quality: Formality, Grammar/Spelling/Tense, Consistency Used formal, concise language, was free of any grammar or spelling errors, used a consistent style			
Total:			/20
Notes:			

Poster Assessment Process

Each semester, a committee member (who does not take part in rating the posters) uses a list of research methods posters identified by IRB numbers and instructors to randomly select 10 posters for assessment using a random number generator. The number of posters selected is proportional to the number of sections of research methods each instructor teaches. For example, if there are 10 sections of the course in each semester, and an instructor teaches two of them, then two posters will be randomly drawn from all the posters from their sections. This ensures that all instructors are proportionally represented in the posters that the committee assesses. Instructors are asked to download posters from their classes into a folder on a shared network drive with student and instructor names removed. Only this committee member knows from which instructor the posters were drawn.

Each semester, a committee member (who does not take part in rating the posters) uses a list of research methods posters identified by IRB numbers and instructors to randomly select 10 posters for assessment.

The assessment committee members independently evaluate each poster using the rubric and enter their scores into online survey software. The results are downloaded and the raters then meet to adjudicate their scores. Inter-rater reliability has improved over time as the rubric has improved. Average scores are calculated for the posters and each of the elements. The program sets minimum scores for successful average poster ratings (14/20 points) which reflects a satisfactory grade. Thus far, poster ratings have exceeded the cut-off, with an average score of 16.3/20 over four semesters of poster assessment (prior to the COVID-19 pandemic). This information is provided in the department's program assessment report, as required by the institution, and is reported to faculty who teach the research methods course so that they may adjust course content and teaching practices as necessary.

Considerations for CURE Implementation and Assessment

Undergraduate academic programs are different in their vision, mission, and student learning outcomes. Therefore, there can be no one-size-fits-all strategy for implementing a program-wide CURE. However, it is useful to identify the barriers and facilitators for the successful administration of these experiences so that programs may tailor practices to meet

their needs. In addition, given that there are limited examples of direct assessment of CURE, it may be useful to identify how these considerations may impact the assessment process.

As is the case in many universities, the Institutional Review Board (IRB) requires that all research projects with human or animal subjects be reviewed for a preliminary determination of review status (i.e., exempt, expedited, or full board review) (United States Department of Health and Human Services, 2021). Completing approximately 90 undergraduate research projects involving over 400 students annually requires pronounced efficiency of implementation. While nearly all projects typically meet the exempted or expedited review levels, the proposal is lengthy, requires specificity and advanced knowledge of terminology, can only be completed by one student in the group, and may impact the IRB turnaround time. Given the one-semester timeline constraint, the number of projects submitted simultaneously to the IRB as a result of a CURE may result in delayed feedback for some student groups given institutional capacity. To reduce the need for extensive edits, it is suggested that each faculty member assist in the revision process. Not surprisingly, inconsistency across IRB reviewers' comments and suggestions occurred within and across semesters, which created additional challenges in students receiving timely approval. Some student groups failed to grasp the importance of timely and thorough revisions, which delayed approval and limited the time available for data collection and analysis. As a result, it is suggested to have an open line of communication with one's IRB to facilitate this process, especially if the volume of applications will increase drastically. Furthermore, it is important to become familiar with the IRB review process at one's institution and determine whether a course-wide application is permitted and feasible.

The CURE is tied to a program level assessment; therefore, students must acquire certain skills from pre-requisite courses to be able to plan, propose, and conduct a research project within one semester.

The CURE is tied to a program level assessment; therefore, students must acquire certain skills from pre-requisite courses to be able to plan, propose, and conduct a research project within one semester. This is an important consideration in developing an assessment of a CURE at the program level. Adding or modifying content and skills within pre-requisite courses may require buy-in from all program faculty (Rawle et al., 2017). Depending on the class size of pre-requisite courses (e.g., ~ 45 students per section), fostering the development of writing skills may be a challenge. While the research methods course is offered as a three-credit course, it would more ideally be offered as a two-semester course sequence or for four credits with a lab component. If the course is not adequately resourced, faculty who teach the course will incur unofficial loads of work during office hours or additional one-on-one student/group meetings. It is important to ensure the pre-requisite skills are included in the early curriculum and to appropriately resource the CURE course to ensure high-quality mentoring from faculty. This helps to ensure the assessment of the CURE maps to the program curriculum and not just to the one course.

As a major without a secondary admissions process, gating option, or progression standards, there are significant differences in preparation, interest, and motivation to conduct research among students in the program. This may pose challenges in using the CURE as a program-level assessment if students do not have buy-in to the major and the need to understand research in this particular discipline. Additionally, students have a wide array of health topics that interest them, some of which are less adaptable to the one-semester timeline and available methodologies. As team-based projects, the differences, particularly in motivation, have resulted in tension between some students within groups that have necessitated intervention by the faculty member (Wallace & Walker, 2017). The team formation process is essential to the success and effectiveness of the team-based learning experience (Connerley & Mael, 2001) and offers the potential to prepare students to collaborate in diverse teams in their future careers (Lang et al., 2017). The students in the presented CURE are not assigned specific roles within the group. All work is completed cooperatively (with the exception of required individual research ethics training) and thus students must sometimes use conflict resolution skills such as communication and compromise. Naturally, some students within groups informally step-up into a leadership role by reminding others of due dates, reviewing all work for completeness and accuracy, or taking responsibility for turning work in on time.

There are several ways to address problematic team dynamics. For instance, the faculty member may have students complete a self-assessment regarding their individual strengths, weaknesses, and personality to assist in identifying partners who may work well together (Parmelee & Michaelsen, 2010; Steger et al., 2011). Randomly assigning students to teams (rather than self-assignment) has been found to positively impact group dynamics, attitudes toward the overall experience, and performance outcomes (Chapman et al., 2006; Parmelee & Michaelsen, 2010). An alternative approach is student-selected teams with significant instructor guidance in identifying necessary skills needed for an assignment and suggesting those students who may have the right fit of personality and talent (Steger et al., 2011). Such team creation can result in diversity of gender, age, function, culture, and ethnicity (Stahl et al., 2010; Troster et al., 2014; Watson et al., 2002).

Finally, it is important for instructors to recognize that sometimes not all students within a group will participate at the same level. For example, one group member may not complete their required work leaving it to others to complete or correct so that the group is not penalized during grading. As such, instructors may want to consider employing rules about appropriate group engagement. For instance, some instructors require that students submit an author contributions summary which professors use to determine which students did not contribute adequately and should be penalized.

Covid-19 Considerations

Two primary challenges to offering inclusive, rigorous HIP opportunities are both the resources to support writing-intensive courses as well as the student perception of the difficulty of such courses.

Several changes were made to the research methods course project in response to the barriers imposed by the COVID-19 pandemic. Because of university policies limiting in-person meetings, most research courses were administered with online instruction for the 2020-2021 academic year. IRB limited in-person human subjects data collection, necessitating that data be virtually collected. Rather than allow students to develop their own surveys and independently collect data (which was prohibited by IRB), faculty instructors modified the project by developing a common survey covering many health-related topics that all students in the research methods course, as well as students in other courses, completed for extra credit. Students then had access to this anonymous data on which to base the development of their topics, research questions, and analyses. All other research procedures remained the same (literature review, methodology write-up, analysis, interpretation). Though the symposium was canceled, students were still required to create a poster for their projects, allowing for the continued assessment of the relevant program objectives. This may be an option for programs to consider where traditional data collection may not be feasible.

Given the success of the most recent assessment within this program, two primary procedures will be maintained for future iterations of the assessment. First, student posters will be required to contain all necessary sections of the project (i.e., introduction/literature review, methodology, results, and discussion/conclusion). For instance, during previous poster assessments, some faculty members required students to include references while other faculty members did not impose this requirement. Consistency is key in ensuring an effective assessment of student learning outcomes (Gosselin & Golick, 2020; Summers, 2005). The second established procedure involves the use of a standard rubric in the evaluation of randomly assigned student posters from various sections of the research methods course, which again, is vital to ensuring consistency in poster assessment (Gosselin & Golick, 2020; Kishbaugh et al., 2012).

Two primary challenges to offering inclusive, rigorous HIP opportunities are both the resources to support writing-intensive courses as well as the student perception of the difficulty of such courses. Students frequently do not appreciate their experiences until they progress into their careers or graduate school. In addition, the potential negative impact such a course may have on student evaluations of teaching (SET) is yet another challenge (Vevere & Kozlinskis, 2011), particularly when college students typically do not enjoy working in teams with other students, largely due to collective grading and the perceptions of unequal distribution of effort (LaBeouf et al., 2016; Shimazoe & Aldrich, 2010). The intensity of

teaching effort, delayed student appreciation, and the potential impact on SETs warrant a further discussion about the benefits and challenges of offering an applied research course.

In terms of student appreciation of the course, data collected to support the program review (n=91) suggests a third of responding alumni (34%) listed the research course experience as one of the most meaningful educational experiences of their time in college. Asked specifically about the utility of the course via an online survey, approximately 50% of responding alumni indicated they had actively used skills developed in the course after graduation, 42% indicated improved information literacy (including understanding literature and the research process), 34% reported skill improvements they felt directly contributed to their success, and 29% indicated that the team skills helped them in their career and graduate school pursuits. The following are relevant reflective quotes from students pertaining to the course:

Course-based undergraduate research experiences provide numerous benefits to students including research, writing, and presentation skills.

> "*As a graduate student, I am beyond grateful for the experience I had in [research methods]. I feel far ahead of my classmates in my cohort who never had an experience of carrying out their own research project*."

> "*Research methods pushed me to look for a career outside of the typical health provider role that I was originally working towards, and I am very grateful for this exposure. The research project was extremely helpful and gave me a head start on my training once I was hired in clinical research*!"

To facilitate the potential immediate appreciation and application of the course, instructors frequently remind students of the utility of research skills in their future careers and health literacy. In addition, the instructors also developed a handout to guide students in listing research skills on their resumés to facilitate job-seeking opportunities. The handout contains language translating the project into skills that can be listed on a resumé, typical keywords to search for jobs that require research skills, and suggested common graduate school and job interview questions where the research project might be a suitable example.

Conclusion

Course-based undergraduate research experiences provide numerous benefits to students including research, writing, and presentation skills. In addition, CURE can positively impact students' view of the sciences and, therefore, increase interest in pursuing graduate education, especially among underrepresented student groups. However, there are many barriers to implementing program-wide CURE experiences, especially among high-volume departments. Further, many programs may have difficulty in developing direct assessments of learning for such courses. This paper discussed how one program implemented and assessed such an experience by focusing on assessment of the demonstration of specific research-related skills, rather than the subjective evaluation of the quality of the overall research project. Programs that wish to develop an assessment of CURE must understand that developing an assessment process and tools is an iterative process, which should include the collaboration of course instructors, department chairs, and assessment and evaluation experts, if available. A successful assessment of CURE may guide further development of course content and teaching strategies.

References

Allen, M. J. (2003). *Assessing academic programs in higher education*. John Wiley & Sons.

Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational leadership*, *57*(5), 13-19.

Arter, J. A., & Chappuis, J. (2007). *Creating & recognizing quality rubrics*. Pearson Merrill Prentice-Hall.

Auchincloss, L. C., Lauresen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., Lawrie, G., McLinn, C. M., Pelaez, N., Rowland, S., Towns, M., Trautmann, N. M., Varma-Nelson, P., Weston, T. J., & Dolan, E. L. (2014). Assessment of course-based undergraduate research experiences: A meeting report. *CBE—Life Sciences* Education, 13, 29-40. https://doi.org/10.1187/cbe.14-01-0004

Bangera, G., & Brownell, S. E. (2014). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE—Life Sciences Education*, *13*, 602-606. https://doi.org/10.1187/cbe.14-06-0099

Bloom, B. S. (Ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives*: *The classification of educational goals*. David McKay Publications.

Chapman, K. J., Meuter, M., Toy, D., & Wright, L. (2006). Can't we pick our own groups? The influence of group selection method on group dynamics and outcomes. *Journal of Management Education*, *30*(4), 557-569. https://doi.org/10.1177/1052562905284872

Collins, T. W., Grineski, S. E., Shenberger, J., Morales, X., Morera, O. F., & Echegoyen, L. E. -(2017). Undergraduate research participation is associated with improved student outcomes at a Hispanic-serving institution. *Journal of College Student Development*, *58*(4), 583-600. doi:10.1353/csd.2017.0044

Connerley, M. L., & Mael, F. A. (2001). The importance and invasiveness of student team selection criteria. *Journal of Management Education*, *25*(5), 471-494. https://doi.org/10.1177/105256290102500502

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

Gosselin, D. C., & Golick, D. (2020). Posters as an effective assessment tool for a capstone course. *Journal of Environmental Studies and Sciences*, *10*, 426-437. https://doi.org/10.1007/s13412-020-00612-x

/cbe.Hanauer, D. I., Graham, M. J., & Hatfull, G. F. (2016). A measure of college student -persistence in the sciences (PITS). *CBE—Life Sciences Education*, *15*(4), ar54. https://doi.org/10.1187/cbe.15-09-0185

Huitt, W. (2011). Bloom et al.'s taxonomy of the cognitive domain. *Educational Psychology Interactive*. Valdosta State University. http://www.edpsycinteractive.org/topics/cognition/bloom.html

Hunter, A., Laursen, S. L., & Seymour, E. (2007). Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development. *Science Education*, *91*, 36-74. https://doi.org/10.1002/sce.20173

James Madison University. (2021). JMU plans: Mission, vision, and values. https://www.jmu.edu/jmuplans/mission-vision-values.shtml

Kinkead, J. (2003). Learning through inquiry: An overview of undergraduate research. *New Directions for Teaching and Learning*, *93*, 5-17. https://doi.org/pdf/10.1002/tl.85

Kinzie, J. (2012). High-impact practices: Promoting participation for all students. *Diversity & Democracy*, *15*(3).

Kishbaugh, T. L. S., Cessna, S., Horst, S. J., Leaman, L., Flanagan, T., Graber Neufeld, D., & Siderhurst, M. (2012). Measuring beyond content: A rubric bank for assessing skills in authentic research assignments in the sciences. *Chemistry Education Research and Practice*, *13*, 268-276. https://doi.org/10.1039/C2RP00023G

Kuh, G. D., & Association of American Colleges & Universities. (2008). High-impact educational practices: What they are, who has access to them, and why they matter. Association of American Colleges & Universities. https://www.aacu.org/node/4084

Kuh, G. D., & Ewell, P. T. (2010). The state of learning outcomes assessment in the United States. *Higher education management and policy*, *22*(1), 1-20. http://dx.doi.org/10.1787/hemp-22-5ks5dlhqbfr1

LaBeouf, J. P., Griffith, J. C., & Roberts, D. L. (2016). Faculty and student issues with group work: What is problematic with college group assignments and why? *Journal of Education and Human Development*, *5*(1). https://doi.org/10.15640/jehd.v5n1a2

Lang, D. L., Reisinger Walker, E., Steiner, R. J., & Woodruff, R. C. (2017). Implementation and mixed-methods evaluation of team-based learning in a graduate public health research methods course. *Pedagogy in Health Promotion*, *4*(2), 140-150. https://doi.org/10.1177/2373379917707222

Linn, M. C., Palmer, E., Baranger, A., Gerard, E., & Stone, E. (2015). Undergraduate research experiences: Impacts and opportunities. *Science*, *347*, 1-8. http://doi.org/10.1126/science.1261757

Lopatto, D. (2010). Undergraduate research as a high-impact student experience. *Peer Review*, *12*(2). Retrieved from https://www.aacu.org/publications-research/periodicals/undergraduate-research-high-impact-student-experience

Moskal, B. M. (2002). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment*, *Research*, *and Evaluation*, *8*(1), 14. https://doi.org/10.7275/jz85-rj16

National Survey of Student Engagement (NSSE). (2007). Experiences that matter: Enhancing student learning and success-annual report 2007. Indiana University Center for Postsecondary Research. https://nsse.indiana.edu/research/annual-results/past-annual-results/nsse-annual-report-2007.html

National Survey of Student Engagement (NSSE). (2019). Engagement insights: Survey findings on the quality of undergraduate education-annual results 2019. Indiana University Center for Postsecondary Research. https://scholarworks.iu.edu/dspace/bitstream/handle/2022/25321/NSSE_2019_Annual_Results.pdf?sequence=1&isAllowed=y

Parmelee, D. X., & Michaelsen, L. K. (2010). Twelve tips for doing effective team-based learning (TBL). *Medical Teacher*, *32*(2), 118-122. https://doi.org/10.3109/01421590903548562

Peachey, A. A., & Baller, S. L. (2015). Ideas and approaches for teaching undergraduate research methods in the health sciences. *International Journal of Teaching and Learning in Higher Education*, *27*(3), 434-442.

Rawle, F., Bowen, T., Murck, B., & Hong, R. (2017). Curriculum mapping across the disciplines: Differences, approaches, and strategies. *Empowering Learners*, *Effecting Change*, *10*, https://doi.org/10.22329/celt.v10i0.4765

Russell, S. H., Hancock, M. P., & McCullough, J. (2007). Benefits of undergraduate research experiences. *Science*, *316*, 548-549. http://doi.org/10.1112/science.1140384

Shimazoe, J., & Aldrich, H. (2010). Group work can be gratifying: Understanding & overcoming resistance to cooperative learning. *College Teaching*, *58*(2), 52-57. https://doi.org/10.1080/87567550903418594

Shortlidge, E. E., & Brownell, S. E. (2016). How to assess your CURE: A practical guide for instructors of course-based undergraduate research experiences. *Journal of microbiology & biology education*, *17*(3), 399-408. https://dx.doi.org/10.1128%2Fjmbe.v17i3.1103

Stahl, G. K., Makela, K., Zander, L., & Maznevski, M. L. (2010). A look at the bright side of multicultural team diversity. *Scandinavian Journal of Management*, *26*(2010), 439-447. https://doi.org/10.1016/j.scaman.2010.09.009

Steger, R. A., Mankin, J. A, & Jewell, J. J. (2011). How to organize a real life problem-based learning project in a business class using strength assessment to determine team assignment. *Journal of Higher Education Theory and Practice*, *11*(1), 45-55.

Stiggins, R. J. (2001). *Student-involved classroom assessment (3rd ed)*. Prentice-Hall.

Summers, K. (2005). Student assessment using poster presentations. *Paediatric Nursing*, *17*(8), 24-26. https://doi.org/10.7748/paed2005.10.17.8.24.c1008

Suskie, L. (2018). *Assessing student learning*: *A common sense guide*. John Wiley & Sons.

Troster, C., Mehra, A., & van Knippenberg, D. (2014). Structuring for team success: The interactive effects of network structure and cultural diversity on team potency and performance. *Organizational Behavior and Human Decision Processes*, *124*(2014), 245-255. https://doi.org/10.1016/j.obhdp.2014.04.003

United States Department of Health and Human Services. (2021). Regulations, Policy & Guidance. *Office for Human Research Protections*. https://www.hhs.gov/ohrp/regulations-and-policy/index.html

Vevere, N., & Kozlinkskis, V. (2011). Students' evaluation of teaching quality. *US-China Education Review*, *B5,* 702-708.

Wallace, V., & Walker, L. (2017). Team-based learning. In G. Kayingo, & V. McCoy Hass (Eds.), *The health professions educator*: *A practical guide for new and established faculty* (pp. 67-78). Springer Publishing.

Watson, W. E., Johnson, L., & Zgourides, G. D. (2002). The influence of ethnic diversity on leadership, group process, and performance: An examination of learning teams. *International Journal of Intercultural Relations*, 26(1), 1-16. http://dx.doi.org/10.1016/S0147-1767(01)00032-3

Weston, T. J., & Laursen, S. L. (2015). The undergraduate research student self-assessment (URSSA): Validation for use in program evaluation. *CBE—Life Sciences Education*, *14*(3), ar33. https://doi.org/10.1187/cbe.14-11-0206

Abstract

In the United States, an increasing number of teacher education programs are using coteaching as a model for student teaching. Coteaching occurs when teacher candidates work collaboratively with their clinical educator(s) to share responsibility for students' learning, develop teaching practices and skills, and coevaluate instruction. Currently, there are no psychometrically validated instruments that assess teacher candidates' and clinical educators' coteaching experiences. This study documents the development and validation of a coteaching instrument that used thematic content analysis and Confirmatory Factor Analysis (CFA) to identify eight subscales. The subscales are *Equality in the Classroom*, *Learning Opportunities for Students*, *Connecting Theory to Practice*, *Coteacher Collaboration*, *Professional Development*, *Personal Pedagogical Skill Development*, *Types of Teaching*, and *General Coteaching Practices*. The results of this study demonstrate that the coteaching survey is a valid and reliable instrument to measure perspectives and experiences of coteaching across a variety of research settings.

AUTHORS

Andrea Drewes, Ph.D.
Rider University

Kathryn Scantlebury, Ph.D.
University of Deleware

Elizabeth Soslau, Ph.D.
University of Deleware

Evaluating Coteaching as a Model for Pre-Service Teacher Preparation: Developing an Instrument Utilizing Mixed Methods

Teacher education programs incorporate a variety of field experiences to expose pre-service teachers to the nuances and challenges of teaching. Field experiences are a mainstay in teacher education programs and re-emphasize the importance and value of student teaching as the "practice turn" (Cochran-Smith et al., 2016). Traditional models of student teaching have three common characteristics to support teacher candidates'[1] learning: (1) observation of clinical educator; (2) feedback from clinical educators and university field instructors; and (3) teacher candidates' reflection on practice. However, research suggests that a lack of collaboration limits teacher candidates' opportunities for reflection on learning and can often result in the mimicking of teaching practices without developing an understanding of the pedagogical reasonings underpinning teacher decision-making (Drewes et al., 2021; Soslau, 2012; Soslau et al., 2018).

Increasingly, teacher education programs are implementing a coteaching model for student teaching (Bacharach et al., 2010; Drouin et al., 2020; Strobaugh & Everson, 2019). Coteaching occurs when teacher candidates work collaboratively with their clinical educator to coplan, coteach, and coevaluate their instruction. These actions are employed to reduce the theory to practice gap, share responsibility for students' learning (Soslau et al., 2018), and develop teaching practices and skills (Gallo-Fox & Scantlebury, 2016; Murphy & Martin, 2015).

[1] Teacher candidate refers to students enrolled in teacher education programs and clinical educators are teachers who host teacher candidates during field experiences.

CORRESPONDENCE

Email
adrewes@rider.edu

There are currently no psychometrically rigorous, validated survey instruments to collect and report on teacher candidates' and clinical educators' perceptions of their coteaching classroom experiences.

Several qualitative studies have documented the positive outcomes of the coteaching model for all stakeholders–students, teacher candidates, and clinical educators alike. Meaningful learning outcomes for students include improved student achievement (Bacharach et al., 2010) and attitudes (Murphy et al., 2004). Coteaching models also provide professional development for clinical educators (Milne et al., 2011; Gallo-Fox & Scantlebury, 2016) and have been shown to encourage a willingness in beginning teachers to seek collaborative professional relationships (Murphy & Scantlebury, 2010).

The increasing number of teacher education preparation programs using coteaching as a model for student teaching has expanded the quantitative options for studying this model (Drouin et al., 2020; Guise et al., 2017; Strobaugh & Everson, 2019). There is also a need for reliable and valid instruments to gather information from various stakeholders within the coteaching relationship (Drewes et al., 2020). Yet, there are currently no psychometrically rigorous, validated survey instruments to collect and report on teacher candidates' and clinical educators' perceptions of their coteaching classroom experiences. This poses a problem as accreditation for teacher education programs from U.S.-based organizations, such as the Council for the Accrediation of Educator Preparation (CAEP), require the use of statistically reliable and valid instruments to assess teacher candidate preparation and performance (CAEP, 2013). Beyond a rationale related to the ubiquitous need for teacher preparation programs to attend to accreditation requirements, the ability to assess the usefulness of coteaching is of critical importance to any program improvement efforts. As teacher preparation programs across the United States take up the Blue Ribbon Panel report's recommendation to implement coteaching models (NCATE, 2010), teacher education researchers are apt to require assessment tools to evaluate program improvement efforts and study the quality of coteaching initiatives. The aim of our work was to address this critical need. More specifically, this study's goal was to develop a reliable and valid instrument to help teacher education program administrators, education researchers, and other stakeholders ascertain teacher candidates' (student teachers) and clinical educators' (mentor teachers) perceptions of the implementation, effectiveness, and use of coteaching.

Research Context

The research participants for this study were enrolled in one of three teacher preparation programs across two colleges within the same university located on the mid-Atlantic coast of the United States. The programs included teacher candidates studying to earn certifications in early childhood, elementary teacher education, special education, middle grades content areas, and secondary science. Candidates all sought a four-year undergraduate degree, completed full time student teaching for at least one semester (15 weeks), were aged 18-21, and the majority were female. Coteaching was used as the model for student teaching across all three programs and the researchers and teacher educators were interested in learning more about the experiences of coteachers. All three programs were also in need of a valid and reliable instrument to collect data for accreditation approval as evidence of continuous program improvement efforts. Though the study took place in the United States, the work is applicable in an international context as coteaching is gaining popularity across the globe as a viable approach for clinical practice.

Coteaching as a model for student teaching allows for teacher candidates and clinical educators to share responsibility for all aspects of student (pupil) learning including instructional planning, teaching, assessment, and evaluation (Martin, 2009). Both teacher candidates and experienced teachers share expertise in content and pedagogy as they coplan, coteach, and coevaluate student learning and their professional practice (Soslau et al., 2018). Coteaching experiences also offer a number of avenues for teacher candidates and clinical educators to improve the classroom learning environment through the equality of teacher voices, increased learning opportunities for students, occasions for teacher collaborations, connecting educational theory to practice, varied avenues for professional development, and opportunities to employ a diverse array of instructional approaches. These outcomes framed the survey development. The following section expands upon each of these outcomes as the foci of the eight scales developed in this survey.

Equality of Voices in the Classroom is evident when coteachers share ideas, demonstrate mutual respect, view each other as colleagues, take coresponsibility for student learning, and share authority in the classroom. The four items on this scale were drawn from the literature supporting teacher learning during student teaching, particularly during conditions when clinical educators and teacher candidates work together in the same classroom. Clinical educators have more power in the student teaching practicum site (Anderson, 2007) and teacher candidates generally assume additional responsibility only during later stages of the practicum (Hoy & Woolfolk, 1990). However, research has shown that more equal power distribution increased teachers' opportunities for learning (Nguyen, 2009; Smith, 2007; Zeichner, 1992). Coteaching is a reduced hierarchical model of student teaching, emphasizing the sharing of power and responsibility along with respecting and valuing all teachers' voices (Drewes et al., 2021; Scantlebury et al., 2008). Because candidates are expected to assume power and responsibility immediately, and clinical educators share control of all classroom aspects related to planning, instruction, management, and assessment from the first day of student teaching, this scale examines participants' perceptions on sharing power and input on teaching decisions and practices.

Student Learning Opportunities items are built upon studies that examined benefits to students in classrooms with two teachers, including a teacher candidate and clinical educator or two classroom teachers such as in a special education setting, or in a classroom with a high population of English language learners. Though some lament the lack of empirical data related to student outcomes and perspectives (Drewes et al., 2020; Friend et al., 2010), researchers of the coteaching model have begun to show that coteaching leads to increased learning opportunities for students (Bacharach et al., 2010; Badiali & Titus, 2010; Dove & Honigsfeld, 2010), more effective student learning (Rice & Zigmond, 2000), increases in students' positive attitudes toward science (Murphy & Beggs, 2010), increased access to help for students (Magiera et al., 2005), and increase of students' exposure to a variety of instructional approaches (Kamens, 2007).

Coteaching experiences also offer a number of avenues for teacher candidates and clinical educators to improve the classroom learning environment through the equality of teacher voices, increased learning opportunities for students, occasions for teacher collaborations, connecting educational theory to practice, varied avenues for professional development, and opportunities to employ a diverse array of instructional approaches.

Connecting Theory to Practice and *Coteacher Collaboration* are the next two scales and are interconnected because teachers' learning opportunities are the result of strong teacher-to-teacher collaboration. There are several benefits of clinical educators and teacher candidates working together closely throughout the practicum experience. Since the coteaching model requires coteachers to engage in discussions of practice and to develop justifications for their instructional decisions about a shared teaching experience, candidates and clinical educators have opportunities to make theory to practice connections in the conversations (Soslau et al., 2018).

Research has shown that when justifications and rationales are shared between clinical educators and teacher candidates, not only are these connections possible, but developing a shared understanding of the characteristics of effective teaching is more likely (Soslau, 2012; Zeichner, 2010). Research around collaboration between teacher candidates and clinical educators shows that a strong relationship via mutual respect is critical to enabling coteachers to collaborate and to resolve instructional problems such as issues related to classroom management, student motivation, and interactions with parents (Austin, 2001; Parsons & Stephenson, 2005; Phelan et al., 1996).

Professional Development and *Personal Pedagogical Skill Development*, the next two scales, focused on whether teachers were aware of and receptive to these opportunities for personal growth and if they viewed planning and teaching episodes as sites for their own professional development. Proponents of coteaching often cite the coteaching experience as a form of professional development for clinical educators (Bacharach et al., 2010; Gallo-Fox & Scantlebury, 2016). The idea that clinical educators improve their own professional practice when hosting a candidate is not new (Koskela & Ganser, 1999; Landt, 2004); however, coteaching deliberately places both teachers in the role of learner. Shifting the roles of the clinical educator from mentor to colearner of teaching provides opportunities for growth (Gallo-Fox & Scantlebury, 2016).

Personal Pedagogical Skill Development scale included small grain size skills such as learning formative assessment techniques or building understandings of how to integrate

literacy in the classroom. Similarly, we were interested to learn if participants perceived the experience as good preparation for candidates' future practice as independent practitioners. Opponents of coteaching may claim that candidates do not have enough opportunities for independent practice. Yet, many coteaching approaches provide for lead roles which likely reflect similar independent practice conditions in more traditional student teaching models (Gallo-Fox et al., 2006).

Types of Teaching and General Coteaching Practices are the final two scales on the survey and they focused on the approaches used during the student teaching experience, such as whether or not teacher candidates completed independent practice, led instruction, engaged in "stepping up" and "stepping back" (Tobin & Roth, 2006) during coinstruction, and took active roles across coplanning, coteaching, and coevaluation. These oft-cited components of coteaching are hallmarks of successful partnerships (Bacharach et al., 2010; Scantlebury et al., 2008; Soslau et al., 2018). By examining the frequency of specific types of teaching practice (Types of Teaching) and the prevalence of coteaching activities beyond instruction (General Coteaching Practices), researchers may be able to better describe coteaching contexts, their efficacy, and areas for improvement.

Method

Participants

Survey respondents were clinical educators and teacher candidates who recently participated in a coteaching student teaching placement. The clinical educators represented teachers from all grade levels from infants to high school and the teacher candidates accordingly were placed in a diverse collection of school settings from early childhood through secondary levels. In the initial pilot, 147 responses were collected and for the testing of the revised survey instrument, 590 responses were collected over the following four semesters. Further details on the background information on these survey respondents will be presented in the results section.

Data Collection

The research team, comprised of university personnel, clinical educators, and teacher candidates, sought to develop, to empirically validate, and to implement a widely applicable survey to measure clinical educator and teacher candidate beliefs regarding their experiences with coteaching.

This study was completed within a larger on-going research study on the impact of coteaching on teacher candidates, clinical educators, and their students. These pilot and development phases were encompassed under established Institutional Review Board protocols. Participants were recruited to complete this survey via email at the completion of a coteaching placement. Completion of the survey was optional and therefore this sample may suffer from volunteer bias. Additionally, no extra credit nor compensation was offered due to the anonymous nature of the online survey format.

Measure Development Process

Our initial research interests were to explore what avenues were available for clinical educators and teacher candidates to provide feedback on the overall coteaching program and also to share their experiences and the relevant impacts on their teaching that resulted from the coteaching model. Supported by a collection of research literature and anecdotally from years, actually decades, of experience with this coteaching model on a small scale, the research team knew that clinical educators often reported benefits like integration of new pedagogies, innovation in instructional techniques and management, improved self-efficacy toward their own teaching practice, and increases in student learning outcomes. Our research team had numerous discussions early in the process regarding the best way to explore these two-way interactions between clinical educators and teacher candidates: how the clinical educator impacted the teacher candidate's professional expertise and also how candidate's presence influenced the clinical educator's pedagogical practice as well. We sought a research approach that would best reflect the dialogic nature of the coteaching model for teacher preparation.

From this research impetus, the research team, comprised of university personnel, clinical educators, and teacher candidates, sought to develop, to empirically validate, and to implement a widely applicable survey to measure clinical educator and teacher candidate

beliefs regarding their experiences with coteaching. The goal of this type of survey was to provide feedback from both perspectives which would be used to modify the overall coteaching model for teacher preparation at the university. Additionally, the survey offered teacher candidate and clinical educator respondents an opportunity to reflect on their own development as a teacher and also as a teacher educator for clinical educator respondents. In sum, the driving force behind the development and validation of this survey is that our research team wanted to develop a better conceptual understanding of what coteaching looks like from the perspectives of the participants.

This study focuses on the survey development process using both qualitative thematic analysis and quantitative CFA methods over two phases: Phase 1 Pilot Instrument Analysis and Phase 2 CFA.

The survey development process was influenced by the meta-framework presented by Onwuegbuzie et al. (2010) for a mixed methods development process and the four-step procedure established for developing and validating measures (Crocker & Algina, 1986; Sax, 1997). These frameworks guided our approach to the creation of possible survey items, testing, and refinement of these items, all while utilizing cyclic qualitative and quantitative approaches to develop a measure that would allow the clinical educators and teacher candidates to reflect on the coteaching experience as part of the path to their own professional development. While taking this approach, our research team sought to describe accurately this experiential learning setting for clinical educators and teacher candidates and in doing so, we worked to develop this survey to operationalize the practices and outcomes that we could expect these stakeholders to experience across a variety of coteaching settings.

In the first cycle of item development, our research team qualitatively reviewed existing surveys related to student teaching already in use by our home institution and other similarly purposed surveys from other institutions. In the next phase of literature review [Step 1 in Table 1], the lead author collected and reviewed numerous coteaching related articles through a literature review to create an initial list of 88 survey items for clinical educators and 73 items for teacher candidates. This initial collection was then also qualitatively reviewed by the research team for face validity and content validity based on their collective expertise in this research field [Step 2 in Table 1]. All survey items were structured as Likert-type responses with 5-point scale with a not applicable or unclear option. Additionally, after each grouping of eight to ten items, there were open response spaces included in the online survey with a prompt to encourage coteachers to indicate unclear items or provide additional comment if desired. Coteachers rarely used open response spaces; but when used, the comments provided useful insight to specific circumstances within the coteaching placement. These items were written to encompass the prevalent themes of the body of coteaching related research described previously. With full realization of the negative impact on responses from such a long survey (Galesic & Bosnjak, 2009; Schwarz et al., 1998), we undertook the next round of analytic review to pare down these items to devise an effective and practical survey instrument.

Overview of Analytic Approach

Factor analysis is a category of statistical techniques that examine patterns of variance and correlation (covariance) within participant responses on a survey instrument. Exploratory factor analysis (EFA) starts with all items and works to uncover related latent variables and to group these items into subsets based on participants' patterns of responses. The main goal of EFA is to identify these sets of items and does not base the organization of survey items to relevant theory. Confirmatory factor analysis (CFA) emphasizes the testing of hypothetical groupings of items based on an appropriate theoretical framework to determine how well patterns of responses fit with the proposed model. Since prior research in coteaching as a teacher preparation model was used to develop an a *priori* framework to classify survey items in the pilot phase, CFA is most appropriate to employ for these theory testing survey development efforts (Stevens, 1996). As such, this study focuses on the survey development process using both qualitative thematic analysis and quantitative CFA methods over two phases: Phase 1 Pilot Instrument Analysis and Phase 2 CFA. See Table 1 for an overview of the development and analytic steps.

Table 1
Overview of Development and Analysis Procedures

Initial Survey Development Phase	**Step 1:** Qualitatively focused item writing based on themes present in literature review of coteaching studies
	Step 2: Qualitatively focused item review by expert panel and research team for face and content validity
Analysis Phase 1: Pilot Instrument Analysis	**Step 3:** Pilot data collected from one semester of coteaching placements
	Step 4: Quantitively driven item analysis of pilot data
	Step 5: Qualitatively focused content analysis of items by research experts
	Step 6: Qualitatively focused and consensus driven thematic analysis by research team to create subscales
Analysis Phase 2: Factor Analysis of Revised Survey Instrument	**Step 7:** Data collected from four additional semesters of coteaching placements
	Step 8: Quantitative item analysis of full data set
	Step 9: Quantitatively driven CFA to investigate and confirm construct validity
	Step 10: Qualitatively focused final review by research team to confirm content validity

Overview of Analytic Approach in Phase 1. As the pilot sample from the first round of data collection was too small for traditional EFA approaches (MacCullum et al., 1999), we conducted an item analysis of this pilot data [Step 4 in Table 1]. We examined the correlation matrix for items to remove that had numerous very low (<.4) or many very high correlations (>.8) with other items (Field, 2013). Additionally, we reviewed the item-total correlations and identified items for removal that were also very low (<.4) (Field, 2013). Next, utilizing the open response feedback and the research team's professional expertise in coteaching, the remaining items were reviewed to ensure there were no further items that were redundant or unclear in meaning to the survey respondents [Step 5 in Table 1]. These extraneous items were removed. Lastly, the research team, along with additional experienced clinical educators and teacher candidates, evaluated the remaining items qualitatively with a thematic analysis of the content of each item. Through a consensus driven approach, items were categorized to create hypothesized subscales [Step 6 in Table 1]. Cronbach's alpha was calculated for the hypothesized subscales to determine the initial reliability. A final draft of the survey was then employed in the second phase of this development and validation project.

Overview of Analytic Approach in Phase 2. After the preliminary mixed methods analysis of the pilot survey, the revised draft of the survey was used for data collection [Step 7 in Table 1]. There is no one rule for acceptable or minimum sample sizes to conduct a confirmatory factor analysis (MacCullum et al., 1999). The acquired sample size met the wide array of diverse guidelines for sample sizes for CFA, including the absolute sample size (DiStefano & Hess, 2005), ratio of sample size and number of items (N/p; Benson & Nasser, 1998), ratio of number of items to factor (p/f; Marsh et al., 1998), evaluation of factor loading values (Wolf et al., 2016), calculations of maximal reliability and construct validity (ω and H; Gagne & Hancock, 2006). The sample size for the current CFA met all of the aforementioned guidelines and was deemed appropriate.

The first step of the next round of analysis was to examine again the correlation matrix [Step 8 in Table 1] for the items with too low or too high correlations (Field, 2013). Next, the statistical software package AMOS was used to represent the model graphically with each of the six hypothesized latent variables being illustrated in a 6-factor model [Step 9 in Table 1]. This CFA presents a Chi squared statistic for determining model fit. However, numerous researchers have determined that solely judging a CFA model by the Chi squared statistic is problematic (Brown, 2006; Hu & Bentler, 1999). Instead, we reviewed several fit statistics to determine how well the factor model explains the observed data.

Bentler (1994) and Thompson (2004) identified a problem with only interpreting one model fit index and instead support the evaluation of multiple indices to gain a more in-depth understanding of the overall model fit. Other model fit indices were also examined as there are established problems with interpreting the χ^2 statistic (Dickey, 1996; Schumacker & Lomax, 1996; Stevens, 1996) as it can be strongly influenced by sample size. The Normative Fit Index (NFI) and Comparative Fit Index (CFI) are less likely to be influenced by sample size (Hu & Bentler, 1999). Root mean square error of approximation, or RMSEA, is another recommended index to indicate good model fit (Arbuckle, 2005; Fan et al., 1999). These fit indices guidelines are summarized in Table 2. Lastly, to determine the internal reliability of each of the examined subscales from the CFA model, the mean, standard deviation, and Cronbach's alpha for each subscale were calculated.

Table 2
Fit Indices Guidelines for Confirmatory Factor Analysis

Fit Index	Guidelines
χ^2 p value	> .05
CMIN/DF	< 5.0
NFI	> .90
CFI	> .90
RMSEA	< .10
SRMR	< .08
AGFI	> .90
PCFI	> .50

Table 3
Certifications Held or Pursued for All Survey Respondents

Certification Pursued or Held by Coteacher	Percent of Pilot Sample (Phase 1)	Percent of Validation Sample (Phase 2)
Early Childhood	19.7%	19.7%
Elementary School	86.4%	81.7%
Middle Grades	27.9%	44.2%
Secondary Grades	2.7%	7.8%
Sample Size	147	590

Results

Pilot Data Analysis

Analysis Phase 1. We conducted the initial pilot study of the survey with teacher candidates and clinical educators participating in coteaching experiences in the Fall, 2014 semester [Step 3 in Table 1]. We collected electronic surveys from 60 teacher candidates and

87 clinical educators. The teaching certifications held by clinical educators or being pursued by teacher candidates were largely elementary grade levels, which is reflective of the teacher preparation program at the university research site. The percentages of each certification type are presented in Table 3. Clinical educators may hold and teacher candidates may pursue more than one certification so the percentages do not total 100%. Additionally, all the clinical educators reported holding a graduate degree and had between one and over 21 years of experience teaching. Further demographic data, such as years teaching for clinical educators and teacher candidates' program affiliations, is presented in Tables 4 and 5.

This phase 1 pilot data was first analyzed via item analysis of the means, standard deviations, and item correlations of the 73 parallel items for clinical educators and teacher candidates and the 15 additional items only presented to clinical educators. This first quantitative review [Step 4 in Table 1] identified 30 items that met the guidelines established for removal as described in the methods section (e.g., very low item-total correlation, <.4). Next, the research team reviewed the open response sections for items that the survey respondents indicated were unclear. The research team worked to edit these items to improve the clarity and readability or decided to remove the item due to redundancy. From this qualitative content review [Step 5 in Table 1], we deleted 19 additional items and rewrote four other items.

Table 4
Years of Teaching Experience in Clinical Educator Survey Respondents

Years of Teaching Experience	Percent of Pilot Sample (Phase 1)	Percent of Validation Sample (Phase 2)
1-5 years	5.4%	5.9%
6-10 years	19.0%	16.6%
11-15 years	9.5%	11.5%
16-20 years	15.6%	8.1%
21 years or more	9.5%	9.5%
Sample Size	87	306

Table 5
Program Affiliations of Teacher Candidate Survey Respondents

Teacher Preparation Program Affiliation	Percent of Pilot Sample (Phase 1)	Percent of Validation Sample (Phase 2)
Early Childhood Education	13.3%	17.3%
Elementary Teacher Education	86.4%	80.6%
Secondary Science Education	0.0%	2.1%
Sample Size	60	284

Lastly, during this first analytic phase of the project

[Step 6 in Table 1], the research team thematically evaluated the remaining 40 items that appear on both the clinical educator and teacher candidate parallel versions and the three additional items that were only presented to clinical educators. Using a consensus building approach among experts, we grouped items that referred to similar content or theory to devise eight hypothetical or proposed subscales of the survey. The first six subscales were thematically grouped by relevant research topics. The seventh subscale is made up of items that relate to the various types of teaching approaches that can occur during coteaching (i.e., stepping up versus stepping back during instruction; or solo teaching compared to assisting instruction or coteaching). This group of items is purposefully diverse in nature to understand better the frequency of use for these different teaching approaches across coteaching settings. The eighth subscale is comprised of items that ask the respondent to reflect more generally on the coteaching experience and its primary components. Again, this last group of items is a purposefully diverse collection. Due to the intentionally broad scope of the last two subscales, we did not include these items in the next phase of analysis as their underlying group variable will not be represented by a theoretically driven latent variable in the factor model and will be described as survey subsections moving forward to delineate from the first six instrument subscales.

After this sequential, mixed methods analytic review, the revised survey instrument was comprised of 40 parallel items for teacher candidates and clinical educators and three additional items only for clinical educators across eight subscales.

After this sequential, mixed methods analytic review, the revised survey instrument was comprised of 40 parallel items for teacher candidates and clinical educators and three additional items only for clinical educators across eight subscales. The subscales included the following topical collections of items: *Equality in the Classroom*; *Learning Opportunities for Students*; *Connecting Theory to Practice*; *Coteacher Collaboration*; *Professional Development*; *Personal Pedagogical Skill Development*; *Types of Teaching*; and *General Coteaching Practices*. Each subscale has between four and six items. Sample items from each subscale are shown in Table 6. [Authors' Note: Researchers interested in deploying this coteaching survey in research settings should contact the first author for a copy of the entire survey instrument.] Internal reliabilities of each of the eight subscales are presented in Table 7.

Confirmatory Factor Analysis

Analysis Phase 2. During the subsequent rounds of data collection over the following four semesters of coteaching placements [Step 7 in Table 1], we sought to collect enough survey responses to have a robust sample for CFA to validate the model of the proposed subscales, or factors, devised in Analysis Phase 1. Employing the 40 parallel items plus three additional items version of the survey, we gathered responses from 284 teacher candidates and 306 clinical educators for a total of 590 responses, satisfying the recommended sample size for CFA per the various guidelines described earlier. Demographic details of the Analysis Phase 2 sample are found in Tables 4 and 5.

Subsections 7 and 8 (*Types of Teaching and General Coteaching Practices*) are purposefully diverse for the collection of more logistical data related to the frequency of particular teaching approaches and more general reflections on the coteaching experiences, and as such, are not included for the following item analysis and CFA. As in Analysis Phase 1, we first reviewed the Analysis Phase 2 data via item analysis, especially the item-total correlations [Step 8 in Table 1]. No items were identified for possible removal using the established guidelines. See Table 8 for all item-total correlations from Analysis Phase 2.

Finally, a CFA was conducted on the full data set [Step 9 in Table 1]. The initial model fit indices showcase poor fit for the data across many indices (N=590; $\chi^2(390)$ = 2776.7; p = .001; CMIN/DF = 7.12; normed fit index (NFI) = .808; comparative fit index (CFI) = .830; root mean square error approximation (RMSEA) = .102 (90 percent confidence intervals of .098 and .106); SRMR = .0765; AGFI = .668; PCFI = .744. These initial findings are summarized in Table 9. To improve the model fit, correlations were added between several items on the same subscale (e.g., between item 1 & 6; between 15 & 17; and between 20 & 21) for a total of 19 correlations allowed according to the modification indices from AMOS (Kline, 2005). The model fit improved and the fit guidelines were deemed acceptable to great for all indices.

Table 6
Sample Items of Survey by Subscale

Subscale #1 – Equality in the Classroom
7. A mutual sense of respect was developed between my coteacher and me.
15. I viewed my coteacher as a colleague.
8. My coteacher and I developed a coresponsibility for meeting our students' needs.
Subscale #2 – Learning Opportunities for Students
1. Coteaching provided more opportunities for students to learn.
6. Coteaching helps the students learn content more effectively.
18. Coteaching allowed the students to get the help they needed.
Subscale #3 – Connecting Theory to Practice
5. Coteaching allowed me to link educational theory to practice.
31. We discussed what we learned about ourselves and our teaching practice.
32. We shared the reasons behind instructional decisions.
Subscale #4 – Coteacher Collaboration on classroom issues
10. My coteacher and I discussed issues that impacted our teaching.
29. We decided together to change upcoming lessons because they weren't working as desired.
30. We collaborated to determine student needs.
Subscale #5 – Professional Development
2. Coteaching provided opportunities for my coteacher to grow as a teacher.
3. Coteaching provided opportunities for me to grow as a teacher.
16. My coteacher provided insight and knowledge that improved my own teaching.
Subscale #6 -- Personal Pedagogical Skill Development (specific)
36. The coteaching experience showed me new ways to integrate literacy into my classroom.
39. I improved my understanding of how to utilize technology in my classroom.
40. Coteaching has shown me new ways to build student engagement.
Purposefully diverse sections:
Subsection #7 – Types of Teaching
22. I solo taught.
34. I stepped up to take the lead instructional position.
35. I stepped back to take a supportive instructional position.
Subsection #8 – General Coteaching Practices
26. We coplanned instruction.
27. We coreflected on the effectiveness of lessons for student learning.
28. We coevaluated our own teaching practices.

Authors' Note: Researchers interested in deploying this coteaching survey in research settings should contact the first author for a copy of the entire survey instrument.

Table 7
Phase 1 Statistics and Internal Reliability by Scale

Subscale	Mean	Standard Deviation	Cronbach Alpha
1	4.34	.74	.888
2	4.38	.63	.867
3	3.91	.69	.795
4	4.24	.61	.839
5	4.17	.62	.837
6	3.64	.87	.879
7*	3.43	.45	.594
8*	4.03	.53	.716

Note: *Subscales 7 and 8 are intentionally diverse in scope.

Table 8
Phase 2 Item-Total Correlations

Item #	Item-Total Correlation
Q01	.637
Q02	.603
Q03	.692
Q04	.705
Q05	.693
Q06	.647
Q07	.655
Q08	.719
Q09	.752
Q10	.637
Q11	.685
Q12	.715
Q13	.638
Q14	.722
Q15	.620
Q16	.692
Q17	.594
Q18	.633
Q19	.634
Q20	.626
Q21	.664
Q29	.450
Q30	.610
Q31	.590
Q32	.584
Q33	.550
Q36	.600
Q37	.593
Q38	.592
Q39	.526
Q40	.605

Finally, a CFA was conducted on the full data set [Step 9 in Table 1]. The initial model fit indices showcase poor fit for the data across many indices (N=590; $\chi^2(390) = 2776.7$; p = .001; CMIN/DF = 7.12; normed fit index (NFI) = .808; comparative fit index (CFI) = .830; root mean square error approximation (RMSEA) = .102 (90 percent confidence intervals of .098 and .106); SRMR = .0765; AGFI = .668; PCFI = .744. These initial findings are summarized in Table 9. To improve the model fit, correlations were added between several items on the same subscale (e.g., between item 1 & 6; between 15 & 17; and between 20 & 21) for a total of 19 correlations allowed according to the modification indices from AMOS (Kline, 2005). The model fit improved and the fit guidelines were deemed acceptable to great for all indices.

The final model fit indices upon a preliminary review displayed mixed findings (N=590; $\chi^2(398) = 1647.7$; p = .001; CMIN/DF = 4.14; NFI = .901; CFI = .912; RMSEA = .073 (90 percent confidence intervals of .069 to .077); SRMR = .066; AGFI = .806; PCFI = .781. These indices disagree as a significant χ^2 value indicates poor fit, but the other indices fall within acceptable, good, or excellent ranges. Based on the majority of model fit indices, it was determined that the proposed six-factor model is a good representation of the data analyzed. These initial and final model fit indices are summarized in Table 9.

Table 9
Phase 2 CFA Model Fit Indices

Fit Index	Initial Model	Interpretation	Final Model	Interpretation
χ^2 p value	.001	Poor	.001	Poor
CMIN/DF	7.12	Poor	4.14	Good
NFI	.808	Poor	.901	Good
CFI	.830	Poor	.912	Good
RMSEA	.102	Poor	.073	Acceptable
SRMR	.077	Acceptable	.066	Good
AGFI	.668	Poor	.806	Acceptable
PCFI	.744	Good	.781	Good

The results of the two phases of analysis demonstrate that the coteaching survey is a valid and reliable instrument to measure perspectives and experiences of coteaching with following scales: Equality in the Classroom; Learning Opportunities for Students; Connecting Theory to Practice; Coteacher Collaboration; Professional Development; Personal Pedagogical Skill Development; Types of Teaching; and General Coteaching Practices.

Examination of the model more deeply shows that each item of the survey has a statistically significant loading onto its relevant construct. Most factor loadings, or regression weights, are at least .60, with many weights in the .75-.85 range. Regression weights for Q29 and Q33 are slightly lower; however, they are still statistically significant. In future analyses, the inclusion of these items may be revisited, but there is enough evidence to continue to include in this factor. The generally large regression weights indicate there is a strong theoretical connection between each of the items and the related theoretical construct. Overall, the use of confirmatory factor analysis shows strong support for the six subscales, or latent variables, present in the portion of the instrument analyzed.

The covariances for the current model also were consulted. AMOS labels the critical ratio as C.R., but this is synonymous with the t-statistic or Wald statistic. Any parameter that has an absolute value of less than 2 for its C.R. indicates that it lacks statistical significance (Stevens, 1996). All values for the current model are above 2; however, some of the covariances between disturbances are approaching this value.

Lastly, a possible threat to this model is the high correlations between a few of the examined subscales as seen in Table 10. When latent variables are so highly correlated, this may indicate the need for an advanced second order factor model (Brown, 2006). This consideration may be taken into account for future analyses to eliminate this high correlation.

Discussion

The results of the two phases of analysis demonstrate that the coteaching survey is a valid and reliable instrument to measure perspectives and experiences of coteaching with following scales: *Equality in the Classroom*; *Learning Opportunities for Students*; *Connecting Theory to Practice*; *Coteacher Collaboration*; *Professional Development*; *Personal Pedagogical Skill Development*; *Types of Teaching*; and *General Coteaching Practices*. The last two sections are purposefully diverse to collect information on the frequency of relevant coteaching activities.

Our analyses and findings contribute to the existing knowledge base in coteaching by developing a set of scales as part of a valid and reliable measure of coteaching, which to date, does not exist in the literature. Though teacher education programs across the globe have introduced coteaching as a model for student teaching, in part because it promotes collaboration between teachers and emphasizes reflective practice (Guise et al., 2017), there

Table 10
Phase 2 Statistics and Internal Reliability by Scale

Subscale	Mean	Standard Deviation	Cronbach Alpha
1	4.20	.77	.869
2	4.28	.69	.863
3	3.91	.75	.835
4	4.15	.65	.852
5	4.13	.74	.876
6	3.56	.93	.925
7*	3.49	.47	.648
8*	3.93	.56	.766

Note: *Subscales 7 and 8 are intentionally diverse in scope.

are no psychometrically developed survey instruments to evaluate teacher candidates' and clinical educators' coteaching experiences. Researchers have found that fundamental to coteaching is the expectation that coteachers will plan and implement instruction together and reflect upon how instruction has impacted student learning (Badiali & Titus, 2010; Tobin & Roth, 2006). Our instrument directly relates to the need to assess whether these essential components are existent in the model. For example, the two subscales *Connecting Theory to Practice* and *Coteacher Collaboration* ask coteachers for their perceptions of whether they discussed their pedagogical and curricular choices, reflected upon how theory can influence practice, and if they shared decision making about student learning and instruction during the coteaching placement. The coteaching model assumes that teachers will engage in these activities, yet we have limited empirical evidence to determine whether these practices actually occur during a student teaching placement. The future use of our instrument will allow researchers to make more valid claims regarding the presence of such activities during coteaching placements.

Over a decade ago, Scantlebury et al. (2008) identified corespect and coresponsibility as critical components for successful coteaching experiences, yet even today no researchers have posited approaches to assess corespect or coresponsibility. This new instrument addresses the dearth of tools for further examination of the coteaching model. For example, the *Equality in the Classroom* scale addresses this aspect of coteaching by asking coteachers whether they shared the teaching space, the responsibility for planning and implementing instruction, and their perception of the professional relationship between coteachers. The insights gleaned from these items will enable researchers to determine if the model is functioning as expected and if coteaching is providing optimal opportunities for the development of collaborative expertise (Soslau et al., 2018) through the use of shared responsibility across all aspects of coteaching.

Teacher education programs cite the potential of coteaching as an avenue to improving student learning outcomes because clinical educators remain in the class with the teacher candidate. Thus, coteaching reduces the student to teacher ratio, takes advantage of all the human capital in the classroom, and thus increases students' learning opportunities (Hartnett et al., 2014). Again, this tool is the first of its kind to actually explore if this intended outcome is coming to fruition in cotaught classrooms. The scale titled *Learning Opportunities for Students* scale focused on whether coteachers perceived that their students had increased learning opportunities through a variety of teaching practices and access to more than one instructor which may not occur in a traditional student teaching arrangement. Positive results on this scale would work toward confirming *all* available teaching resources are being leveraged to attend to individualized students' needs in ways that would prove difficult for a single teacher.

The future use of our instrument will allow researchers to make more valid claims regarding the presence of such activities during coteaching placements.

One possible utilization of the coteaching survey could identify teachers with fewer positive perspectives on coteaching for targeted intervention and professional development to improve their readiness to act as an effective coteacher. Additionally, if deployed early in the program, the coteaching survey could identify confusion within coteachers' understandings related to the goals of using coteaching in a student teaching arrangement.

Qualitative studies (Gallo-Fox & Scantlebury, 2015; Scantlebury et al., 2008) have documented that successful coteaching provides professional development for teacher candidates and clinical educators. Teacher candidates can bring subject matter expertise to facilitate the teaching of science in primary schools (Murphy & Beggs, 2010), knowledge of new technologies or curricular innovations, or by having more human resources in the classroom. Teachers are in a position to take 'risks' in implementing new methodologies or pedagogical approaches (Scantlebury et al., 2008). Through these avenues, coteachers report on the value of having a colleague with whom they can discuss questions of teaching and learning in a local context. Thus, coteachers have *Professional Development* (subscale #5) experiences while engaged in coteaching (Gallo-Fox & Scantlebury, 2015). These experiences can also lead to teachers' noting an increase in their *Personal Pedagogical Skill Development* (subscale #6) as a result of the collaborative learning environment for teacher candidates and clinical educators alike.

The subsection *Types of Teaching* addresses whether coteachers are engaged in different roles during the coteaching experience. A coteacher may take the lead in instructing a class, assume a peripheral role by stepping aside and working with a group of students, be a spectator, or engage as an expert (Tobin, 2006). The *Types of Teaching* subscale also asked coteachers to indicate if they had any of these different teaching experiences. The *General Coteaching Practices* subsection asked teachers to indicate if they shared in evaluating aspects of their coteaching experiences such as lesson planning and implementation. Future studies employing this survey instrument might explore comparisons between responses with high and low frequency of different coteaching practices highlighted in subsection 7 and 8, such as the prevalence of coevaluation (item 28). A hypothetical study could investigate broader patterns of responses across the established subscales #1 to #6 using coevaluation frequency (item 28) as a predictor or independent variable. Use of this survey instrument in such a method could inform a deeper understanding of the impact of coevalution within the coteaching model–a stated need in the coteaching research literature (e.g., Drewes et al., 2020).

A limitation of this study is that the sample parameters of this specific university context resulted in a majority of respondents being elementary and middle school teachers with fewer high school teachers. We agree with the belief stated by Andrews and colleagues (2017) that "survey validation is a continuous process" (p. 16) and, as such, this survey could benefit from additional validity evidence that encompasses more high school coteaching respondents. Future studies should expand the scope to include a more diverse target population across grade levels and content areas.

Another limitation is that this survey instrument does not incorporate student (K-12 pupil) learning outcomes or student beliefs. Students are experts in their own classrooms and can provide important insights on the classroom learning environment (Bayne, 2012). We recommend future work along this path to better incorporate perspectives of all coteaching stakeholders (e.g., Drewes et al., 2020) and to connect analysis of this survey's findings to other data such as student achievement or teachers' performance criteria.

The implications of the practical application and use of these scales, and the instrument they constitute, are manyfold. One possible utilization of the coteaching survey could identify teachers with fewer positive perspectives on coteaching for targeted intervention and professional development to improve their readiness to act as an effective coteacher. Additionally, if deployed early in the program, the coteaching survey could identify confusion within coteachers' understandings related to the goals of using coteaching in a student teaching arrangement.

We also posit that if open response spaces were continued to be included, the survey could serve as a reflective space to initiate ongoing conversations between teacher educators, clinical educators, and university personnel involved with teacher education programs and field experiences. Both coteachers (candidate and clinical educator) could use the survey items as a form of reflective self-assessment throughout the student teaching experience to judge how well they are implementing the model. These self-assessments could be shared as a way to collaborate around improving the model and to scaffold individual and collaborative introspection on problems of practice. Pairing discussions of survey responses

with a framework such as the Guide for Reflective Practice (Greenberger, 2020) may also be particularly generative for documenting and improving teachers' reflective practice.

The survey items can also be introduced during professional development sessions with teacher candidates and clinical educators as a way to inform participants about the intended functions, features, and outcomes of the coteaching model. These are several ways that teacher educators can practically apply the instrument and avenues for future research, which could incorporate the coteaching survey to improve coteaching experiences and teacher preparation models more broadly.

References

Anderson, D. (2007). The role of cooperating teachers' power in student teaching. *Education*, *128*(2), 307-312.

Andrews, S. E., Runyon, C., & Aikens, M. L. (2017). The Math-Biology Values Instrument: Development of a tool to measure life science majors' task values of using math in the context of biology. *CBE–Life Sciences Education*, *16*, 1-12. https://doi.org/10.1187/cbe.17-03-0043

Arbuckle, J. L. (2005). *AMOS 6.0 user's guide*. AMOS Development Corporation.

Austin, V. (2001). Teachers' beliefs about coteaching. *Remedial and Special Education*, *22*(4), 245-255. https://doi.org/10.1177/074193250102200408

Bacharach, N., Heck, T., & Dahlberg, K. (2010). Changing the face of student teaching through coteaching. *Action in Teacher Education*, *32*(1), 3-14. https://doi.org/10.1080/01626620.2010.10463538

Badiali, B., & Titus, N. E. (2010). Co-teaching: Enhancing student learning through mentor-intern partnerships. *School-University Partnerships*, *4*(2), 74-80.

Bayne, G. U. (2012). Capturing essential understandings of the urban science learning environment. *Learning Environments Research*, *15*(2), 231-250. https://doi.org/10.1007/s10984-012-9112-8

Benson, J., & Nasser, F. (1998). On the use of factor analysis as a research tool. *Journal of Vocational Education*, *23*, 13-33.

Bentler, P. (1994). On the quality of test statistics in covariance structure analysis: Caveat emptor. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 237-260). Plenum Press. https://doi.org/10.1007/978-1-4757-9730-5_11

Brown, T. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

Cochran-Smith, M., Villegas, A. M., Abrams, L., Chavez Moreno, L., & Mills, T. (2016). Research on teacher preparation: Charting the landscape of a sprawling field. In D. Gitomer & C. Bell (Eds.), *Handbook of Research on Teaching* (pp. 439-547). American Educational Research Association. https://doi.org/10.3102/978-0-935302-48-6_7

Council for the Accrediation of Educator Preparation. (CAEP). (2013). *CAEP Accrediation Standards*. Washington, DC: Council for the Accrediation of Educator Preparation. http://caepnet.org/accreditation/standards/

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

Dickey, D. (1996). Testing the fit of our models of psychological dynamics using confirmatory methods: An introductory primer. In B. Thompson (Ed.), *Advances in COUNCIL social science methodology*, *4*, (pp. 219-227). JAI Press.

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*, 225-241. https://doi.org/10.1177/073428290502300303

Dove, M., & Honigsfeld, A. (2010). ESL Coteaching and collaboration: Opportunities to develop teacher leadership and enhance student learning. *TESOL Journal*, *1*(1), 3-22. https://doi.org/10.5054/tj.2010.214879

Drewes, A., Soslau, E., & Scantlebury, K. (2020). Listening to the Missing Voices: Students' Perspectives on Coteaching. *Research & Practice in Assessment*. *14*, 5-18.

Drewes, A., Soslau, E., & Scantlebury, K. (2021). Striving towards an ideal: Coevaluation of student coteaching experiences. *Journal of Education for Teaching*, *47*(1), 60-74. https://doi.org/10.1080/02607476.2020.1845954

Drouin, S., Karathanos-Aguilar, K., & Lehmkuhl-Dakhwe, V. (2020). Affordances and constraints: Pre-service science educators co-teaching in support of ELLs. *Journal of Education and Culture Studies*, *4*(1), 1-18. https://doi.org/10.22158/jecs.v4n1p1

Fan, X., Thompson, B., & Wang, L. (1999). The effects of sample size, estimation methods, and model specification on SEM fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 56-83. https://doi.org/10.1080/10705519909540119

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. SAGE

Friend, M., Cook, L., Hurley-Chamberlain, D., & Shamberger, C. (2010). Coteaching: An illustration of the complexity of collaboration in special education. *Journal of Educational & Psychological Consultation*, *20*(1), 9-27. https://doi.org/10.1080/10474410903535380

Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*(1), 65-83. https://doi.org/10.1207/s15327906mbr4101_5

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey, *Public Opinion Quarterly*, *73*(2), 349-360. https://doi.org/10.1093/poq/nfp031

Gallo-Fox, J., & Scantlebury, K. (2015). "It isn't necessarily sunshine and daisies every time": Coplanning opportunities and challenges when student teaching. *Asia-Pacific Journal of Teacher Education*, *43*(4), 324-337. https:/doi.org/10.1080/1359866X.2015.1060294

Gallo-Fox, J., & Scantlebury, K. (2016). Coteaching as professional development for cooperating teachers. *Teaching and Teacher Education*, *60*, 191-202. https://doi.org/10.1016/j.tate.2016.08.007

Gallo-Fox, J., Wassell, B., Scantlebury, K., & Juck, M. (2006). Warts and all: An ethical struggle with disseminating research on coteaching. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *7*(4). https://doi.org/10.17169/fqs-7.4.183

Greenberger, S. W. (2020). Creating a guide for reflective practice: Applying Dewey's reflective thinking to document faculty scholarly engagement. *Reflective Practice*, *21*(4), 458-472. https://doi.org/10.1080/14623943.2020.1773422

Guise, M., Habib, M., Thiessen, K., & Robbins, A. (2017). Continuum of co-teaching implementation: Moving from traditional student teaching to co-teaching. *Teaching and Teacher Education*. *66*, 370–382. https://doi.org/10.1016/j.tate.2017.05.002

Hartnett, M. J., McCoy, A., Weed, R., & Nickens, N. (2014). A work in progress: Unraveling the lessons learned in a co-teaching pilot. *The Renaissance Group*, *3*(1), 33-54.

Hoy, W. K., & Woolfolk, A. E. (1990). Socialization of student teachers. *American Educational Research Journal*, *27*(2), 279-300. https://doi.org/10.3102/00028312027002279

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Kamens, M. W. (2007). Learning about coteaching: A collaborative student teaching experience for preservice teachers. *Teacher Education and Special Education*, *30*(3), 155-166. https://doi.org/10.1177/088840640703000304

Kline, R. B. (2005). *Principles and practices of structural equation modeling* (2nd ed.). Guilford.

Koskela, R., & Ganser, T. (1999). The cooperating teacher role and career development. *Education*, *119*(1), 106-125.

Landt, S. M. (2004). Professional development of middle and secondary level educators in the role of cooperating teacher. *Action in Teacher Education*, *26*(1), 74-84. https://doi.org/10.1080/01626620.2004.10463315

MacCullum, R. C., Widaman, K., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84-99. https://doi.org/10.1037/1082-989X.4.1.84

Magiera, K., Smith, C., Zigmond, N., & Gebauer, K. (2005). Benefits of coteaching in a secondary mathematics classroom. *Teaching Exceptional Children*, *37*(3), 20-24. https://doi.org/10.1177/004005990503700303

Marsh, H. W., Hau, K., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181-200. https://doi.org/10.1207/s15327906mbr3302_1

Martin, S. (2009). Learning to teach science. In K. Tobin & W.-M. Roth (Eds.), *World of science education: North America* (pp. 567-586). Sense Publishers.

Milne, C., Scantlebury, K., Blonstein, J., & Gleason, S. (2011). Coteaching and disturbances: Building a better system for learning to teach science. *Research in Science Education*, *41*, 414-440. https://doi.org/10.1007/s11165-010-9172-7

Murphy, C., & Beggs, J. (2010). A five-year systematic study of coteaching science in 120 primary schools. In C. Murphy & K. Scantlebury (Eds.), *Coteaching in international contexts: Moving forward and broadening perspectives*. (pp. 11-34). Springer. https://doi.org/10.1007/978-90-481-3707-7_2

Murphy, C., Beggs, J., Carlisle, K., & Greenwood, J. (2004). Students as 'catalysts' in the classroom: The impact of co-teaching between science student teachers and primary classroom teachers on children's enjoyment and learning of science. *International Journal of Science Education*, *26*(8), 1023-1035. https://doi.org/10.1080/1468181032000158381

Murphy, C., & Martin, S. N. (2015). Coteaching in teacher education: research and practice. *Asia-Pacific Journal of Teacher Education*, *43*(4), 277-280. https://doi.org/10.1080/1359866X.2015.1060927

Murphy, C., & Scantlebury, K. (Editors). (2010). *Coteaching in international contexts: Research and practice*. London: Springer.

National Council for the Accreditation of Teacher Education [NCATE]. (2010). *Transforming teacher education through clinical practice: A national strategy to prepare effective teachers*. Report of the Blue-Ribbon Panel on Clinical Preparation and Partnerships for Improved Student Learning. National Council for Accreditation of Teacher Education. Retrieved from: http://caepnet.org/~/media/Files/caep/accreditation-resources/blue-ribbon-panel.pdf

Nguyen, H. T. (2009). An inquiry-based practicum model: What knowledge practices, and relationships typify empowering teaching and learning experiences for student teachers, cooperating teachers and college supervisors? *Teaching and Teacher Education*, *25*(5), 655-662. https://doi.org/10.1016/j.tate.2008.10.001

Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research*, *4*(1), 56-78. https://doi.org/10.1177/1558689809355805

Parsons, M., & Stephenson, M. (2005). Developing reflective practice in student teachers: Collaboration and critical partnerships. *Teachers and Teaching*, *11*(1), 95-116. https://doi.org/10.1080/1354060042000337110

Phelan, A., McEwan, H., & Pateman, N. (1996). Collaboration in student teaching: Learning to teach in the context of changing curriculum practice. *Teaching and Teacher Education*, *12*(4), 335-353. https://doi.org/10.1016/0742-051X(95)00044-K

Rice, D., & Zigmond, N. (2000). Co-teaching in secondary schools: Teacher reports of developments in Australian and American classrooms. *Learning Disabilities Research & Practice*, *15*(4), 190-197. https://doi.org/10.1207/SLDRP1504_3

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation*. Wadsworth.

Scantlebury, K., Gallo-Fox, J., & Wassell, B. (2008). Coteaching as a model for preservice secondary science teacher education. *Teaching & Teacher Education*, *24*, 967–981. https://doi.org/10.1016/j.tate.2007.10.008

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Lawrence Erlbaum Associates.

Schwarz, N., Groves, R. M., & Schuman, H. (1998). Survey methods. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 2, 4th ed., pp. 143-179). McGraw-Hill.

Smith, E. R. (2007). Negotiating power and pedagogy in student teaching: Expanding and shifting roles in expert-novice discourse. *Mentoring & Tutoring: Partnership in Learning*, *15*(1), 87-106. https://doi.org/10.1080/13611260601037405

Soslau, E. (2012). Opportunities to develop adaptive teaching expertise during supervisory conferences. *Teaching and Teacher Education*, *28*(5), 768–779. https://doi.org/10.1016/j.tate.2012.02.009

Soslau, E., Gallo-Fox, J., & Scantlebury, K. (2018). The promises and realities of implementing a coteaching model of student teaching. *Journal of Teacher Education*. https://doi.org/10.1177/0022487117750126

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Erlbaum.

Strobaugh, R., & Everson, K. (2019). Student teacher engagement in co-teaching strategies. *Educational Renaissance*, *8*(1), 30-47. https://doi.org/10.33499/edren.v8i1.137

Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. American Psychological Association.

Tobin, K. (2006). Learning to teach through coteaching and cogenerative dialogue. *Teaching Education*, *17*(2), 133-142. https://doi.org/10.1080/10476210600680358

Tobin, K., & Roth, W.-M. (2006). *Teaching to learn: A view from the field*. Sense Publishers.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2016). 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

Zeichner, K. (1992). Rethinking the practicum in the Professional Development School partnership. *Journal of Teacher Education*, *43*(4), 296-307. https://doi.org/10.1177/0022487192043004009

Zeichner, K. (2010). Rethinking the connections between campus courses and field experiences in college- and university-based teacher education. *Journal of Teacher Education*, *61*(1-2), 89-99. https://doiorg/10.1177/0022487109347671

Abstract

While scholarship on assessment and evaluation has grown significantly over the past forty years, writing tends to focus on the "how-to" implementation of assessment practices within a classroom or programmatic context. While individual case studies and practical manuals offer valuable contributions for implementation, there is a need for assessment research that supports practices that can highlights interventions to inform practice and positively impact student learning outcomes. To this end, we reviewed scholarly literature to explore the degree to which assessment research discusses and informs student learning. We then performed a content analysis examining how academic research on assessment discusses, analyzes, and evaluates student learning and student success. We identify five specific categories of assessment scholarship and offer implications for future assessment practice and research.

AUTHORS

Marjorie L.
Dorimé-Williams, Ph.D.
University of Missouri

Cindy Cogswell,
Director of Data Strategy
New York University

Gianina Baker, Ph.D.
University of Illinois

Assessment in Use: An Exploration of Student Learning in Research and Practice

Current literature on assessment is full of examples of practice. Less common are writings exploring the philosophical or theoretical basis of assessment or their direct impact on student learning. The field needs writings connecting theory to practice: we need to know why we are doing what we are doing, whether what we are doing works, and for whom it works. If we can document and communicate a philosophy of assessment-or philosophies of assessment-we can then support informed ways of framing and doing assessment that more effectively meets students' needs. There is a growing need for individuals in postsecondary education to understand what assessment is and develop influential theories, practices, and expectations for assessment to positively impact student outcomes.

Limited research critically examines the impact of assessment practices on students learning and outcomes (Cogswell, 2016). This paper explores the use and understanding of assessment research and its impact on and relationship to student learning through a content analysis of scholarly literature on assessment. Specifically, we examine to what degree academic research on assessment discusses, analyzes, and evaluates student learning and student success. The following questions guided the analysis:

a) What themes are present within various forms of postsecondary assessment in scholarly assessment-focused journals?

b) How does scholarly assessment research attend to student learning?

a. Are there gaps in assessment research with respect to student learning?

CORRESPONDENCE

Email
dorimewilliamsm@missouri.edu

Due to its interdisciplinary nature, the term assessment refers to various processes and purposes in post-secondary education.

Our rationale for this study and approach is two-fold. First, recent reviews of assessment literature have examined scholarship beginning around 2006 (i.e., Pereira et al., 2016). However, these reviews have not focused on student learning. Second, since the Spellings Commission report's release in 2006 (U.S. Department of Education, 2006), political and public stakeholders in postsecondary education have increasingly called on institutions to demonstrate positive student outcomes (Fuller et al., 2012; Zumeta & Kinne, 2011). This shift has increased institutional attention on assessment practices and student learning. This study presents a meaningful contribution to the broader scholarly and practical discourse on student learning and student success by presenting findings on assessment research and student outcomes. Our study offers an assessment typology that can inform praxis and research as the field moves toward a more student focused approach.

Assessment and Student Learning

Due to its interdisciplinary nature, the term assessment refers to various processes and purposes in postsecondary education. Focusing on a student-centered approach[1] to assessment, Suskie (2009) defines it as follows:

Assessment is the ongoing process of establishing clear, measurable expected outcomes of student learning; ensuring that students have sufficient opportunities to achieve those outcomes; systematically gathering, analyzing, and interpreting evidence to determine how well student learning matches our expectations; using the resulting information to understand and improve student learning. (p. 4).

This framing definition of assessment highlights the centrality of student learning and student outcomes to institutional functioning. Institutions should use assessment of student learning to improve processes (e.g., classroom curriculum, student programming, resource allocation), inform efforts to improve student learning, and respond to regional accreditation requirements (Jankowski et al., 2018). This context also positions student learning at the center of the assessment process. Through the examination of assessment scholarship, we seek to examine the extent to which this perceived relationship is present in the literature and in what ways.

Methods

The purpose of this paper is to examine the results of a content analysis of scholarly literature on assessment and student outcomes. Specifically, we sought to examine the degree to which assessment scholarship directly attended to student learning and student outcomes. The following section defines the data sources, procedures, and data analysis process. We approached this inquiry from a broad perspective to obtain a more thorough and representative sample of research on assessment. We utilized a collaborative approach to improving the reliability and validity of the findings. The original articles were narrowed by reviewing abstracts and keywords to eliminate those that were not specific to assessment in higher education, those that included assessment in the context of testing diagnostic tools, and those that focused on program evaluation rather than students.

Data Sources

Existing literature from assessment and education-related journals were reviewed to define and develop typologies of assessment. The reviews' search parameters included assessment scholarship related to four-year and two-year public and private colleges and universities within the United States. While there is a robust context of assessment scholarship within an international context, there is significantly less literature that examines assessment within the United States context. The journals included in this review were selected because they are (a) recognized as top-tier journals in the field of assessment or postsecondary

[1] The authors acknowledge that there are varied definitions and conceptualizations of "student-centered assessment". For the purposes of this paper, we operationalize student-centered as what supports students best or what students need. We use student-centered as a way to articulate practices that should be focused on students and not simply the institution (e.g., McNair et al., 2016).

education and (b) are used by both researchers and practitioners. With these criteria in mind, we selected the following nine journals for our analysis:

- American Journal of Evaluation
- Assessment and Evaluation in Higher Education
- New Directions for Institutional Research
- The Journal of Case Studies in Accreditation and Assessment
- The Journal of Higher Education
- The Journal of Assessment and Evaluation in Higher Education
- The Journal of Assessment and Institutional Effectiveness
- Research and Practice in Assessment
- Review of Higher Education

Articles were restricted to those published between 2005 and present. Contemporary reviews of assessment literature begin close to 2005 (i.e., Pereira et al., 2016). We searched each of the journals for empirical or scholarly discussions of assessment practices that focused on students in higher education using the search word "assessment." A total of 1,950 articles were included in the initial electronic search. Book reviews, editorials, and other non-scholarly content were not included in the analysis and eliminated from the initial results.

Data Analysis

After the narrowing process, the remaining articles were read and significant themes were articulated based on article topics. The analysis procedures included: (a) establishing summaries of the articles in each journal that fit the criteria; (b) establishing coding categories independently then as a group; (c) individually and collectively revising categories based on each set of articles; and (d) taking steps to improve validity and reliability through triangulation (Bowen, 2009; Eisner, 1991).

By allowing categories to develop throughout the analysis process, we were able to build a more exhaustive list of categories that discussed how assessment is addressed within the scholarly literature.

Our analysis focused on identifying emergent themes rather than investigating the articles with predetermined categories. We entered this analysis with the goal of using an inductive process to develop themes based on the articles, rather than determining a fixed number of themes. First, article abstracts, or if necessary full articles, were read and re-read to generate initial categories codes (Miles & Huberman, 1994). First-cycle coding methods are the processes that happen during an initial coding of data (Saldaña & Omasta, 2016). Through the first cycle of coding and moving into the second cycle, we utilized exploratory codes. As a check for interrater reliability, we met regularly throughout the coding process to discuss, compare, and contrast preliminary codes and findings. We also reviewed the articles collectively to confirm our individual interpretations of the research focus and codes of each.

After the first round of open coding, we came together to discuss preliminary findings. Next, codes were collapsed by grouping categories that seemed to relate to each other while leaving intact those that stood independent from all others. This process supported the convergence of emerging themes and results. In discussions and check-ins, we explored and defined the parameters of code categories. By allowing categories to develop throughout the analysis process, we were able to build a more exhaustive list of categories that discussed how assessment is addressed within the scholarly literature. Notes were made on articles read by each researcher independently and then were shared to compare themes.

Lastly, themes were compared and contrasted to understand the degree to which they were similar; closely related themes were then further collapsed. At the forefront of this process was the lens of whether the work was student-centered or not. If it was not, we asked what audience or practice was being attended to by the article.

This collaborative reviewing process facilitated a discussion on how we each analyzed the articles and decided on codes and themes. Emerson et al. (1995) assert that "…choice of method reflects researchers' deeper assumptions about social life and how to

understand it" (p. 10). Taylor and Bogdan (1998) posit that "as a qualitative researcher, your role is to capture how people define their world or construct their reality" (p. 52). While each researcher looked at separate journals, we worked collaboratively throughout the inquiry process to improve interrater reliability about what was emerging from the coding analysis. Therefore, each round of independent article review was followed by a collaborative discussion. Conversations with each other and about our emerging findings from the data strengthened the emerging categories.

The first distinct category was assessment for measurement… A second category that emerged was the use of assessment as a mechanism to drive policy.

Positionality

We also recognize that while our collaborative process yielded specific results that we define as categories, future research may produce different conclusions. We acknowledge that our analysis is shaped and informed by our own research and practice in the field of assessment. We each see ourselves as scholar-practitioners. Each author has served in various administrative roles related to assessment in higher education and engaged in the process of research, providing professional development, and scholarly writing related to assessment, accreditation, and student success.

Goodness

Qualitative researchers have used the term *goodness* to indicate quality in qualitative research, similar to *trustworthiness* and *validity* in quantitative research. We aligned our process with elements of goodness as defined by Jones et al. (2013). For consistency, our study was designed around our research questions which guided our data collection and analysis process. Throughout this process, we were informed by our methodological training as well as input and feedback from recognized senior scholars in assessment.

Findings

Five thematic domains emerged from our analysis. These themes were not necessarily aligned to individual articles, but instead focused on what we saw throughout the articles overall. Therefore, more than one theme may be present within the articles we reviewed although our findings use distinct articles to highlight examples present in the themes.

1. Assessment for Measurement

The first distinct category was assessment for measurement. The term assessment was used to encompass ways to measure perceived gains, ability, and demonstrated knowledge–a direct connection to student learning. Articles within this category used the word assessment as a proxy for evaluating performance, knowledge, or gains. Many of the journals included articles that discussed assessment as a tool for measurement or a means of assessing individuals, organizations, or processes. For example, Mansilla et al. (2009) presented research on rubric development and used it to assess student writing. Freed and Mollick (2010) measured students' performative knowledge. *Research and Practice in Assessment* was founded in 2006 as a newsletter. It is now a peer-reviewed publication with a plethora of articles on assessment as measurement. Studies from this journal included a focus on surveys and scales used to measure student outcomes (Pastor et al., 2018), augmenting standardized testing (Gray et al., 2017), measuring students' efforts on assessments (Smiley & Anderson, 2011), and measuring the relationship between student assessment outcomes and other academic performance indicators (Pieper et al., 2008).

2. Assessment for Policy

A second category that emerged was the use of assessment as a mechanism to drive policy. According to Merriam-Webster, policy is "a definite course or method of action selected from among alternatives and in light of given conditions to guide and determine present and future decisions" and "a high-level overall plan embracing the general goals and acceptable procedures especially of a governmental body" (2020). Educational policy refers to the set of guidelines, rules, and principles that are enforced and adopted by campus, local, state, and federal agencies to meet set standards and goals (Adams, 2014; Araya, 2015; Mitchell et al., 2018). Assessment policy, therefore, is described as a set of principles related to any facet of assessment, including but not limited to survey protocol and administration, expectations for

collecting and presenting evidence of student learning, requirements for course evaluation, and accreditors and governmental requirements for transparency (Leathwood, 2005; McDonnell, 1994; Warburton, 2018).

This category also included publications on policy through national, state, and local lenses. It focused on policy from multiple stakeholder perspectives such as the government and educational advocacy groups within higher education. Briggs (2007) discussed how the Association of American Colleges and Universities has pushed back on assessment mandated policy and called for increased collaboration and input from faculty in assessment practices and policy. Price (2019) used policy narratives to explore prior learning assessments and how various groups advocate for or use policy to further specific and sometimes competing agendas. Across articles and journals, we found that scholars presented perspectives that often advocated for an increased agency for postsecondary institutions. They also challenged what faculty and staff may perceive to be overbearing assessment mandates (e.g., periodic, ongoing course-level assessment, alignment between co-curricular and curricular assessments, continuous documented change in response to assessment). In this sense, articles that addressed assessment for policy researched, documented, and defended assessment practices and policies but did not explicitly focus on student learning.

3. Assessment for Improvement

The third category to emerge was assessment as a practice to improve outcomes. *Outcomes* have many definitions; therefore, we use the following, "outcomes can be defined as participant-centered, desired effects of a program, a service, or an intervention. In other words, an outcome is a result you want to achieve following a given activity" (Henning & Roberts, 2016 p. 85). Outcomes relate to both statements of student learning, namely what students will know, be able to do, or what changes will be made to their behavior as a result of the impact of attending postsecondary education. When considering long-term outcomes, the focus attends to what students do after graduation, personally and professionally.

Scholars whose work fell into this category described assessment as a practice for teaching and learning and were often based in specific disciplines. For example, Lusher (2010) examined the practice of improving curriculum design in accounting programs at 102 colleges and universities. Lusher's work, like many in the field, centered on individual course change influencing student performance. Similarly, Barrett (2012) discussed writing in the humanities, improving the ways students demonstrate competency, and examining how students are graded. In addition to focusing on assessment in traditional educational contexts such as classrooms, courses, and majors, several articles also discussed co-curricular and student affairs practice of assessment. This theme also directly links to our broader discussion on student-centered learning. We found that this literature focused on ways to use assessment data and results to improve teaching and learning practices.

4. Assessment for Equity

Another category that emerged from several articles describing assessment efforts, most noticeably with studies focused on minoritized populations, is that of assessment for equity. Minoritized populations are those who, due to historical, social, economic, cultural, and other forms of bias, discrimination, and oppression, are excluded from dominant social norms and beliefs and, as a result, are believed to be deficient, different, and inferior to the "dominant groups" in society (Harley et al., 2002). The use of power and privilege often results in unequal outcomes for these groups (e.g., racial gaps in college graduation rates). In this context, we define equity as the processes and practices that ensure all students have what they need to successfully access, navigate, and graduate from college.

The third category to emerge was assessment as a practice to improve outcomes… Another category that emerged from several articles describing assessment efforts… is that of assessment for equity.

In these articles, the authors focused on assessment as a method to support student learning by examining differences in learning outcomes for various student populations (i.e., female students, Black students, Latinx students) (Ching, 2018; Jaeger et al., 2017; Ro & Loya, 2015; White & Lowenthal, 2011). While articles in this category share common perspectives on the importance of incorporating diversity and intercultural competency in assessment scholarship, their approaches varied. In many ways, socially just outcomes of student learning are often not the focus when engaging in assessment processes and practices

centered on specific student populations. Similar to our previous theme on student outcomes, the scholarship within this theme focuses primarily on improving the educational experiences and outcomes for students across various communities.

The final theme that emerged included scholarship focused on assessment to support change management processes.

5. Assessment for Change Management

The final theme that emerged included scholarship focused on assessment to support change management processes. Change management is a cycle, including data-informed decision making, implementing policies and practices, and examining the impact of implemented policies and practices (Kotter, 1996). Change management is well aligned with a student-centered assessment process. Again, the assessment cycle includes defining student learning, providing students with learning opportunities to achieve these goals, assessing how well students have achieved those goals, and using assessment results to improve (Suskie, 2009). Institutional leaders who prioritize assessment, articulate its institutional purpose, provide resources and training for faculty and staff, and incorporate assessment into all institutional practices, have had success and demonstrate the relationship between effective change management and successful assessment practices (Lane et al., 2014).

Scholarship in this area covered a variety of educational and institutional practices that illustrate various aspects of Kotter's (1996) change management process. For example, one article discussed a longitudinal analysis of the retention and matriculation of students who completed a first-year seminar course at one institution (Ben-Avie et al., 2012). The course served as an intervention and the researchers *assessed* its impact. This illustrates step seven of Kotter's change management process, both change and a commitment to using assessment to improve on changes and to continue this process as necessary. In another study, Hora et al., (2017) explored the use of educational data by faculty and whether this data use had implications for their practice. Their findings on barriers and supports that influence faculty use of data in their teaching practices can help institutions empower faculty and staff to use assessment data to improve student outcomes. This study demonstrates how step five, empowering action, and other aspects of Kotter's model are present in scholarship on assessment and change management as well as how this research can inform teaching and student learning.

Discussion

We frame the following discussion as both a response to our research questions and an opportunity to examine the current understanding of assessment and student learning from our findings. The purpose of this study was to explore the scholarly use and understanding of assessment and its relationship and impact on student learning. We sought to examine the dialogue on assessment within scholarly journals and to identify potential gaps in the literature with respect to student outcomes. The analysis resulted in five themes that existed across and within the journals in this investigation: assessment for a) measurement, b) policy, c) outcomes improvement, d) equity, and e) change management were common subjects throughout the literature. We discuss the implications of how these themes can or should connect to student learning in postsecondary education. While some assessment scholarship does indeed attend to issues of student learning, there are areas of assessment approaches and practices where scholars and practitioners can more intentionally focus on students. The following discussion articulates our findings, connects them to the assessment literature writ large, and concludes with limitations of the work and opportunities for future research.

1. Assessment and Measurement

Scholarship reviewed on assessment and measurement demonstrated how choices made about assessment tools and methodological decisions could significantly impact the utility of collected data and the ability of faculty and staff to improve practices and student learning. When appropriately planned, assessment tools that are valid, reliable, and created with student learning experiences in mind can support the measurement of student learning and provide information that allows us to respond in meaningful ways (Cumming & Miller, 2017)

Assessment should be more than Measurement

While the measurement of student learning is essential to improving future outcomes, it is crucial to align measurement appropriately with teaching and learning activities (Biggs & Tang, 2011). Since assessment of student outcomes is not one dimensional, data collection instruments and practices should be informed by the outcomes and students they are intended to measure, not the other way around (Henning & Roberts, 2016). Too often, approaches to assessment center on the tool or instrument that measure students in some way (e.g., survey questions, interview protocols, national surveys) instead of an intentional focus on teaching and learning practices that influence student success. Divorcing the assessment process from the behaviors that guide and shape the student learning experience is evident in much of the literature that discusses assessment as a tool for measurement. However, assessment is not a neutral process. Stakeholders must take intentional steps to ensure that measurement-related issues in assessment are implemented and contextualized appropriately (Dorimé-Williams, 2018; Leathwood, 2005). Future research in this area should attend to how methodological and other choices about measurement can influence what we infer about students and their learning.

Assessment policy should serve the best interest of students. Unfortunately, institutions often fail to provide students with a meaningful seat at the assessment and policy table.

2. Assessment Policy

Assessment used to drive policy is an essential topic within the scholarly literature. The articles featured in this study create a space to push back on and critically examine assessment policies and their impact on institutions. Given the expectations, priorities, and goals are from a wide range of internal and external stakeholders, there are numerous and varied forms of policy that can inform and shape assessment practices. While the articles reviewed presented differing voices and perspectives across settings-local, state, and federal-there continues to be a lack of consideration for the real-world impact of competing educational and assessment policy changes on students. Without structure and intentional collaborative (not competitive) planning, policy can develop rapidly, unpredictably, and incoherently when informed by underlying principles or frameworks that are divergent and uninformed (Araya, 2015). Current educational policy, and as a result assessment policy, has become centralized at the state and federal levels. Institutions must respond to political and public pressure to meet policy goals and increasingly rigorous demands (Adams, 2014; Araya, 2015).

Bolstering Student Learning through Policy

Education policy has shaped educational systems and assessment by centralizing control of finances and governance, shifting decision-making to legislators, and championing one-size-fits-all, test-based accountability and assessment for improving student outcomes (Mitchell, 2017). These factors influence postsecondary institutions and highlight how scholarship in this area can better attend to conducting student-centered assessment at our institutions. While policy can be a useful tool for promoting student success, institutions may not always prioritize students' experiences and instead focus on compliance.

Assessment policy should serve the best interest of students. Unfortunately, institutions often fail to provide students with a meaningful seat at the assessment and policy table. Further, including students' needs in our discussions on assessment and assessment policy can shift us from passive instructional to active teaching institutions. Future research should examine the role students' learning needs to play in assessment and policy-making processes. Practitioners should continue to explore how to center students in institutional conversations about assessment and subsequent policies. Assessment policy developed intentionally can have a significant and positive impact on student learning (Moutsis, 2010).

3. Assessment and Outcome Improvement

Significant research is focused on using assessment to improve student outcomes. The term "outcomes" can refer to many different aspects of an institution's efforts across an array of departments and units. While some articles discussed student learning in classroom settings or from a disciplinary perspective, articles on outcomes assessment often fell short of fully completing the assessment cycle. Specifically, they examined initial interventions for student learning but failed to discuss changes to the student environment that would require acting

on their findings. When reflecting on the breadth of potential student outcomes that can be assessed in an institutional setting (e.g., course learning outcomes, co-curricular learning outcomes, career development outcomes), scholars and practitioners need to consider how to evaluate these outcomes, effect change, and put assessment results to good use (Henning & Roberts, 2016; Suskie, 2009).

Increased Focus on Outcome Improvement

By formulating and assessing learning outcomes, we can: create improved learning environments at the course, program, unit, and institutional level; provide increased direction for how to improve teaching activities; inform internal and external stakeholders of our intentions for students; and continue to foster a student-centered institutional process that prioritizes student learning and development (Huba & Freed, 2000). Outcomes assessment provides a tool that allows scholars and practitioners to focus on learning that should result from a specific experience or activity rather than on the activity alone. This approach distinguishes outcomes assessment from more common forms of evaluation in postsecondary education such as course evaluations or satisfaction surveys (Huba & Freed, 2000). Improving student outcomes requires institutions to be explicit about their mission and values. Alignment between the institutional, unit (academic and co-curricular), program, and course levels can again assist in the shift from passive instructional to active learning organizations. This process also contributes to an institution's ability to articulate to external stakeholders and the general public the value and importance of what students achieve through participation in postsecondary education at a specific institution.

4. Assessment and Equity

A more recent area of discussion within the field of assessment focuses on equity and inclusion. Assessment policies and practices can increase access, foster student retention, and contribute to improved persistence to graduation. As was previously discussed, scholarship on equity in assessment examines diverse and marginalized student populations (e.g., Ching, 2018; Jaeger et al., 2017; Ro & Loya, 2015; White & Lowenthal, 2011). This scholarship also reminds us to be mindful of the differential experiences' students have in postsecondary education due to their various identities. Racial, ethnic, gendered, religious, and disability identities are only some of the ways students differ in how they experience their learning environment. Equitable assessment requires scholars and practitioners to recognize that students come to institutions with varied needs and that improving teaching and learning means improving our cultural competency, even in assessment (Dorimé-Williams, 2018; Leathwood, 2005). Assessment scholarship related to issues of equity and diversity calls on us to recognize how the social, cultural, political, and historical norms and practices within an institution shape each student's experience a little differently.

Promoting Equity in Postsecondary Education

Equity in assessment can support improved outcomes for all students in postsecondary education. First, recognizing how each aspect of the assessment cycle can promote or hinder equitable student participation and outcomes can improve the design and administration of assessment tools and practices (Dorimé-Williams, 2018). By being mindful of differences in student populations, assessment can also inform practices and policy that create a better environment for student success (McArthur, 2016). Promoting equity in assessment can also help shift institutional cultures from instructional to learning organizations. By putting all students' needs and learning at the forefront of assessment practices, we can encourage institutions to use assessment activities to foster inclusive learning environments rather than only using assessment for accreditation or compliance purposes (Jankowski et al., 2018).

Promoting equity in assessment can also help shift institutional cultures from instructional to learning organizations.

5. Assessment within Change Management

The steps of change management, when applied to assessment, can offer direction for institutional leaders to improve their cultures of assessment and engage faculty and staff in processes that shift from an instructional to a teaching paradigm initially mentioned by Barr & Tagg (1995). Considering the alignment between change management and assessment practices, scholarship in this area can contribute to positive organizational change centered on student learning. The steps within change management and well-designed assessment process

call for engagement from institutional leaders, using data to inform decisions, implementing policies and practices from those decisions, and continuously evaluating the impact of those policies and practices, specifically for student learning and development (Henning & Roberts, 2016; Kotter, 1996; Suskie, 2009). This vein of scholarship, while not always explicitly stated, can be a tool for those looking to promote learner-centered practices within their institution. Institutional leaders in postsecondary education may not always be equipped with the tools and knowledge to understand the practical importance of assessment for student learning. However, by using change management scholarship, we can speak to the needs of senior leaders while also engaging in intentional, ongoing, learner-centered assessment practices.

Our analysis illustrates the need for assessment scholarship and practice to be informed by theoretical and conceptual frameworks that prioritize students as learners.

Change Management, Assessment, & Institutional Culture

Scholarship on assessment and change management can also provide institutional leaders with information about how to promote institutional change. The steps to change management, just like the assessment process, require support, action, and change, publicizing good work, continuous and persistent improvement, and dedicating resources to enable faculty and staff to engage in these processes meaningfully (Henning & Roberts, 2016; Kotter, 1996; Suskie, 2009). In addition to promoting sustainable assessment practices that contribute to student learning and success, senior leaders can use assessment change management scholarship to drive institutional conversations that encourage investment from faculty and staff. Through this approach, institutional leaders can reduce the conflict, fear, complacency, and apprehension that often accompany change, particularly assessment–informed, institution-wide change (Kotter, 1996).

Limitations

There are several limitations that should be noted for this study. First, content analyses can be limited by the a) research reviewed, b) research that was missed, and c) personal biases and experiences of the researchers. As demonstrated through our positionality statements, including over thirty-six years combined experience in the field, our assumptions were acknowledged and recognized throughout the analysis. However, we recognize that other researchers may draw different conclusions. Second, our research was specific to postsecondary education focused on four-year institutions within the United States. As we previously acknowledge, there is robust literature of assessment research within an international context. Therefore, the generalizability of these findings may be limited to the United States.

Conclusion

Our analysis illustrates the need for assessment scholarship and practice to be informed by theoretical and conceptual frameworks that prioritize students as learners. Without a solid grounding in such theoretical and conceptual frameworks, approaches to assessing student learning can become reactionary, administratively burdensome, and removed from teaching and learning practices. Some critics believe that assessment activities take place at the expense of other efforts focused on individual student learning and achievement and the improvement of teaching (Gilbert, 2019; Gilbert, 2018; Worthen, 2018). We argue that assessment practices should always center on students; and if engaged holistically by informed stakeholders, can lead to institutional improvement that contributes to student learning and success (Ludvik, 2018). Engaging with scholarship on assessment and analyzing practices can help administrators, educators, and practitioners better understand and implement quality assessment across institutions and improve learning within postsecondary education.

Scholarship on assessment must continue to evolve. We hope that the field of higher education moves toward a more student-centric framework that prioritizes teaching and learning in all institutional aspects. Further, equity in assessment scholarship also means that practitioners and educators must recognize their role in advocating for a quality educational experience for students. While articulating and documenting student learning outcomes has been the expected, in some areas required, practice in postsecondary education for over twenty years, many institutional assessment practices are still in nascent stages at colleges and universities across the country. Scholarly research on the assessment of student learning, and its association with accreditation, accountability, and promoting student success, is an

essential foundation for comprehending the evolution of modern assessment practices. This foundation can provide the context for new assessment practices and frameworks that center student learning and support for an environment that fosters student success for all.

References

An Adams, P. (2014) *Policy and Education*. [VitalSource Bookshelf]. Routledge. https://bookshelf.vitalsource.com/#/books/9781136492990/

Araya, D. (2015). [VitalSource Bookshelf]. Palgrave Macmillan. https://bookshelf.vitalsource.com/#books/9781137475565/

Barr, R. B., & Tagg, J. (1995). From teaching to learning-A new paradigm for undergraduate education. *Change*, *27*(6), 12. https://doi.org/10.1080/00091383.1995.10544672

Barrett, J. M. (2012). Writing assessment in the humanities: Culture and methodology. *Journal of Assessment and Institutional Effectiveness*, *2*(2), 171-195. Penn State University Press. https://www.jstor.org/stable/10.5325/jasseinsteffe.2.2.0171

Ben-Avie, M., Kennedy, M., Unson, C., Li, J., Riccardi, R. L., & Mugno, R. (2012). First-year experience: A comparison study. *Journal of Assessment and Institutional Effectiveness*, *2*(2), 143-170. https://doi.org/10.5325/jasseinsteffe.2.2.0143

Biggs, J., & Tang, C. (2011). Teaching for quality learning at university. Open University Press.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2), 27-40. https://doi.org/10.3316/QRJ0902027

Braskamp, L. A., & Engberg, M. E. (2014, February). *Guidelines to consider in being strategic about assessment*. National Institute for Learning Outcomes Assessment. https://www.learningoutcomesassessment.org/wp-content/uploads/2019/08/Viewpoint-BraskampEngberg.pdf

Briggs, C. L. (2007). Curriculum collaboration: A key to continuous program renewal. *Journal of Higher Education*, *78*(6), 676-711. https://doi-org.proxy.library.ohio.edu/10.1353/jhe.2007.0036

Ching, C. D. (2018). Confronting the equity "learning problem" through practitioner inquiry. *The Review of Higher Education*, *41*(3), 387-421. https://doi.org/10.1353/rhe.2018.0013

Cogswell, C. A. (2016). *Improving our improving: A multiple case study analysis of the accreditor-institution relationship* [Doctoral dissertation, Indiana University]. ProQuest Dissertations Publishing.

Cumming, T., & Miller, M. D. (2017). *Enhancing assessment in higher education: Putting psychometrics to work*. Stylus Publishing.

Dorimé-Williams, M. L. (2018). Developing socially just practices and policies in assessment. *New Directions for Institutional Research*, *177*, 41-56. https://doi.org/10.1002/ir.20255

Eisner, E. W. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. Collier Macmillan Canada.

Emerson, R., Fretz, R., & Shaw, L. (1995). *Writing ethnographic fieldnotes*. University of Chicago Press.

Freed, R., & Mollick, G. M. (2010). Using prior learning assessment in adult baccalaureate degrees in Texas. *Journal of Case Studies in Accreditation and Assessment*, *1*, 1-14. https://www.aabri.com/manuscripts/08081.pdf

Fuller, C., Lebo, C., & Muffo, J. (2012). Challenges in meeting demands for accountability. In R. Howard, W. Knight, & G. McLaughlin (Eds.), *The handbook of institutional research*, (pp. 299-309). Jossey-Bass.

Gilbert, E. (2019, March). Assessment is an enormous waste of time. *The Chronicle of Higher Education*. https://www.chronicle.com/article/Assessment-Is-an-Enormous/245937

Gilbert, E. (2018, January). An insider's take on assessment: It may be worse than you thought. *The Chronicle of Higher Education*. https://www.chronicle.com/article/An-Insider-s-Take-on/242235

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/examining-construct-validity-of-the-quantitative-literacy-value-rubric-in-college-level-stem-assignments/

Harley, D. A., Jolivette, K., McCormick, K., & Tice, K. (2002). Race, class, and gender: A constellation of positionalities with implications for counseling. *Journal of Multicultural Counseling and Development*, *30*(4), 216-238. https://doi.org/10.1002/j.2161-1912.2002.tb00521.x

Henning, G. W., & Roberts, D. (2016). *Student affairs assessment: Theory to practice*. Stylus Publishing, LLC.

Hora, M. T. Bouwma-Gearhart, J., & Park, H. J. (2017). Data driven decision-making in the era of accountability: Fostering faculty data cultures for learning. *Review of Higher Education*, *40*(3), 391-426. https://doi.org/10.1353/rhe.2017.0013

Huba, M. E., & Freed, J. E. (2000). *Learner-centered assessment on college campuses: Shifting the focus from teaching to learning*. Allyn and Bacon.

Jaeger, A. J., Hudson, T. D., Pasque, P. A., & Ampaw, F. D. (2017). Understanding how lifelong learning shapes the career trajectories of women with STEM doctorates: The life experiences and role negotiations (LEARN) model. *The Review of Higher Education*, *40*(4), 477-507. https://doi.org/10.1353/rhe.2017.0019

Jankowski, N. A., Timmer, J. D., Kinzie, J., & Kuh, G. D. (2018). *Assessment that matters: Trending toward practices that document authentic student learning*. National Institute for Learning Outcomes Assessment. https://eric.ed.gov/?id=ED590514

Jones, S. R., Arminio, J. L., & Torres, V. (2013). *Negotiating the complexities of qualitative research in higher education: Fundamental elements and issues* (2nd ed.). Routledge.

Kotter, J. P. (1996). *Leading change*. Harvard Business School Press.

Lane, M. R., Lane, P. L., Rich, J., & Wheeling, B. (2014). Improving assessment: Creating a culture of assessment with a change management approach. *Journal of Case Studies in Accreditation and Assessment*, *4*, 1-11. https://www.aabri.com/manuscripts/141949.pdf

Leathwood, C. (2005). Assessment policy and practice in higher education: Purpose standards and equity. *Assessment & Evaluation in Higher Education*, *30*(3), 307-324. https://doi.org/10.1080/02602930500063876

Ludvik, M. J. B. (2018). *Outcomes-based program review: Closing achievement gaps in and outside the classroom with alignment to predictive analytics and performance metrics*. Stylus Publishing, LLC.

Lusher, A. L. (2010). Assessment practices in undergraduate accounting programs. *Journal of Case Studies in Accreditation and Assessment*, *1*, 1-20. https://www.aabri.com/manuscripts/10550.pdf

Mansilla, V. B., Duraisingh, E. D., Wolfe, C. R., & Haynes, C. (2009). Targeted assessment rubric: An empirically grounded rubric for interdisciplinary writing. *Journal of Higher Education*, *80*(3), 334-353. https://doi-org.proxy.library.ohio.edu/10.1080/00221546.2009.11779016

Merriam-Webster. (n.d.). *Merriam-Webster.com dictionary*. Retrieved May 15, 2020, from https://www.merriam-webster.com/dictionary/policy

McArthur, J. (2016). Assessment for social justice: The role of assessment in achieving social justice. *Assessment & Evaluation in Higher Education*, *41*(7), 967-981, DOI: 10.1080/02602938.2015.1053429

McDonnell, L. M. (1994). Assessment policy as persuasion and regulation. *American Journal of Education*, *102*(4), 394-420. https://www.jstor.org/stable/1085462

McNair, T. B., Albertine, S., Cooper, M. A., McDonald, N., & Major, T., Jr. (2016). Becoming a student-ready college: A new culture of leadership for student success. John Wiley & Sons.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: A sourcebook of new methods* (2nd ed.). Sage.

Mitchell, D. E., Shipps, D., & Crowson, R. L. (Eds.). (2018). *Shaping education policy: Power and process.* (2nd ed.). Routledge.

Moutsios, S. (2010). Power, politics, and transnational policy-making in education. *Globalisation, Societies and Education*, *8*(1), 121-141. https://doi.org/10.1080/14767720903574124

Pastor, D. A., Ong, T. Q., & Orem, C. D. (2018). Categorizing college students based on their perceptions of civic engagement activities: A latent class analysis using the social agency scale. *Research & Practice in Assessment*, *13*, 5-21. http://www.rpajournal.com/categorizing-college-students-based-on-their-perceptions-of-civic-engagement-activities-a-latent-class-analysis-using-the-social-agency-scale/

Pereira, D., Flores, M. A., & Niklasson, L. (2016). Assessment revisited: A review of research in Assessment and Evaluation in Higher Education. *Assessment & Evaluation in Higher Education*, *41*(7), 1008-1032. https://doi.org/10.1080/02602938.2015.1055233

Pieper, S. L., Fulcher, K. H., Sundre, D. L., & Erwin, T. D. (2008). "What do I do with the data now?": Analyzing assessment information for accountability and improvement. *Research & Practice in Assessment*, *3*, 4-10. http://www.rpajournal.com/what-do-i-do-with-the-data-now-analyzing-a-comparison-of-testing-conditions-and-the-implications-for-validity/

Price, M. H. (2019). Strategic stories: Analysis of prior learning assessment policy narratives. *The Review of Higher Education*, *42*(2), 511-535. https://doi.org/10.1353/rhe.2019.0005

Ro, H. K., & Loya, K. I. (2015). The effect of gender and race intersectionality on student learning outcomes in engineering. *The Review of Higher Education*, *38*(3), 359-396. https://doi.org/10.1353/rhe.2015.0014

Saldaña, J., & Omasta, M. (2016). *Qualitative research: Analyzing life*. Sage Publications.

Smiley, W., & Anderson, R. (2011). Measuring students' cognitive engagement on assessment tests: A confirmatory factor analysis of the short form of the cognitive engagement scale. *Research & Practice in Assessment*, *6*, 17-28. http://www.rpajournal.com/measuring-students-cognitive-engagement-on-assessment-tests-a-confirmatory-factor-analysis-of-the-short-form-of-the-cognitive-engagement-scale/

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

Taylor, S., & Bogdan, R. (1998). *Introduction to qualitative research methods: A guidebook and resource*. Wiley.

U.S. Department of Education. (2006). *A Test of Leadership: Charting the Future of U.S. Higher Education*. https://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/pre-pub-report.pdf

Warburton, E. C. (2018). Toward trust: Recalibrating accreditation practices for postsecondary arts education. *Arts Education Policy Review*, *119*(1), 36-41. https://doi.org/10.1080/10632913.2016.1189864

White, J. W., & Lowenthal, P. R. (2011). Minority college students and tacit "codes of power": Developing academic discourses and identities. *The Review of Higher Education 34*(2), 283-318. https://doi.org/10.1353/rhe.2010.0028

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

Zumeta, W., & Kinne, A. (2011). Accountability policies: Directions old and new. In D. E. Heller (Ed.), *The states and public higher education policy: Affordability, access, and accountability*, (pp. 173-199). Johns Hopkins University Press.

Abstract

Data from program assessment in higher education are often used for accreditation purposes and are less focused on decision-making and program improvement. This article illustrates how data-informed decisions were made in a teacher education program. It details how a framework of assessment, Plan Do Study Act, was used to identify areas in need of attention and how the program made feasible incremental changes to its curriculum and assessment process over a three-year period to improve students' learning.

AUTHORS

Ya-Chih Chang, Ph.D.
California State University, Los Angeles

Holly M. Menzies, Ph.D.
California State University, Los Angeles

Data-Informed Decision-Making in Higher Education: Lessons from a Teacher Education Program

As knowledge and skill in assessment practice matures, universities are fostering processes better aligned for use at the program and instructor level. There is a long-standing tension between institutional assessment for improvement and its use for accreditation purposes. While these purposes should be the same, they are more often seen as a "contradiction" (Ewell, 2009, p. 7) because of the accountability mandates inherent in accreditation requirements. However, there has been a shift toward what Ewell (2009) characterized as the "improvement paradigm" and away from the "accountability paradigm."

The focus of an improvement paradigm is for faculty to identify and collect evidence of student work to examine whether students are mastering course and programmatic outcomes and determine whether changes are needed to improve student learning. This contrasts with an accountability stance where the purpose is to signal to an external audience the worthiness of the institution, typically through standardized measures or institutional-level metrics (Ewell, 2008). Blaich and Wise (2010) were among the first to assert that assessments do not necessarily lead to improved learning. They emphasized the importance of assessment but also noted that assessment processes in higher education are frequently more political than data-driven; thus, assessment leaders must understand the social, political, historical, and budgetary context of the institutions to make pragmatic choices about which assessments to administer. This may mean collecting data that are most interesting to faculty, even if it does not directly result in student learning improvement.

Other researchers have proposed assessment models that address the importance of student learning outcomes, including "closing the loop" (Banta & Blaich, 2011), learning improvement (Stitt-Berg et al., 2018), Program Learning Assessment, Intervention, Re-assessment (PLAIR; Fulcher et al., 2014), and Plan, Do, Study, Act

CORRESPONDENCE

Email
ychang27@calstatela.edu

(PDSA; Moen, 2009). These different methods are suited for use at the local level to improve student outcomes. Each foregrounds a practical approach to collecting evidence, making changes, and evaluating the impact of those changes, although each method employs slightly different tactics. Additionally, they all emphasize direct assessment of student learning and a formative approach to instructional and program improvement. For example, the PLAIR model focuses on change and intervention, instead of only on the assessment methodology (Fulcher et al., 2014). Implementing this model may take a few years as programs must assess, identify the area that needs improvement, develop and implement the appropriate intervention, and then reassess to determine whether student learning improved. The PDSA framework is a similar model that also uses a "closing the loop" approach (Moen, 2009). However, it operates on a faster timeline than the PLAIR model. For example, the PDSA allows programs to make changes to the program before the first cohort of students graduates, depending on what is assessed.

A major benefit of a PDSA cycle is the focus on rapid and iterative change. If a change does not result in the desired outcome, another change can be tested.

This case study reports program-level efforts to use a PDSA framework to improve student learning in an Early Childhood Special Education (ECSE) program where students earn a state-issued credential to teach in an early childhood setting and work with young children with disabilities. As faculty adopted formative processes to make programmatic decisions, they faced measurement issues and implementation challenges in making meaningful changes to courses. This paper details the assessment process and decisions made based on program-level information collected about student learning.

Plan, Do, Study, Act

The PDSA cycle was originally used in business to emphasize continuous improvement and was popularized by Charles Deming in the mid-twentieth century (Moen, 2009). It consists of four elements for making iterative changes as part of an ongoing information collection and analysis cycle. The first step is to *Plan* a change that will improve an outcome. The change can be based on formal quantitative data collection, but it can also be from qualitative data collected from faculty or students. While it is important that data have integrity, it is also important not to wait for perfect information as the data or tools used to collect it will invariably have some flaws (Berwick, 1996). The second step, *Do*, is to put the change into place. Once information or data have been collected on the results of the change, you Study to determine whether a positive change has occurred in the targeted outcome. *Act* is deciding whether to permanently adopt the change or try a different change based on the analysis. It can also mean the continuation of data collection for progress monitoring.

A major benefit of a PDSA cycle is the focus on rapid and iterative change. If a change does not result in the desired outcome, another change can be tested. The cyclical nature of the model capitalizes on the use of ongoing information collection and analysis for monitoring improvement. The PDSA model can help programs plan and implement beneficial changes, and the process itself is straightforward.

Context and Process

Institutional Context

The university is located in a dense urban area in the West. It serves a diverse population of primarily first-generation college students. After the university's 2010 accreditation by the Accrediting Commission for Schools, Western Association of Schools and Colleges (WASC), additional resources were directed to assessment practices. Assessment efforts and resources included the establishment of a university-wide committee to promote the use of assessment, a variety of training offered to faculty (which is ongoing), and the formation of an assessment committee within each college. A member of each college committee sits on the university committee to facilitate assessment work across the institution.

The College of Education assessment committee is comprised of two representatives from each of the three departments in the college, the chair of each department, and the associate dean of the college, who serves as the committee chair. The committee predates the 2010 WASC visit as there are over 25 accredited programs in the college, and assessment work was integral to maintaining accreditation status. While compliance concerns were

undoubtedly a driver of instituting assessment practices, programs were encouraged to develop systems tailored to their specific needs. Also, programs had different assessment requirements depending on their accreditation body, so flexibility in assessment practices was crucial.

A major focus of the committee is to guide programs in using data-informed decision-making to improve teaching and learning and address accreditation requirements. Committee members attended assessment-related trainings and read and discussed a series of books and articles on assessment. Additionally, the state university system provided the committee and the associate dean with coaching in improvement science to enhance their ability to improve student outcomes. The PDSA model was one method the committee had explored together. Programs had encountered various challenges using data to make curricular or instructional changes as faculty tended to focus on the quality and quantity of the data instead of "closing the loop." While validity is an important consideration, it had become a barrier to substantive improvement. PDSA was a low-stakes method for focusing on actionable and formative information. Accreditation requirements are typically focused on global indicators such as pass rates on state-mandated exams and the percentage of students successfully completing fieldwork, but these did not provide the more granular information needed for program improvement. However, accreditation for teacher credentialing did require programs to demonstrate how they collected data to make programmatic decisions.

Department Assessment Processes

The Department of Special Education and Counseling includes several special education credential options (e.g., ECSE, visual impairments, mild/moderate support needs, and extensive support needs). Each of the special education program options evaluated candidates during their fieldwork practicum using similar formative and summative measures to allow comparison across options while still evaluating competencies specific to the credential area. The program coordinator for each option facilitated data collection and then aggregated and analyzed the data. Results from each program were discussed annually in a department meeting. Next, we describe how one of these options, ECSE, used the PDSA cycle to make data-informed changes for program improvement.

Early Childhood Special Education Program

Overview of the Program

The ECSE Program is a two-year program that prepares teacher candidates to serve young children (age 0-5) who are at risk or with a disability. The program follows an intentional sequence of coursework to first introduce candidates to foundational knowledge in disability, characteristics of children with special needs, special education law, first and second language acquisition in the context of cognitive development, social emotional development, and classroom management and positive behavior support. Subsequent coursework uses this grounding as context for developing knowledge for assessing, planning, and providing learning opportunities for infants, toddlers, and preschoolers. Candidates demonstrate knowledge and teaching competencies in a final student teaching experience during their last term in the program where they are placed in an early childhood setting with young children (ages 0-5) with and without disabilities.

Programs had encountered various challenges using data to make curricular or instructional changes as faculty tended to focus on the quality and quantity of the data instead of "closing the loop."

Student Teaching Fieldwork Course

Approximately 25-30 candidates enroll in the student teaching fieldwork course each year. Most fieldwork placements are in high-needs schools (e.g., low SES, Title I). At the end of the course, candidates are expected to demonstrate teaching competencies in the areas of Assessment, Curriculum, Managing the Teaching and Learning Environment, and Collaboration and Professionalism to be recommended for a credential to the state credentialing authority.

University Supervisors

All university clinical supervisors hold an ECSE credential or have experience in ECSE. The program coordinator meets individually with newly hired university supervisors to review the fieldwork requirements and explain how to administer the assessment measures. There is a cadre of experienced supervisors who are assigned to candidates each semester and occasionally a new supervisor is hired. Most clinical supervisors are adjunct faculty, but also include tenure line faculty. University supervisors meet with each candidate a minimum of six times over the course of the term. When conducting an observation, the university supervisor completes a formative assessment measure to evaluate and provide feedback to the candidate on their teaching. This measure is completed electronically, making it simple to collect, aggregate, and analyze the data at the end of each term. A single summative measure is completed at the end of the term to indicate the candidate's proficiency level for each of the competency domains. Candidates also receive structured feedback from their cooperating teacher, or supervising administrator, if they are interns.

A new observation form was developed that included the same global domains as the summative measure but comprised discrete items that could be rated to provide more specificity about candidate performance.

Measures

The Early Childhood Special Education (ECSE) program uses both summative and formative assessment tools to evaluate the program, provide feedback, and determine whether candidates meet program competencies.

Summative measure

The summative rubric assesses skills in four competency domains: Assessment, Curriculum, Managing the Teaching and Learning Environment, and Collaboration and Professionalism. A five-point scale ranging from 1 (Preliminary) to 5 (Mentor Level) is used to describe performance. Each level includes a detailed narrative description of performance. Expected performance at the end of the semester is a score of 3 in each domain with a total summed score of 12. Mentor level is included because many candidates are interns and have considerable teaching experience by the time they complete their program.

Formative measure

Originally, supervisors provided only written feedback after each observation; however, this was cumbersome to aggregate and report and challenging to track over time. A new observation form was developed that included the same global domains as the summative measure but included discrete items that could be rated to provide more specificity about candidate performance. For example, in the assessment domain, one item is "Provides timely and high-quality feedback to students about lesson content/material." Supervisors rate performance on each item using a scale ranging from 1 (does not meet standard) to 4 (exceeds standard). Items can also be rated as not applicable (N/A). The formative measure includes areas for narrative comments and collects demographic information, such as whether the candidate is an intern or a traditional student teacher. This allows the program to analyze data with more precision (See Appendix 1 for Sample Items from Formative Assessment).

Data Analysis

We report retrospectively on the program improvement process and describe the steps taken for program improvement. This endeavor was not originally conceived as a research study, so an a priori data analytic plan was not created. However, the study details the processes in how we implemented a new model to inform programmatic changes.

Each term, the formative measure data uploaded by an individual university supervisor was retrieved and stored on the department's SharePoint site. The data from the summative measure were entered into an Excel spreadsheet as the form was completed on a paper copy. It, too, was stored on SharePoint. At the end of each academic year, the data are aggregated and examined by the tenure line faculty in the program.

Data described in this case study were collected from five different supervisors who observed two to three candidates each term (approximately 25-30 candidates per academic year). Supervisors completed the formative measure at each observation and the summative

measure at the end of the term. Descriptives, including means and standard deviations, were calculated for approximately 100 observations yearly over the course of three years.

The ECSE PDSA Cycle

Below we describe the three-year iterative process of using fieldwork assessment data to initiate and evaluate programmatic changes (See Table 1). These included the development of a new clinical course, modification of assignments and readings in existing courses, and refresher training for university supervisors.

Table 1. PDSA Cycle

	Year 1	**Year 2**	**Year 3**
Plan	Assess how well the program prepared candidates in working with young children with and without disabilities in naturalistic classroom settings.	Assess how well the program prepared candidates in working with young children with and without disabilities, specifically in two areas: 1. Candidates' competency in assessments 2. Candidates' understanding and effective use of technology in early childhood classrooms	Assess how well the program prepared candidates in working with young children with and without disabilities. 1. Continue to monitor candidates' competency in assessments 2. Establish university supervisors' reliability and consistency in scoring students' use of technology in classrooms
Do	Data collection using both formative and summative assessments during final fieldwork.	Data collection using both formative and summative assessments during final fieldwork.	Data collection using both formative and summative assessments during final fieldwork.
Study	1. Assessments had relatively low scores compared to the three other domains that were evaluated during fieldwork. 2. University supervisors rated "N/A" in student teachers' use of technology.	1. Candidates demonstrated an increase in their competencies in the domain of Assessments. 2. University supervisors continued to rate "N/A" in student teachers' use of technology.	1. Candidates continued to demonstrate competencies in the domain of Assessments. 2. There was a significant decrease in "N/A" ratings on student teachers' use of technology.
Act	Modify two courses in the program to increase candidates' competency in assessment practices and effective use of technology.	Redesign supervision training to improve supervisors' administration of the formative assessment.	No modifications were made. Continue to monitor candidates' competency in all domains.

Year 1

Plan

In Year 1, the faculty decided to triangulate the results of both the formative and summative measures to determine how well the program prepared candidates to work with young children with and without disabilities in naturalistic classroom settings.

Do

Each term, university supervisors used the newly created formative measure to provide feedback to candidates during each in-person observation. They also used the summative assessment rubric to determine whether candidates met all teaching competencies at the end of the term. The data for the formative measure were entered electronically during each visit, and the summative measure was completed by hand using a paper form and later entered into an Excel database. At the end of the academic year, the program coordinator aggregated and analyzed both datasets.

Study

Summative data indicated candidates scored relatively low in the domain of Assessment. This domain included: (a) the selection and use of multiple, appropriate, formal and informal non-biased assessment tools with consideration of cultural, linguistic, and ability status across developmental and educational domains; (b) monitoring of student's progress regularly with data-based, anecdotal, and authentic input from all team members; and (c) appropriate adaptation of student programs in response to regular assessment of progress across developmental and academic domains. Average scores ranged between 3.7-4.0 (Advanced/ Independent Level) for all domains except in *Assessment*, where they were at approximately 3.2 (Proficient/ Beginning Teacher - Advanced/ Independent Level), a noticeable difference in contrast to average scores in the other domains.

The results helped faculty identify two areas for program improvement for the following year: 1) Competency in assessment practices and 2) Effective use of technology in early childhood classrooms.

The formative measure indicated lower scores in using formative data to develop lesson plans aligned to the Preschool Learning Foundations and providing specific feedback to children. It was also evident that university supervisors frequently used the "N/A" descriptor on items related to technology in the classrooms. These items were "Effectively uses varying levels of technology (low tech/high tech) to meet student needs specific to classroom management and whole class participation" and "Integrates technology (low tech/high tech) to enhance student engagement and address learner needs specifically to lesson/content learning." This was a concern as it signaled that technology was either not being adequately used by candidates or supervisors were having difficulty distinguishing what constituted technology use.

Act

The results helped faculty identify two areas for program improvement for the following year: 1) Competency in assessment practices and 2) Effective use of technology in early childhood classrooms. The faculty in the ECSE program decided to make modifications to two courses in the program to increase candidates' competency in the identified areas of concern.

Year 2

Plan

In Year 2, the program continued to collect assessment data on how well the program prepared candidates. In addition, they made modifications to two courses as decided at the end of Year 1. The first modification was made to the assessment course which focuses on understanding how different assessments are used in early childhood, including standardized and formative assessments. The final assignment for the course is to write an assessment report of a young child (ages 0-5) with disabilities and include three clear goals based on data collected during the term. Candidates are required to use three different types of assessments (e.g., parent interview, classroom observation, standardized assessment) and include progress monitoring data in the appendix of the report. This assignment was modified in Year 2 to include a reflection on the type and usefulness of the data collected with the goal of making candidates more intentional about their assessment practice.

The second change was the addition of a new clinical course. Its purpose was to bridge knowledge and clinical practice by providing candidates with additional hands-on experience working with children with and without disabilities in a diverse inclusive community enrichment program. Candidates were to enroll in it during the first year of the program

and it was to be taken in conjunction with a methods course that included topics such as classroom management, routine building, early language and literacy, play, and technology use in early childhood settings. To address the issues identified about assessment and technology use, candidates were asked to write developmentally-appropriate lesson plans and implement evidence-based strategies under the supervision and coaching of course instructors. Candidates were required to assess children's understanding of the lessons and monitor children's progress throughout the term. Candidates were also expected to use technology (e.g., short videos) in their lessons each week. Readings on technology use in early childhood settings were assigned to increase candidates' understanding and effective use of technology and provide candidates with different examples of technology use in early childhood settings.

Do

University supervisors continued to use formative and summative assessments to determine whether candidates met teaching competencies, the new course was offered, and the proposed curricular changes were made.

Study

Data from the summative assessment showed that on average, the candidates' scores in Assessment were higher. Average scores were now in the 4.4 (Advanced/ Independent Level), comparable to those in the other three domains. This suggested that the course modifications increased candidates' understanding and practice in using assessments for planning instruction.

However, there was no change in average scores for the technology items on the formative measure. Despite the additional course readings and practice of technology use in the Early Intervention Lab course, university supervisors continued to frequently rate "N/A" in candidates' technology use.

Act

Candidates demonstrated an increase in their scores in the domain of Assessment; therefore, the course changes were made permanent. Candidates' use of technology remained an area of concern. It was hypothesized that candidates may have been effectively using technology, especially after the curricular modifications from the previous year, but the problem may be one of measurement. Therefore, the faculty decided to refresh university supervisors' knowledge of how to score candidates' use of technology to improve their administration of the formative measure.

Year 3

Plan

In Year 3, the program faculty continued to collect and monitor assessment data. They redesigned the supervisor training process to improve reliability and consistency in scoring candidates' use of technology in the classrooms.

Data from the summative assessment showed that on average, the candidates' scores in Assessment were higher. Average scores were now in the 4.4 (Advanced/ Independent Level), comparable to those in the other three domains.

Do

University supervisors were invited for refresher training on using the formative assessment rubric for final student teaching. The program coordinator led the training which was approximately two hours long. University supervisors were provided with a small stipend to attend the training. Prior to the training date, supervisors were sent two 15-minute videos of an ECSE lesson. Each supervisor was asked to evaluate the two videos using the formative measure and provide their ratings.

The goal was to re-calibrate scoring across all university supervisors in the program. At the first meeting, program coordinators emphasized the items where there were substantial differences in the use of the N/A category. Supervisors were reminded that any use of technology, both low tech (e.g., individualized communication boards) or high tech (e.g., use of iPad for short videos), should be scored. Additionally, specific descriptions for each rating

level (e.g., 1, 2, 3, and 4) were added for each item based on the discussion at the meeting. See Appendix for examples of descriptors added. Using the revised descriptions, the group re-scored the first two videos and came to a consensus on their ratings.

The PDSA cycle was a valuable process for making meaningful and consequential changes for program improvement.

After the meeting, the university supervisors were provided with another set of two training videos and were again asked to rate the videos using the revised rubric. The program coordinator reviewed the ratings and found the ratings to be more reliable across university supervisors.

Study

At the end of Year 3, candidates' average scores on the summative assessment in Assessment continued to be in the range of 4 (Advanced/ Independent Level), as did the other three domains. There was a significant decrease in the rating of "N/A" on the formative assessment for technology use. Instead, students were rated highly for using appropriate technology to advance the quality of their lessons.

Act

No curricular or program modifications were planned for the following year because average scores had increased to an acceptable level. Data collection and analyses were retained to examine program quality on an ongoing basis.

Discussion

Using the PDSA model, program faculty were able to identify areas in need of change and make improvements over the course of three years. This process was useful for accreditation purposes because it documented how programmatic changes were data-informed, but more importantly, it enabled incremental and feasible changes that improved the *quality* of the program.

In the first two years, course improvements were made that addressed program competencies in using assessments, including collecting and analyzing data on young children's skills in various developmental domains. Assignments and courses were modified or added to provide more contextual opportunities for candidates to practice these specific skills. For example, an early intervention lab was instituted that required candidates to collect data about their students and monitor progress over the course of the term. Student teachers identified areas of need, developed lessons, and implemented them with embedded learning opportunities to meet their young students' individual needs. The early intervention lab was offered during the first year of the program offering candidates the opportunity to practice assessment skills before being evaluated for competency in the fieldwork practicum. In the third cycle, a measurement issue was identified. After two years of curriculum modifications to address the use of technology in early childhood classrooms, university supervisors were still frequently using the N/A category instead of rating candidates' level of proficiency on the technology items. The program decided to retrain university supervisors in identifying and evaluating candidates' use of technology. The PDSA process made it possible to bridge the divide that sometimes occurs between content learned in coursework and evaluation of its application in practice. In this case, once university supervisors received additional training on the topic of technology aligned to the coursework and understood what to look for, they were far less likely to use the N/A category.

The PDSA cycle was a valuable process for making meaningful and consequential changes for program improvement. It can be challenging to know where to start with assessment, especially when faced with the task of making changes to a program that has several courses, a variety of fieldwork experiences, and many instructors. However, using the PDSA cycle made the endeavor both manageable and productive.

References

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

Berwick, D. M. (1996). A primer on leading the improvement of systems. *BMJ*, *312*(7031), 619-622.

Blaich, C. F., & Wise, K. S. (2010). Moving from assessment to institutional improvement. *New Directions for Institutional Research*, *2010*(S2), 67-78.

Ewell, P. T. (2008). *Assessment and accountability in America today: Background and context*. In V. M. H. Borden & G. Pike (Eds.), Assessing and accounting for student learning: Beyond the Spellings Commission (New Directions for Institutional Research, Assessment Supplement 2007, pp. 7-18). Jossey-Bass.

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

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 (NILOA).

Moen, R. (2009). Foundation and history of the PDSA cycle. https://deming.org/wp-content/uploads/2020/06/PDSA_History_Ron_Moen.pdf

Stitt-Bergh, M., Kinzie, J., & Fulcher, K. (2018). Refining an approach to assessment for learning improvement. *Research & Practice in Assessment*, *13*, 27-33.

Appendix

Appendix 1. Sample items from the formative assessment

1=does not meet standard 2=approaching standard 3=meets standard 4=exceeds standard NA=no lesson plan

	1	2	3	4	NA
Integrates technology (low tech/ high tech) to enhance student engagement and address learner needs, ***specifically to lesson/ content learning.*** **should have conversation with student candidate if they are not using technology to enhance lesson* 1= no use of technology, but should have 2= have technology but did not use appropriately 3= have technology AND used it appropriately 4= Appropriate use of technology that advanced the quality of the lesson and was accessible to all students	( )	( )	( )	( )	( )
Provides timely and high quality feedback to students about ***lesson/content material.*** 1= None 2= Responds to student 3= Responds AND embeds strategies 4= Consistently responding and embedding opportunities throughout activities AND provides additional content information	( )	( )	( )	( )	( )

Abstract

Many universities shifted how students were assessed during the COVID-19 pandemic. This movement to online learning altered the format of some assessments that were previously administered in-person and proctored. Since the start of COVID-19 in 2020, James Madison University (JMU) shifted some assessments to an unproctored internet testing (UIT) format. The bi-annual, university-wide Assessment Day was one such set of assessments that underwent the change to UIT at JMU. As we interpret scores from those UIT administrations and contemplate future changes, it is important to understand what the experience was like for the students. At the end of their battery of assessments, students were asked to share their thoughts and suggestions. The current study employed a conventional content analysis to code responses to this item for two recent Assessment Days. About 20% of students responded to the item, of which many of the comments were generally positive and said something positive about UIT specifically. Few comments were negative. This study highlights the positive impact of UIT on our campus. We aim to continue incorporating the student perspective into our assessment process.

AUTHORS

Katarina E. Schaefer, M.A.
James Madison University

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

Samantha N. Harmon, M.A.
James Madison University

Examinee Perspectives on Unproctored Internet Testing

COVID-19 prompted assessment professionals in higher education to make quick decisions; decisions that would typically take months or years to finalize happened in a matter of weeks. Moreover, rather than making a single change, many universities and other higher education programs had to make multiple, drastic changes or completely reconstruct their traditional instruction and assessment processes altogether. Unprecedented times became the norm. Faculty began teaching online with varying comfort levels, assignments and assessments were modified, deadlines were extended, alternative assessments were assigned, previously proctored assessments were unproctored, and empathy toward students increased (Jankowski, 2020a, 2020b; Pastor & Love, 2020). Large-scale university admissions testing began to offer at-home testing, which had been an impossible thought only weeks before (Camara, 2020). Looking back on the 2020-2021 school year, we wonder: how do we interpret the data that came from such a hectic time?

Assessment professionals voiced their worry about the shift to online instruction and testing. Many worried that cheating would increase, students would be less motivated academically, or students would perform poorly due to increased anxiety (Jankowski, 2020b). Some speculated that the validity of the results of assessments administered after this quick transition would be lowered. However, Fulcher and Leventhal (2020) and Busby (2020) stressed that testing can and should go on despite these fears. They emphasized that it is still important to track student knowledge. Without continued testing, we would be unable to understand whether students gain, maintain, or lose knowledge due to the drastic changes that have taken place since COVID-19 began. Continued testing also

CORRESPONDENCE

Email
schae2ke@jmu.edu

provides the opportunity to explore the effects of the pandemic on the validity of assessment scores and whether the potential effects are the same for all students.

Without continued testing, we would be unable to understand whether students gain, maintain, or lose knowledge due to the drastic changes that have taken place since COVID-19 began.

Recognizing the benefits of continued assessment during the pandemic, many higher education academic and non-academic programs alike made changes to their assessment procedures. Whether assessment was administered for academic degree programs, student affairs programs, campus initiatives, etc., many institutions chose to administer assessments in a new, UIT format. At James Madison University (JMU), one such set of assessments that shifted to UIT was the biannual, low-stakes Assessment Day (Pastor & Love, 2020). These assessments have been administered both proctored and in-person for over 30 years. About 4,000 students are assessed on a typical Assessment Day throughout two three-hour sessions. Assessments are "low stakes" for students because they have no direct personal consequences to the student. Although university-wide Assessment Days are not common, smaller scale assessment of student learning is routine in higher education. Similar to other higher education programs navigating the pandemic, our university knew that these assessments could not be administered in-person and proctored.

All assessments were administered remotely and unproctored during the Fall 2020 and Spring 2021 Assessment Days to reduce exposure to COVID-19, a change from previous years. Test length and content during these administrations was identical to in-person administrations from previous years. Given the numerous differences in administration format (e.g., remote vs. in-person; proctored vs. unproctored) and context (pre-COVID-19 vs. COVID-19), we anticipated that the results from Fall 2020 and Spring 2021 would differ from those in the past. Indeed, initial reports of first-year students' scores from Fall 2020 Assessment Day reveal that overall, students seem to have much more varied scores compared to in-person administrations (Alahamadi & DeMars, 2021). The tests considered in the study included a history test, a global issues test, and a test of scientific reasoning. On the scientific reasoning assessment, which contains more items and is more cognitively demanding than the other tests, Alahamadi & DeMars (2021) reported that first-year students did much worse than expected in Fall 2020 (during COVID-19) compared to the four previous years' students.

Looking only at the numbers, we know scores were affected for at least some students, with a more pronounced effect for one assessment. However, though we might speculate how students were affected given the data, the only people who know the entire story are the students who experienced those assessments.

What students thought about the use of UIT for Assessment Day was particularly important because it was a considerable departure from the norm and results from previous studies were mixed.

Before COVID-19, higher education assessment professionals had already considered what it would mean for their programs to integrate the student perspective into their practice. The leading voices of diversity, equity, and inclusion in assessment have emphasized that the student perspective must be considered (Jankowski, 2020a, 2020b; Montenegro & Jankowski, 2020). Jankowski (2020a, 2020b) has emphasized that it is even more important to consider the student perspective during these unprecedented times. Such calls motivated us to obtain the students' perspective on their assessment experience in general and, more specifically, their take on the remote administration format. What students thought about the use of UIT for Assessment Day was particularly important because it was a considerable departure from the norm and results from previous studies were mixed. Some research shows that students have generally had positive online testing experiences (e.g. Milone et al., 2017), although some report negative experiences with proctors in online testing (Karim et al., 2014). At JMU, assessments were unproctored so we expected students to have a generally positive experience, but did not know for sure. We also did not know how COVID-19 would affect their experience without asking them – so we did.

Method

Procedures & Sample

Data were collected during the Fall 2020 and Spring 2021 Assessment Days, which were forced to use UIT due to COVID-19. Incoming first-year students were assessed in Fall 2020, and sophomore students[1] with 45-70 credit hours were assessed in Spring 2021.

[1]Although some students in this credit hour range are juniors, we refer to students who completed the spring assessment as sophomores throughout this article.

Students were assigned to a battery of online assessments during both Assessment Days. In a video describing the content of the last assessment, examinees were informed that the testing format differed from the typical in-person, proctored experience. Additionally, examinees were told they would be asked to describe their Assessment Day experience and provide suggestions for improvement during the last assessment. Examinees were asked to respond to the following questions at the end of the assessment: "Want to tell us about your Assessment Day requirement experience? Have suggestions for how to improve the Assessment Day requirement? If so, please share your experience and/or suggestions below." The item did not inquire about UIT specifically to avoid leading students to mention something about UIT.

A little less than 20% of examinees (including first-years[2] and sophomores[3]) responded to the question, yielding 1,421 responses. The first-year and sophomore samples were 63% female and 77% White, with all other races and ethnicities representing less than 10% of the sample. These distributions align with those for undergraduates at the university overall during the 2020-2021 academic year (58% female, 75% White).

Analysis

Meaning was extracted from the responses through a conventional content analysis (Hsieh & Shannon, 2005), which is appropriate when the goal is to allow themes to emerge from the data. Although we anticipated some generally positive responses, we did not want to constrain the categorization of responses to our preconceived notions; instead, we wanted to allow themes to "flow from the data" (Hsieh & Shannon, 2005, p. 1279). Two authors read two different sets of comments from 50 randomly selected first-year students and separately created initial codes to begin the analysis. After discussing the initial codes, the final set of codes and their descriptions were created. Example responses for each code were identified along with responses for training purposes. The remaining author and two additional raters were trained to use the codes, with each of the five raters assigned an equal number of responses. Although not ideal, to make the workload manageable, all first-year responses were rated first (before collecting the sophomore data) and all sophomore responses were rated second. Thus, raters were aware of the class level of the students during rating.

All four raters independently coded 100 of the same randomly selected first-year student comments to compute intercoder reliability. O'Connor and Joffe (2020) report that all raters typically code between 10%-25% of the same data to estimate intercoder reliability. For this study, all raters independently coded roughly 7% of the same comments. After all responses were coded, the responses associated with each code were reviewed by the study authors, resulting in the creation of subcategories and the merging of two initial codes. The number of responses classified into each code and subcategory was then tallied.

Table 1 contains the codes, representative examples of text to describe each code, and two indices of intercoder reliability calculated using all four raters. These code descriptions were used to train all raters. Although "Assessment Content" and "Assessment Format" were merged during the review process, estimates of intercoder reliability were calculated separately for these codes. In addition to the percent agreement index, Gwet's AC1 is provided. Gwet's AC1 differs from percent agreement because it corrects for chance agreement and is preferable to many alternatives (Gwet, 2014). Intercoder reliability indices were favorable, with values > .92 for 11 of the 13 codes.

[2]The responses from only those examinees who completed testing by the extended deadline were used in this study. Of the 3,847 incoming first-years required to participate in Assessment Day and assigned to complete the assessment used in this study, 3,408 completed the assessment on which the item was administered by the extended deadline. Out of these 3,408 examinees, 718 provided responses to the item. Thus, 21% (718/3408) of the incoming first-years who completed the assessment by the extended deadline provided a response.

[3]The responses from only those examinees who completed testing by the Assessment Day deadline were used in this study. Out of the 3,524 sophomores required to participate in ADay, 3,174 completed the assessment on which the item was administered by the deadline. Of these 3,174 examinees, 703 provided responses to the item. Thus, 22% (703/3174) of the sophomores who completed the assessment by the deadline provided a response.

Table 1
Codes, Definitions, and Average Intercoder Reliability Estimates

Code	Representative Examples of Text	Gwet's AC_1		%
Positive	Overall liked their experience. Felt excited to start school. Felt like they knew what to anticipate. Had a predominately good experience. Didn't feel overwhelmed.		0.71	85.2%
Neutral	Overall didn't have strong feelings one way or another about the assessment. Felt like they had enough time to complete it. Said "It was ok." It was uneventful.		0.82	87.0%
Negative	Overall didn't like something about ADay. Had a mainly bad experience. Their assessments took too long. Felt overwhelmed. They didn't care about this.		0.93	94.2%
Online Positive	Liked the online format. Liked that they could spend as much time as they wanted on the assessments. Didn't feel overwhelmed specifically because it wasn't in person. They don't have to explicitly mention the online or remote format.		0.92	94.8%
Online Negative	Didn't like the online format. Would rather be in person.		0.98	98.2%
Communication/ A-Day Purpose	Felt like they didn't receive enough information about ADay. Something they said could have been changed if they'd read the emails/received more emails. Would like to know more about why it's important. Would want to know why they should feel motivated to do the assessments.		0.96	96.2%
Low Motivation	Didn't feel motivated to do well. Didn't try their hardest.		0.96	96.2%
Stressed	Said they had a lot going on at the time. This added a lot to their plate. They were dealing with lots of stress (COVID-19 related or not).		0.98	98.0%
Performance concern	Don't think they did well. They think something affected their performance today. Didn't feel prepared. They want to know their scores.		0.94	94.7%
Assessment Format/Content	Comment on a specific aspect of the test. Offers suggestion to the format. Said something was too long. Wished there were less of a type of question (multiple choice, short answer, etc.). Comment on the content of the text related to how questions were asked, what questions were asked, or the difficulty. Mentioned grammar or spelling mistakes.	Format	0.92	92.8%
		Content	0.97	97.8%
Flag	Student brings up something concerning.		0.96	96.2%
Other	Noteworthy information in response not captured by other codes.		0.92	94.8%

Note. % = percent agreement.

The percentage of comments classified as conveying something positive about UIT was 15.6% and 28.7% for first-year and sophomore students, respectively.

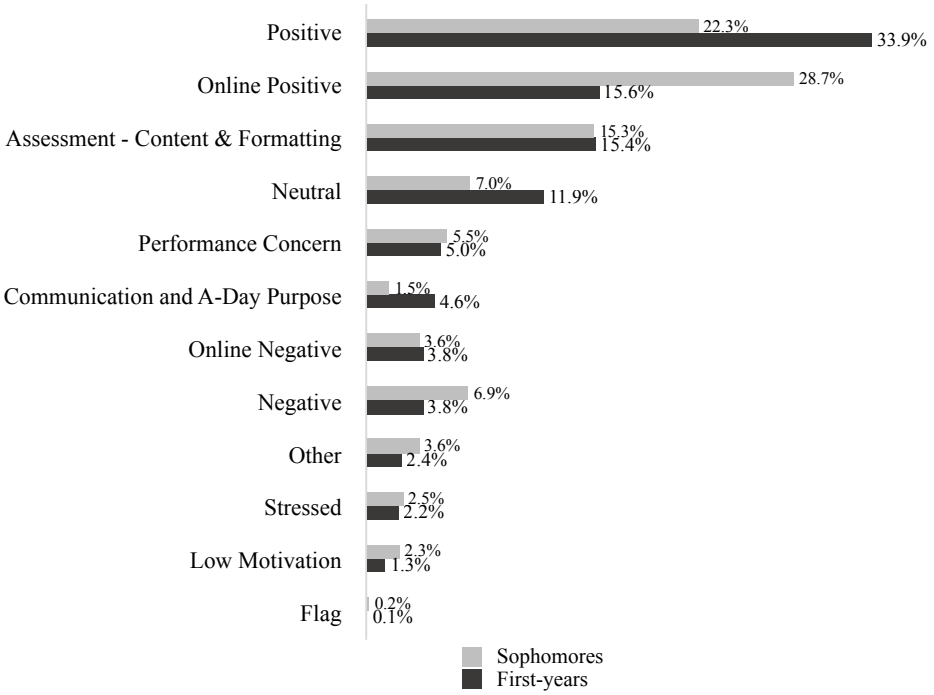
Although all codes and subcategories are informative, the most relevant for understanding students' experiences with UIT is "Online Positive" and "Online Negative." For this reason, we begin by considering the percentage of comments classified according to these two codes and whether these percentages differed across first-year and sophomore students. Additionally, general "Positive" and "Negative" codes are discussed for comparison. We then consider the subcategories of "Online Positive" and "Online Negative" to better understand students' specific positive and negative comments regarding UIT.

The percentage of comments classified by code is shown in Figure 1 separately for first-year and sophomore students. Comments classified as "Online Positive" said something generally positive concerning the remote testing format. For example, a comment that was coded as "Online Positive" for a sophomore student stated, "Taking the Assessment Day requirement remotely was stress free and more impactful." The percentage of comments classified as conveying something positive about UIT was 15.6% and 28.7% for first-year and sophomore students, respectively. The only other code capturing a largerpercentage of student responses was the "Positive" code, which captured general positive comments about the testing experience (not necessarily related to UIT). A comment that was coded as "Positive" for one sophomore student stated, "It went well." In contrast, very few comments (approximately 4%) were classified as "Online Negative" across first-year and sophomore

student comments. Comments classified as "Online Negative" mentioned feeling displeased with a remote Assessment Day or mentioned that Assessment Day should be in-person in the future. For example, one sophomore student said, "Having Assessment Day online is not good. in person [sic] is better." Additionally, only 3.8% of first-years and 6.9% of sophomores said something that fell into the general "Negative" code. These students said something generally negative about Assessment Day (not necessarily related to UIT). For example, one first-year student said, "It was boring and tedious, there are much better ways to spend time on campus, like studying our courses or making friends, rather than sitting in a room answering an assessment survey."

In contrast, very few comments (approximately 4%) were classified as "Online Negative" across first-year and sophomore student comments.

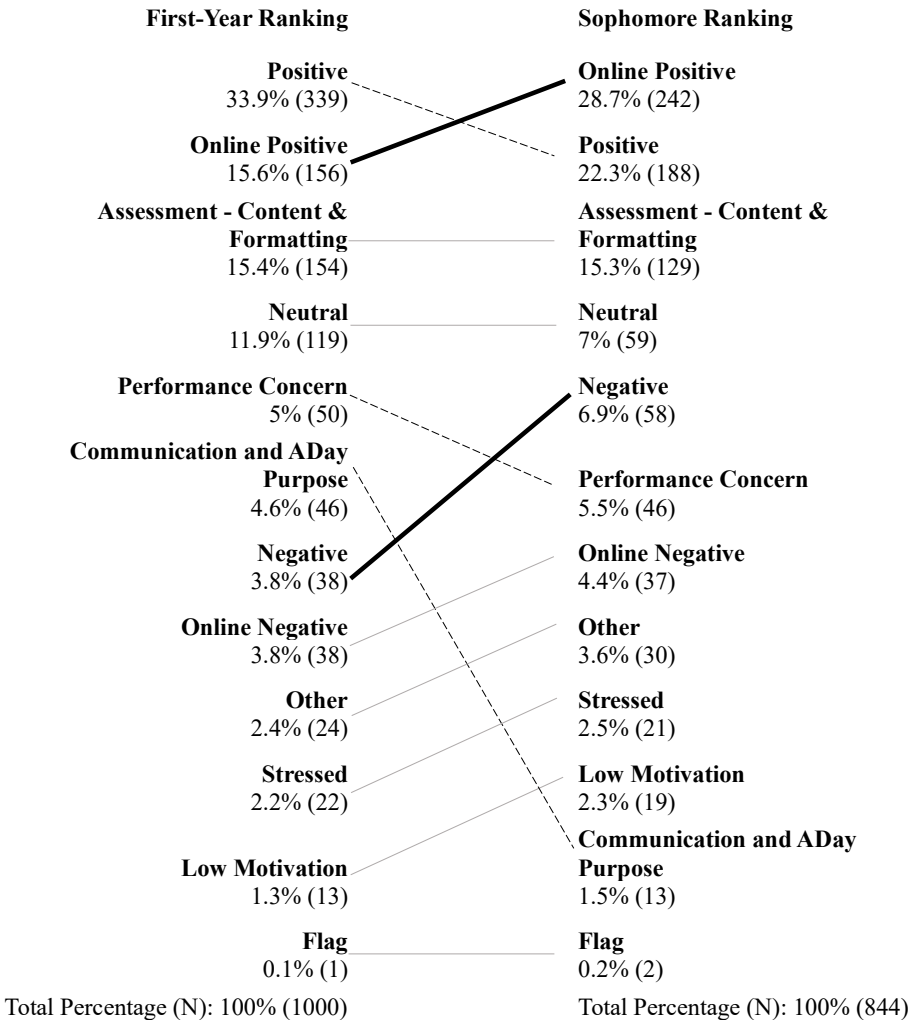
Figure 1.
Percentage of Comments Classified by Code for First-year and Sophomore Students



Although most comments were classified into each code at similar rates for first-year and sophomore students, there were some differences. Figure 2 presents the rank-ordering of codes separately for first-year and sophomore students according to the percentage of comments classified by each code. Lines are provided within the Figure to illustrate differences in rank-ordering of codes between the two groups. Notable differences include the rank ordering of the "Positive" and "Online Positive" codes. A larger percentage of comments were classified as "Online Positive" for sophomore students (28.8%) relative to first-year students (15.6%), while "Positive" was higher for first-year students (33.9%) relative to sophomore students (22.3%). Another notable difference was the rank ordering of "Negative." This code was higher for sophomore students (6.9%) compared to first-year students (3.8%). The rank-order and percentage of comments classified as "Online Negative" stayed relatively similar for the two groups.

Table 2 contains the subcategories for comments coded "Online Positive" for both first-year and sophomore students. Only subcategories that were larger than 10% are listed. Recall that out of the total comments, about 16% (156) of first-year student comments and about 29% (242) of sophomore student comments contained text that was classified as "Online Positive." Across both groups, most students fell into the top two subcategories - "online eases stress/anxiety" and "ease of online use." Students who reported "online eases stress/anxiety" said something about the online aspect of testing that helped them feel less stressed or anxious. Students who reported "ease of online use" mentioned how testing was easy to do online. For example, one first-year student said, "The virtual Assessment Day test ran smoothly and I enjoyed being able to complete this task on my own time. In addition, the questions were

Figure 2
Ranking of Codes by Group



Note. Codes are rank ordered for the first-year student group and sophomore group separately.

clear and easy to understand. Based of my experience, I would recommend Assessment Day to be virtual permanently." Of the "Online Positive" comments, a higher proportion of first-year students (27.4%) compared to sophomore students (20.4%) fell into "online eases stress/anxiety" over "ease of online use." In contrast, a higher proportion of sophomore students (27.6%) compared to first-year students (19.0%) fell into "ease of online use" over "online eases stress/anxiety." Both first-year students (14.2%) and sophomore students (18.1%) appreciated having extended time to complete their assessments and/or complete these assessments independently. First-year students (11.1%) and sophomore students (15.8%) said they would prefer online to in-person testing.

Table 2
Online Positive subcategories

	First-Years			Sophomores		
Subcategory	**Count**	**%**	**Rank**	**Count**	**%**	**Rank**
Being online eases stress/anxiety	62	27.4%	1	62	20.4%	2
Ease of online use	43	19.0%	2	84	27.6%	1
Extended time/on their own time	32	14.2%	3	55	18.1%	3
Prefer online to in-person	25	11.1%	4	48	15.8%	4

Note. Some comments fell into more than one subcategory.

Table 3 contains the subcategories for text coded "Online Negative" for both first-year and sophomore students. Only subcategories that were larger than 10% are listed. Recall that out of the total comments, only about 4% (38) of first-year student comments and about 4% (37) of sophomore student comments contained text that was classified as "Online Negative." The majority of comments within this code fell into four subcategories: "prefer in-person," "had an issue with the test," "ability to focus," and "motivation issues." For example, one first-year student said: "I know that Covid [sic] has had a big impact on the way assessments are taken, however, I feel that the online environment is not a great way to conduct the assessments because it is much easier to just skip through and not put forth your best effort. I had trouble remaining focused and motivated to complete the assessment. I feel that being in person would have been better." Of the "Online Negative" comments, the largest proportion of first-year student comments in this category said that assessments should be conducted in-person instead of online (27.7%), followed by comments that described testing issues completing the assessments online (19.1%). These subcategories were ranked differently for sophomores. The largest percentage of sophomores (31.7%) felt they had issues focusing with assessments being conducted online, followed by those saying they struggled with motivation to complete the assessment due to the online versus in-person administration (24.4%).

Of the roughly 20% of students who elected to provide feedback on their assessment experience, substantially more students said something positive about the online administration format than something negative.

Table 3
Online Negative subcategories

	First-Years			Sophomores		
Subcategory	**Count**	**%**	**Rank**	**Count**	**%**	**Rank**
Prefer in-person	13	27.7%	1	8	19.5%	3
Had an issue with the test	9	19.1%	2	6	14.6%	4
Ability to focus	6	12.8%	3	13	31.7%	1
Motivation issues	6	12.8%	4	10	24.4%	2

Note. Some comments fell into more than one subcategory.

Discussion

Of the roughly 20% of students who elected to provide feedback on their assessment experience, substantially more students said something positive about the online administration format than something negative. Specifically, almost 30% of sophomore student comments and 16% of first-year student comments conveyed something positive about UIT. We were encouraged to see the large number of positive statements surrounding UIT, particularly because students were not explicitly asked to address the online administration format in their feedback. Additionally, we were satisfied with the amount of generally positive comments we received about Assessment Day from first-year (34%) and sophomore (22%) students coupled with the low amount of generally negative feedback from first-year (4%) and sophomore (7%) students.

Further inspecting the responses coded "Online Positive" revealed several reasons for students' favorable attitudes toward UIT. Both first-year and sophomore students said they had lower stress or anxiety due to UIT. Additionally, both first-year and sophomore students cited the ease of the online assessments as a positive aspect of UIT. These subcategories are meaningful because they represent a substantial number of students. The few students who commented negatively about the online experience cited technical difficulties, difficulties focusing on the test, or trouble feeling motivated. It is essential to understand why students had negative comments about the online format. While positive comments may support the continued use of this type of remote testing, negative comments identify areas for improvement to the assessment process. Still, we must keep in mind that few students provided negative comments about UIT or general negative comments, and the reasons for those comments were not unanimous.

When reflecting on the comments about online testing we received, it is also worthy to note the kinds of comments we did not receive. No student mentioned trouble with internet connectivity or lack of access to a device to use for testing. The lack of such comments might also be specific to our university and a function of the characteristics of our students (e.g., socioeconomic status) and campus (e.g., availability of on-campus testing lab). Additionally, a lack of technology or technology issues may have prevented students from completing their assessments altogether. Typically students are required to complete their Assessment Day tests; however, due to COVID-19, students who had not completed their tests by the final deadline were not forced to complete them. However, noncompletion may not be a major issue because most students (around 90% for both groups) completed their assessments.

Because the codes were developed for the first-year student group, we considered whether the number of comments in each category and subcategory differed between the two groups. We felt that it was important to do so because the context for the two administrations differed and because previous research indicates test-taking motivation differs by class level (Pastor et al., 2019; Thelk et al., 2009; Wise, 2006; Wise & DeMars, 2010). In general, the results were similar between the two groups: both first-year and sophomore students reported more positive than negative comments about UIT. A noteworthy difference was that although positive UIT comments outnumbered negative UIT in both groups, the proportion of sophomore students that specifically cited UIT as a positive experience was higher than the proportion of first-year students. In other words, sophomore students particularly liked the online format, more so than first-year students. A critical difference between the two groups is that sophomore students had previously experienced an in-person Assessment Day. Contrary to first-year students, sophomores were able to compare in-person Assessment Day to remote Assessment Day. This difference could be why they seem more likely to cite UIT as the reason for their positive Assessment Day experience.

There were several limitations to this study. As mentioned previously, the codes were created using only first-year student comments so the raters were aware of the year of the students. This process is not ideal because raters' coding may be biased by knowing the year of the student. However, splitting comments this way eased the weight of creating codes for all the comments at one time. Another limitation is that only 20% of all students responded to the open-ended item. Those who chose to respond may have had a different perspective than those who chose not to respond in ways that limit our ability to generalize these results to all students who participated in Assessment Day. Second, social desirability may or may not have been a factor in these results. Although some students may have provided less than genuine positive responses in an attempt to "look good," the number is likely small because the assessments were low-stakes for students and answering the question was optional. Additionally, Caputo (2017) noted that social desirability might account for less than 10% of the variance in self-report measures. For that reason, we are not too worried about social desirability in this study. Finally, these results may be specific to our university, our students, and our UIT procedures. The generalizability of these results may be limited due to these settings.

Although this study focuses on comments pertaining to UIT, the collection of comments will help us understand what the experience was like for students and inform improvements to future assessment in higher education.

Despite the limitations, these results are encouraging for the continued use of UIT for assessment. Although this study focuses on comments pertaining to UIT, the collection of comments will help us understand what the experience was like for students and inform improvements to future assessment in higher education. The act of asking and sifting through

the responses brought us closer to the students, allowing us to see things from their point of view – a perspective we value but do not always actively seek out. Decisions about UIT use may still need to consider the effects of COVID-19, cost, accessibility, and the quality of the data (Jankowski, 2020b; Montenegro & Jankowski, 2020). Further, decisions of its continued use should also weigh the student's perspective. Incorporating the student perspective at our institution revealed a positive experience, which others have cited as a vital aspect of a high-quality UIT program (Beaty et al., 2009). This finding is encouraging for our institution and others who would like to use UIT. We aim to continue incorporating the student perspective into the assessment process to ensure that UIT continues to facilitate a positive experience for all students at JMU.

AUTHORS NOTE

Special thanks are given to our interns, Bree Pifer and Tanna Walters, who assisted in the long process of reading and coding hundreds of student comments. This paper would not be possible without them.

References

Alahmadi, S. & DeMars, C. E. (2022). Large-scale assessment during a pandemic: Results from James Madison University's remote Assessment Day. *Research and Practice in Assessment*, *12*(1), 5-15.

Beaty, J. C., Dawson, C. R., Fallaw, S. S., & Kantrowitz, T. (2009). Recovering the scientist-practitioner model: How IOs should respond to unproctored internet testing. *Industrial and Organizational Psychology*, *2*(1), 58-63. https://doi.org/10.1111/j.1754-9434.2008.01109.x

Busby, A. K. (2020). Resilient assessment during COVID-19. *Assessment Update*, *32*(6), 1-16. https://doi.org/10.1002/au.30231

Camara, W. (2020). Never let a crisis go to waste: Large-scale assessment and the response to COVID-19. *Educational Measurement: Issues and Practice*, *39*(3), 10-18. https://doi.org/10.1111/emip.12358

Caputo, A. (2017). Social desirability bias in self-reported well-being measures: Evidence from an online survey. *Universitas Psychologica*, *16*(2). https://doi.org/10.11144/Javeriana.upsy16-2.sdsw

Fulcher, K. H., & Leventhal, B. C. (2020). James Madison University: Assessing and planning during a pandemic. *Assessment Update*, *32*(6), 4-5. https://doi.org/10.1002/au.30233

Gwet, K. L. (2014). Handbook of inter-rater reliability, 4th edition. Gaithersburg, MD: Advanced Analytics.

Hsieh, H.-F. & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9),1277-1288. https://doi.org/10.1177/1049732305276687

Jankowski, N. A. (2020a). Guideposts for assessment during COVID-19. *Assessment Update*, 32(4), 10-11. https://doi.org/10.1002/au.30222

Jankowski, N. A. (2020b, August). *Assessment during a crisis: Responding to a global pandemic*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Karim, M. N., Kaminsky, S. E., & Behrend, T. S. (2014). Cheating, reactions, and performance in remotely proctored testing: An exploratory experimental study. *Journal of Business and Psychology*, *29*, 555-572. https://doi.org/10.1007/s10869-014-9343-z

Milone, A. S., Cortese, A. M., Balestrieri, R. L., & Pittenger, A. L. (2017). The impact of proctored online exams on the educational experience. *Curr Pharm Teach Learn*, *9*(1), 108-114. https://doi.org/10.1016/j.cptl.2016.08.037

Montenegro, E., & Jankowski, N. A. (2020, January). *A new decade for assessment: Embedding equity into assessment praxis* (Occasional Paper No. 42). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*, 1-13. https://doi.org/10.1177/1609406919899220

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).

Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment*, *24*(3), 189-212. https://doi.org/10.1080/10627197.2019.1615373

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. *Journal of General Education*, 58(3), 129-151. https://doi.org/10.1353/jge.0.0047

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95-114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*, 27-41. https://doi.org/10.1080/10627191003673216