RESEARCH & PRACTICE IN ASSESSMENT

VOLUME SIXTEEN | ISSUE 2 | RPAjournal.com | ISSN # 2161-4120

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.

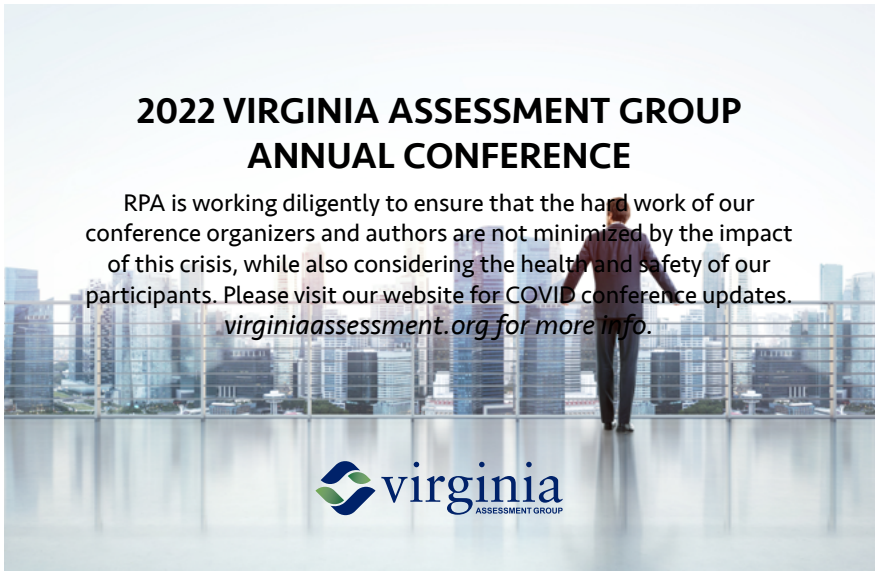
Ms. Katie Busby served as RPA Editor from November 2014-January 2019 and focused her attention on the growth and sustainability of the journal. During this time period, RPA explored and established collaborative relationships with other assessment organizations and conferences. RPA readership and the 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



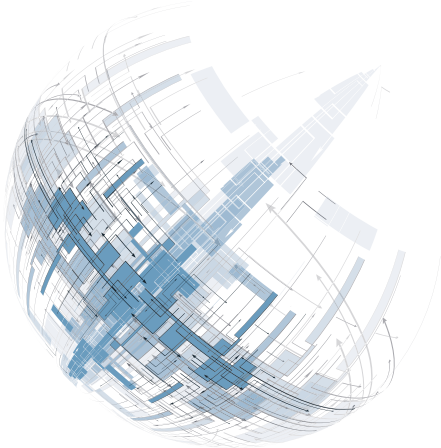
Assessment as a Question

"The way a question is asked limits and disposes the ways in which any answer to it-right or wrong-may be given." - Susanne Katherina Langer

"At its heart, assessment is indeed just a series of targeted questions with which we endeavor to not only find answers, but also the next questions. This issue of RPA provides readers with a variety of both! We hope that this issue of Research & Practice in Assessment finds you well and looking forward to the next questions of your assessment work.

Volume 16, Issue 2 of RPA includes four articles that cover a variety of topics. 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.

We hope that the questions, answers, and subsequently new questions posed in this issue provide many discussion points for you and your colleagues.



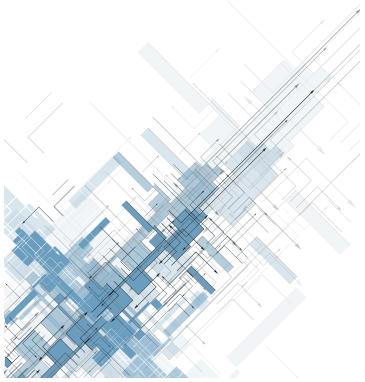
Best Wishes,

Nicholas Curtis

Editor-in-Chief,
Research & Practice in Assessment

Abstract

The intersection in 2020 of the new COVID-19 pandemic with the ongoing pandemic of anti-black racism exacerbated existing injustices as well as caused and revealed new inequities in US higher education. Because inequities in assessment in particular were intensified by these twin pandemics, faculty at several US colleges revised assessment approaches as part of their pedagogical partnership work over the last year. This paper describes the one-on-one, semester-long, pedagogical partnerships these faculty undertook with undergraduates not enrolled in the faculty members' courses. It reviews the commitments of such partnership work to equity and justice, offers examples of how four faculty-student pairs across the disciplines at three US colleges revised their approaches to assessment, and analyzes how these examples work toward equity and justice. The paper concludes with a discussion of the implications of such work not only at the intersection of twin pandemics but under all circumstances.



AUTHORS

Alison Cook-Sather, Ph.D.
Bryn Mawr & Haverford Colleges

Responding to Twin Pandemics: Reconceptualizing Assessment Practices for Equity and Justice

A few months into 2020, the COVID-19 pandemic put colleges and universities around the world into lockdown. Most US institutions responded by pivoting to remote and hybrid teaching and learning, and many continued with these modes through the Fall-2020 and Spring-2021 terms. The intersection of this pivot with the worldwide uprisings against anti-black racism threw into stark relief long-standing socio-economic injustices and inequities in US higher-education contexts and revealed new ones (Fain, 2020). The double disadvantaging—and, in some cases, devastation—at the intersection of the life-threatening pandemic and the life-affirming uprisings added urgency to the need to reconceptualize practices in US colleges and universities. This article focuses on the efforts of four pairs of student-faculty pedagogical partners at liberal arts colleges in the northeast, Mid-Atlantic, and southern regions of the US to revise assessment practices as part of their work to address injustices and inequities in higher education.

The widest context in which these efforts unfolded is that of anti-black systemic racism—a "transnational phenomenon" born of global white supremacy (Busey et al., 2020). As Tometi (2017), co-founder of Black Lives Matter, argues, "anti-black racism is everywhere—globalized in large part by the legacy of the enslavement of people of African descent, the colonial legacy and the current neo-colonial relations" (para. 4). The effects of anti-black racism in US higher education include high mental health costs (Anderson,

CORRESPONDENCE

Email

acooksat@brynmawr.edu

2020) and low completion rates for black and Hispanic students (Shapiro et al., 2017). These outcomes are not manifestations of students' failures but rather "of our broader, historical social system of privilege and oppression" (Williams, 2018, p. 2; Malcom-Piqueux, 2018). In the spring of 2020 in the US, these existing injustices were compounded by new inequities, such as higher rates of job loss and of mortality among black and Latino workers (Fain, 2020), many of whom were college students or members of their families.

Research has documented that every student does not have an equal opportunity to succeed in higher education (Cahalan et al., 2018; Singer-Freeman & Robinson, 2020). The twin pandemics revealed and exacerbated the ways in which socio-economic disparities intersected with race-based inequities in students' experiences. As Casey (2020) documented, while one student retreated to a vacation home to learn remotely, another struggled "to keep her mother's Puerto Rican food truck running while meat vanished from Florida grocery shelves." The shift to remote learning, one faculty member asserted, "made visible realities [students] were previously contending with, although there had not been an occasion to bring them to light until then" (Labridy-Stofle, 2020, p. 3). The intersection of the pandemic, the systemic racism in the US, and racial inequities in higher education has, according to Clayton (2021), "prompted a clarion call for more effective strategies that will result in more equitable outcomes for underrepresented populations" (para. 6).

Faculty and students participating in pedagogical partnership programs at a number of colleges saw the necessity of revising assessment as the intersection of the twin pandemics made them newly or more deeply aware of long-standing injustices and inequities.

Inequities in assessment have consistently been a concern in higher education (Leathwood, 2005; Montenegro & Jankowski, 2017), and these too have been exacerbated by the intersection of the twin pandemics. Most approaches insist on "assessing students in the same way without paying attention to their differences" (Montenegro & Jankowski, 2017, p. 16). Furthermore, many methods of assessment, like much else in higher education, both consciously embrace and unconsciously manifest characteristics of white supremacy culture (Jones & Okun, 2001), such as only one right way, either/or thinking, and objectivity. These characteristics contribute to the erroneous conflation of equity and sameness, to the failure to recognize multiple ways of problem solving and creating, and to the discounting of alternative logics and pathways to those privileged by those in power.

Characteristics of white supremacy culture inform the very structures of our educational systems. They are embodied in practices such as grading, which, as undergraduate student Nordstrom-Wehner argues, constitutes "a scale that inhibits learning and perpetuates existing inequalities" (Del Rosso & Nordstrom-Wehner, 2020, p. 7). Inoue (2015) has noted that, "Racism seen and understood as structural…reveals the ways that systems, like the ecology of the classroom, already work to create failure in particular places and associate it with particular bodies" (p. 4). Montenegro and Jankowski (2020) argue that equitable assessment practices are those that afford all learners an equal and unbiased opportunity to demonstrate their knowledge and achievements. The twin pandemics have revealed that historical patterns, institutional structures, and individual practices militate against all learners having such opportunities. Refusing the characteristics of white supremacy culture and creating assessments that are equitable—that take into account how students and institutional structures influence ways of knowing—involve, according to Montenegro and Jankowski (2020), providing opportunities for students to demonstrate knowledge in different ways.

Faculty and students participating in pedagogical partnership programs at a number of colleges saw the necessity of revising assessment as the intersection of the twin pandemics made them newly or more deeply aware of long-standing injustices and inequities. This paper begins with definitions of pedagogical partnership offered in current literature and highlights commitments of partnership work to equity and justice. It then presents the revisions to assessment faculty-student pairs across four disciplines developed during late 2020 and early 2021 at Bryn Mawr College, Davidson College, and Vassar College, and it analyzes how these examples work toward equity and justice. The paper concludes with a discussion of the implications of such work not only at the intersection of the twin pandemics but under all circumstances.

Student-Faculty Pedagogical Partnerships for Equity and Justice

Through pedagogical partnerships, academic and professional staff, administrators, and other students "engage students as co-learners, co-researchers, co-inquirers, co-developers, and co-designers" (Healey et al., 2016, p. 2) in and of approaches to learning and teaching. Pedagogical partnerships constitute "a collaborative, reciprocal process" whereby "all participants have the opportunity to contribute equally, although not necessarily in the same ways, to curricular or pedagogical conceptualization, decision making, implementation, investigation, or analysis" (Cook-Sather, Bovill, & Felten, 2014, pp. 6-7). In all four of the examples featured in this paper, faculty and student pairs worked in semester-long, one-on-one partnerships through which the student partners: visited their faculty partners' classrooms weekly; took observation notes focused on pedagogical questions and practices they and their faculty partners agreed to analyze; met weekly with their faculty partners to discuss the observation notes and both affirmations and potential revisions of practice; and met regularly with the partnership program facilitator and other student partners. In each case, the student partners earned monetary compensation or course credit for the time they spent.

This kind of partnership work has been shown to deepen engagement and enhance learning and teaching for all participants (Cook-Sather et al., 2014; Matthews, Mercer-Mapstone, Dvorakova, et al., 2019; Mercer-Mapstone, Dvorakova, Matthews, et al., 2017). Of particular importance to the present discussion, pedagogical partnership work has the potential to foster more equitable and inclusive practices (Cates, Madigan, & Reitenauer, 2018; Cook-Sather & Agu, 2013; Cook-Sather, Krishna Prasad, Marquis, et al., 2019; Cook-Sather, Signorini, Dorantes, et al. 2020) and redress some of the epistemic, affective, and ontological harms caused by the structures and practices of higher education (de Bie et al., 2019; 2021). A participant in Curtis and Anderson's (2021a) study noted that "[assessment in the classroom is one of the] most highly guarded and protected aspects of higher education and one of the last holdouts of sole faculty ownership" (p. 56). And yet, like the pedagogical partnership work described above, the co-creation of assessment by instructors and enrolled students can "empower and improve perceptions of the classroom, toward the end of fostering a more equitable learning environment for all students" (Chase, 2020, p. 11; see also Deeley & Bovill, 2017; Deeley & Brown, 2014).

Increasingly, pedagogical partnership programs name inclusion, belonging, equity, and justice as foundational commitments. In the US, for instance, Smith College (Cook-Sather, Bahti, & Ntem, 2019), Berea College (Cook-Sather, Ortquist-Ahrens, & Reynolds, 2019), and Florida Gulf Coast University (Cook-Sather, Ortquist-Ahrens, et al., 2019; Cook-Sather, Bahti, et al., 2019; Gennocro & Straussberger 2020) all named equity goals as foundational to their advent. Partnership programs beyond the US also explicitly embrace such commitments, including those at Victoria University of Wellington in Aotearoa / New Zealand (Leota & Sutherland 2020; Lenihan-Ikin et al. 2020), Kaye Academic College of Education in Beer-Sheva, Israel (Cook-Sather, Bahti, et al., 2019; Narkiss & Naaman 2020), and McMaster University in Ontario, Canada (Marquis, Carrasco-Acosta, et al., 2019).

Supporting the Development of Assessment that Moves Toward Equity and Justice

As the creator of a long-standing pedagogical partnership program at Bryn Mawr and Haverford Colleges, I am often asked to support other institutions in developing such programs, including at Davidson College and Vassar College. In March of 2020, at the suggestion of a student partner at Vassar College, she and I invited student partners from all institutions participating in Pairing Student Partners: An Intercollegiate Collaboration (a support structure she had created with my guidance) to participate in a Zoom conversation about how best to support their faculty partners when colleges pivoted to remote teaching and learning. Student partners at nine different institutions generated a set of recommendations (see linked resource) that was published on Haverford College's website as well as on other institutions' websites with the goal of reaching as wide an audience as possible. These recommendations included four overarching considerations and detailed approaches under each: (1) start with and sustain the human; (2) embrace practices that

Increasingly, pedagogical partnership programs name inclusion, belonging, equity, and justice as foundational commitments.

are equitable and accessible; (3) offer students choices; and (4) create regular opportunities to assess learning goals.

Hoping to showcase the work student and faculty partners were doing at these institutions, I contacted program directors at all nine institutions. I asked them to extend an invitation to all faculty participating in their partnership programs to share examples of developing more equitable practices of assessment. Four faculty members and their student partners responded, sending the detailed examples included below. Each of the examples was drafted and revised by the faculty and student partners and approved by them for inclusion in this discussion.

Assessment for Equity and Justice in a Psychology Course at Bryn Mawr College

Students as Learners and Teachers (SaLT) was conceptualized in 2006 and piloted in 2007 at Bryn Mawr and Haverford Colleges, two liberal arts colleges approximately 14 miles outside Philadelphia. SaLT developed in response to faculty desire to engage in more culturally responsive and inclusive practices (Cook-Sather, 2019; 2018b) and was supported by several grants from The Andrew W. Mellon Foundation. Since its advent, each semester SaLT has included between 50% and 75% student partners who identify as belonging to under-represented and under-served groups. All student partners are paid by the hour for the time they spend on partnership activities.

Since its advent, each semester SaLT has included between 50% and 75% student partners who identify as belonging to under-represented and under-served groups

In the Fall-2020 term, one faculty participant in SaLT, Ariana Orvell, Assistant Professor in the Psychology Department at Bryn Mawr, and her student partner, Sarah Phillips, Class of 2022 and a psychology major, worked together through the SaLT program on Orvell's course, Introduction to Psychology. This course was taught remotely, and Orvell used a flipped classroom (asynchronous lectures followed by synchronous Zoom sessions that addressed student questions, fostered discussion, checked for understanding, and extended concepts from lecture). In thinking through assessment, Orvell set up exams so that they would not feel quite as 'high stakes' and so that students could learn how to improve their studying/learning of the material and be rewarded for that when it comes to assessment. For example, she introduced an option for students to weigh the lowest grade on any of the three exams less heavily. Students also completed written responses after viewing the lectures, which gave them the opportunity to engage in deeper processing through synthesis, asking questions, and making connections between the course content and their own lives.

Feedback from her students and from Phillips informed Orvell that students appreciated being able to participate in this course in a variety of ways (e.g. chat, polls, discussions). Orvell therefore modified and expanded opportunities for students to engage in the course material. These modifications to respond to pandemic conditions intersected with uprisings in protest of anti-black racism. For instance, Orvell received emails from approximately one third of the students enrolled in her course expressing their intention to engage in the student-led strikes for racial justice that took place at Bryn Mawr and Haverford Colleges in the Fall-2020 term. In collaboration with a colleague, Laura Grafe, Orvell responded to students' desire to engage with content related to issues around racism by modifying an existing form of assessment—a 3-5-page reaction paper in which students synthesize and comment on an article—to focus on a particular article: "The Psychology of American Racism," written by Steven Roberts and Michael Rizzo (2020).

With Phillips' input and support, Orvell developed additional alternative assignments and readings, integrated language on DEI and anti-racism into her syllabus, strengthened her commitment to integrating perspectives from psychologists from diverse social identities and cultural contexts, and extended to students an invitation to question the implicit (or explicit) norms of the white hegemony that underlie many of the theories/studies covered in Introductory Psychology.

The changes described above were implemented at different points throughout the academic school year, in response to different types of student feedback, contextual factors, and discussions between Orvell and Phillips. For example, being intentional about giving students multiple ways to participate (e.g., chat, polls, discussion) was informed through

feedback and observations that Phillips shared with Orvell, as well as mid-semester feedback that Orvell and Phillips gathered from the class through an anonymous online survey. The Reaction Paper assignment was adapted during the student strike, in response to the strikers' call for classes to integrate learning about race into coursework (previously, students would have been given a choice between several articles that covered different topics in Introductory Psychology). The decision to allow students to weigh exams less heavily was largely informed by the recognition that the pandemic introduced severe mental health burdens for large swaths of the population, particularly adults 18-29 (see linked resource), in addition to Orvell's belief that assessment should reflect and reward students' growth and progress. This was built into the course from the onset. Orvell made changes to the syllabus (e.g., inviting students to question norms, DEI statement) after the Fall-2020 semester to promote a more inclusive classroom. Finally, Orvell received feedback from several students after teaching Introductory Psychology in the Fall-2020 semester indicating that students appreciated changes that were made to the course and evaluated it as inclusive.

Assessment for Equity and Justice in a Religion Course at Vassar College

The Student Teacher Engaged Pedagogical Partnership (STEPP) program was piloted in the Spring-2020 semester at Vassar, a small, liberal arts college in the Hudson Valley, New York. The program was an outgrowth of the Engaged Pluralism Initiative (EPI) Inclusive Pedagogies Working Group (Bala, 2021; Bala & Kahn, forthcoming). Through STEPP, Professor of Religion, Jonathon Kahn, and his student partner, Ananya Suresh, Class of 2021, undertook what they called "an experiment in student self-assessment during covid" in a 100-level course Kahn was teaching in a hybrid format. There were 28 mostly first-year students enrolled. In-person meetings were in an outdoor tent classroom, with 20 students in person and eight fully remote. Suresh had previously taken the course, but she was partnered with Kahn because of her involvement in the EPI working group. She was a two-year veteran of EPI and was involved in the development of the pilot, and she received academic credit for the partnership (.5 credit).

Kahn and Suresh worked together to revise grading procedures to follow a self-assessment structure. The emphasis was on encouraging students to take a more active role in their learning experience by reflecting on their goals, hopes, and effort for the semester. Furthermore, the revised grading procedures emphasized the role of collaborative learning in the classroom, prompting students to contemplate their extended engagement with one another during class (small partner groups for 40-45 minutes at times). The students were asked to give qualitative descriptions of what it was like to spend time in class together during Covid; with that description as a prompt, students were asked to give accounts of how and in what ways they got to know their classmates. In an effort to promote these interactions, Suresh and Kahn structured the final writing exercise as an interview; each student was assigned a class partner to write a profile of in terms of their experience in the class; students were encouraged to ask their partner how the experience of the class material newly shaped their experiences as a member of the Vassar community.

The approach gave them more flexibility during a time (a pandemic) when life was exceedingly unpredictable (at any time they could test positive and have to quarantine for 10 days) and precarious. It allowed them to continue to learn…

The self-assessment strategy Kahn and Suresh developed was a response to Kahn's discomfort with grading a class during a pandemic. Because interaction was circumscribed, and because he normally weighed class participation 20%, he was uncomfortable basing a grade on so much work that would go unseen at best. Self-assessment became a way that he could engage the students in their own learning process, prompt them to reflect on what they valued and how they wanted to develop over the course of the semester, and then have them see if they accomplished what they set out to do. For students, the benefits were several fold. The approach gave them more flexibility during a time (a pandemic) when life was exceedingly unpredictable (at any time they could test positive and have to quarantine for 10 days) and precarious. It allowed them to continue to learn—at least this is what they reported: they learned—while not feeling as though the pace and demands were backbreaking. Students reported a high degree of satisfaction both in terms of what they learned and their enjoyment of class.

Reflecting on this work, Kahn acknowledged that any time we attempt more equitable practices, we have only inequitable experience to pull from. He also noted,

though, that this is true with any grading scheme. But over time and through dialogue that addresses the norms we have in place for assessing work, students and faculty can become better at assessment—including students assessing themselves. Through such an evolution, self-assessment represents a type of work through which we transform the inequitable experiences we pull from.

Kahn found that talking with students about the norms they use to assess themselves, and offering his perspective on their work without the authority of determining their grade, led to students' growing understanding of why they work, what they like to work on, and what counts for them as fulfilling work—outcomes that are consistent with Kahn's course goals. He also found that students' self-assessment allowed him to engage more fully with the students' writing. His comments on their work were not aimed at justifying a grade. Instead, they were more directly tied to pointing out what was working well in a paper, what wasn't working, and what could get better. Not having to append a grade at the end of such comments made the experience of grading much less burdensome and more fulfilling for him, too. He has continued student self-assessment in subsequent semesters, both refining the self-assessment questions and planning to continue the evolution.

Assessment for Equity and Justice in a Chemistry Course at Davidson College

More frequent low-stakes assessments helped to encourage a growth mindset by checking comprehension and allowing for opportunities for clarification before the next assessment.

Fostering Inclusivity and Respect in Science Together (FIRST) is an initiative supported by a grant from the Howard Hughes Medical Institute to Davidson College. Davidson is situated in Davidson, North Carolina, a small town north of Charlotte somewhat at the suburban/rural divide of the region. As part of this initiative, the More Inclusive Learning Environments (MILE) was created in 2019 to improve the state of inclusivity and leadership in its science education (Hernandez Brito, 2021; Hossain, 2021). Student partners, identified by the FIRST Program coordinators, were chosen for their passion and interest related to inclusivity initiatives. They were then matched with faculty partners based on whether they had already taken a course with the faculty member (not allowed) as well as the likelihood of the student taking a future course with their faculty partner (the less likely, the better, and ultimately highly discouraged). The students in this cohort typically are victims of microaggressions, marginalization, racism, sexism, and other forms of oppression. The student and faculty partners participated in training, both student- or faculty-only and with student and faculty partners together. The students were encouraged by the program leads to communicate with the faculty partner about any and all observations and use the program leads as another outlet for observations. These positions were funded for both the faculty and student partners. The mantra throughout the experience was that the students were experts in their own experience.

Through the FIRST program, an Assistant Professor of Chemistry, Mitch Anstey, and his student partner, Claire Tobin, Class of 2021 and a Physics and Economics major who would not need to take the Inorganic Chemistry course (and the associated pre-requirements) that was the focus of the partnership, worked together in the context of one of Anstey's courses. Upon the shift to students moving off campus, the course converted to synchronous/asynchronous, and lectures were recorded in real time for students to view later for studying or for a first viewing if they couldn't attend. Attendance was typically greater than 90% in the fully remote setting. The class had 32 students, which is the maximum at Davidson College (total student population of 1,983). The course is both a requirement for chemistry majors as well as an elective for pre-health students.

Upon the shift to fully remote learning, all assessments (tests and problem sets) were divided into smaller portions to decrease study time and lower grade impact of any one assignment. This change resulted in more frequent assessments that were shorter in duration and smaller in terms of student effort. The changes aimed to break down the units, so students were responsible for less material on each assessment. More frequent low-stakes assessments helped to encourage a growth mindset by checking comprehension and allowing for opportunities for clarification before the next assessment.

Representing not so much changes *per se*, but a reinforcement of existing methods, Anstey and Tobin made a number of adjustments. They included small, low-stakes assignments due roughly each class period. These assignments were only graded for completion and could be completed collaboratively, and answer keys were provided. The class itself was conducted using the process-oriented guided instructional learning (POGIL) pedagogy (Farrell et al, 1999). This approach to group work has advantages in learning to support and debate claims, learning to give space and make space for others, using multiple viewpoints to understand topics/issues, and building community within the classroom that persists outside of the classroom. Because working in a group is often met with unease and leads to negative feelings around the activity (as supported by student course evaluations over several years in previous iterations of this course), group composition and support play a large role in how well the group functions, especially as many students are not skilled in working group dynamics.

Williamson [faculty] noted that his conversations with Weiler [student] always made him think differently about his teaching, so he rarely assumed that his first idea for how to solve a problem would be the optimal one.

Through MILE, Anstey and Tobin were able to work together to make even more observations about how well groups were functioning, and they developed strategies for choosing future groups that would ultimately facilitate the best outcome for all involved. In one instance, a student was often seen observing but not directly contributing to their group due to the presence of two students who knew each other previously and were already comfortable interacting. Additionally, this student self-identified as black and later mentioned that they felt the group was dominated by the other two, who did not make efforts to ask for others' contributions or thoughts. Even before this information was offered by the student, Tobin had identified the dynamic, alerted Anstey, and worked to find a new group where the dynamic was more equitable. Additionally, the two close friends were separated in future groups to enable more discussion among all parties.

Anstey and Tobin received a lot of positive feedback. The final student feedback specifically about the use of MILE in the classroom was positive, and Anstey and Tobin often heard throughout the semester that even the presence of the MILE student partner was a signal that inclusivity and equity were valued in the classroom.

Assessment for Equity and Justice in a Biology Course at Bryn Mawr College

Immunologist and Assistant Professor of Biology, Adam Williamson, and his student partner, Kate Weiler, Class of 2020, worked in partnership for two semesters through the SaLT program during Weiler's senior year at Bryn Mawr College. Weiler and Williamson were paired based on scheduling compatibility, as is the case with virtually all student-faculty pairings through SaLT. Weiler completed an independent major in education and was paid for her work as a SaLT student partner.

In Spring 2020, Williamson and Weiler worked together in a senior thesis seminar. The course enrolled eight senior biology majors and met in person for the first six weeks of the term before a shift to a remote-only format. At the end of the term, students were required to submit a thesis to meet their major requirements. During the transition to remote learning, Williamson and Weiler, in collaboration with students in the seminar, reconfigured the course as a sequence of twice-weekly meetings dedicated to student support and accountability opportunities for Williamson and the enrolled students.

Specifically, Williamson and Weiler made the following three revisions. First, they moved to student-set (rather than faculty-determined) deadlines for draft sections of the thesis. After a sudden transition to remote learning, students were working under difficult circumstances. For instance, many students in the course took on new job or childcare responsibilities at home that made working to the schedule on the syllabus impossible. Williamson and Weiler encouraged students to work towards self-set deadlines to complete draft sections of their thesis.

Second, they de-emphasized student peer-review of other students' work. Williamson and Weiler had planned for students to review one another's work and provide critical feedback, but they removed this requirement for students because peer review was impossible when

students were working on different, self-set schedules. Instead, Williamson provided timely feedback on student work. Finally, they removed a required student-led seminar meeting. They had planned for students to lead a seminar meeting during the semester about their thesis work for discussion with the class. They removed this component of the course so students could focus time and energy on the time-sensitive thesis work required to graduate.

The revisions Williamson and Weiler made were directly influenced by the students in class. They asked students to complete a brief set of questions about changes that would best support their learning and offered a set of proposed changes rather than a set of new rules. Students offered suggestions about these changes during their first remote meeting. Thus, Williamson and Weiler developed the course revisions as part of an iterative process in collaboration with their students, not as unilateral decisions about what they assumed their students required.

In reflecting on this work, Williamson noted that his conversations with Weiler always made him think differently about his teaching, so he rarely assumed that his first idea for how to solve a problem would be the optimal one. Weiler was instrumental in communicating to Williamson the importance of regular weekly contact as a full group. While Williamson's initial instinct had been to switch to individual meetings to help students complete their thesis work and graduate on time, Weiler's concise, convincing argument about the importance of class community and student-led mutual support networks was an important factor in building their revised seminar structure. Williamson has adopted the revised structure of the course (twice-weekly meetings, with a full class meeting dedicated to build seminar community) as the new format in which he teaches this class (most recently in the Spring-2021 term), and students have voiced appreciation of a community-focused, full-class meeting once per week supplemented by "writing workshops" that serve as spaces for individual meetings and conversations about student research.

Implications

The examples included here emerged in response to a particular crisis and intersection. The heightened awareness, care, willingness to rethink, and specific revisions these faculty-student partners co-created reject characteristics of white supremacy culture

The examples included here emerged in response to a particular crisis and intersection. The heightened awareness, care, willingness to rethink, and specific revisions these faculty-student partners co-created reject characteristics of white supremacy culture (Jones & Okun, 2001). They move toward affording all students equal and unbiased opportunities to demonstrate their knowledge and achievements (Montenegro & Jankowski, 2020). And they respond to the student partner recommendations to start with and sustain the human, offer students choice, and create regular opportunities to assess learning goals.

In the psychology course at Bryn Mawr College, Orvell and Phillips developed alternative assignments and assessments that responded to student desire for content related to issues around racism, afforded students more choice, and more explicitly prioritized their learning. In the religion course at Vassar, Kahn and Suresh revised grading procedures in ways that shifted the sole locus of control from faculty to students and, like Orvell and Phillips' revisions, shifted the focus from performance of what faculty expect to engagement in what deepens student learning.

In the chemistry course at Davidson College, Anstey and Tobin created shorter, more frequent assessments that, like Orvell's and Kahn's revisions, encouraged a growth mindset. They also built class community, linking to the refusal of one right way, since different students take different approaches. Finally, in the biology course at Bryn Mawr College, Williamson and Weiler reconfigured the structure of the course, shifting to student-set (rather than faculty-determined) deadlines for draft sections of student theses, de-emphasizing student peer-review of other students' work to lower pressure, and reducing requirements. All of these changes, prompted by the pandemic-necessitated shift to remote teaching and learning, also reflected, according to Weiler and Williamson (2020), "necessary and overdue conversations about white supremacy and what to do to create a sustained anti-racist academy" (p. 6).

Through their approaches, across disciplines and institutions, these faculty refused the conflation of equity and sameness, recognized multiple ways of problem solving and creating, and embraced students' alternative logics and pathways in demonstrating knowledge.

Furthermore, they did this work in partnership with students not enrolled in their courses, which led, in turn, to greater partnership with enrolled students—a phenomenon that has been demonstrated across student-faculty partnerships (Cook-Sather, 2014; Cook-Sather, Hong, Moss, et al., 2021).

Faculty and student partners alike note that the changes made in response to a crisis are actually important to consider under all circumstances. Williamson notes that "partnerships are uniquely positioned to help faculty build and sustain trauma-informed learning spaces, respond to mistakes in content and facilitation quickly in a student-centered way, and avoid making blunders in the first place" (Weiler & Williamson, 2020, p. 6). Weiler notes that Williamson's caring approach "was present before we shifted to remote learning and continued through the disruption caused by COVID-19" (Weiler & Williamson, 2020, p. 3). She asserts that Williamson's "care-centered pedagogy exemplifies that showing care towards students should be prioritized always, not only during unprecedented circumstances" (Weiler & Williamson, 2020, p. 3).

In the context of long-standing inequities and injustices made (more) apparent by the intersection of the global pandemic and the protests against anti-black racism in the US, reconceptualizing assessment practices is more important than ever.

These reflections are consistent with what other participants in pedagogical partnership have argued. Reflecting on the partnership she developed not only with her formal student partner but also with all the students enrolled in her literature course, Labridy-Stofle (2020) anticipates:

> When we return to in-person teaching (one day), I will keep with me this new understanding of my students. How I can continue to make room for their multiplicity in a face-to-face setting and to think in terms of 'becoming' rather than 'being' is something I will keep striving for. In truth, however, as a Caribbean-born person, I already carried notions of multiplicity, intersectionality, and the rhizome within me, but I am more determined than ever to infuse them more consistently in my teaching (p. 3).

Labridy-Stofle (2020) credits her work with her student partner, Parker Matias, for helping her achieve this clarity: "My partnership with Parker made me realize the possibility of such collaborations becoming the norm, rather than isolated experiments, and how they could be deployed in the as-yet-incomplete project of social justice" (Labridy-Stofle, 2020, p. 4). Such collaborations "becoming the norm" might contribute to student-faculty partnership becoming part not only of one-on-one partnerships, as discussed here, but also program-level assessment in higher education (Curtis & Anderson, 2021a, 2021b).

In the context of long-standing inequities and injustices made (more) apparent by the intersection of the global pandemic and the protests against anti-black racism in the US, reconceptualizing assessment practices is more important than ever. There is both opportunity and imperative to ensure that this focus on humane consideration, equity, and justice not get lost in the overwhelm (for many people) of engaging in remote and hybrid teaching or in the rush to return to in-person modes. If enough faculty prioritize the creation of equitable and just approaches to assessment, we can begin to dismantle the structures, not only the practices, that sustain inequity and injustice.

ACKNOWLEDGEMENTS

Many thanks to Mitch Anstey, Ariana Orvell, Jonathan Kahn, Sarah Phillips, Ananya Suresh, Claire Tobin, Kate Weiler, and Adam Williamson for doing important equity work and for sharing their approaches, rationales, and analyses of both for inclusion in this discussion.

References

Anderson, G. (2020, October 23). The emotional toll of racism. *Inside Higher Ed*. https://www.insidehighered.com/news/2020/10/23/racism-fuels-poor-mental-health-outcomes-black-students

"Anti-blackness is global": An interview with Opal Tometi, Black Lives Matter (2017). https://unhumanrights.medium.com/anti-blackness-is-global-an-interview-with-opal-tometi-black-lives-matter-307e485287e6

Bala, N. (2021). A STEPP into uncertainty: Pursuing passions to embrace pedagogical risks. *Teaching and Learning Together in Higher Education*, *33*. https://repository.brynmawr.edu/tlthe/vol1/iss33/4/

Bala, N. & Kahn, J. (forthcoming). Exploring belonging in pedagogical partnership programs. *International Journal for Students as Partners*.

Busey, C. L., & Coleman-King, C. (2020). All around the world same song: Transnational anti-black racism and new (and old) directions for critical race theory in educational research. *Urban Education*. https://doi.org/10.1177/0042085920927770

Cahalan, M., Perna, L. W., Yamashita, M., Wright, J. & Santillan, S. (2018). 2018 indicators of higher education equity in the United States: Historical trend report. Washington, DC: The Pell Institute for the Study of Opportunity in Higher Education, Council for Opportunity in Education (COE), and Alliance for Higher Education and Democracy of the University of Pennsylvania (PennAHEAD).

Casey, N. (2020, May 5). College made them feel equal. The virus exposed how unequal their lives are. *The New York Times*. https://www.nytimes.com/2020/04/04/us/politics/coronavirus-zoom-college-classes.html

Cates, R. M., Madigan, M. R., & Reitenauer, V. L. (2018). "Locations of possibility": Critical perspectives on partnership. *International Journal for Students as Partners* 2(1), 33-46. https://doi.org/10.15173/ijsap.v2i1.3341

Chase, M. K. (2020). Student voice in STEM classroom assessment practice: A pilot intervention. *Journal of Research & Practice in Assessment*, *15*(2). https://www.rpajournal.com/student-voice-in-stem-classroom-assessment-practice-a-pilot-intervention/

Clayton, T. B. (2021). Refocusing on diversity, equity, and inclusion during the pandemic and beyond: Lessons from a community of practice. *Higher Education Today*. https://www.higheredtoday.org/2021/01/13/refocusing-diversity-equity-inclusion-pandemic-beyond-lessons-community-practice/

Cook-Sather, A. (2019). Increasing inclusivity through pedagogical partnerships between students and faculty. *Diversity & Democracy*. https://www.aacu.org/diversitydemocracy/2019/winter/cook-sather

Cook-Sather, A. (2018b). Developing "Students as Learners and Teachers": Lessons from ten years of pedagogical partnership that strives to foster inclusive and responsive practice. *Journal of Educational Innovation, Partnership and Change*, 4(1). https://journals.studentengagement.org.uk/index.php/studentchangeagents/article/view/746/pdf

Cook-Sather, A. (2014). Multiplying perspectives and improving practice: What can happen when undergraduate students collaborate with college faculty to explore teaching and learning. *Instructional Science*, 42, 31–46.

Cook-Sather, A., Bahti, M., & Ntem, A. (2019). *Pedagogical partnerships: A how-to guide for faculty, students, and academic developers in higher education*. Elon University Center for Engaged Teaching Open Access Series. https://www.centerforengagedlearning.org/books/pedagogical-partnerships/

Cook-Sather, A., Hong, E., Moss, T., & Williamson, A. (2021). Developing new faculty voice and agency through trustful, overlapping, faculty-faculty and student-faculty conversations. *International Journal for Academic Development*. https://www.tandfonline.com/doi/abs/10.1080/1360144X.2021.1947296?src=&journalCode=rija20

Cook-Sather, A., Ortquist-Ahrens, L., & Reynolds, W. (2019, November 13-17). *Building belonging through pedagogical partnership: Connecting within and across institutions*. [Conference presentation]. POD Network conference, Pittsburgh, PA.

Cook-Sather, A., Signorini, A., Dorantes, S., Abuan, M., Covarrubias-Oregel, G., & Cribb, P. (2020). "I never realized…": Shared outcomes of different student-faculty partnership approaches to assessing student learning experiences and evaluating faculty teaching. *Journal of Higher Education Theory and Practice*, *20*(7). https://www.articlegateway.com/index.php/JHETP/article/view/3160/3004

Curtis, N., & Anderson, R. (2021b, May). A framework for developing student-faculty partnerships in program-level student learning outcomes assessment (Occasional Paper No. 53). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). https://www.learningout comesassessment.org/wp-content/uploads/2021/05/OccPaper53_Partnership.pdf

Curtis, N. A., & Anderson, R. D. (2021a). Moving toward student-faculty partnership in systems-level assessment: A qualitative analysis. *International Journal for Students as Partners*, *5*(1), 57-75. https://doi.org/10.15173/ijsap.v5i1.4204

de Bie, A., Marquis, E., Cook-Sather, A., & Luqueño, L. P. (2021). *Promoting equity and justice through pedagogical partnership*. Stylus Publishers. https://www.centerforengagedlearning.org/books/promoting-equity-and-justice-through-pedagogical-partnership/

Deeley, S. J., & Bovill, C. (2017). Staff student partnership in assessment: Enhancing assessment literacy through democratic practices. *Assessment & Evaluation in Higher Education* 42 (3), 463-477. https://www.tandfonline.com/doi/abs/10.1080/02602938.2015.1126551?journalCode=caeh20

Deeley, S. J., & Brown, R. A. (2014). Learning through partnership in assessment. *Teaching and Learning Together in Higher Education*, *13*. http://repository.brynmawr.edu/tlthe/vol1/iss13/3

Del Rosso, J., & Nordstrom-Wehner, B. (2020). Team grade anarchy: A conversation about the troubled transition of grading. *Teaching and Learning Together in Higher Education*, 30. https://repository.brynmawr.edu/tlthe/vol1/iss30/5

Fain, P. (2020, June 17). Higher education and work amid crisis. *Inside Higher Ed*. https://www.insidehighered.com/news/2020/06/17/pandemic-has-worsened-equity-gaps-higher-education-and-work

Farrell, J.J., Moog, R.S., & Spencer, J.N. (1999). A Guided Inquiry Chemistry Course. J. *Chem. Educ*., 76, 570–574.

Gennocro, A., & Straussberger, J. (2020). Peers and colleagues: Collaborative class design through student-faculty partnerships. In A. Cook-Sather & C. Wilson (Eds.), *Building courage, confidence, and capacity in learning and teaching through student-faculty partnership: stories from across contexts and arenas of practice* (pp. 29-37). Lanham, MD: Lexington Books.

Healey, M., Flint, A., & Harrington, K. (2016). Students as partners: Reflections on a conceptual model. *Teaching & Learning Inquiry*, 4(2), 1–13. https://doi.org/10.20343/teachlearninqu.4.2.3

Hernandez Brito, C. (2021). Creating more inclusive learning environments at Davidson College. *Teaching and Learning Together in Higher Education*, *33*. https://repository.brynmawr.edu/tlthe/vol1/iss33/3/

Hossain, S. (2021). Embracing the risk and responsibility of starting a pedagogical partnership program focused on fostering inclusivity and respect in science. *Teaching and Learning Together in Higher Education*, *33*. https://repository.brynmawr.edu/tlthe/vol1/iss33/2/

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. https://wac.colostate.edu/docs/books/inoue/ecologies.pdf

Jones, K., & Okun, T. (2001). The characteristics of white supremacy culture. In *Dismantling Racism: A Workbook for Social Change Groups*. https://www.showingupforracialjustice.org/white-supremacy-culture-characteristics.html

Labridy-Stofle, C. (2020). Uprooted rhizomes: Collaborating in times of troubling transitions. *Teaching and Learning Together in Higher Education*, 30. https://repository.brynmawr.edu/tlthe/vol1/iss30/4

Leathwood, Carole. (2005). Assessment policy and practice in higher education: purpose, standards and equity. *Assessment & Evaluation in Higher Education*, 30(3), 307-324. https://www.tandfonline.com/doi/abs/10.1080/02602930500063876

Lenihan-Ikin, I., Olsen, B., Sutherland, K.A., Tennent, E., & Wilson, M. (2020). Partnership as a civic process. In L. Mercer-Mapstone & S. Abbot (Eds.), *The power of partnership: Students, staff, and faculty revolutionizing higher education. Center for Engaged Learning Open Access Series*. https://www.centerforengagedlearning.org/books/power-of-partnership/section-one/chapter-5/

Leota, A., & Sutherland, K. (2020). 'With your basket of knowledge and my basket of knowledge, the people will prosper': Learning and leading in a student-staff partnership program. In A. Cook-Sather & C. Wilson (Eds.), *Building courage, confidence, and capacity in learning and teaching through student-faculty partnership: stories from across contexts and arenas of practice* (pp. 93-102). Lanham, MD: Lexington Books.

Loglie, C. (2019). Culturally Responsive Assessment 2.0: Revisiting the Quest for Equity and Quality in Student Learning. *Journal of Research & Practice in Assessment*, 14(2). https://www.rpajournal.com/culturally-responsive-assessment-2-0-revisiting-the-quest-for-equity-and-quality-in-student-learning/

Malcom-Piqueux, L. (2018). Making sense of data in equity-minded ways (p. 52). In AAC&U (Eds.). A Vision for Equity: Results from AAC&U's Project Committing to Equity and Inclusive Excellence: Campus-Based Strategies for Student Success. Washington, DC: AAC&U.

Marquis, E., Carrasco-Acosta, E., de Bie, A., Krishna Prasad, S., Wadhwani, S., & Woolmer, C. (2019). Pedagogical partnerships and equity in the classroom: Insights from one partnership program. Presentation at the 2019 Symposium on Scholarship of Teaching and Learning, Banff, AB., November 7-9.

Marquis, B., de Bie, A., Cook-Sather, A., Krishna Prasad, S., Luqueño, L., & Ntem, A. (forthcoming). "I saw a change": Enhancing classroom equity through pedagogical partnership. *The Canadian Journal for the Scholarship of Teaching and Learning*.

Matthews, K. E., Mercer-Mapstone, L., Dvorakova, S. L., Acai, A., Cook-Sather, A., Felten, P. Healey, M. Healey, R., & Marquis, E. (2019). Enhancing outcomes and reducing inhibitors to the engagement of students and staff in learning and teaching partnerships: Implications for academic development. *International Journal for Academic Development*, 24(3), pp. 246-259. https://www.tandfonline.com/doi/abs/10.1080/1360144X.2018.1545233?journalCode=rija20

Mercer-Mapstone, L., Dvorakova, L. S., Matthews, K. E., Abbot, S., Cheng, B., Felten, P. Knorr, K., Marquis, E., Shammas, R., & Swaim, K. (2017). A systematic literature review of students as partners in higher education. *International Journal of Students as Partners*, 1(1), 1-23. https://mulpress.mcmaster.ca/ijsap/article/view/3119

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). www.learningoutcomesassessment.org

Montenegro, E., & Jankowski, N. A. (2017, January). Equity and assessment: Moving towards culturally responsive assessment. (Occasional Paper No. 29). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). https://learningoutcomesassessment.org/wp-content/uploads/2019/02/OccasionalPaper29.pdf

Narkiss, D., & Naaman, I. (2020). Voicing and reflecting in a pedagogical partnership. In *Building Courage, Confidence, and Capacity in Learning and Teaching through Student-Faculty Partnership: Stories from across Contexts and Arenas of Practice*, edited by Alison Cook-Sather and Chanelle Wilson. Lanham, MD: Lexington Books.

Shapiro, D., Dundar, A., Huie, F., Wakhungu, P., Yuan, X., Nathan, A & Hwang, Y., A. (2017, April). Completing College: A National View of Student Attainment Rates by Race and Ethnicity – Fall 2010 Cohort (Signature Report No. 12b). Herndon, VA: National Student Clearinghouse Research Center.

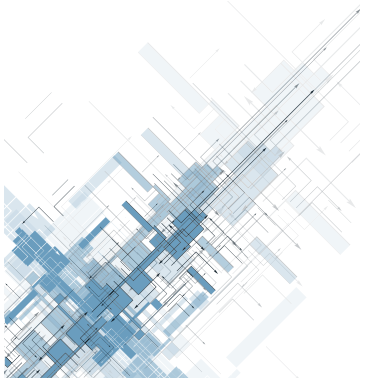
Singer-Freeman, K., & Robinson, C. (2020, November). Grand challenges in assessment: Collective issues in need of solutions (Occasional Paper No. 47). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Weiler, K., & Williamson, A. (2020). Partnering to build responsive learning communities that support students in crisis. *Teaching and Learning Together in Higher Education*, 30. https://repository.brynmawr.edu/tlthe/vol1/iss30/3

Williams, E. (2018, February). First thing's first: Privilege, power, and pedagogy the antecedent of assessment (Equity Response). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). Retrieved from https://www.learningoutcomesassessment.org/wp-content/uploads/2019/08/EquityResponse-EWilliams.pdf

Abstract

Professional standards related to outcomes assessment call for student affairs professionals to use research to inform programming. If professionals are to rely on research to build programs that positively impact student learning outcomes, the research should be credible. We examined the quality of program effectiveness research available for programming decisions. We reviewed five years of quantitative and mixed methods program effectiveness studies published in four student affairs journals. Despite frequent assertions of program effectiveness, the research designs and analyses did not often support such claims due to plausible threats to the validity of those claims. Articles claiming that programming is effective without credible evidence to support such a claim can result in professionals offering ineffective programming and engaging in inefficient assessment efforts. To address the credibility of effectiveness claims, we call for increased training in research methods, careful review of authors' claims by editors, and assistance from assessment practitioners.



AUTHORS

S. Jeanne Horst, Ph.D.
James Madison University

Sara J. Finney, PhD.
James Madison University

Caroline O. Prendergast, MEd
James Madison University

Andrea M. Pope, PhD.
James Madison University

Morgan Crewe, MA
James Madison University

The Credibility of Inferences from Program Effectiveness Studies Published in Student Affairs Journals: Potential Impact on Programming and Assessment

Faculty and student affairs professionals strive to offer programming (e.g., activities, pedagogies, strategies) that results in students achieving intended learning and development outcomes. Professionals are then expected to assess the programming for its level of effectiveness. If effectiveness is not achieved, professionals are expected to use assessment results to inform programming changes that improve learning and development. However, there are few examples of such improvement efforts resulting in greater student learning (Banta & Blaich, 2011; Jankowski, et al., 2018). In turn, assessment practitioners have considered strategies to address this issue and increase learning improvement (e.g., Fulcher & Prendergast, 2019; Smith, et al., 2018).

One strategy is to implement evidence-based programming (Finney & Buchanan, 2021). Building programming based on evidence is also referred to as "evidence-based practice (EBP): instructional approaches shown by high-quality research to result reliably in generally improved student outcomes" (Cook et al., 2011, p. 493). Student affairs' professional competencies and standards call for programming to be intentionally built using current research that indicates what effectively impacts particular outcomes (e.g., ACPA & NASPA, 2015; Finney & Horst, 2019a, 2019b). Additionally, the *Assessment Skills Framework* indicates that the ability to identify literature domains to inform program development is necessary for high-quality assessment practice (Horst & Prendergast,

CORRESPONDENCE

Email

jeanne.horst@gmail.com

2020). When discussing the Grand Challenge in Assessment of "Driving Innovation", Singer-Freeman and Robinson (2020) noted that professionals must "identify evidence-based solutions from the research literature" (p. 5) to improve students' outcomes. In short, consulting existing research increases the probability that programming will impact intended student learning and development outcomes (Pope et al., 2019; Smith & Finney, 2020).

Carpenter (2001) likened evidence-based program development to evidence-based medicine, arguing that in the absence of rigorous evaluations of effectiveness "student affairs may be doomed to repeating past mistakes in the name of tradition and convention" (p. 302). Although often not computed, the cost of implementing ineffective programs can be quite high (Bickman & Reich, 2015). Students engaging in ineffective programs may not achieve desired outcomes, which may prompt additional programming and increased time to degree completion. Thus, professionals should strive to identify and implement effective programs that have credible evidence of impacting desired outcomes. Subsequent outcomes assessment is still necessary and is used in a confirmatory manner to assess if the evidence-based programming is effective in the specific institutional context (Finney et al., 2021). This confirmatory approach is efficient. Less time and resources are needed to improve programming because it is likely to be effective. Thus, fewer iterations of the assessment cycle are needed to inform changes to programming to obtain desired levels of student learning and development.

The Need to Evaluate the Quality of Published Effectiveness Studies

High-quality evidence of program effectiveness requires carefully designed research studies. Put simply, studies that are methodologically suspect do not provide compelling evidence for making programming and assessment decisions.

To implement evidence-based programming, professionals must locate evidence and appraise the evidence for its validity, effect size, and applicability to the population in question. This paper focuses on evidence provided by published program effectiveness studies, which are used in the development of evidence-based programs. An early definition described program effectiveness as "… the extent to which pre-established objectives are attained as a result of activity" (Deniston et al., 1968, p. 324). Analogous to the logic underlying the outcomes assessment process, program effectiveness studies are conducted to evaluate if programming impacted intended outcomes. Thus, in this paper, we refer to "program effectiveness studies" as those that explicitly state student learning or development outcomes for a program and then report findings to evaluate whether students have met those outcomes. The ideal inference from a program effectiveness study is that the program led to or "caused" student learning or development. Not all program effectiveness studies can support this inference; however, this inference is foundational to developing evidence-based programming and merits scrutiny.

High-quality evidence of program effectiveness requires carefully designed research studies. Put simply, studies that are methodologically suspect do not provide compelling evidence for making programming and assessment decisions. "[T]he methodological rigor of a piece of research dictates directly the 'credibility' (Levin, 1994; Murnane & Willett, 2011) of its evidence, or the 'trustworthiness' (Jaeger & Bond, 1996) of the research findings and associated conclusions" (Levin & Kratochwill, 2013, p. 469). Whether the evidence influences program-related decisions "depends in part on the judgements that people make of its credibility, as credibility judgements precede processes of persuasion, influence and use" (Miller, 2015, p. 41). Due to lack of training in appraising the credibility of empirical evidence (Cooper et al., 2016; Muller et al., 2018), professionals may rely on journal editors and reviewers to judge research quality (Miller, 2015). By virtue of publication in peer-reviewed journals, professionals may believe evidence is credible and, in turn, trust inferences and implications provided by the study's authors (Hilligoss & Rieh, 2008).

When consulting with student affairs colleagues who were using published program effectiveness research to inform programming, we observed variability in the credibility of evidence found in the journals they referenced. Moreover, concerns about the quality of research design and credibility of inferences have been voiced by student affairs professionals (e.g., Grace-Odeleye & Santiago, 2019; Valentine et al., 2011). These concerns prompted calls for more rigorous designs that afford trustworthy claims, thereby facilitating successful engagement in programming and assessment efforts.

Program effectiveness studies typically infer that programming *caused* or *did not cause* an outcome. Whenever causal inferences are stated, they should be held to standards and assessed for common threats to the validity of causal inferences (e.g., Shadish et al., 2002). Unjustified statements that programming "led to," "caused," or "influenced" student learning or development outcomes can lead professionals to implement ineffective programs. Moreover, if misleading causal statements are prevalent in the literature, professionals may believe these statements are justified and may offer unsubstantiated interpretations of their findings when engaging in outcomes assessment. Ultimately, when studies with poor methodological quality and incorrect inferences are routinely published, it "not only reduces the faith placed in the findings from studies examining the effectiveness of a specific intervention, but it undermines the faith that policymakers, practitioners, and the public at large place in the educational research enterprise" (Robinson et al., 2018, p. 12). Despite this reality, a formal review of the credibility of inferences about program effectiveness published in student affairs journals has not been conducted. Thus, in the current study, we systematically examined the quality of program effectiveness studies published in four student affairs journals. Using a rigorous approach, we appraised the validity of causal inferences made about program effectiveness (i.e., extent to which programming impacts intentional outcomes) given the studies' designs, data, and analyses.

It is important to note that not all assessment endeavors can be expected to support causal inferences, nor are we calling for the abandonment of outcomes assessment that does not meet the criteria of program effectiveness research.

It is important to note that not all assessment endeavors can be expected to support causal inferences, nor are we calling for the abandonment of outcomes assessment that does not meet the criteria of program effectiveness research. Instead, we are calling for honesty and transparency in the inferences made from program effectiveness studies, as these inferences may influence programming, implementation, and assessment decisions. In fact, Upcraft and Schuh (2002) noted that professionals assessing programming, particularly in published assessment studies, must describe any limitations, stating

> Failure to take this step is not only unethical, it leaves readers to assume that because the investigators did not identify limitations, they must not know them (or worse yet, they made a conscious decision to leave them out), and therefore both the investigators and the study itself lack credibility. (p. 20)

Although they show many commonalities, there are differences between "assessment" and "research", including the purpose, context, use, audience, and role of the researcher or assessment professional (Grey, 2002; Henning & Roberts, 2016; Yousey-Elsener, 2019). The intended generalizability of the findings is another distinction between assessment and research (Upcraft & Schuh, 2002). Assessment reports are intended to represent the local institution, rather than provide broadly generalizable findings. Moreover, data produced via the outcomes assessment process do not typically afford inferences about program or curriculum effectiveness. Thus, it is critical to build programming that *should* be effective based on previous research, often found in the form of program effectiveness studies (Finney et al., 2021). Recognizing this need, we focused on inferences stated in published program effectiveness studies, which student affairs professionals may read and use to build programming on their campuses.

Previous Reviews of Published Articles

Previous reviews of the methodological characteristics of research published in higher education and student affairs journals are limited. Moreover, these reviews tallied study characteristics rather than appraised the credibility of inferences given the characteristics. Common themes among the reviews were frequent use of quantitative techniques, such as regression analysis (Ferrao 2020; Hutchinson & Lovell, 2004; Johnson et al., 2016; Volkwein et al., 1988; Wells et al., 2015) and infrequent use of rigorous experimental or quasi-experimental designs (Hutchinson & Lovell, 2004; Volkwein et al., 1988; Wells et al., 2015). Non-probability sampling (Langrehr et al., 2015), descriptive research (Kuh, Bean, Bradley, & Coomes, 1986; Kuh, Bean, Bradley, Coomes, & Hunter, 1986) and cross-sectional designs (Kuh, Bean, Bradley, & Coomes, 1986; Kuh, Bean, Bradley, Coomes, & Hunter, 1986; Langrehr et al., 2015) were common in student affairs journals and journals focused on understanding college students. Moreover, one review reported that only one third of the

studies used theory to guide the research, resulting in weak to non-existent connections between the current study and prior research (Langrehr et al., 2015).

None of the reviews evaluated the credibility of the authors' inferences given the research design, sampling, and analyses. None of the reviews summarized threats to the validity of inferences (Murnane & Willett, 2011; Shadish et al., 2002). Thus, we undertook this task for published studies in several student affairs journals to provide insight into the trustworthiness of claims regarding program effectiveness. Given our aim was to inform outcomes assessment and learning improvement practice, a summary of findings, didactic explanation, and call to action follow.

Method

Position Statement

We position ourselves as assessment specialists and higher-education researchers with a primarily post-positivist research orientation. While valuing other paradigms, we acknowledge the methods and results below promote a quantitative research methods paradigm when evaluating program effectiveness. This is intentional, given historical dialogue about causality (Shadish et al., 2002). Similar to other methodologists and interventionists (e.g., Robinson et al., 2018), we believe the best evidence upon which to base recommendations for programming is that which allows for causal claims.

It is important to emphasize that, despite the choice of quantitative studies as the focus of this manuscript, we value qualitative approaches as useful, legitimate, and sound approaches to assessment (Suskie, 2018) that we also use in practice. However, although the logic underlying causality does not differ across quantitative and qualitative approaches, the way in which data are viewed and interpreted does differ (Shadish et al., 2002). Therefore, to keep the study within a manageable scope and within our personal areas of expertise, we chose to focus on quantitative studies of program effectiveness. Additional studies that review effectiveness inferences based on qualitative data would be useful but were not included in this study.

Article Sources

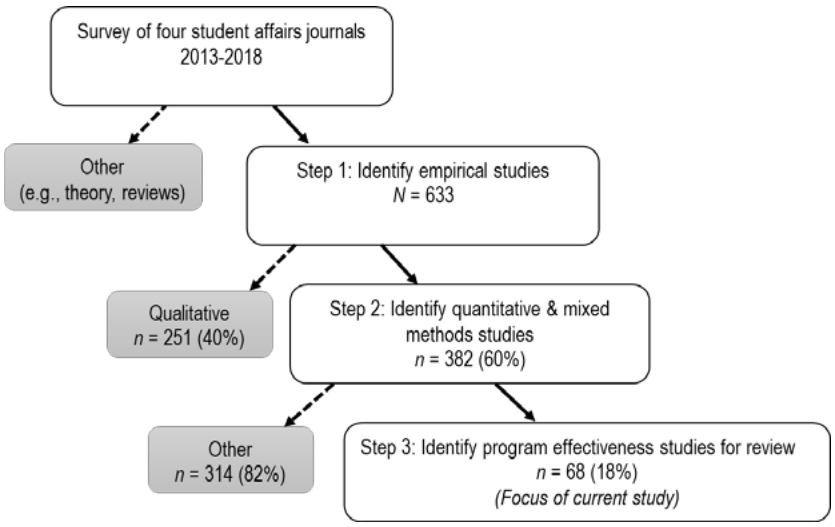
It is important to emphasize that, despite the choice of quantitative studies as the focus of this manuscript, we value qualitative approaches as useful, legitimate, and sound approaches to assessment (Suskie, 2018) that we also use in practice.

We reviewed articles published in four journals: *College Student Affairs Journal*, *Journal of College Student Development*, *Journal of Student Affairs Inquiry*, and *Journal of Student Affairs Research and Practice*. Three of the four journals are perceived as prestigious (Bray & Major, 2011), with the fourth (JSAI) being a new outlet. They are affiliated with student affairs organizations (e.g., ACPA, NASPA, and Student Affairs Assessment Leaders), have editorial boards, and conduct double-blind peer review. In our decades of experience working with student affairs colleagues, these are the journals they often reference, which aligns with studies of readership (Bray & Major, 2011). We reviewed five years (2013-2018) of articles for two reasons. First, research suggests statistical techniques tend to be stable over five years (Goodwin & Goodwin, 1985), and methodological approaches tend to be stable over 15 years (Volkwein et al., 1988). Second, the journals published issues between one and six times per year, resulting in an adequate sample of articles.

Article Selection

All 2013 to 2018 issues of the journals were examined. The process of selecting articles to review is shown in Figure 1. There were 633 published empirical studies (Step 1), comprised of 382 (60%) quantitative or mixed methods studies and 251 (40%) qualitative studies (Step 2). Of the quantitative or mixed methods studies, 68 (18%) reported an effectiveness study (Step 3). We retained quantitative studies in this step because our goal was to evaluate use of quantitative methods, analytic tools, and inferences. Although qualitative studies are essential to answering many questions about programming (e.g., implementation issues), the purpose of this study focused solely on the evaluation of quantitative program effectiveness studies.

Figure 1
Procedure for Selecting Program Effectiveness Studies for Review



The 68 program effectiveness studies were the focus of our review (references available upon request). A study was classified as an effectiveness study if it included *both* a program *and* an intentional student learning or development outcome. For example, a study of an alternative break program that evaluated program effectiveness with respect to influencing students' openness to diversity (intentional outcome) would be included in the current study. All 68 studies included a purpose statement or research question articulating that effectiveness was evaluated in terms of whether or not student learning or development outcomes were met. Some studies involved specific interventions on a single campus (n = 36, 53%), whereas others involved general interventions (e.g., alternative spring break) on multiple campuses (n = 32, 47%). Both were deemed effectiveness studies when effectiveness was considered relative to specific outcomes that were assessed. We did not review articles describing experiences (e.g., living on campus) that were not explicitly linked to intended student outcomes.

A study was classified as an effectiveness study if it included both a program *and* an intentional student learning or development outcome.

Rating Process

Five higher-education assessment professionals (two faculty members, two doctoral-level graduate students, one masters-level graduate student) rated the articles. The faculty members are formally trained in and teach quantitative methods and research design. The doctoral students each completed terminal master's degrees and multiple years of doctoral-level quantitative and research methods coursework. The masters-level graduate student completed multiple statistics and research methods courses and was completing an empirically-focused thesis. Combined, the raters have 50 years of experience in outcomes assessment.

Rating criteria (see Table 1) were based on recommendations in classic research methods texts (e.g., Shadish et al., 2002). During the initial two weeks of rating, all raters evaluated the same articles. Doing so permitted group discussion about the interpretation of rating criteria. Following the initial calibration weeks, each remaining article was evaluated by at least two raters (faculty-student or faculty-faculty pairing). Each of the raters individually rated their assigned articles and then met with another rater to adjudicate ratings, which then were combined into one spreadsheet for analysis. Quantitative analyses were conducted using SPSS 24. Study limitations noted by the 68 studies' authors and the open-ended rater comments were coded using NVivo 12 Pro.

Table 1
Rating Criteria for Published Program Effectiveness Studies

Criterion	Response Option
Citation Information	Journal, Volume, Issue, Year, Pages, Author, Title
General Information	
Type of study	Quantitative/mixed methods
Purpose of study	Description from article
What is (are) the measured outcome(s)?	Open-ended description
General intervention or specific program	General/Specific
Description of program or intervention	Open-ended description
Primary or secondary data source?	Primary/secondary
If secondary, what data were used?	Description of secondary data source
Information about Research Design	
Was there a comparison group?	Yes, No, Not clear (and open-ended description)
Was group membership self-reported?	Yes, No, Not clear
Was there random assignment to groups?	Yes, No, Not clear
Number of measurements of outcome	Number and description (e.g., pre-and post-test)
Additional details about research design	Open-ended description
What limitations did authors note?	Open-ended description
What limitations *should* be noted?	Open-ended description
Information about Sampling	
Sample size	Open-ended description
Was there random sampling?	Yes, No, Not clear
Was attrition noted?	Yes, No
Was attrition problematic? (and describe)	Yes, No, Not clear (and open-ended description)
What limitations did the authors note?	Open-ended description
What limitations *should* be noted?	Open-ended description
Information about Analyses	
Analysis	Open-ended description
Covariates	Open-ended description
Was analysis appropriate given *data collected?*	Yes, No (If no, then explanation)
Was analysis appropriate given *purpose of study?*	Yes, No (If no, then explanation)
Were inferential tests appropriately interpreted?	Yes, No (If no, then explanation)
Were effect sizes reported?	Yes, No (If yes, then description of type)
Were effect sizes appropriately interpreted?	Yes, No (if no, then explanation)
What limitations *should* be noted?	Open-ended description
Overall Conclusions	
What was the inference?	Open-ended description
Was the inference appropriate (given purpose, design, and analyses)?	Yes, No (if no, then explanation)
What was the inference?	Open-ended description
Was the inference appropriate (given purpose, design, and analyses)?	Yes, No (if no, then explanation)

Results

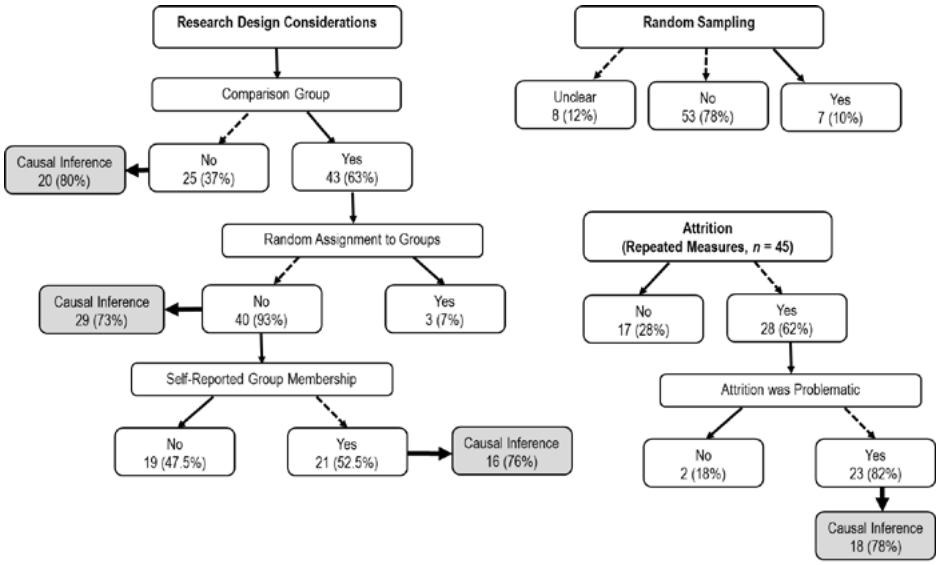
Of the 68 studies, 41 (60%) reported on data obtained from primary sources (i.e., new data collected for that particular study), whereas 27 (40%) reported on data from secondary sources (i.e., existing data collected by others).

Research Design

The sections that follow provide a summary of the research designs in the 68 reviewed studies. Figure 2 provides an overview of the findings that are described below.

Comparison group. A comparison group permits researchers to evaluate whether changes in learning or development may be attributed to causes other than the programming. In many

Figure 2
Design, Sampling, Attrition, and Causal Inferences for 68 Studies Reviewed



Note. Dashed arrows indicate threats to causal inferences. Gray boxes indicate the prevalence of unjustified causal inference.

cases, these changes may be just as feasibly due to natural development of students (i.e., maturation threat) or some other event that occurred at the same time as the intervention (i.e., history threat). Of the 68 studies, 43 (63%) included a comparison group, whereas 25 (37%) did not. Thus, for one-third of the studies, causal inferences about program effectiveness cannot be drawn as many threats to validity cannot be ruled out (see Table 2).

Acknowledging that within educational research it may not be feasible nor ethical to randomly assign students to groups, we expected the number of RCTs to be low. Of the 43 studies using a comparison group, three studies randomly assigned students to groups.

Random assignment. When interested in causal conclusions about a program's effectiveness, random assignment to groups (RCTs) is preferred, and otherwise are prone to self-selection bias (e.g., Shadish et al., 2002). For example, students who self-select into a service-learning program may be more apt to gain skills related to the outcome (e.g., cultural competence) than students who did not self-select due to pre-existing differences between groups in other variables (e.g., appreciation for diversity). Larger gains for the service-learning participants may be misinterpreted as positive program effects, when in fact the gains may have occurred with no programming.

Acknowledging that within educational research it may not be feasible nor ethical to randomly assign students to groups, we expected the number of RCTs to be low. Of the 43 studies using a comparison group, three studies randomly assigned students to groups. Of the 40 studies lacking random assignment to groups, 21 (53%) operationalized group membership through student self-report, frequently through retrospective self-reporting at post-test. If causal inferences are drawn from these non-RCT studies, they are tenuous, and it is necessary to note internal validity threats.

Number of time points. Collecting data at multiple time points permits evaluation of change over time. For program effectiveness studies, this typically means collection of data prior to and following programming, at a minimum. However, any inference that this change was due to programming is prone to validity threats associated with history, maturation, testing, and instrumentation. The addition of a comparison group aids with investigating these threats.

The number of time points for outcome measurements varied across studies. The most common design was pre-post (48%), followed by single-time point (34%) designs. The remainder included 3 (12%) or 4 time points (6%). Notably, the three RCTs included multiple time points. These RCTs, unlike single-group designs with multiple timepoints, directly address history and maturation effects.

Table 2
Appropriate Inferences Related to Specific Design Features

	Is __ a plausible threat?					
Description of Design	Selection	History	Maturation	Testing	Instrumentation	Appropriate Inference
RCT (random assignment) with a. random sampling b. no attrition c. pre- & post-test	No	Explore	Explore	Explore	Explore	Can infer cause-effect.
RCT with a. NO random sampling b. no attrition c. pre- & post-test	No	Explore	Explore	Explore	Explore	Results may not be generalizable to the population of interest given lack of random sampling. Otherwise, can infer cause-effect.
RCT with a. random sampling b. attrition is present c. pre- & post-test	Yes	Explore	Explore	Explore	Explore	If causal claims are desired, the plausibility of attrition as a threat must be considered. If attrition is non-random, characteristics of students remaining in the sample may lead to the appearance of an effect, when there is none.
RCT with a. random sampling b. no attrition c. No pre-test	No	Explore	Explore	No	No	Causal claims about the effect of the program should be made cautiously. Random assignment and control group data help to strengthen the claim, but there is no record of participants' outcomes scores prior to the program.
Two-Group Pre- & Post-Test Quasi-Experiment (same as first design, but no *random assignment*)	Yes	Explore	Explore	Explore	Explore	Variables related to selection into the program need to be considered as plausible threats to the accuracy of cause-effect claims.
One-Group Pre- & Post-Test (nothing else)	Yes	Yes	Yes	Yes	Yes	Causal claims should not be made without consideration of the plausibility of all threats. Exceptions may be when the information taught is *so specific or unusual* that the students would not have learned the information elsewhere.
One-Group Post-Test (nothing else)	Yes	Yes	Yes	No	No	Causal claims should not be made. Exceptions may be when the information taught is *so specific or unusual* that the students would not have learned the information elsewhere.

Note. "Yes" = a plausible threat. "No" = not a threat. "Explore" = threat may be plausible, but can be ruled out through group comparison. Selection = students selecting into or assigned to program differ on some variable related to outcome. History = event occurring at the same time as program may influence outcome. Maturation = students naturally develop or grow on outcome. Testing = changes in students' approach to completing an outcome measure (e.g., social desirability). Instrumentation = changes in test administration (e.g., modality, stakes).

Design limitations noted by authors. Transparency about validity threats is key to maintaining the credibility of program effectiveness studies (Levin & Kratochwill, 2013). Authors need to scrutinize causal claims and address limitations to the validity of those claims. Therefore, we recorded limitations noted in each of the studies. Of the 68 studies, 26 (38%) did not mention limitations related to research design. The remaining 42 (62%) studies' design limitations were coded into the broad themes of threats to validity, data, design, and operationalization of independent and dependent variables. The most mentioned threat to validity was selection bias (n = 14). Instrumentation, history, directionality of effect, and contamination were each mentioned once. Data limitations noted by studies' authors included lack of comprehensive sets of covariates, self-reported data, small sample size, and archival, retrospective, or secondary data. The most commonly mentioned design limitation was lack of random assignment (n = 14), cross-sectional/single-time point design (n = 10), and lack of control group (n = 5). Finally, 11 articles cautioned about the operationalization of treatment condition, particularly self-reported group membership.

Sampling and Attrition

If interested in representativeness of a population, then random sampling (or census data) without differential attrition is critical (Shadish et al., 2002). We reviewed characteristics related to sampling and attrition.

Random sampling. Of the 68 studies, 7 (10%) reported random sampling. For the remaining studies, 53 (78%) did not employ random sampling and 8 (12%) were unclear about sampling method. Some of the large secondary data sources reported initial random sampling, but data were retrieved only for specific subgroups that were not randomly sampled. Most reported high rates of attrition, which negated benefits of random sampling.

Attrition. Of the 45 repeated measures studies, 28 (62%) reported attrition. Of these, we rated 23 (82%) instances as problematic, based upon the percent of attrition and lack of acknowledgement of attrition as an issue. For example, it was common for 26% to 50% of students to provide data at time-point 1 but not time-point 2. Of the three RCT studies, one reported random sampling with non-problematic attrition.

Sample size. When tabulating sample size, we included the final sample size reported for analyses. When there were multiple samples, we recorded the size of the largest sample. Sixty-five studies reported sample sizes, ranging from a minimum of 8 participants to a maximum of 15,847. The median sample size was 436 (25th percentile = 100; 75th percentile = 1,502).

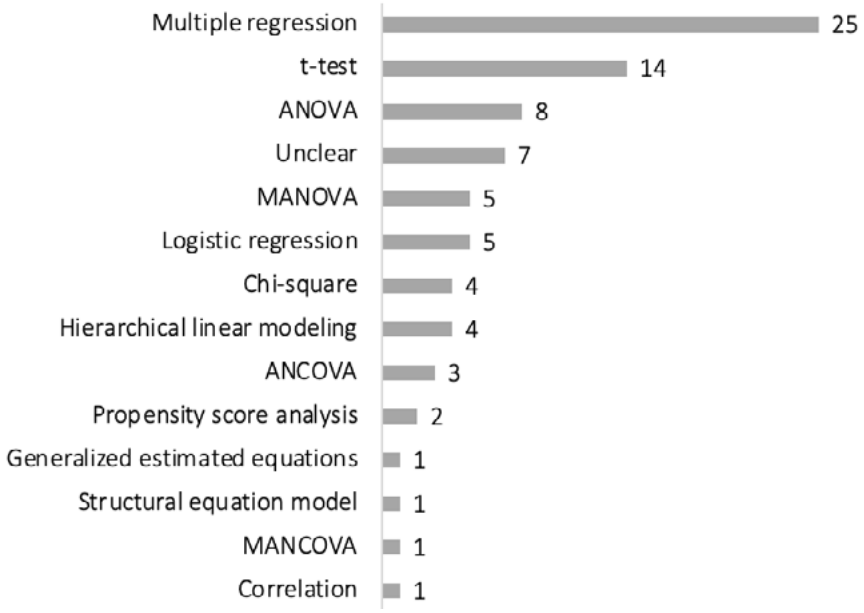
Sample limitations noted by authors. Of the 68 studies, 27 (40%) did not note limitations regarding sampling or attrition. Commonly mentioned limitations were generalizability (n = 15) or the sample composition was not representative of the population (n = 23). Eleven noted issues related to response rate or attrition, and eight noted small sample size. Other limitations included non-random sampling (n = 5), convenience sampling (n = 2), inadequate demographic information (n = 4), and clustered data (n = 1).

Notably absent for most of these studies were tests of the assumption of homogeneity of regression slopes. Without reporting of this critical assumption, we could not determine if the analyses or their interpretation were appropriate.

Analysis

Type. Types of analyses varied across the 68 studies (see Figure 3). Note, the numbers reported in Figure 3 total to greater than 68, because some studies involved more than one type of analysis. Notably absent from many studies were descriptive statistics (e.g., means, standard deviations). Of the 25 studies employing multiple regression, 23 were ANCOVA-type analyses that included a grouping variable (e.g., intervention versus comparison group) and covariates employed as "control" variables. These studies included 3 to 22 covariates (median = 8). Common choices for covariates were pre-test scores on the outcomes, demographic characteristics (e.g., gender), students' pre-college characteristics (e.g., high school involvement), and student- and institution-level college characteristics (e.g., students' major, type of college). Notably absent for most of these studies were tests of the assumption of homogeneity of regression slopes. Without reporting of this critical

Figure 3
Analyses reported in the 68 studies reviewed



assumption, we could not determine if the analyses or their interpretation were appropriate. Consequently, we assigned ratings of "inappropriate" to these analyses and interpretations of their inferential tests.

Many studies did not report descriptive statistics (e.g., distributions of scores, standard deviations). Therefore, readers cannot self-assess if the data align with statistical assumptions and whether the presence of floor or ceiling effects explain the lack of change in the outcome from pre- to post-intervention.

Appropriateness of analysis. Proper statistical analysis is necessary to achieve statistical-conclusion validity (Shadish et al., 2002). We evaluated the appropriateness of the analysis by examining if it (and its associated statistical assumptions) aligned with the type of data being modeled and if it aligned with the purpose of the study (i.e., research questions).

Of the 68 studies, 32 (47%) clearly aligned the analyses to the type of data collected. Twenty-five (37%) studies reported analyses that were misaligned to the data (e.g., ANOVA with continuous predictors requiring artificial categorization of predictors). An additional 6 studies (9%) clearly aligned some analyses to the type of data collected, while at the same time misaligning other analyses. For 5 (7%) studies, it was unclear from the studies' description whether the analyses were aligned or not.

Assumption testing was seldom reported. In addition to the lack of testing the homogeneity of regression assumption for ANCOVA, there was infrequent discussion of variability- or distribution-related assumptions. Analyses conducted to explain variability in continuous outcomes (e.g., regression, ANOVA) lack utility if there is little variability in outcomes to explain. Many studies did not report descriptive statistics (e.g., distributions of scores, standard deviations). Therefore, readers cannot self-assess if the data align with statistical assumptions and whether the presence of floor or ceiling effects explain the lack of change in the outcome from pre- to post-intervention.

Of the 68 studies, 27 (40%) appropriately aligned their statistical analyses with the purpose of the study (i.e., research question posed). Fourteen studies (16%) were unclear. However, 30 (44%) studies reported analyses misaligned with the research question (e.g., a research question about differential change in the outcome across intervention and comparison groups without testing the hypothesized interaction). Thus, for 30 studies, the results presented could not provide answers to the research questions posed.

Interpretation. Of the 68 studies, 31 (46%) appropriately interpreted the inferential statistical tests. Eight (11%) studies were unclear. Common themes among the remaining 29 (43%) included implying or misinterpreting main effects in the presence of interactions and

noting "significance" of results without conducting inferential tests. Results sections with unclear terminology and a mismatch between text and table information led to difficulty in interpretation.

Effect size (ES). Measures of ES aid in understanding the practical significance of findings. Of the 68 studies, 58 (85%) reported ES measures. Ten (15%) studies reported ES measures for some but not all analyses or did not report any ES measure.

The findings of our review are a result of curiosity about the quality of program effectiveness evidence published in student affairs journals, given expectations that professionals use research to identify programming that impacts desired outcomes.

As expected, type of ES varied by type of analysis. For 15 of the 25 multiple regression analyses that included measures of ES, 12 reported R^2 for the model, 1 reported R^2-change,10 reported standardized coefficients, and 2 reported unstandardized coefficients. When reporting t-test findings, out of the 14 studies, 1 reported Cohen's *d* and 4 reported raw mean differences. When reporting ANOVA findings, several reported eta-squared and partial eta-squared.

Although most authors provided effect size values, few authors interpreted those values for readers. Of the 68 studies, 24 (35%) both presented and explained ES values, thereby providing an interpretation of the practical significance of their findings. The remainder (65%) did not report ES, did not interpret ES, or inaccurately interpreted ES.

Causal Inferences

Given the focus of these studies, authors made inferences from results regarding program effectiveness. We evaluated the appropriateness of causal inferences. An inference was flagged for review if the discussion of, or implications from, the findings were reported with wording, such as Program X "*impacted*," "*affected*," or "*led to gains in*" outcome Y. Implications sections commonly included program suggestions informed by the study's results, implying a causal relation between programming and outcomes. If authors uncovered non-significant results and inferred a non-causal relation, we evaluated the inference regarding a lack of causality for alignment to the research design and analyses.

Of the 68 studies evaluated, one study (Thatcher, 2016) was able to make an appropriate causal inference given its design, data, and analyses. Of the remaining 67 studies, 12 (18%) drew appropriate non-causal inferences from the findings, remaining tentative about the causal impact of programming on the outcome. The remaining 55 (82%) studies included a causal claim in the results, discussion, or implication sections of the article.

To better understand when inappropriate causal inferences were made, we examined the extent to which authors drew causal inferences when employing research designs that did not support such inferences. Of the 23 single-time-point design studies, 12 (52%) made causal inferences. Of the 25 studies with no comparison group, 20 (80%) included causal inferences. Of the 40 studies with a comparison group, but non-random assignment, 29 (73%) included causal inferences. Of the 21 studies with non-random assignment and for which students self-reported group membership, 16 (76%) included causal inferences.

We recorded statements from the 68 studies' results, discussion, and implications sections that led to the rating of "inappropriate causal claim." The statements were coded for themes. The most commonly identified theme was "effect of Program X on outcome Y." Other common phrasings included "benefits of," "influenced," "improved/promoted," "result of participation in," "efficacy/effectiveness," "fosters," "successful program," "transformative," and "reduced." Another common theme was the implication of no program effect on the outcome given non-significant results (e.g., "Program X has *no impact* on outcome Y"). Nonetheless, the results were often used inappropriately to argue for or against future or additional programming to impact the particular outcome.

Discussion

When discussing peer review, Carpenter (2001) noted: "This is not a call to be critical of each other as people, but to be very critical of our work and our results. Scholars evaluate each other's work" (p. 305). The findings of our review are a result of curiosity about the quality of program effectiveness evidence published in student affairs journals, given expectations that professionals use research to identify programming that impacts

desired outcomes. Using criteria in Table 1, we examined 68 program effectiveness studies published between 2013 and 2018.

Similar to previous methodological reviews (Hutchinson & Lovell, 2004; Johnson et al., 2016; Wells et al., 2015), statistics such as multiple regression, *t*-tests, and ANOVA were the most common analyses. Unlike Wells and colleagues (2015), who noted frequent reporting of descriptive statistics, the studies we reviewed did not typically report descriptive statistics. Notably absent was reporting of assumptions testing, threatening the validity of conclusions drawn from analyses. Despite frequent claims of program effectiveness, the research design and analyses did not often support such claims due to highly-plausible threats to the validity of those claims. This finding is not new. In their review of 21 years of research, Reinhart and colleagues (2013) noted an increase in causal inferences drawn from correlation studies.

Moreover, many causal claims were unaccompanied by acknowledgment of limitations or threats to the validity of these inferences. To provide credible and trustworthy evidence of program effectiveness, at a minimum, professionals need to acknowledge plausible threats to the validity of causal claims (Levine & Kratochwill, 2013; Shadish et al., 2002). Consider a hypothetical service-learning course with the following outcome: "As a result of participation, students will demonstrate increased openness to diversity". Openness to diversity is assessed before and after the course for students who opt to participate and is found to increase. The following are plausible threats to the validity of the causal statement that the service-learning course caused (or "led to") increased openness to diversity: 1) selection bias (e.g., students interested in diversity enroll in the course), 2) attrition (e.g., uninterested students drop out of the course or skip the post-test), 3) history (e.g., another event on campus influenced the outcome), 4) maturation (e.g., students naturally develop openness to diversity), 5) testing (e.g., students respond differently to the post-test because they realize the focus on diversity), and 6) instrumentation (e.g., instructors communicate greater importance of the post-test than the pre-test, resulting in higher scores at post-test). It is essential to critically evaluate evidence and report plausible threats to the accurate interpretations of findings if program effectiveness studies are to be trustworthy (Upcraft & Schuh, 2002). Table 2 provides a concise guide to evaluate studies for threats to validity.

Call for Action

Research-to-practice efforts require being able to understand research. One course in statistics and research methods is not enough if professionals are expected to evaluate the credibility of inferences in effectiveness studies.

We understand that gathering rigorous evidence of effective programming is challenging. Random selection or random assignment of students to programs may not be feasible. The collection of pre-post data with a comparison group may not be feasible. Moreover, given the low-consensus nature of the student affairs profession (Torres et al., 2019) and higher education in general (Wells et al., 2015), limiting "evidence" to RCT studies risks over-narrowing the information available to professionals. We are *not* advocating for RCT studies as the only way to assess program effectiveness. Instead, we are advocating for professionals to 1) acknowledge threats to the validity of causal inferences, 2) draw appropriate inferences given the plausibility of threats for a specific research design, and 3) consider quasi-experimental designs that can support causal inferences in the absence of RCTs (regression discontinuity designs, interrupted time series designs, propensity score matching; Murnane & Willett, 2011). All of these support research-to-practice efforts called for in the domains of student affairs (Finney & Horst, 2019a, 2019b) and outcomes assessment (Horst & Prendergast, 2020; Singer-Freeman & Robinson, 2020).

We also echo calls for changes in graduate school training, journal review practices, and professional organization practices (Carpenter, 2001; Malaney, 2002; Wells et al., 2015). In the 68 studies reviewed, lack of clarity when describing designs and analyses suggested that some authors were not familiar with the methods they were using. Professionals must have a repertoire of research techniques not only to conduct research but to evaluate its quality (Schroeder & Pike, 2001). Research-to-practice efforts require being able to understand research. One course in statistics and research methods is not enough if professionals are expected to evaluate the credibility of inferences in effectiveness studies. Without increased training in methodology, assessment professionals will need to provide support to colleagues who are unable to independently evaluate the quality of research. With that said, assessment

professionals themselves may need to acquire additional knowledge of statistics and research methods (Curtis et al., 2020), in order to fulfill this role.

Journal editors and reviewers can also contribute to an increase in the quality of evidence. Through the double-blind peer review process, journals aim to provide content and inferences that are scrutinized and shared to improve practice (Liddell, 2019). Professionals with methodological expertise must volunteer time to the review process and hold the profession to high standards. If published studies include misinformation, the burden then falls on readers to evaluate research credibility. Rigorous review processes can reduce this burden.

Moreover, when journals require an implications section, researchers face conflicting roles, in which they need to accurately convey the limited inferences from their single study and yet are asked to speculate about broad implications for practice (Robinson et al., 2013). In doing so, the temptation is to fall into causal language. Consequently, if readers skip over methods and results sections and head straight to the discussion section, they are likely to believe the causal implications. To address this issue, Robinson and colleagues (2013) suggested the following be added to education research journal policies: "Contributors should restrict their discussion and conclusions to their data and not offer recommendations for educational practice nor speculate about the educational policy implications of their research" (p. 291). Instead, they recommended that implications from research be developed via conversation among practitioners. Professional organizations are a venue for such conversations. Organizations can also influence the quality of implications from these conversations by providing training on causal inferences.

Professionals creating programming must work to become fluent in their critiques of published literature. These skills can be developed through critical reading of published research.

Professionals creating programming must work to become fluent in their critiques of published literature. These skills can be developed through critical reading of published research. Methodological review articles, such as the current study, expose readers to the variable quality of published research. These critiques also provide useful training in identifying and understanding links between design, results, and interpretations.

Finally, assessment professionals should ask fundamental questions about program rationale when supporting colleagues engaged in outcomes assessment. Simple questions such as "What evidence supports the belief that this strategy/program will result in that student learning outcome?" may reveal that no credible evidence exists to support a programming decision (Finney & Buchanan, 2021). This awareness may provide insight into disappointing assessment results (i.e., no student learning) and struggles with learning improvement efforts. This awareness may also spur frustration for professionals who spent years implementing and assessing programming they believed would be effective given published claims. Assessment professionals can help colleagues process this frustration, frame this realization as an opportunity, and locate credible evidence of effectiveness to build should-be-effective programming.

References

American College Personnel Association & National Association of Student Personnel Administrators (2015). *ACPA/NASPA professional competency areas for student affairs educators*. Authors.

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

Bickman, L. & Reich, S. (2015). Randomized controlled trials: A gold standard or gold plated? In S.I. Donaldson, C.A. Christie, & M.M. Mark (Eds.), *Credible and actionable evidence: The foundation for rigorous and influential evaluations* (2nd ed., pp. 83-113). Sage. https://dx.doi.org/10.4135/9781483385839

Bray, N., & Major, C. (2011). Status of journals in the field of higher education. *The Journal of Higher Education*, *82*, *479-503*. https://www.jstor.org/stable/29789535?seq=1

Carpenter, S. (2001). Student affairs scholarship (re?)considered: Toward a scholarship of practice. *Journal of College Student Development*, *42*, 301-318.

Cook, B. G., Smith, G. J., & Tankersley, M. (2011). Evidence-based practices in education. In K. R. Harris. S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook*, volume 1 (pp. 495–528). Washington, DC: American Psychological Association. https://doi.org/10.1037/13273-017

Cooper, J., Mitchell, D., Eckerle, K., & Martin, K. (2016). Addressing perceived skill deficiencies in student affairs graduate preparation programs. *Journal of Student Affairs Research and Practice*, *53*, 107-117. https://doi.org/10.1080/19496591.2016.1121146

Curtis, N.A., Anderson, R. D., & Van Dyke, R. (2020). A field without a discipline? Mapping the uncertain and often chaotic route to becoming an assessment practitioner. *Research and Practice in Assessment*, *15*(2), 1-8. https://www.rpajournal.com/dev/wp-content/uploads/2020/08/A-Field-Without-A-Discipline.pdf

Deniston, O., Rosenstock, I., & Getting, V. (1968). Evaluation of program effectiveness. *Public health reports*, *83*(4), 323.

Ferrao M. E. (2020). Statistical methods in recent higher education research. *Journal of College Student Development*, *61*, 366-371. https://doi.org/10.1353/csd.2020.0033

Finney, S. & Buchanan, H. (2021). A more efficient path to learning improvement: Using repositories of effectiveness studies to guide evidence-informed programming. *Research & Practice in Assessment*, *16*, 36-48. https://www.rpajournal.com/a-more-efficient-path-to-learning-improvement-using-repositories-of-effectiveness-studies-to-guide-evidence-informed-programming/

Finney, S. & Horst, S. (2019a). Standards, standards, standards: Mapping professional standards for outcomes assessment to assessment practice. *Journal of Student Affairs Research and Practice*, *56*, 310-325. https://doi.org/10.1080/19496591.2018.1559171

Finney, S. & Horst, S. (2019b). The status of assessment, evaluation, and research in student affairs. In V. L. Wise & Z. Davenport (Eds.), *Student affairs assessment, evaluation, and research: A guidebook for graduate students and new professionals* (pp. 3-19). Charles C. Thomas, Publisher, Ltd. https://ebookcentral.proquest.com/lib/jmu/detail.action?docID=5722940

Finney, S., Wells, J., & Henning, G. (2021). *The need for program theory and implementation fidelity in assessment practice and standards* (Occasional Paper No. 51). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2021/03/Occ_Paper_51-1.pdf

Fulcher, K., & Prendergast, C. (2019). Lots of assessment, little improvement? How to fix a broken system. In S. Hundley & S. Kahn (Eds.), *Trends in assessment: Ideas, opportunities, and issues in higher education* (pp. 157-174). Stylus.

Grace-Odeleye, B., & Santiago, J. (2019). A review of some diverse models of summer bridge programs for first-generation and at-risk college students. *Administrative Issues Journal: Education, Practice & Research*, *9*, 35-47. https://doi.org/10.5929/9.1.2

Goodwin. L., & Goodwin, W. (1985). Statistical techniques in *AERJ* articles, 1979-1983: The preparation of graduate students to read the educational research literature. *Educational Researcher*, *14*, 5-11. https://www.jstor.org/stable/1174902

Gray, P. (2002). The roots of assessment: Tensions, solutions, and research directions. In T. W. Banta and Associates (Eds.) *Building a scholarship of assessment* (pp. 49-66). San Francisco, CA: Jossey Bass.

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

Hilligoss, B. & Rieh, S. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing and Management*, *44*, 1467-1484. https://doi.org/10.1016/j.ipm.2007.10.001

Horst, S.J. & Prendergast, C. (2020). The assessment skills framework: A taxonomy of assessment knowledge, skills and attitudes. *Research & Practice in Assessment*, *15*, 1-25. https://www.rpajournal.com/dev/wp-content/uploads/2020/05/The-Assessment-Skills-Framework-RPA.pdf

Hutchinson, S. & Lovell, C. (2004). A review of methodological characteristics of research published in key journals in higher education: Implications for graduate research training. *Research in Higher Education*, *45*, 383-403. https://link.springer.com/content/pdf/10.1023/B:RIHE.0000027392.94172.d2

Jankowski, N., Timmer, J., Kinzie, J., & Kuh, G. (2018). *Assessment that matters*: *Trending toward practices that document authentic student learning*. University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://files.eric.ed.gov/fulltext/ED590514.pdf

Johnson, M., Wagner, N., & Reusch, J. (2016). Publication trends in top-tier journals in higher education. *Journal of Applied Research in Higher Education*, *8*, 439-454. http://dx.doi.org/10.1108/JARHE-01-2015-0003

Kuh, G., Bean, J., Bradley, R., & Coomes, M. (1986). Contributions of student affairs journals to the literature on college students. *Journal of College Student Personnel*, *27*, 292–304.

Kuh, G., Bean, J., Bradley, R., Coomes, M., & Hunter, D. (1986). Changes in research on college students published in selected journals between 1969 and 1983. *Review of Higher Education*, *9*, 177-192.

Langrehr, K., Phillips, J., Melville, A., & Eum, K. (2015). Determinants of nontraditional student status: A methodological review of the research. *Journal of College Student Development*, *56*, 876-881. http://dx.doi.org/10.1353/csd.2015.0090

Levin, J. & Kratochwill, T. (2013). Educational/psychological intervention research circa 2012. In I. B. Weiner (Series Ed.), W. M. Reynolds & G. E. Miller (Volume Eds.), *Handbook of psychology*: *Vol. 7. Educational psychology* (2nd ed., pp. 465–492). Wiley.

Liddell, D. (2019). 60 years of scholarship. *Journal of College Student Development*, *60*, 641-644. http://dx.doi.org/10.1353/csd.2019.0059

Malaney, G. (2002). Scholarship in student affairs through teaching and research. *NASPA Journal*, *39*, 132-146. https://doi.org/10.2202/1949-6605.1795

Miller, R. (2015). How people judge the credibility of information: Lessons for evaluation from cognitive and information sciences. In S. Donaldson, C. Christie, & M. Mark (Eds.), *Credible and actionable evidence*: *The foundation for rigorous and influential evaluations* (2nd ed., pp. 83-113). Sage.

Muller, K., Grabsch, D., & Moore, L. (2018). Factors influencing student affairs professionals' attainment of professional competencies. *Journal of Student Affairs Research and Practice*, 55, 54-64. https://doi.org/10.1080/19496591.2017.1345755

Murnane, R. & Willett, J. (2011). *Methods matter*: *Improving causal inference in educational and social science research*. Oxford University Press.

Pope, A., Finney, S., & Bare, A. (2019). The essential role of program theory: Fostering theory-driven practice and high-quality outcomes assessment in student affairs. *Research & Practice in Assessment*, *14*, 5–17. https://files.eric.ed.gov/fulltext/EJ1223397.pdf

Reinhart, A., Haring, S., Levin, J., Patall, E., & Robinson, D. (2013). Models of not-so-good behavior: Yet another way to squeeze causality and recommendations for practice out of correlational data. *Journal of Educational Psychology*, *105*, 241-247. https://doi.org/10.1037/a0030368

Robinson, D., Levin, J., Graham, S., Schraw, G., Fuchs, L., & Vaughn, S. (2018). Improving the credibility of educational intervention research. In A. M. O'Donnell (Ed.), *Handbook of educational psychology*. Oxford University Press.

Robinson, D., Levin, J., Schraw, G., Patall, E., & Hunt, E. (2013). On going (way) beyond one's data: A proposal to restrict recommendations for practice in primary educational research journals. *Educational Psychology Review*, *25*, 291-302. https://doi.org/10.1007/s10648-013-9223-5

Schroeder, C., & Pike, G. (2001). The scholarship of application in student affairs. *Journal of College Student Development*, *42*, 342-355.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.

Singer-Freeman, K., & Robinson, C. (2020). *Grand challenges in assessment*: *Collective issues in need of solutions* (Occasional Paper No. 47). University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://files.eric.ed.gov/fulltext/ED612032.pdf

Smith, K., & Finney, S. (2020). Elevating program theory and implementation fidelity in higher education: Modeling the process via an ethical reasoning curriculum. *Research & Practice in Assessment*, *15*, 1-13. https://www.rpajournal.com/dev/wp-content/uploads/2020/09/Elevating-Program-Theory-and-Implementation-Fidelity-in-Higher-Education.pdf

Smith, K., Good, M., & Jankowski, N. (2018). Considerations and resources for the learning improvement facilitator. *Research & Practice in Assessment*, *13*, 20-26. http://www.rpajournal.com/dev/wp-content/uploads/2019/02/W18_A4.pdf

Suskie, L. (2018). *Assessing student learning*: *A common sense guide* (3rd ed.) San Francisco CA: Jossey-Bass.

Thatcher, W. G. (2016). FREAKS - A new program for student life and success. *College Student Affairs Journal*, *34*, 48-55. https://doi.org/10.1353/csj.2016.0002

Torres, V., Jones, S., & Renn, K. (2019). Student affairs as a low-consensus field and the evolution of student development theory as foundational knowledge. *Journal of College Student Development*, *60*, 645-658. https://doi.org/10.1353/csd.2019.0060

Upcraft, M., & Schuh, J. (2002). Assessment vs. research. *About Campus*, *7*, 16-20. https://doi.org/10.1177/108648220200700104

Valentine, J.C., Hirschy, A.S., Bremer, C.D., Novillo, W., Castellano, M., & Banister, A. (2011). Keeping at-risk students in school: A systematic review of college retention programs. *Educational Evaluation and Policy Analysis*, *33*, 214 – 234. https://doi.org/10.3102/0162373711398126

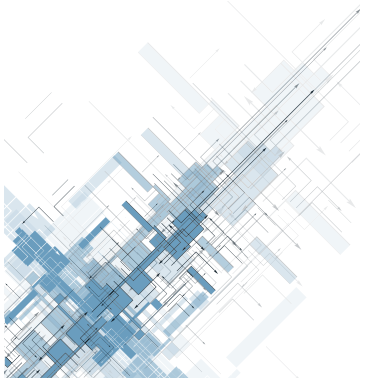
Volkwein, J., Carbone, D., & Volkwein, E. (1988). Fifteen years of scholarship. *Research in Higher Education*, *28*, 271-280. https://www.jstor.org/stable/40195866?seq=1

Wells, R., Kolek, E., Williams, E., & Saunders, D. (2015). "How we know what we know": A systematic comparison of research methods employed in higher education journals, 1996-2000 v. 2006-2010. *The Journal of Higher Education*, 86, 171-198. https://doi.org/10.1080/00221546.2015.11777361

Yousey-Elsener, K. (2019). Development of competencies in assessment, evaluation, and research, with terms and concepts. In V. L. Wise & Z. R. Davenport (Eds.), *Student affairs and assessment*, *evaluation*, *and research*: *A guidebook for graduate students and new professionals* (pp. 20-47). ProQuest Ebook Central https://ebookcentral.proquest.com

Abstract

Meta-assessment is a useful strategy to document assessment practices and guide efforts to improve the culture of assessment at an institution. In this study, a meta-assessment of undergraduate and graduate academic program assessment reports evaluated the maturity of assessment work. Assessment reports submitted in the first year (75 undergraduate and 35 graduate programs) provided baseline data. As part of implementation of revised reporting processes, the authors facilitated faculty workshops to promote effective assessment practices and increase the clarity of communication in assessment reports. Review of assessment reports submitted the following year (69 undergraduate and 41 graduate programs) evaluated the impact of institutional efforts to develop a more mature culture of assessment. Reviewers used a rubric to score assessment reports on reporting compliance, assessment maturity, and evidence of impact on student learning. Findings indicate reliable improvements in compliance and assessment maturity, but no evidence of efforts to evaluate impact on learning.



AUTHORS

Claudia J. Stanny, Ph.D.
University of West Florida

Angela A. Bryan, PhD.
University of West Florida

Meta-Assessment of the Assessment Culture: Using a Formal Review to Guide Improvement in Assessment Practices and Document Progress

Although higher education institutions have been engaged in the assessment of educational programs for several decades, they continue to struggle to meet the expectations for program-level assessment set by accreditors. Early standards for assessment emphasized sustained assessment efforts rather than episodic assessment (American Association for Higher Education, 1992, cited in Hutchings, Ewell, & Banta, 2012). However, institutional accreditors have shifted their focus to emphasize assessment work that "provides evidence of seeking improvement based on analysis of the results" (Southern Association of Colleges and Schools Commission on Colleges, 2020, p. 66). In addition, recent conversations around program-level assessment of student learning (in the United States) and evaluation of academic programs (in Europe and the UK) raise concerns about what impact (if any) these efforts have had on the quality of academic programs and student learning (e.g., Banta & Blaich, 2011; Blaich & Wise, 2011; Fulcher, Smith, Sanchez, & Sanders, 2017; Kuh, Jankowski, Ikenberry, & Kinzie, 2014).

The history of program-level assessment of student learning can be characterized by a continuing conflict between assessment for compliance and assessment for improvement (e.g., Blumberg, 2018; Stitt-Bergh, Kinzie, & Fulcher, 2018; Suskie, 2015, 2018; Walvoord, 2014). Assessment critics have argued that assessment processes represent little value beyond compliance with external mandates (Gilbert, 2018; Worthen, 2018). In contrast,

CORRESPONDENCE

Email
abryan@uwf.edu

professional organizations aligned with assessment advocate that mature assessment cultures should focus on the use of results to improve teaching, learning, and assessment (e.g., Association for the Assessment of Learning in Higher Education, Association of American Colleges & Universities, National Institute for Learning Outcomes Assessment). Although the requirements of external stakeholders such as government agencies and accrediting bodies can motivate efforts to assess student learning, external mandates tend to focus attention and effort on assessment for compliance (Stanny & Halonen, 2011; Suskie, 2015). However, institutions should nurture an assessment culture that focuses on improvement because this strategy can enhance the quality of academic programs (Isabella & McGovern, 2018; O'Neill, Slater, & Sapp, 2018; Lattuca, Terenzini, & Volkwein, 2006; Magruder, McManis, & Young, 1997).

That is, when assessment is done in the right way for the right reasons, accountability should take care of itself.

Fulcher and his colleagues describe a culture of assessment as one in which academic programs define learning outcomes, map outcomes to the curriculum, select an assessment instrument, collect assessment data, analyze and report the results, and communicate their findings to stakeholders (Fulcher, Good, Coleman, & Smith, 2014; Fulcher, Swain, & Orem, 2012). In a mature culture of assessment, the institution uses assessment processes and evidence as opportunities for self-reflection and identification of actions that might promote student learning (Fulcher et al., 2014; Fulcher et al., 2017; Lending, Fulcher, Ezell, May, & Dillon, 2018; Maki, 2010; Stanny, 2015, 2018, 2020; Suskie, 2015). Programs should assess and analyze student learning data, identify and implement changes to the curriculum and/or instructional methods (if needed), and then reassess to evaluate the impact of the implemented changes on student learning (Lending, et al., 2018; Stanny 2021). Discussions of changes implemented and how learning changed following implementation, grounded in an analysis of follow-up assessment findings, are the two most vital components of a culture of improvement and are often missing in assessment reports (Reder & Crimmins, 2018; Stitt-Bergh et al., 2018; Suskie, 2015, 2018).

How can institutions maintain accountability to external stakeholders yet still foster a culture of improvement? Wehlburg (2008, 2013) argues that programs can best meet accountability expectations when they assess with the goal of increasing program effectiveness. These programs focus on using assessment findings to identify promising areas to improve student learning and conduct follow-up assessments to determine whether implemented changes made a difference for student learning. When programs document these activities, they can meet expectations for accountability set by external stakeholders (Souza & Rose, 2021). That is, when assessment is done in the right way for the right reasons, accountability should take care of itself.

Meta-assessment, such as a formal review of assessment, can yield a wealth of useful information that serves multiple goals (Stanny, 2020). It can provide a broad description of the types of assessment practices in use, including evidence of efforts to use assessment results to improve programs. Findings can be used to guide future professional development efforts and campus interventions that promote the adoption of mature assessment practices. Dissemination of review findings communicates to faculty and administrators that assessment reports are read by multiple individuals. Because the findings provide formative feedback on both the quality of assessment processes and how well assessment reports communicate the program's assessment story to reviewers, a formal review can improve the quality of reporting. Walvoord (2014) offers general suggestions for how to "tell the story of how you are assessing and improving," (p. 45). Stitt-Bergh et al. (2018) identify five elements required to connect and align assessment activities and improvement initiatives to tell a compelling assessment story: clearly identify the learning targeted, specify the scope of the initiative (course, program, institution), identify specific changes and actions implemented, collect multiple types of evidence from two points in time to evaluate whether improvements occurred, and reflect on and interpret the assessment evidence.

Efforts to promote a mature culture of assessment have been guided by the framework of an assessment cycle, presented in guidelines for assessment (e.g., Maki, 2010; Suskie, 2018; Walvoord, 2014) and discussions of the characteristics of "mature" assessment, and expectations for documentation for institutional accreditation reports. Rubric elements articulate these goals in concrete language intended for the campus audience as part of

a pragmatic effort to transparently communicate expectations for assessment work and assessment reporting to chairs and faculty assessment committees.

The rubric tries to balance two points of view in language that will be understandable to the campus community. First, the compliance items reflect documentation needs established through prior experience preparing reports for external audiences (such as institutional and disciplinary accrediting bodies). Second, the maturity of assessment items reflect best practices in the literature and describe assessment processes that move beyond compliance and motivate efforts to improve student learning. The rubric is microscopic because we wanted to track the emergence of specific practices and document the number of departments that adopted each practice over time. Written in the spirit of rubrics for specifications grading advocated by Nilson (2015), these detailed, specific rubric items connect mature practices to unambiguous, concrete characteristics of assessment work that could appear in an assessment report. An added advantage of these concrete criteria is that the rubric elements can be scored as present or absent, which promoted more reliable scoring and simplified on-the-fly computation of inter-rater agreement.

When departments receive feedback from reviewers that describe problem areas and see a score that can be compared to a mean of their college or the university as a whole, we can refocus the conversation on improvement, even if initial changes are directed at improving the assessment report itself (Stanny, Stone, & Mitchell-Cook, 2018). Evaluations of the clarity of reporting can guide decisions about the design of report templates and guiding instructions prepared by an Assessment Office or Office of Institutional Effectiveness (IE). Together, the findings and follow-up interventions can both provide evidence that programs comply with accreditor expectations and shift the culture toward a focus on efforts to seek improvement.

Audit of the Assessment Reporting Process

As part of preparation for an impending compliance report to an institutional accreditor, the Office of IE conducted an audit of the assessment process and reviewed three years of programmatic assessment reports, the reporting template, and the submission process (Walvoord, 2014). The audit revealed strengths and weaknesses in institutional assessment processes. The good news was that the institution could document a systematic and ongoing culture of assessment. Nearly all departments had reported assessment activities annually for each of their educational programs, with few departments failing to participate in the process. The audit also revealed areas for improvement. Specifically, the report template included question prompts and instructions for several reporting fields that were vague, ambiguous, or did not elicit narratives that fully documented the assessment work completed by faculty. Because reports were submitted as responses to questions in a Qualtrics survey that had limited text fields, narratives frequently lacked the level of detail needed to understand the work reported. In addition, the submission process, which required completing a new form for each learning outcome, was awkward, repetitive, and cumbersome. As a result, most departments reported assessment work for only one or two student learning outcomes, although evidence from recent disciplinary accreditations and program reviews indicated that several programs engaged in more extensive assessment activity. In addition, few departments had created a multi-year assessment plan. There was scant evidence that any department had reassessed a student learning outcome to determine the impact of changes implemented in a prior year. Thus, the structure of the reporting process encouraged departments to treat each assessment cycle as a snapshot of work from the current reporting year, with no thought given to assessments that could evaluate the impact of changes made in a prior year (Suskie, 2018).

Evaluations of the clarity of reporting can guide decisions about the design of report templates and guiding instructions prepared by an Assessment Office or Office of Institutional Effectiveness (IE).

Information from the audit motivated us to modify assessment processes. First, IE staff designed a new report template based on an Excel spreadsheet with revised prompts that more clearly communicate expectations about the information requested. Second, each department was asked to develop a five-year assessment plan for each educational program that described how the department planned to conduct a full, multi-year cycle of assessment for each program-level student learning outcome within a five-year period. A full cycle of assessment was defined as a two to three year process. In the first year of an assessment

cycle, the program collects baseline assessment data. Then, the program should reflect on the findings and make decisions about possible implementation of an improvement initiative. In the final year of the cycle, the program conducts follow-up assessments to either evaluate the impact of the implemented change or document the stability of student performance on the targeted learning outcome.

Rubric elements describe "best practices" and hallmarks of a mature assessment process. These best practice elements contribute to assessment work that is likely to produce meaningful information and guide faculty decisions about curriculum and instruction.

In addition to changing the assessment reporting process, the Director of IE and the Director of the Center for Teaching and Learning (CTL) facilitated a series of workshops designed to educate faculty and administrators in assessment leadership positions on how to write assessment reports that would clearly document an actionable use of assessment results toward seeking improvement in student learning (Fulcher, et al., 2017; Walvoord, 2014). Workshops included a half-day mini-conference on assessment, an annual peer review of assessment events (described in Stanny, et al., 2018), workshops on effective assessment practices, targeted workshops on specific assessment skills (writing measurable learning outcomes, creating a five-year assessment plan), presentations to disseminate findings from the current formal review, and one-on-one consultations with chairs and members of assessment and curriculum committees.

The institution had adopted an annual formal review of assessment reports, in which trained reviewers used a rubric to evaluate the quality of assessment work described in assessment reports. Although the previous four formal reviews had documented improvements in assessment reporting (Stanny, 2020), the audit confirmed the need for extensive changes to the reporting process, which had emerged from a series of conversations during the peer review event and in one-on-one consultations. The formal review was extended to evaluate the impact of changes made to the new report template and other initiatives to promote a more mature assessment culture. The rubric was revised to reflect the new reporting fields and guiding language in the new Excel template. The review continued our evaluation of submitted assessment reports as a meta-assessment of the impact of these changes on the quality of assessment reporting. In addition, examination of the types of assessment practices documented in these reports enabled us to describe the ongoing evolution toward a more mature culture of assessment.

Method

Rubric

The rubric used for the review is comprised of three major sections: *Reporting Compliance Criteria*, *Maturity of Assessment*, and *Evidence of Impact*. A list of the rubric elements is presented in Table 1. Rubric elements were scored as a 0 (evidence is weak, missing, or the criterion is not applicable to the reporting program, as when no evidence is provided for an optional item) or 1 (evidence that a report meets expectations).

Scores for the *Reporting Compliance* and *Evidence of Impact* sections are based on the number of rubric elements that describe best practices for this section (two – six rubric elements). *Maturity of Assessment* produced scores on six dimensions of maturity, based on the number of rubric elements that described best practices for this dimension (two – five rubric elements). Summary findings report the scores for each dimension as diagnostic feedback and report an overall score for Maturity of Assessment (23 elements). Rubric elements describe "best practices" and hallmarks of a mature assessment process. These best practice elements contribute to assessment work that is likely to produce meaningful information and guide faculty decisions about curriculum and instruction. Composite scores, based on the rubric elements included in a section or dimension of a section, create global measures of the quality of reporting and maturity of the assessment culture.

Reporting Compliance Criteria. This score was based on the sum of 6 rubric elements that evaluate key elements that should appear in every assessment report to adequately document the program's compliance with expectations for reporting assessment processes with clear and compelling narratives. The elements evaluated the following characteristics: (1) report documents assessment on at least 20% of program student learning outcomes (SLOs), (2) completion of the summary tab portion of the Excel template for assessment reports, (3) clear description of program delivery, including locations and modalities of

Table 1
Rubric used for Scoring Annual Assessment Reports

Each rubric element scored a criterion as present/met (1) or absent/not met (0).
Department reports assessment for at least 20% of identified SLOs for the program
Summary narrative of assessment activity
Clear description of the delivery mode of the program
Evidence of faculty engagement and reflection on the assessment findings
Curriculum Map is available (posted on the IE website)
5-Year Assessment Plan is available (posted on the IE website)
Maturity of Assessment (6 dimensions)
Quality of Measures (4 criteria)
At least one measure aligns with the SLO(s) assessed
Assessments include at least one direct measure for each SLO
At least one SLO was assessed with multiple measures
Discussion of the reliability or validity of at least one measure used to assess an SLO
Credible Data Collection Processes and Representative Sampling (4 criteria)
Measures used for assessment have face validity for and align with the SLO assessed
Data analysis includes disaggregation by locations and delivery modes as appropriate
Report includes the number of course sections that provided data
Report includes the number of students assessed
Report of Results (5 criteria)
Report identifies a benchmark and description of criteria for meeting the benchmark
Report includes the number of students that meet or exceed expectations
Narrative compares current findings with evidence from previous assessments
Narrative summarizes results that appear in another document
Department provides the meeting date(s) where faculty discussed assessment findings
Department documents the attendance of faculty at the meeting
Department submitted the meeting minutes as supporting evidence
Narrative describes a logical relationship between decisions and assessment findings
Use of Results for Improvement (2 criteria)
Department describes an ***actionable*** use of results to improve student learning that is clearly related to the assessment evidence
Narrative provides convincing evidence of a concrete plan to implement
Faculty Engagement with Assessment Processes (4 criteria)
Evidence of broad faculty engagement
Narrative describes how assessment findings and decisions are communicated
Evidence that findings were disseminated to all appropriate faculty
Evidence that findings were disseminated to other relevant stakeholders
Evidence of Impact on Student Learning (2 criteria)
Narrative includes an evaluation of the impact of any changes implemented during a prior academic year on student learning
Evidence provided about the impact (either positive or negative) of a new initiative

instruction, (4) documentation of faculty engagement and reflection on assessment evidence for program improvement, (5) curriculum map posted to the IE website, and (6) five-year assessment plan posted to the IE website.

Maturity of Assessment. An overall score for maturity of assessment was based on the sum of 23 rubric elements, which described six dimensions or characteristics of a mature assessment process: (1) quality of measures (four rubric elements), (2) credible data collection processes and representative sampling (four rubric elements), (3) report of results (five rubric elements), (4) interpretation of findings (four rubric elements), (5) use of results for improvement (two rubric elements), and (6) faculty engagement with assessment processes (four rubric elements).

Evidence of Impact. This metric identifies programs that provide concrete examples of tangible changes in student learning that can be attributed to teaching and learning initiatives motivated by assessment findings. The metric was based on two rubric elements: (1) evidence that the program assessed and evaluated impact and (2) evidence presented for the impact of changes implemented was compelling.

Sample

The sample included assessment reports submitted to the Office of IE during two cycles of assessment reporting (ending in 2019 and 2020). The 2018-2019 assessment cycle included 75 reports for undergraduate programs and 35 reports for graduate programs. The 2019-2020 assessment cycle included 69 reports for undergraduate programs and 41 reports for graduate programs. Departments submitted assessment reports using an Excel spreadsheet template prepared by the Office of IE. Departments were encouraged to supplement information in their report narratives by uploading supporting documents (such as meeting minutes, examples of assignments or rubrics, and reports summarizing large data analyses). Reviewers examined the narratives and all supporting documents when they scored each report.

Procedure for training and maintaining inter-rater reliability

Serving as a reviewer of programmatic assessment reports is regarded as intensive professional development for faculty. Although faculty may serve as reviewers more than once, we encourage applications from new reviewers each year to increase assessment expertise across the university.

Reviewers. Each year, the CTL issues a call for faculty reviewers. Faculty are invited to submit letters of interest that include information regarding their full-time status, their department and college, and their availability to meet during the spring semester. The CTL and IE collaborate to review the applications. Four reviewers are selected based on their application responses and availability with the constraint that the four reviewers come from different colleges. This ensures that no reviewer scores assessment reports submitted by departments from the college in which they teach (except during initial training, when all reviewers score all reports in the training sample).

Serving as a reviewer of programmatic assessment reports is regarded as intensive professional development for faculty. Although faculty may serve as reviewers more than once, we encourage applications from new reviewers each year to increase assessment expertise across the university. For both years included in this study, four reviewers were selected for both 2019 and 2020, for a total of eight reviewers over the two-year period. Reviewers received formal training on how to apply the rubric to score the assessment reports. The reliability of scoring was evaluated and monitored continuously during the review.

Reviewer training and reliability. Reviewers completed an initial training and discussed how to score the assessment reports based on the rubric elements. Next, reviewers scored a training sample of assessment reports (six reports in 2019, seven reports in 2020). Reports were read and scored by all four reviewers. To compute inter-rater agreement, each reviewer was first paired with every other reviewer and we computed individual rater agreement scores (pair-wise) for each rubric element. We then computed the average agreement score across all possible pair-wise comparisons for each rubric element. Thus, agreement scores are the percentage of pair-wise comparisons that produced identical scores for a rubric element. We also computed the average percent agreement across all rubric elements.

After computing the initial reliability data, reviewers discussed areas of disagreement on individual rubric elements. Reviewers developed guidelines to help them apply the rubric consistently. Reviewers then independently rescored the reports in the training sample. The second calculation of reliability scores established acceptable levels of reliability (82% agreement, averaged over all rubric elements for the 2019 review and 81% agreement for the 2020 review).

Scoring procedures. After achieving an acceptable level of consensus (exceeding the target of 75% average agreement), reviewers scored the remaining reports. The review was completed as a series of assignments (three assignments for undergraduate reports, with 19-28 reports per assignment; three assignments for graduate reports, with 10-13 reports per assignment). Each reviewer was paired with every other reviewer for a subset of the reports included in an assignment. Two reviewers independently scored each report. Thus, percent agreement scores reflect the scoring consistency of each reviewer with every other reviewer and the average rater agreement score reflects the collective judgment of all four reviewers. No reviewer scored reports submitted by a department from his or her college.

The most problematic rubric elements entailed judgments about the maturity of assessment, especially practices that either did not apply to all programs … or did not have an obvious location or prompt in the reporting template…

Scoring consistency was maintained by computing the rater agreement metrics for scores submitted for each assignment (percent agreement for individual rubric elements, average agreement across all rubric elements). In addition, we computed cumulative percent agreement scores (individual rubric elements and average across rubric elements) for all reports scored to date. Reviewers discussed the reliability data and developed consensus about problem areas they encountered in the most recent assignment before they scored reports in the next assignment. Reviewers added notes to the scoring guidelines as needed to resolve emerging challenges and maintain consistency throughout the review. For the few instances when the scores submitted by two reviewers were not identical, differences were resolved by computing the average of the submitted scores.

Results and Discussion

Reliability

Reviewer agreement was monitored for each assignment and for the population of reports reviewed. We monitored agreement for individual rubric elements and for the agreement averaged across all rubric elements, with the goal of maintaining aggregate agreement above 75%. Final reliability metrics were based on the entire population of assessment reports in a given year, disaggregated by program (undergraduate or graduate).

The average percent agreement for the 2019 review was 90% for undergraduate reports (n = 75) and 87% for graduate reports (n = 35). Similarly, the average percent agreement for 2020 was 81% for undergraduate reports (n = 69) and 84% for graduate reports (n = 41). Agreement scores for individual rubric elements (31 elements) ranged from 58% to 100%. During the 2019 review, only 3 of the 31 rubric elements (10%) produced percent agreement scores that were less than 75% agreement (values were 63%, 69%, and 72%) when reviewing undergraduate reports. Among the graduate reports (when scoring pivoted to remote work), seven rubric elements (22.6%) fell below 75% agreement (four elements ranged between 70% and 74% agreement; the remaining three elements ranged between 66% and 69% agreement). The review of the 2020 reports, completed entirely through remote work, was a bit more variable: eight rubric elements (26%) for the undergraduate reports produced percent agreement scores that were less than 75% agreement (values ranged from 59% to 74% agreement) and eight rubric elements (26%) for the graduate reports fell below 75% agreement (values ranged from 58% to 74% agreement).

Examination of the rubric elements that posed the greatest challenges for reliable scoring reflected as much about the quality of the template and prompts as the judgment of reviewers. The most problematic rubric elements entailed judgments about the maturity of assessment, especially practices that either did not apply to all programs (e.g., *data analysis includes disaggregation by locations and delivery methods as appropriate*) or did not have an obvious location or prompt in the reporting template (e.g., *comparison of current findings with evidence from previous assessments*, *summaries of results in a supporting*

document, *discussions of how findings and decisions were communicated*, *evidence that findings were disseminated to all appropriate faculty*).

The most difficult rubric elements were two criteria that concerned the use of results for improvement (*description of actionable use of results* and *description of a concrete plan to implement*). Reviewers said they had difficulty seeing a distinction between these two aspects of use of results. Future reviews might merge these items because we found that when reviewers disagreed, they usually scored one element as present and the other as absent, but chose different elements to score as present (versus one reviewer scoring both elements as present while the other reviewer scored both elements as absent). The items became more reliable when rescored as a single item (scoring one if at least one of the original two elements had been scored one and zero only when both elements were scored as zero). In addition to the challenge of attempting to capture a nuanced characteristic of mature assessment, reliable scoring of these two elements was further hampered by ambiguities inherent in the way the template requested information about decisions and actions (either implemented or planned for the coming year). This illustrates the multi-layered value of a formal review. Difficulties establishing reliability for some rubric elements often surfaced problems with the reporting template and ambiguous communications from IE to faculty responsible for reporting assessment work.

An interesting observation during this review was related to the impact of COVID-19 and the pivot to remote work. In 2019, reviewers had completed their work on undergraduate program reports by the end of February. In March, we shifted to remote work and continued weekly meetings via web conference software. The following year, the entire review, including initial training and weekly meetings, was implemented via web conferences. The data on inter-rater agreement reflect the challenges associated with clear communication via web meetings to maintain calibration and consensus. These challenges were compounded by schedule conflicts that prevented all reviewers from meeting at the same time. Based on these observations, we conclude that although it is possible to maintain better than 75% agreement among reviewers under these conditions, reviewers will reach higher levels of consensus if they can meet in person at the same time. It is unclear whether meeting via web conferencing software or meeting as two groups contributed to the lower agreement values observed during remote work.

Analysis of rubric scores

Difficulties establishing reliability for some rubric elements often surfaced problems with the reporting template and ambiguous communications from IE to faculty responsible for reporting assessment work.

Reporting compliance. The sum of the first six rubric elements served as a global measure of compliance with reporting expectations. Values could range from 0 (no report filed, no documents posted to the IE website) to 6 (all reporting criteria met expectations). In 2019, the mean reporting compliance score was 1.94 for undergraduate reports (SD = 1.222) and 2.64 for graduate reports (SD = 1.579). In 2020, reporting compliance scores increased to 4.88 for undergraduate reports (SD = 1.192) and 4.65 for graduate reports (SD = 1.744). Analysis of the reporting compliance composite scores produced a significant main effect of year ($F(1, 216) = 158.090$, $MSe = 1.914$, $p < .001$, partial $\varepsilon^2 = .423$) as well as a significant interaction of year by type of program report (undergraduate, graduate) ($F(1, 216) = 5.681$, $MSe = 1.914$, $p < .02$, partial $\varepsilon^2 = .026$).

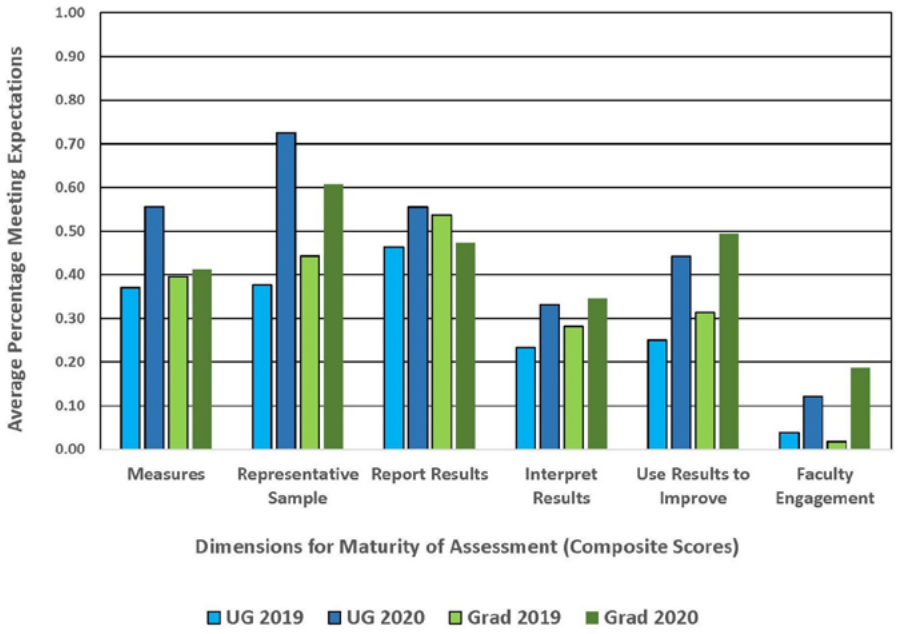
Maturity of assessment. Because each dimension of assessment maturity was based on two to five rubric elements, we computed an average of the contributing rubric elements instead of a sum to create composite scores with the same range of values (0 to 1, representing the average proportion of rubric elements in a dimension that met expectations). A 2 X 2 X 6 repeated measures analysis of variance was computed on composite scores in which report year (2019, 2020) and type of report (undergraduate, graduate) were between subjects factors and the six composite scores were repeated measures: *quality of measures* (four rubric elements), *credible data collection processes and representative sampling* (four rubric elements), *report of results* (five rubric elements), *interpretation of findings* (four rubric elements), *use of results* to improve student learning (two rubric elements), and faculty engagement with assessment processes (four rubric elements). A parallel statistical analysis, based on the raw scores produced by sums of rubric elements, produced the same pattern of findings. Only one analysis is reported here.

Average maturity of assessment improved from the first review (M = .31, SE = .020) to the second review (M = .437, SE = .019), producing a significant main effect of year of report ($F(1, 216) = 20.644$, MSe = .232, $p < .001$, partial ε^2 = .087). Although reports received significantly different scores for the six dimensions of assessment maturity ($F(5, 1080) = 153.899$, MSe = .035, $p < .001$, partial ε^2 = .416), this factor produced significant two-way interactions with the year of report and type of program as well as a significant three-way interaction between maturity scores, year of report, and type of report. As a result, this discussion focuses on the significant three-way interaction ($F(5, 1080) = 4.189$, MSe = .035, p = .001, partial ε^2 = .019). Other comparisons (the main effect of type of report and the interaction between year of review and type of report) were not statistically reliable.

Differences among dimensions of assessment maturity reflect strengths and weaknesses in the culture of assessment.

Mean composite scores (average proportion of rubric elements in a dimension that met expectations) are presented in Figure 1 as a function of the year of report (2019, 2020) and type of report (undergraduate and graduate program reports). Consistent with the significant main effect of year of report, with only a few exceptions, scores for both undergraduate and graduate reports improved from 2019 to 2020. The exceptions were that graduate programs showed no change on the *quality of measures* metric and showed a small decline on the *report of results metric*. Differences among dimensions of assessment maturity reflect strengths and weaknesses in the culture of assessment. Undergraduate programs showed pronounced improvements in the *quality of measures* gathered and the *collection of assessment evidence from a representative sample of student work*. The findings also suggest areas for further growth and maturation in the areas of *interpretation of findings* and *breadth of faculty engagement*.

Figure 1
Two-year comparison (2019 versus 2020) of the proportion of rubric elements within each of six dimensions of assessment maturity that met expectations for undergraduate and graduate programs.



Note: Number of rubric elements varied across dimensions: Measures (four elements), Representative Sample (four elements), Report of Results (five elements), Interpret Results (four elements), Use Results to Improve (two elements), Faculty Engagement (four elements).

Evidence of impact. Two rubric elements generate the composite score for evidence of impact. However, this metric produced no evidence for change across reports for either graduate or undergraduate programs, with few reports submitting documentation of the impact of an implemented change on assessments of student learning. Although few departments currently meet expectations on these rubric elements, they remain part of the

review to enable the institution to capture and document when departments reach this level of assessment maturity.

Conclusions

Our findings clearly indicate positive changes in the culture of assessment. Building on improvements documented in previous years (Stanny, 2020), data generated by the new rubric and reporting process document additional advances in both compliance with reporting expectations and adoption of assessment practices that characterize a more mature assessment culture. Strengths included widespread use of direct measures of student learning, improved alignment of assessment measures with targeted learning outcomes, collection of artifacts from a representative sample of students, more complete documentation of faculty discussions and reflections on assessment findings, and increases in the breadth of faculty engagement. Although the absolute value of scores for rubric elements related to mature assessment practices indicate substantial room for additional improvement, the changes from year one (baseline use of the new reporting template in 2019) to year two (2020) unambiguously document movement in the desired direction.

Institutional change often occurs at a glacial pace (Halonen, Ellenberg, Stanny, El-Sheikh, 2011). Assessment professionals charged with leading an initiative to promote a culture of assessment might feel they are making little progress from year to year. This project illustrates the value of meta-assessment to monitor progress on these large-scale efforts. Systematic monitoring of the maturity of assessment helped make incremental changes in the culture of assessment visible. The findings, along with informal observations from reviewers, suggested opportunities where small modifications could drive ongoing change. For example, during training, reviewers sometimes commented that they were unsure where in the assessment report they should look to find evidence for a given assessment practice. Reviewers also identified ambiguous language in report instructions. These observations identified shortcomings in template prompts and instructions that interfered with our ability to gather the information needed to document assessment activities. Revision of the reporting template was informed by the various observations gleaned from reviewer comments. Reviewer feedback also informed the design of professional development workshops to guide faculty charged with writing assessment reports and help them "tell their assessment story" to reviewers outside their discipline (Stitt-Berg, et al., 2018).

Assessment professionals charged with leading an initiative to promote a culture of assessment might feel they are making little progress from year to year. This project illustrates the value of meta-assessment to monitor progress on these large-scale efforts.

In summary, this formal review of assessment reports supported assessment efforts in several ways. The data provided tangible evidence of the quality of assessment work on campus, creating a year-to-year snapshot that proved useful as documentation of the institution's compliance with accreditation standards for assessment. The review provided formative feedback to the Office of IE and the CTL about the progress made toward achieving unit operational goals. The findings informed decisions about how to structure assessment reporting, such as the format of templates, how we framed requests for assessment information, and the logistics of reporting (timelines and interfaces with software and other reporting technology). These structural changes helped eliminate unintended obstacles to effective reporting. The data provided formative feedback to academic departments about their assessment practices and identified areas where small, realistic changes could produce tangible improvements in the quality of their assessment work. Dissemination of the findings helped allay a common misconception among faculty and critics of assessment: the belief that assessment reports are simply not read (Stanny, 2021).

An additional, serendipitous benefit emerged while the institution prepared a major accreditation report for its institutional accreditor. Scores on rubric elements for mature assessment practices served as an index to the population of assessment reports. When the authors of the accreditation report wanted to locate examples from assessment reports to include as evidence in the report narrative, they consulted the data file of rubric scores to identify programs that submitted relevant documentation with their assessment report. Rubric scores accurately identified relevant examples of departments that had disaggregated data, uploaded a rubric or description of an embedded assessment assignment, submitted minutes of a faculty meeting in which faculty reflected on assessment results and discussed curriculum changes or other initiatives intended to improve an aspect of student learning.

AUTHORS NOTE:
The authors thank the faculty reviewers for their contributions to scoring assessment reports for this project: Christopher L. Atkinson, Eric Bostwick, Jasara Norton, Pamela Meyers, Mizanoor Rahman, Bhuvaneswari Ramachandran, April Schwartz, and Jacqueline Thomas.

References

Association for the Assessment of Learning in Higher Education (AALHE). https://www.aalhe.org/

Association of American Colleges & Universities (AAC&U). https://www.aacu.org/

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

Blaich, C. F., & Wise, K. S. (2011, January). *From gathering to using assessment results: Lessons from the Wabash National Study* (Occasional Paper No. 8). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). https://www.bu.edu/provost/files/2015/09/From-Gathering-to-Using-Assessment-Results_Lessons-from-the-Wabash-Study-C.-Blaich-K.-Wise1.pdf

Blumberg, P. (2018, Summer/Fall). Two underused best practices for improvement focused assessments. *Research & Practice in Assessment*, *13*, 78-84. http://www.rpajournal.com/dev/wp-content/uploads/2019/01/RPA_Summer_Fall_Issue_2018_NIB.pdf

Fulcher, K. H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014, December). *A simple model for learning improvement:weigh pig, feed pig, Weigh pig*. (Occasional Paper No. 23). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://in.ewu.edu/facultycommons/wp-content/uploads/sites/129/2016/12/A-Simple-Model-for-Learning-Improvement_Weigh-Pig-Feed-Pig-Re-Weigh-Pig.pdf

Fulcher, K. H., Smith, K. L., Sanchez, E. R. H., Sanders, C. B. (2017). Needle in a haystack: Finding learning improvement in assessment reports. *Professional File*, *Article 141*. http://doi.org/10.34315/apf1412017

Fulcher, K. H., Swain, M. S., & Orem, C. D. (2012, January-February). Expectations for assessment reports: A descriptive analysis. *Assessment Update*, *24* (1), 1-2, 14-16. https://uncw.edu/assessment/documents/fultcherswainandorem2012.pdf

Gilbert, E. (2018, January 12). An insider's take on assessment: It may be worse than you thought. *The Chronicle of Higher Education*. https://www.chronicle.com/sectin/Commentary/44

Halonen, J. S., Ellenberg, G. B., Stanny, C. J., & El-Sheikh, E. (2011). First things first: Attending to assessment issues, accountability, and accreditation. In D. S. Dunn, M. A. McCarthy, S. C. Baker, & J. S. Halonen, *Using quality benchmarks for assessing and developing undergraduate programs* (pp. 46-70). Jossey-Bass.

Hutchings, P., Ewell, P., & Banta, T. (2012, May). *AAHE principles of good practice: Aging nicely*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/08/Viewpoint-Hutchings-EwellBanta.pdf

Isabella, M., & McGovern, H. (2018). Identity, values, and reflection: Shaping (and being shaped) through assessment. *New Directions for Teaching and Learning*, *2018* (155), 89-96. https://doi.org/10.1002/tl.20307

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. (2014). *Knowing what students know and can do: Current state of student learning outcomes assessment in U.S. colleges and universities*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/2013AbridgedSurveyReport.pdf

Lattuca, L. R., Terenzini, P. T., & Volkwein, J. F. (2006). *Engineering change: A study of the impact of EC2000 - Executive summary*. Accreditation Board for Engineering and Technology. https://www.abet.org/wp-content/uploads/2015/04/EngineeringChange-executive-summary.pdf

Lending, D., Fulcher, K., Ezell, J. D., May, J. L., & Dillon, T. W. (2018, Winter). Example of a program-level learning improvement report. *Research & Practice in Assessment*, *13*, 34-50. http://www.rpajournal.com/dev/wp-content/uploads/2019/02/W18_A6.pdf

Magruder, J., McManis, M. A., & Young, C. C. (1997). The right idea at the right time: Development of a transformational assessment culture. *New Directions for Higher Education*, *100*, 17-29. https://doi.org/10.1002/he.100002

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

National Institute for Learning Outcomes Assessment (NILOA). https://www.learningoutcomesassessment.org/

Nilson, L.B. (2015). *Specifications grading: Restoring rigor, motivating students, and saving faculty time*. Stylus.

O'Neill, M., Slater, A., & Sapp, D. G. (2018). Writing and the undergraduate curriculum: Using assessment evidence to create a model for institutional change. *New Directions for Teaching and Learning*, *2018*(155), 97-104. https://doi.org/10.1002/tl.20308

Reder, M., Crimmins, C. (2018, Winter). Why assessment and faculty development need each other: Notes on using evidence to improve student learning. *Research & Practice in Assessment*, *13*, 15-19. http://www.rpajournal.com/dev/wp-content/uploads/2019/02/W18_A3.pdf

Southern Association of Colleges and Schools Commission on Colleges (2020). *Resource manual for the principles of accreditation: Foundations for quality enhancement (Third Edition)*. SACSCOC. https://sacscoc.org/app/uploads/2019/08/2018-POA-Resource-Manual.pdf

Souza, J.M. & Rose, T.A. (Eds.) (2021). *Exemplars of assessment in higher education: Diverse approaches to addressing accreditation standards*. Stylus Publishing, LLC.

Stanny, C. J. (2015). Assessing learning in psychology: A primer for faculty and administrators. In D. S. Dunn (Ed.), *The Oxford handbook of undergraduate psychology education* (pp. 813-831). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199933815.013.065

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

Stanny, C. J. (2020, June). *Promote change by assessing the maturity of your assessment culture*. Single paper focus session presented at the Association for the Assessment of Learning in Higher Education (AALHE) conference. Conference proceedings: https://www.aalhe.org/2020-conference-proceedings

Stanny, C. J. (2021). Overcoming obstacles that stop student learning: The bottleneck model of structural reform. In S. A. Nolan, C. M. Hakala, & R. E. Landrum (Eds.), *Assessing undergraduate learning in psychology: Strategies for measuring and improving student performance*, (pp. 77-93). APA Books. https://doi.org/10.1037/0000183-007

Stanny, C. J., & Halonen, J. S. (2011). Accreditation, accountability, and assessment: Addressing multiple agendas. L. Stefani (Ed.), *Evaluating the effectiveness of academic development: A professional guide* (pp. 169-181). Routledge.

Stanny, C. J, Stone, E., & Mitchell-Cook, A. (2018). Evidence-based discussions of learning facilitated through a peer review of assessment. *New Directions for Teaching and Learning, 2018* (155), 31-38. https://doi.org/10.1002/tl.20300

Stitt-Bergh, M., Kinzie, J., Fulcher, K. (2018, Winter). Refining an approach to assessment for learning improvement. *Research & Practice in Assessment*, *13*, 27-33. http://www.rpajournal.com/dev/wp-content/uploads/2019/02/W18_A5.pdf

Suskie, L. (2015). *Five dimensions of quality: A common sense guide to accreditation and accountability*. Jossey-Bass.

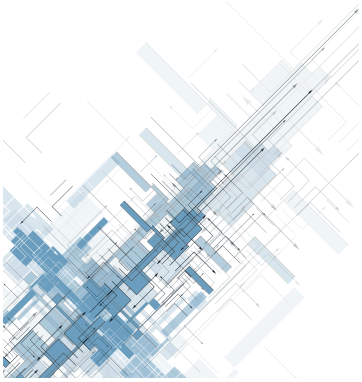
Suskie, L. (2018). *Assessing student learning: A common sense guide* (3rd ed.). Jossey-Bass.

Walvoord, B. E. (2014). *Assessment clear and simple: A practical guide for institutions, departments, and general education (2nd ed)*. Jossey-Bass.

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

Wehlburg, C. M. (2013). "Just right" outcomes assessment: A fable for higher education. *Assessment Update: Progress, Trends, and Practices in Higher Education*. *25* (2). https://doi.org/10.1002/au.252

Worthen, M. (2018, February 22). 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



Abstract

When engaging in outcomes assessment, higher education professionals (i.e., faculty, student affairs educators) are expected to gather reliable data and make valid inferences. Decisions about how to measure student learning and development outcomes impact inferences about the achievement of outcomes and determination of improvement efforts. Professionals may search for existing outcome measures due to lack of experience in the challenging instrument development process and/or the time required to construct a high-quality measure. To support professionals in their search, we created a tool that describes relevant repositories of measures. Given most professionals lack training in psychometrics, we purposefully categorized these repositories by the level of guidance they provide when selecting a measure. That is, in addition to identifying an existing measure and summarizing the measure's psychometric properties, some repositories provide an evaluation of the measure's quality. This resource facilitates the collection of high-quality data that informs valid inferences about student outcomes.

AUTHORS

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

Gabriel R. Gilmore, MA
James Madison University

Sarah Alahmadi, MS
James Madison University

CORRESPONDENCE

Email
finneysj@jmu.edu

"What's A Good Measure Of That Outcome?" Resources To Find Existing And Psychometrically Sound Measures

Student learning and development outcomes assessment is challenging and time consuming. The typical outcomes assessment process involves six general steps. The process begins by specifying measurable student learning and development outcomes—what students should know, value/appreciate, or be able to do (Step 1). These outcomes direct the activities completed during the remaining steps of the process. Faculty and student affairs educators map programming to the outcomes (Step 2). Evidence-informed programming (e.g., activities, pedagogies, strategies) that facilitates students achieving the desired outcomes should be intentionally selected (e.g., Finney & Buchanan, 2021; Finney et al., 2021; Horst, et al., 2021; Pope et al., in press; Pope et al., 2019; Smith & Finney, 2020). Once programming is mapped to outcomes, professionals must decide how to measure the outcomes (Step 3). A measure of an outcome (e.g., test, rubric, inventory, observational protocol) can be selected from existing measures or created. Inferences about student learning and development, and, in turn, program effectiveness are drawn from data gathered using these measures. Therefore, careful attention must be paid to how well measures align with intended outcomes, along with the measures' sensitivity to program impact (Bandalos, 2018; Suskie, 2009). The next steps (Steps 4 and 5) involve collecting implementation fidelity and outcomes data (e.g., Gerstner & Finney, 2013; Smith, et al., 2017, 2019). These data are then integrated, analyzed, interpreted, and reported (Step 6). Educators use the results to guide programming changes (Step 7), as the purpose of

the assessment process is to make data-based program modifications to improve student learning and development (Fulcher, et al., 2014).

Each step of the assessment process can be unpacked into more precise activities that involve particular skills (e.g., analyzing data, distinguishing between related but different outcomes, evaluating evidence of effectiveness). Our focus in the current paper is on determining how to measure outcomes (Step 3). Correct inferences about student ability, attitudes, skills, and behavior necessitate high-quality measures of those outcomes (Bandalos, 2018). Determining whether a high-quality measure exists or should be created is an essential activity at this step. Creation of a measure that allows for valid inferences requires a deep understanding of the outcome domain (e.g., critical thinking, intercultural competence, quantitative reasoning, career decisiveness, writing ability, ethical reasoning); skills to develop instructions, items, rubrics, or tasks that reflect the construct; an understanding of appropriate reliability and validity evidence, how to collect it, and how to interpret it; and pilot testing to improve the measure's psychometric properties. Although selection of an existing measure does not require skills to create a new measure or the study of its functioning, it does require an understanding of the outcome domain, a recognition of the need for relevant psychometric information, and skills to interpret those psychometric properties. In short, creation or selection of psychometrically-sound outcome measures both entail numerous competencies.

Although selection of an existing measure does not require skills to create a new measure or the study of its functioning, it does require an understanding of the outcome domain, a recognition of the need for relevant psychometric information, and skills to interpret those psychometric properties.

Faculty engaging in outcomes assessment are trained in a variety of disciplines (Leaderman & Polychronopoulos, 2019), and many are not formally trained in outcomes assessment via masters or doctoral programs (Hutchings, 2010; Nicholas & Slotnick, 2018). They instead gain knowledge, skills, and appreciation for assessment via workshops, conference presentations, webinars, and self-directed study (Curtis, et al., 2020). Unfortunately, there is rarely intentional, coherent sequencing of training opportunities, and many are targeted to novices. In turn, these trainings may not result in the depth of understanding and skill necessary for measurement-related concepts (e.g., reliability, validity, standard setting, factor analysis).

Unlike faculty landing in assessment positions from various domains across academic affairs (e.g., English, business), student affairs professionals are expected to understand and practice outcomes assessment (Finney & Horst, 2019a, 2019b). Yet, formal preparation programs may offer little training in measurement (e.g., Biddix et al., 2020; Cooper et al., 2016). According to Jablonski and colleagues (2006), "Even students from some of our best [student affairs] programs are inadequately trained in research, evaluation, and assessment." (p. 187).

Nonetheless, there are expectations regarding responsible practice in educational measurement. The preeminent source is *The Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), which applies to anyone creating measures, gathering data, and using scores. Moreover, standards or competencies related to measurement and assessment have been created by education organizations (see Table 1).

In student affairs, several organizations have created documents that detail expectations related to the selection or development of outcome measures: the *Assessment Skills and Knowledge (ASK) Standards* (2006) created by the American College Personnel Association (ACPA), the *Professional Competencies* (2015, 2016) created jointly by ACPA and the Student Affairs Administrators in Higher Education Association (NASPA), and the *CAS Standards* (2019) created by the Council for the Advancement of Standards in Higher Education. The difficulty in measuring outcomes of student affairs and co-curricular programs has been acknowledged (ACPA, 2006, p. 4): "In student affairs, the articulation and assessment of student learning has been especially challenging given the complex psychosocial and cognitive constructs that are the hallmarks of our work with students. Messy constructs such as leadership, citizenship, appreciation for diversity, critical and ethical judgement, and a host of interpersonal and intrapersonal intelligences present unique measurement issues." These "measurement issues" require measurement skills.

In the *Assessment Skills Framework*, Horst and Prendergast (2020) outlined the knowledge, skills, and attitudes important for assessment in higher education. They

categorized each domain by skill level: novice, intermediate, and advanced. Regarding the measurement of student learning and development outcomes, there were six domains: evaluate instruments for alignment, evaluate instruments for context and resource considerations, evaluate instruments for reliability and validity, design selected response measures, design non-cognitive measures, and design performance assessments. In Table 1, we listed novice-level skills (i.e., providing basic explanations of concepts). Professionals at the intermediate and advanced levels (not listed) can provide detailed explanations and apply knowledge to real assessment efforts. More general than the *Assessment Skills Framework*, *The Code of Professional Responsibilities in Educational Measurement* (1995) serves as a guide for anyone engaged in educational assessment, including faculty and staff assessing student learning and development.

In academic affairs, *The Standards for Teacher Competence in Educational Assessment of Students* were developed to guide teacher educators in teacher education programs, to offer a mechanism for self-assessment by teachers, and to serve as a framework for workshop content (Brookhart, 2011). If teachers and teacher educators demonstrated the listed competencies, they may be sought out for consultation by those engaged in higher education outcomes assessment (Kerr et al., 2020). Unfortunately, even the profession of teaching, which involves a tremendous amount of testing and interpretation of scores, does not consistently provide instruction in measurement during formal training (Lukin et al., 2004; Plake et al., 1993; Wise, 1993). If a formal course in measurement is available, the course may not provide instruction on all topics relevant to assessment-related work due to the numerous topics covered in such a course, the diverse needs of students, and the level of preparation of students (Bandalos & Kopp, 2012).

We agree that the competencies listed in Table 1 are necessary to engage in high-quality assessment practice, and, like others (Curtis, et al., 2020), we are concerned that educators practicing outcomes assessment have not engaged in formal training or self-directed study to meet these expectations. Because construction of a new measure is time intensive, requiring training in item writing and measurement prior to creating the measure, it is most efficient to identify existing measures. If no existing measures can be located or none are of sufficient quality, then the time-consuming process of creating a new measure should be pursued. Unfortunately, resources to guide locating and selecting high-quality existing measures are not well-advertised or organized. Thus, to facilitate the assessment of student learning and development outcomes using high-quality measures, we provide a didactic resource to foster the use of repositories of measures.

…to facilitate the assessment of student learning and development outcomes using high-quality measures, we provide a didactic resource to foster the use of repositories of measures.

Our resource differs from previous summaries of available surveys and measures used in post-secondary settings. For example, a 2001 American Council on Education and Association for Institutional Research report summarized the characteristics of 27 national assessments of institutional quality (Borden & Owens, 2001). These assessments include surveys students complete prior to enrollment (e.g., expectations about college), while enrolled in college (e.g., perceptions of college experiences, satisfaction), and after graduation (e.g., reflections on the impact of college). The report also included a few commercial measures of student learning outcomes (e.g., writing, critical thinking). Unlike the measurement repositories we describe below, this report does not discuss the quality of these measures. Although decades old, this report is useful in that it reflects the type of data collected to address accountability and improvement 20 years ago (prevalence of surveys collecting perceptions of college and institutions). Currently, high-quality accountability and improvement efforts emphasize student learning and development outcomes tied to intentional programming, which necessitates high-quality measures of these outcomes.

Description of the New Resource: Organization of Measurement Repositories

To facilitate faculty members', student affairs professionals', and assessment specialists' search for measures, we created a resource that identifies and organizes measurement repositories relevant to higher education outcomes. Repositories of existing measures differ in their utility; thus, we sorted them into three tiers according to the

Table 1
Professional standards and competences related to the development or selection of outcome measures

ACPA ASK Standards	ACPA NASPA Competencies	CAS Standards	Assessment Skills Framework (novice-level only)	Professional Responsibilities in Educational Measurement	Teacher Competencies in Assessment
Identify strengths and weaknesses of existing measures	Select measures that fit with assessment purposes	Ensure measures and methods are rigorous and reflect the characteristics of validity, reliability, and trustworthiness	Describe basic types of instruments and intended uses (e.g., indirect, direct, selected response, constructed response, cognitive, non-cognitive)	Conduct thorough evaluation of available measures that may be valid for intended uses	Skilled in selecting assessment methods appropriate for instructional decisions
Create measure with effective wording, format, and appropriate administration method	Utilize student learning and development research to inform content and design of assessment tools	Employ multiple measures and methods of data collection	Describe pros and cons of selecting an existing measure versus	Inform users of appropriateness of assessment for intended use, protection of examinee rights, costs, known consequences and limitations	Skilled in developing assessment methods appropriate for instructional decisions
Select most appropriate measure for desired outcome	Facilitate appropriate data collection for assessment purposes	Implement assessment process that is culturally responsive, inclusive, and equitable	Describe advantages and disadvantages of using different types of measures	Select measure based on evidence of technical quality not insubstantial claims	Skilled in administering, scoring, and interpreting results of both externally produced and self-produced assessments
Develop rubrics	Assess legitimacy and validity of various methods of data collection	Use methods and measures that allow for the collection of data that reflect intended outcomes	Match instrument to SLO	Comply with security precaution	Skilled in using assessment results when making decisions about students, instruction, developing curriculum, and improvement
Determine manner in which those with disabilities will use measure	Use culturally relevant and culturally appropriate terminology		Describe pros and cons of using commercial versus non-commercial measures	Plan accommodations for test-takers with disabilities when developing assessments	Skilled in communicating assessment results to students, parents, and other educators
Review a measure for inclusive and accessible language			Acknowledge importance of considering reliability and validity when selecting measure	Ensure assessments are developed to meet technical and legal standards	Skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information
Use measure with rigor appropriate for intended use			Describe common types of reliability and validity evidence	Caution users against most likely misinterpretations/ misuses of data	
			Identify components of multiple-choice item (e.g., stem, distractor)	Correct substantive inaccuracies in assessments as soon as feasible	
			Identify best practices for constructing selected response measures (e.g., use test blueprint, pilot items, revise)	Develop assessments free from bias due to characteristics irrelevant to construct being measured, such as gender, age, ethnicity, disability, SES	
			Identify characteristics of non-cognitive measures (e.g., variety of response options)	Develop score reports that promote understanding assessment results	
			Identify best practice for constructing noncognitive measures	Recommend against assessment likely to be administered, scored, and used in invalid manner for reasons of race, ethnicity, age, gender, disability, language background, SES, or religion	
			Identify basic rubric components (e.g., rating scale, scoring criteria)	Make information available about steps to develop and score assessment, including current information regarding reliability, validity, scoring and reporting	
			Distinguish holistic and analytic rubrics (advantages of each)		

information they provide. Some repositories simply identify measures aligned with a particular outcome and indicate where the measures can be found (what we refer to as Tier Three repositories). Other repositories include a summary of the psychometric information associated with the measure (what we refer to as Tier Two repositories). The most helpful repositories, in our opinion, are those that provide a review or rating of the measure's quality given the psychometric information (what we refer to as Tier One repositories).

The most helpful repositories, in our opinion, are those that provide a review or rating of the measure's quality given the psychometric information (what we refer to as Tier One repositories).

For each repository, we provide its name and web link, description of the resource, information provided about the measures' characteristics and quality, and five example measures. These five measures serve simply as exemplars and a mechanism to quickly access and examine the database.

Each repository is further labeled by the CAS Learning Outcomes Domains. CAS "promotes standards to enhance opportunities for student learning and development from higher education programs and services" (CAS, 2015, para. 1). CAS has developed six student outcome domains: knowledge acquisition, construction, integration, and application; cognitive complexity; intrapersonal development; interpersonal competence; humanitarianism and civic engagement; and practical competence. All six domains are listed for each repository, and the specific domains that the repository includes are bolded and *. For example, the database "emerge" has knowledge acquisition, construction, integration, and application; interpersonal competence; and practical competence bolded and *. Hence, in this repository, you will find measures that align with those specific student learning and development domains. For those who do not use the CAS outcome domains, but rather outcomes specified by the Liberal Education and America's Promise (LEAP) initiative, the Degree Qualifications Profile (DQP), Learning Reconsidered, or other organizations, CAS created a useful crosswalk of outcomes by organization to show their overlap.

To create our resource, we independently searched the internet for measurement repositories and concatenated the repositories we each found. We independently studied each repository to contribute to its description and example measures before identifying the appropriate tier and relevant CAS outcomes. We then excluded measurement repositories if they did not include measures relevant for the higher education context and population. Two students (one graduate and one undergraduate) examined the new resource and provided us with feedback (e.g., broken links, incomplete directions to access resource). We then piloted the resource during a week-long professional development session offered to United States and international faculty and student affairs professionals.

To create our resource, we independently searched the internet for measurement repositories and concatenated the repositories we each found.

How to Use the New Resource: Didactic Examples

To facilitate familiarity and use of this resource, we walk through two repositories in each tier and explain the type of information and psychometric evaluation they provide.

Tier One

Repositories in Tier One provide psychometric information (e.g., reliability of scores, validity evidence) as well as their own rating of the quality of the measure. This rating can be in the form of a number, statement, or recommendation for use. Ratings may not be provided for every measure but are available for the majority of measures in the repository. We consider repositories in this tier of the highest utility to select evidence-informed existing measures.

Mental Measurements Yearbook (MMY) Series

Tier One houses the Mental Measurements Yearbook (MMY) series, which is published by the Buros Center for Testing (Carlson et al., 2017). The MMY addresses the need for informed test evaluation by offering expert reviews of existing measures. Typically, detailed descriptions of the measures referenced in the MMY are provided, along with two reviews conducted by volunteer professional measurement experts. Volunteer reviewers are selected for each measure based on their domain-specific knowledge and training in measurement and psychometric evaluation. They also must carry a terminal degree (e.g.,

PhD, PsyD, EdD). To qualify for a review in MMY, a measure must be commercial, available in the English language, new or widely used, and provide psychometric qualities (e.g., reliability estimates, validity evidence). Reviews published in the MMY can be accessed through electronic databases, such as EBSCO or Ovid, to which many academic libraries subscribe. Additionally, the Buros Center for Testing offers a *Test Reviews Online* service, through which reviews for a particular test can be purchased.

To demonstrate the utility of MMY for selecting an existing measure, we searched for measures of critical thinking, a common student learning outcome in higher education. *The Cornell Critical Thinking Test* (CCTT; Ennis et al., 1985) was one of the tests identified by our search. The MMY test entry for the CCTT first provides descriptive information about the measure, including authors of the test, publication date, publisher information, purpose, population, scores, administration mode, testing time, price, name of MMY reviewers, yearbook volume in which the test appears, and relevant references (Carlson et al., 2017). Next, the two professional reviews of the measure are provided. The reviews typically summarize the developmental history of the measure, the norming samples, and evidence of technical quality provided by the test developers in the test manual. Following that, the reviewers provide their own commentary and recommendation for use of the measure. Test entries are often concluded with references to articles, manuals, or books that informed the experts' reviews.

For example, the first reviewer of the CCTT noted that "the Cornell Critical Thinking Test provides an objective method for evaluating critical thinking abilities that have been identified as necessary for individuals to respond appropriately to problems encountered in our complex world" (Porter, 2017). The second reviewer stated that the data presented supports the use of the test, but also noted the need for further empirical evidence to support the inferences made based on the test's scores. Specifically, the CCTT may not be appropriate for individual decision-making (Schafer, 2017). Overall, the reviewers support the use of the measure for the purposes of outcomes assessment or program evaluation, but advise against its use for making critical, person-level decisions. Such comprehensive and insightful appraisal of the measure and its appropriate use affords valuable information for informed measure selection.

Evidence-Based Measures of Empowerment for Research on Gender Equality (EMERGE)

EMERGE (2017a) is another Tier One repository that offers expert evaluation of carefully curated measures that assess knowledge, attitudes, and behaviors relating to gender equality and empowerment. Measures housed in the repository were selected with the help of gender equality and empowerment experts and reviews of available literature. For a measure to be included, it must have the following characteristics: quantitative in nature; published in either a national or international survey, or a peer-reviewed journal with impact factor ≥ 1; and include empirical evidence for reliability and validity.

Such comprehensive and insightful appraisal of the measure and its appropriate use affords valuable information for informed measure selection.

To provide ratings of the psychometric quality and utility of the measures, trained EMERGE staff score each measure (EMERGE, 2017b). The psychometric properties rated include the following aspects: formative research (qualitative research, theoretical framework, expert input, and pilot testing), reliability (internal consistency, test-retest, and inter-rater reliability), and validity evidence (content, face, criterion, and construct forms of validity). The scores for the three psychometrics aspects are aggregated into a total score: "Low" (≤ 33.3%) "Medium" (33.4% - 66.6%), "High" (≥ 66.7%), or "No Data" if the measure could not be scored. Another score utilizes information provided by Google Scholar on the number of citations of the measure's primary source: "Low" (< 20 citations), "Medium" (20 - 49 citations), "High" (≥ 50 citations), or "No Data" if the Google Scholar citation record is not available.

An example of a measure found in this repository is the *Illinois Rape Myth Acceptance Scale* (Payne et al., 1999). This measure may be useful for university bystander intervention programs designed to influence outcomes related to intervening in a potential assault. EMERGE provides a brief description of the measure, its purpose, intended population,

intended age range, list of items, response scale (e.g., Likert, multiple choice), and the measure's primary citation. EMERGE staff's ratings of the measure's psychometric properties and citation frequency is highlighted and explained (i.e., whether each of the scoring aspects received full or partial points, were not assessed, or were not applicable). This measure received a "high" psychometric score (EMERGE, 2017c), which would support its use as an outcome measure. Another valuable resource found on the EMERGE site is a report explaining how to utilize measurement in the field of gender equality and empowerment, how to identify psychometrically sound measures, and how to adapt a measure to different cultural contexts (Bhan et al., 2017).

Tier Two

Tier Two repositories provide psychometric information (e.g., reliability, validity) for the measures, but do not provide their own rating of the quality of the measures. Psychometrics may not be provided for every measure but are available for most measures in the databases. The majority of the repositories in our resource fall in this category. Below we provide two examples.

RAND Educational Assessment Finder

The RAND Educational Assessment Finder (RAND Corporation, n.d.) requires that included measures reflect interpersonal (e.g., empathy, leadership), intrapersonal (e.g., adaptability, perseverance), or higher-order thinking constructs (e.g., critical thinking, creativity), are appropriate for use in educational settings, and are appropriate for populations of students in the United States. To summarize the psychometric quality of a measure, RAND professionals read the publicly available studies that examined the reliability and/or validity of the measure. The psychometric summaries RAND creates are then shared with the measures' developers to provide any corrections before the summaries are published in the repository. The RAND Education Assessment Finder includes both commercial and non-commercial (i.e., free) measures and identifying free measures is facilitated by the "Fee for Use" search filter (Schweig, et al., 2018).

Tier Two repositories provide psychometric information (e.g., reliability, validity) for the measures, but do not provide their own rating of the quality of the measures.

The Cornell Critical Thinking Test (CCTT; Ennis et al., 1985) can be found in the RAND Education Assessment Finder, but the information this repository provides lacks the expert reviews provided by the Tier One MMY repository. The RAND Education Assessment Finder summarizes the following aspects of the measure: purpose, publication year, administration method, number of items, item format, administration time, available languages, fee, scoring, interpretive information, reliability evidence, validity evidence, links to obtain a copy of the measure, and references (RAND Corporation, 2018). Given the lack of review or rating conducted by measurement experts, the user manual of the RAND repository states that "because interpreting validity evidence is complex and generally requires measurement expertise, users are encouraged to seek input from measurement experts to evaluate the adequacy and relevance of the available evidence for a particular assessment purpose" (Hamilton et al., 2018, p. 11).

ETS Research Report Series

Educational Testing Services (ETS) publishes the ETS Research Report Series journal. This journal, which is freely accessible via the Wiley Online Library, includes resources related to psychometric and statistical methods, educational evaluation, and large-scale assessment. Highly relevant to higher education outcomes assessment are the syntheses of current literature on measures of pertinent student learning outcomes (e.g., quantitative literacy, intercultural competence, written communication, critical thinking). The syntheses contain information on the development of the outcome measures, the available reliability and validity information, target populations, and typically conclude with future directions for assessment in that domain. These reports are classified in Tier Two because they detail the psychometric properties of the measures, but do not include conclusions or interpretations regarding the psychometric quality of the measures.

An example report from the ETS Research Report Series provides the current state of assessment of civic competency and engagement in higher education (Torney-Purta et al., 2015). The report includes current definitions and conceptualizations of the construct in addition to (over 25) available measures. Measures are contrasted in terms of themes, test developer, test format, and length. The report discusses implications related to the reliability of scores and the validity inferences for such a multifaceted outcome. The report ends with a proposed framework for future assessments of civic competency and engagement to facilitate better measurement.

… identifying high-quality existing measures promotes common measurement of outcomes and comparison of results across different programming, teaching approaches, and institutions.

Tier Three

Unlike Tier One and Two, repositories in Tier Three do not provide psychometric information (e.g., reliability, validity) for the measures or their own rating of the quality of the measures. Often, the psychometric information can be found in the linked source articles.

PsycTests

PsycTests is a repository produced by the American Psychological Association (2021). It holds more than 60,000 measures, many of which are free to use. The measures are collected from various sources: directly from authors, peer-reviewed journals, books, dissertations, and websites.

Returning to our example of finding an existing measure of critical thinking, we searched PsycTests. *The Halpern Critical Thinking Assessment* (HCTA; Halpern, 2010) emerged as an option. The repository provided a "Master Test Profile" for the HCTA that included a description of the test, its purpose, the developer's contact information, and whether it is commercial, among other basic pieces of information. Typically, no information regarding reliability or validity is provided. If psychometric information is available in the original source of the test, the PsycTest entry will include the information, but no professional review of such information is provided. Thus, the amount of information provided by this Tier Three repository is limited compared to that provided by repositories from Tier One and Tier Two.

Assessment and Curriculum Support Center

Another Tier Three repository is the Assessment and Curriculum Support Center at the University of Hawai'i at Mānoa (Assessment and Curriculum Support Center, 2020). This center specializes in assessment of learning outcomes for improvement, and it includes a collection of rubrics used to assess outcomes such as civic knowledge, collaboration, critical thinking, ethical deliberation, integrative learning, information literacy, intercultural knowledge, and others. The repository contains links to the original sources of the rubrics. As such, it is a collection of performance assessments but does not review their psychometric quality. The user is encouraged to collaborate with an assessment expert to evaluate these measures.

Discussion

Our goal was to create a resource of measurement repositories that supports educators' search for high-quality measures. These repositories can increase efficiency in the outcomes assessment process and the trustworthiness of resulting scores. However, we want to stress that high-quality scores and valid inferences require more than quality measures. Students may have negative attitudes (Zilberberg, et al., 2012; Zilberberg, et al., 2013; Zilberberg, et al., 2014) or emotions (Finney, Perkins & Satkus, 2020; Finney, Satkus, & Perkins, 2020) toward higher education outcomes assessment initiatives. Thus, students may not be motivated to provide valid responses (Barry et al., 2010; Wise & DeMars, 2005) or attend testing sessions (Brown & Finney, 2011; Kopp & Finney, 2013; Swerdzewski et al., 2009). In turn, professionals should engage in strategies to increase examinee motivation (Barry & Finney, 2009; Finney, et al., 2016; Myers & Finney, 2021) or analyses that address the lack of motivation (Swerdzewski et al., 2011; Wise & DeMars, 2010).

Beyond identifying and evaluating existing measures, repositories of measures have additional benefits for faculty, student affairs educators, and assessment specialists. Measurement repositories showcase the various definitions and operationalization of what some professionals assume to be simple outcomes. They force educators to clearly articulate the outcome of interest given the number of different but related outcome measures that exist. They counter vague language describing outcomes, which facilitates alignment between outcomes and effective programming. Moreover, identifying high-quality existing measures promotes common measurement of outcomes and comparison of results across different programming, teaching approaches, and institutions.

> Organizations or individuals responsible for a group of programs could consider using [measurement repositories] to identify and endorse a specific set of outcome measures that are both reliable and valid for the populations served across a variety of domains. Endorsing a specific set of outcome measures could allow for consistency in tracking core outcomes or indicators of effectiveness across an array of programs (Acosta et al., 2014, p. 3).

Measurement repositories also showcase the rigorous process of scale development. By reviewing psychometric evidence, they uncover the need for additional psychometric study before trustworthy inferences can be made about student learning and development on our campuses.

References

Acosta, J., Reynolds, K., Gillen, E., Feeney, K., Farmer, C., & Weinick, R. (2014). The RAND Online Measure Repository for evaluating psychological health and traumatic brain injury programs. *Rand Health Quarterly*, *4*(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5051979/

Assessment and Curriculum Support Center (2020, November). *Rubric Bank*. http://manoa.hawaii.edu/assessment/resources/rubric-bank/

American College Personnel Association. (2006). *ASK standards: Assessment skills and knowledge content standards for student affairs practitioners and scholars*. Washington, DC: Author.

American College Personnel Association & National Association of Student Personnel Administrators. (2015). *ACPA/NASPA professional competency areas for student affairs educators*. Washington, DC: Authors. https://www.naspa.org/images/uploads/main/ACPA_NASPA_Professional_Competencies_FINAL.pdf

American College Personnel Association & National Association of Student Personnel Administrators. (2016). *ACPA/NASPA professional competencies rubrics*. Washington, DC: Authors. https://www.naspa.org/images/uploads/main/ACPA_NASPA_Professional_Competency_Rubrics_Full.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

American Psychological Association (2021). *PsycTests*. https://www.apa.org/pubs/databases/psyctests

Bandalos, D.L. (2018). *Measurement theory and applications for social sciences*. Guilford Press.

Bandalos, D.L., & Kopp, J.P. (2012). Teaching introductory measurement: Suggestions for what to include and how to motivate students. *Educational Measurement: Issues and Practice*, *31*(2), 8-13. https://doi.org/10.1111/j.1745-3992.2012.00229.x

Barry, C.L. & Finney, S.J. (2009). Does it matter how data are collected? A comparison of testing conditions and the implications for validity. *Research & Practice in Assessment*, *3*, 1-15. http://www.rpajournal.com/dev/wp-content/uploads/2012/05/A33.pdf

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

Bhan, N., Jose, R., McDougal, L., & Raj, A. (2017). *EMERGE measurement guidelines report 1: What is measurement and how do we quantitatively measure gender equality and empowerment?* Center on Gender Equity and Health (GEH), University of California, San Diego School of Medicine. San Diego, CA. https://bit.ly/3dzA4Z5

Brown, A.R. & Finney, S.J. (2011). Low-stakes testing and psychological reactance: Using the Hong Psychological Reactance Scale to better understand compliant and non-compliant examinees. *International Journal of Testing*, *11*, 248-270. https://doi.org/10.1080/15305058.2011.570884

Biddix, J.P., Collom, G.D., & Roberts, D.M. (2020). Scholarship, professional development, and community of practice in student affairs assessment. *College Student Affairs Journal*, *38*, 157-171 (EJ1275325). ERIC. https://files.eric.ed.gov/fulltext/EJ1275325.pdf

Borden, V.M. & Owens, J.L. (2001) *Measuring quality: Choosing among surveys and other assessments of college quality*, Washington, DC: American Council of Education and Association of Institutional Research (ED457767). ERIC. https://files.eric.ed.gov/fulltext/ED457767.pdf

Brookhart, S.M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30*, 3-12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Carlson J.F., Geisinger, K.F., & Jonson, J.L. (Eds.). (2017). *The twentieth mental measurements yearbook*. Buros Center for Testing.

Cooper, J., Mitchell, D., Jr., Eckerle, K., & Martin, K. (2016). Addressing perceived skill deficiencies in student affairs graduate preparation programs. *Journal of Student Affairs Research and Practice*, *53*, 107-117. https://doi.org/10.1080/19496591.2016.1121146

Council for the Advancement of Standards in Higher Education. (2019). *CAS professional standards for higher education* (10th ed.). Washington, DC: Author.

Curtis, N.A, Anderson, R.D & Van Dyke, R. (2020). A field without a discipline? Mapping the uncertain and often chaotic route to becoming an assessment practitioner. *Research & Practice in Assessment*, *15*(1), 1-8. https://www.rpajournal.com/dev/wp-content/uploads/2020/08/A-Field-Without-A-Discipline.pdf

Ennis, R.H., Millman, J., & Tomko, T.N. (1985). *Cornell critical thinking tests level X & level Z: Manual*. Pacific Grove, CA: Midwest Publications.

Evidence-based Measures of Empowerment for Research on Gender Equality (EMERGE) (2017a). *Overview*. https://emerge.ucsd.edu/measurement-overview/

Evidence-based Measures of Empowerment for Research on Gender Equality (EMERGE) (2017b). *Scoring methodology*. https://emerge.ucsd.edu/scoring-methodology/

Evidence-based Measures of Empowerment for Research on Gender Equality (EMERGE) (2017c). *Illinois Rape Myth Acceptance (IRMA) Scale*. https://emerge.ucsd.edu/r_ag79krbvxz9zcon/

Finney, S.J. & Buchanan, H.A. (2021). A more efficient path to learning improvement: Using repositories of effectiveness studies to guide evidence-informed programming. *Research & Practice in Assessment*, *16*(1), 36-48. https://www.rpajournal.com/dev/wp-content/uploads/2021/04/Repositories-of-Effectiveness-Studies-to-Guide-Programming.pdf

Finney, S.J., & Horst, S.J. (2019a). Standards, standards, standards: Mapping professional standards for outcomes assessment to assessment practice. *Journal of Student Affairs Research and Practice*, *56*, 310-325. https://doi.org/10.1080/19496591.2018.1559171

Finney, S.J., & Horst, S.J. (2019b). The status of assessment, evaluation, and research in student affairs. In V. L. Wise & Z. Davenport (Eds.), *Student affairs assessment, evaluation, and research: A guidebook for graduate students and new professionals* (pp. 3-19). Springfield, IL: Charles Thomas Publisher.

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

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

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

Finney, S.J., Wells, J.B., & Henning, G.W. (2021). *The need for program theory and implementation fidelity in assessment practice and standards*. (Occasional Paper No. 52). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://www.learningoutcomesassessment.org/wp-content/uploads/2021/03/Occ_Paper_51-1.pdf

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

Gerstner, J.J. & Finney, S.J. (2013). Measuring the implementation fidelity of student affairs programs: A critical component of the outcomes assessment cycle. *Research & Practice in Assessment*, *8*, 15-28. https://www.rpajournal.com/dev/wp-content/uploads/2013/11/SF2.pdf

Halpern, D. (n.d.). *HCTA Halpern Critical Thinking Assessment* [Database record]. APA PsycTests. Retrieved February 25, 2021, from https://doi.org/10.1037/t10940-000

Hamilton, L.S., Stecher, B.M., Schweig, J. & Baker, G. (2018). *RAND Education Assessment Finder: User instructions*. RAND Corporation. https://www.rand.org/education-and-labor/projects/assessments.html

Halpern, D.F. (2010). *Halpern Critical Thinking Assessment*. SCHUHFRIED (Vienna Test System): Möedling, Austria.

Horst, S.J., Finney, S.J., Prendergast, C.O., Pope, A.M. & Crewe, M. (2021). The credibility of inferences from program effectiveness studies published in student affairs journals: Potential impact on programming and assessment. *Research & Practice in Assessment*, *16*(2), 17 – 32. https://www.rpajournal.com/dev/wp-content/uploads/2021/09/The-Credibility-of-Inferences-from-Program-Effectiveness-Studies.pdf

Horst, S.J. & Pendergast, C.O. (2020). The Assessment Skills Framework: A taxonomy of assessment knowledge, skills and attitudes. *Research and Practice in Assessment*, *15*(1), 1-25. https://www.rpajournal.com/dev/wp-content/uploads/2020/05/The-Assessment-Skills-Framework-RPA.pdf

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

Jablonski, M., Mena, S.B., Manning, K., Carpenter, S. & Siko, K.L. (2006) Scholarship in student affairs revisited: The summit on scholarship. *NASPA Journal*, 43, 182-200. https://doi.org/10.2202/1949-6605.1729

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

Kopp, J.P. & Finney, S.J. (2013). Linking academic entitlement and student incivility using latent means modeling. *Journal of Experimental Education*, *81*, 322-336. https://doi.org/10.1080/00220973.2012.727887

Leaderman, E.C., & Polychronopoulos, G.B. (2019). Humanizing the assessment process: How the RARE model informs best practices. Research & Practice in Assessment, 14, 30-40. https://www.rpajournal.com/dev/wp-content/uploads/2020/03/Humanizing-the-Assessment-Process_r.pdf

Lukin, L.E., Bandalos, D.L., Eckhout, T.J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, *23*, 26-32. https://doi.org/10.1111/j.1745-3992.2004.tb00156.x

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

Nicholas, M.C., & Slotnick, R.C. (2018). *A portrait of the assessment professional in the United States: Results from a national survey*. (Occasional Paper No. 34). University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. https://bit.ly/3dGpQX4

Payne, D.L., Lonsway, K.A., & Fitzgerald, L.F. (1999). Rape myth acceptance: Exploration of its structure and its measurement using the Illinois Rape Myth Acceptance Scale. *Journal of Research in Personality*, *33*(1), 27-68. https://doi.org/10.1006/jrpe.1998.2238.

Plake, B.S., Impara, J.C., & Fager, J.J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice*, *12*, 10-12. https://doi.org/10.1111/j.1745-3992.1993.tb00548.x

Pope, A.M., Finney, S.J., & Bare, A. (2019). The essential role of program theory: Fostering theory-driven practice and high-quality outcomes assessment in student affairs. *Research & Practice in Assessment*, *14*, 5-17. https://www.rpajournal.com/dev/wp-content/uploads/2019/07/L1.pdf

Pope, A.M., Finney, S.J. & Crewe, M. (in press). Evaluating the effectiveness of an academic success program: Showcasing the importance of theory to practice. *Journal of Student Affairs Inquiry*.

Porter, J.Y. (2017). [Test review of Cornell Critical Thinking Tests, Fifth Edition]. In Carlson J. F., Geisinger, K. F. & Jonson, J. L. (Eds.). *The twentieth mental measurements yearbook*. Buros Center for Testing.

RAND Corporation (n.d.). *RAND Education Assessment Finder*. https://www.rand.org/education-and-labor/projects/assessments.html

RAND Corporation (2018). *Cornell Critical Thinking Test (Level X) (CCTT)*. https://www.rand.org/education-and-labor/projects/assessments/tool/1971/cornell-critical-thinking-test-level-x-cctt.html

Schafer, W.D. (2017). [Test review of Cornell Critical Thinking Tests, Fifth Edition]. In Carlson J. F., Geisinger, K. F. & Jonson, J. L. (Eds.). *The twentieth mental measurements yearbook*. Buros Center for Testing.

Schweig, J., Baker, G., Hamilton, L. S., & Stecher, B.M. (2018). *Building a repository of assessments of interpersonal, intrapersonal, and higher-order cognitive competencies*. RAND Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RR2500/RR2508/RAND_RR2508.pdf

Smith, K.L., & Finney, S.J. (2020). Elevating program theory and implementation fidelity in higher education: Modeling the process via an ethical reasoning curriculum. *Research and Practice in Assessment*, *15*(2), 1-13. https://www.rpajournal.com/dev/wp-content/uploads/2020/09/Elevating-Program-Theory-and-Implementation-Fidelity-in-Higher-Education.pdf

Smith, K.L., Finney, S.J., & Fulcher, K.H. (2017). Actionable steps for engaging assessment practitioners and faculty in implementation fidelity research. *Research & Practice in Assessment*, *12*, 71-86. http://www.rpajournal.com/dev/wp-content/uploads/2018/02/NIB1.pdf

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

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

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

Swerdzewski, P.J., Harmes, J.C., & Finney, S.J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*, 162 – 188. https://doi.org/10.1080/08957347.2011.555217

Torney-Purta, J., Cabrera, J.C., Crotts Roohr, K., Liu, O.L., & Rios, J.A. (2015). *Assessing civic competency and engagement in higher education: Research background, frameworks, and directions for next-generation assessment* (Research Report No. RR-15-34). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12081

Wise, S.L. (Ed.). (1993). *Teacher training in measurement and assessment skills*. Lincoln, NE: Buros Institute of Mental Measurements.

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

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

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

Zilberberg, A., Anderson, R.A., Swerdzewski, P.J., Finney, S.J., & Marsh, K.R. (2012). Growing up with No Child Left Behind: An initial assessment of the understanding of college students' knowledge of accountability testing. *Research & Practice in Assessment*, *7*, 12-25. https://files.eric.ed.gov/fulltext/EJ1062686.pdf

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