# RESEARCH & PRACTICE IN ASSESSMENT

VOLUME SIXTEEN | ISSUE 1 | 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 © 2021

Editorial Staff

Editor-in-Chief
Nicholas A. Curtis
*Marquette University*

Senior Associate Editor
Robin D. Anderson
*James Madison University*

Associate Editor
Lauren Germain
*SUNY Upstate Medical University*

Associate Editor
Megan Good
*Auburn University*

Associate Editor
Sarah Gordon
*Arkansas Tech University*

Associate Editor
Julie A. Penley
*El Paso Community College*

Associate Editor
Gina B. Polychronopoulos
*George Mason University*

Associate Editor
Megan Shaffer
*Independent Assessment Consultant*

Editorial Board

Laura Ariovich
*Prince George's Community College*

Gianina Baker
*National Institute for Learning Outcomes Assessment*

Kellie M. Dixon "Dr. K"
*North Carolina Agricultural and Technical State University*

Natasha Jankowski
*Higher Education & Assessment Consultant*

Monica Stitt-Bergh
*University of Hawai'i at Manoa*

Ray Van Dyke
*Weave*

Ex–Officio Members

Denise Ridley-Johnston
Virginia Assessment Group President
*College of William & Mary*

Linda Townsend
Virginia Assessment Group President–Elect
*Longwood University*

# TABLE OF CONTENTS

## FROM THE EDITOR



## ASSESSMENT DIALOGUE



## ARTICLES



# Assessment Lessons and Predictions

Prediction is very difficult, especially if it's about the future.

-Niels Bohr

"Yet, despite Dr. Bohr's warning, we endeavor to do so anyway! We hope that this issue of *Research & Practice in Assessment* finds you well and looking forward to the future of your assessment work.

Volume 16, Issue 1 of RPA includes our first 'assessment dialogue' along with five peer-reviewed articles that are sure to pique your interest. Our first 'assessment dialogue' showcases a discussion between two of our colleagues, Fulcher and Eubanks, who provide deep thoughts about the future of assessment practice. Tucker, Moreno, and Jafari make the case for the importance of core competencies as a unifying institutional assessment tool. Walker and Roconni aim to demonstrate construct validity and measurement invariance of pre-experience and postexperience perception surveys. Finney and Buchanan describe systematic review repositories that synthesize high-quality research and include a tool created to organizes relevant repositories. Hobbs, Singer-Freeman, Robinson examine equity gaps including whether certain assignment types were associated with inequitable grade distributions for underrepresented minority (URM) and transfer students. Finally, Hsu, Li, and Acosta share their work on the psychometric properties of A Faculty Encouragement Scale (FES) created to measure students' perception of faculty encouragement.

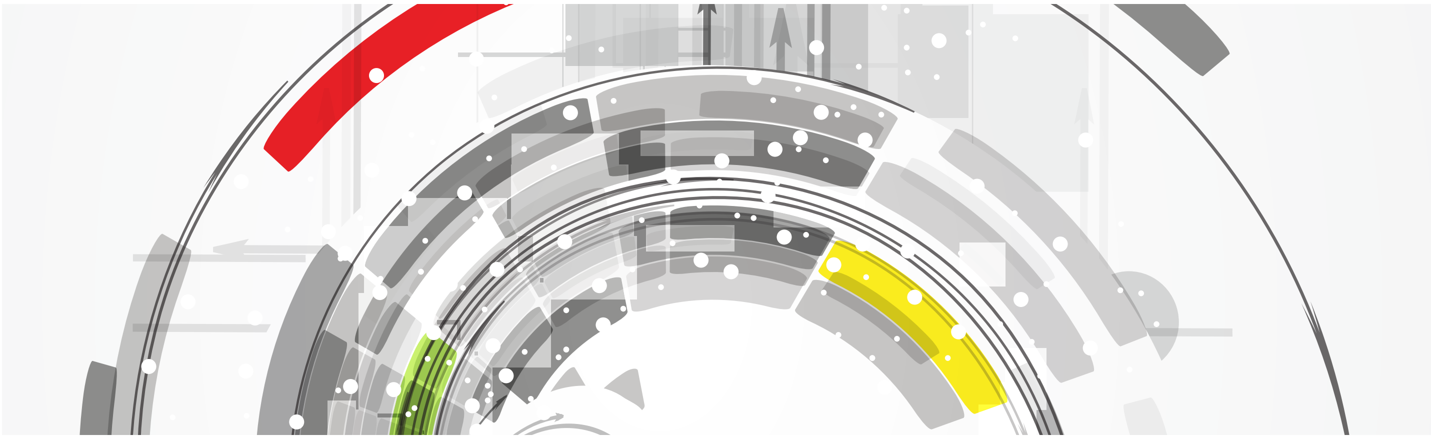
With such a diverse range of topics, I hope this issue of Research & Practice in Assessment holds something to interest everyone!

Best Wishes,

Nicholas Curtis

Editor-in-Chief,
*Research & Practice in Assessment*

# The Next Ten Years: The Future of Assessment Practice?



## Forward

In these times of extraordinary change and hardship, there is perhaps no better time to consider how we as assessment professionals might reimagine our established practices. This is exactly what the RPA editors had in mind when we reached out to two visionary thinkers in the field with the following question: "If you are given unrestricted power to change assessment practice for the better over the next 10 years, what does assessment look like? What changes would you make over the next 10 years and why are those changes needed?" Both David Eubanks and Keston Fulcher responded to this prompt with thoughtful and telling insights into how we might improve assessment. While the two may disagree on some points, they both agree that we can do better. We invite you to explore their visions of the future of assessment along with an integrative response from RPA associate editor, Megan Good. We hope that these thoughts spur you to consider your own vision of assessment.

We also wish to note that this is purposefully an incomplete conversation. *Research & Practice in Assessment* is partnering with the IUPUI Assessment Institute to produce a podcast where David, Keston, and Megan will continue this engaging conversation about the future of assessment.

To listen to the podcast visit: https://assessmentinstitute.iupui.edu/overview/podcast-episodes.html

*Forward by RPA Editor-in-Chief, Nicholas A. Curtis, Ph.D.*

### Authors

David Eubanks, Ph.D
*Furman University*

David Eubanks holds a doctorate in applied mathematics and has served at four colleges and universities over a 29-year career. Since 2015 he has served at Furman University as Assistant Vice President for Institutional Assessment and Effectiveness. His 2016 article "A Guide for the Perplexed" challenged the practical usefulness of standard assessment practices for accreditation reporting and led to an ongoing conversation within the higher education community.

Keston Fulcher, Ph.D
*James Madison University*

Dr. Keston Fulcher is the Executive Director of the Center for Assessment and Research Studies and Professor of Graduate Psychology at James Madison University. Keston's research focuses on structuring higher education to foster learning improvement at scale. Keston's larger work endeavors to help higher education transform from a "culture of assessment" to a "culture of learning improvement."

**David Eubanks, Ph.D.**
**Furman University**

This essay is a summary of my conclusions from working in assessment, accreditation, and institutional research capacities for two decades, as a practitioner, researcher, and peer reviewer. It reflects disillusionment with the rigid, almost dogmatic, restrictions inherent in accreditation-style reports that are intended to demonstrate the quality of academic programs. Criticisms of these reports have been widely reported, most recently starting with my (2017) article about methods, and followed by public statements and articles from luminaries (Lederman, 2019).

Measuring student achievements and using that information to improve academic programs is a perfectly fine idea; it's not that this goal is unreasonable or impossible, merely that it cannot be turned into a checkbox bureaucracy. In attempting to find a middle road between improvement and accountability, compliance standards for institutional accreditors have accomplished neither.

Using the SACSCOC standard 8.2 as an example, assessment reports are explicitly data projects: "The institution identifies expected outcomes, assesses the extent to which it achieves these outcomes, and provides evidence of seeking improvement based on analysis of the results." Accordingly, assessment offices churn out dozens or hundreds of reports a year, most with small samples of poor or untested data that are subjected to rudimentary analysis, so that—at best—the only conclusions that can be drawn are from an average or from a proportion being "too low."

**In attempting to find a middle road between improvement and accountability, compliance standards for institutional accreditors have accomplished neither.**

Regardless of whether a finding is possibly random, action is required, often resulting in anemic changes like "we added more critical thinking content to the syllabus." Accreditors complain about this "checkbox" reporting (CHEA, 2019), but seem unaware that the standards in place practically ensure that efforts won't pay off. This fact is probably due to a cadre of consultants and peer reviewers who continually reinforce the rules. As if the *system* is perfect and our local problems are due to our own lack of perfection. That's certainly what I thought for years: that if I just did exactly what the consultants and accreditors were describing, a flood of insights about student learning would follow. It's embarrassing to admit that it took years to realize that the same principles I was teaching in Statistics 101 applied to assessment reports too.

The way forward is to combine all the information we have and use the best methods available. In particular, it means overcoming the prejudice against grades. Accreditor's rules vary, but for most institutions, course grades are considered invalid as primary measures of student learning, which leads to the need for a whole second set of books: rubric ratings of papers and so on. It's wasteful and ineffective to have two disconnected systems—accreditation reports and course grades—that have the same goal of assessing student achievement. We need a single integrated system with the goals of (1) improving success for all types of students, and (2) ensuring that transcripts are meaningful, both for individual courses and for degrees. For examples of the first of these, see the 2020 webinar put on by the United States Department of Education, "Predictive Analytics to Improve Student Outcomes."

A unified system uses measurement methods, meaning large samples of data gathered under similar conditions and tested for reliability and validity, but only when suitable. For everything else, we should trust faculty, who are the experts on their classes and students. Course grades lie in the intersection of those two sets: they represent a summative faculty judgment after seeing student work over a period of weeks, and there is usually a grade for every class a student takes, connected via student IDs to hundreds of other data points. As such, it is straightforward to evaluate the characteristics of grades, including reliability overall and within programs (Beatty et al., 2015), and instructor or program "leniency" (Millet, 2018). These assess the fairness of grading, which can be improved through feedback (Millet, 2010). One can look for courses that block students from curricular pathways or predict drop-outs, and include demographics or other factors as explanatory variables to identify systemic biases.

Grade validity is more difficult to assess than reliability (as in intra-class correlation), but taking the question seriously breathes new life into the assessment project. Because of the richness and completeness of the data, there are numerous strategies to try. For example, can a factor analysis of grades associated with course prefixes (e.g., BIOL or ENGL) extract dimensions that plausibly associate with domain knowledge? At my institution, the answer seems to be yes—we can distinguish "humanities" skills from "math" skills using grades.

Grades generally only show student development qualitatively, through the courses taken, since per-student grade averages tend not to change significantly over time. That makes assessment measures on a developmental scale attractive as a complement to grades. For example, an institutional study of student writing that can measure growth over time is a project that can benefit all programs. I co-authored a validity study of our work along those lines, where we found evidence of differential growth related to grade averages (Eubanks & Vanovac, 2020). The findings suggest that educational opportunities are not equally accessible to all students. Without the combination of grades and assessment data, these insights would be lost.

The framework I've described here, combining the official records of student achievement (grades) with complementary high-quality research, eliminates the need for assessment reports that attempt to improve learning outcomes one by one. That business of writing down learning outcomes, finding a plausible data element to match, and so on, has a giant plot hole: there are a lot of learning outcomes in a college degree: far too many to treat that way.

By my count, using section headings from a textbook, a first calculus course has 30-40 substantial learning outcomes, just in that one course. Focusing on individual outcomes is the wrong way to go about it; it makes much more sense to increase faculty teaching ability in general through faculty development, including pedagogy and assessment. A faculty member who notices a problem with a learning outcome while a class is going on and fixes it right away raises the level of learning generally, and there's no multi-year lag between noticing the problem and fixing it.

You've probably noticed that my description of fixing assessment means essentially doing the opposite of what consultants and accreditors have been advocating or requiring for decades. Indeed, that's my general rule of thumb by now: if it's considered a "best practice" in accreditation reporting, it's probably the reverse of what you should do to get results. I recommend that you test this for yourself by asking for evidence to back up claims.

**The main purpose of ensuring program quality—the reason for the accreditation requirements—will never be achieved with checkbox-style reporting. Our students and institutions deserve better, and the need is urgent.**

None of the foregoing takes away from the work that assessment staff does to support academic programs by helping with curriculum and course designs, thinking through what students are expected to learn, designing tests and marking processes, and so on. The good that has come from the assessment movement is undoubtedly driven in part by the accreditation mandate. However, the main purpose of ensuring program quality—the reason for the accreditation requirements—will never be achieved with checkbox-style reporting. Our students and institutions deserve better, and the need is urgent.

**Keston Fulcher, Ph.D.**
**James Madison University**

Twenty years ago, as a graduate student, I believed that higher education was a tightly conceptualized, rigorously executed enterprise. From a seat in the classroom, I felt the vast majority of professors were passionate about their subject areas and cared deeply about student success. If asked how I would make higher education better, I would have scratched my head. More parking, perhaps?

Ten years ago, as an early-career assessment professional, I was in a better position to think about the efficacy of higher education. My gaze narrowed in on academe's ability to foster student learning. Shouldn't learning be the most prized outcome of higher education? Don't we want students to have the knowledge, attitudes, and skills that prepare them for a successful career and a meaningful life? My positive view of higher education persisted. Nevertheless, the post-secondary sheen did not appear as bright. I began realizing that programs were not perfect. They could be more effective if tweaked. And, of course, the mechanism for tweaking could be nothing else than solid assessment practice (said from the myopic lens of an assessment professional).

I realized that the requirements for assessment, through accreditors and internal college policies, promoted a checkbox mentality. Assessment often was treated as a bureaucratic chore to accomplish rather than a mechanism for real change. If I were asked a decade ago how to make higher education better, I would have suggested more attention to assessment and more rigorous methodology. My assumption was that if higher education professionals had access to better assessment data, then they would use it to improve the enterprise!

Between then and now, several assessment insiders, including my team, have rejected this assumption. It turns out that assessment, even conducted with pristine methodology, rarely catalyzes improvement efforts. Blaich and Wise's (2011) excellent work on the Wabash Project shined a bright, expensive light on the misunderstanding. They, too, believed learning improvement would be propelled if institutions could access robust assessment methodology. Two million dollars later, the team had helped dozens of institutions gather trustworthy data but found little evidence of use, much less evidence of improved student learning. In other words, Blaich and Wise debunked the Copernican-like-view that assessment lies at the center of the learning improvement universe. Similarly, at James Madison University, we began examining our assessment reports across time. Over the years, we had provided assessment guidance and support to academic programs. And, the work appeared to pay off. Almost all areas of assessment were demonstrably better. The exception: the use of data for improvement. Uggh. It was the Blaich and Wise finding at a smaller scale.

Today, I think about the bigger question. What would it take to improve student learning at scale? I also believe that higher education's treatment of student learning does not need a tweak. It needs an overhaul. My colleagues and I are working on a new model that centers on faculty, staff, and administrators. A model where higher education is structured to make evidence-informed changes to the learning environment.

**It turns out that assessment, even conducted with pristine methodology, rarely catalyzes improvement efforts.**

For the next 10 years, our vision is to improve higher education by inspiring and empowering faculty, staff, and administrators to make evidence-based decisions to enhance student learning and development. Enhancing (i.e., improving) student learning and development is explicit in the vision. If we don't see it, we haven't achieved it. Furthermore, the lever to achieve improvement is empowering faculty and staff to make evidence-based decisions. Any effort to improve student learning flows through these on-the-ground educators. It's this empowerment that we are putting our mouths and our money behind.

While not explicit in the vision statement, our strategy for empowering faculty is through professional development. In other words, we believe that a major obstacle to improved learning at scale is a lack of knowledge and skills related to evidence-based decision making. Professional development will help higher education push through this considerable obstacle. If educators can make better decisions, they can make informed changes to the learning environment that can foster better student learning.

We are not abandoning high-quality assessment designs, far from it. However, we argue that without the skills to make evidence-based decisions, faculty and staff will get little use out of quality data. It's like tossing car keys to an unlicensed driver. Educators need time, guidance, and practice to make evidence-based decisions to enhance student learning and development.

I think it is possible for every institution in the United States to achieve at least one example of program-level learning improvement by 2031. For example, a biology program could show that a future cohort of students could write better than a previous cohort because of targeted writing interventions. Operationally, what would it look like to "improve learning" in 10 years?

First, all of higher education, and particularly assessment professionals, must abandon the notion that the lack-of-improvement problem can be solved merely by developing better assessment methodology. Tweaking rubrics won't get us there. Better sampling designs won't get us there. "Big Data" analyses won't get us there. These are all useful but insufficient tools to improve student learning.

Second, we must think of assessment as part of a larger learning system, a system where assessment is integrated with components that influence the learning environment (e.g., program theory, evidence-based interventions, implementation, and change management).

Third, if the end game is better student learning, we should create professional development around that notion. In addition to developing assessment practitioners/professionals and faculty developers, shouldn't we be developing learning system coordinators?

Fourth, and this is the toughest. We need to flesh out, in great detail, the needed knowledge and skills for higher educators to be evidence-based decision makers in a learning systems framework. In other words, how do we help faculty and student affairs professionals prepare to lead successful learning improvement efforts? I suspect that if leaders in assessment and faculty development put their heads together, we could make great strides in this area.

Finally, and this is not mentioned explicitly in the vision, we need to provide guidance to administrators. Deans, provosts, vice presidents, and presidents are the leaders that set priorities and create infrastructure by which initiatives happen. They, too, need guidance to create an environment by which professional development flourishes for on-the-ground faculty and student affairs professionals.

**I believe that the future of better assessment in higher education ironically calls for its subordination; subordination under a larger learning systems framework.**

In sum, I believe that the future of better assessment in higher education ironically calls for its subordination; subordination under a larger learning systems framework. The path forward includes articulating this new framework and providing professional development around it.

*RESPONSE BY:*
**Megan Good, Ph.D.**
***RPA Associate Editor***

*"If you are working on something exciting that you really care about, you don't have to be pushed. The vision pulls you."* – Steve Jobs

Does assessment have a vision? At best, one might clumsily say the vision is to improve student learning; when pressed on what the vision looks like, one might falter. Two trailblazers among us have defined tangible visions – Dr. Keston Fulcher and Dr. David Eubanks. At first, I thought the two perspectives were wildly different. But on reflection, I see great commonalities. Here, I will note the similarities and comment on each perspective.

The primary similarity between Eubanks and Fulcher is a feeling of disillusionment with our current assessment practices, both mentioning forms of bureaucracy. If the vision is to 'improve' student learning and achievement, our current system essentially isn't working. We need space for more innovative work that may truly impact students. To that end, both authors share a sense of optimism for the future. Finally, each visionary incorporates educational development into their dreams.

**The primary similarity between Eubanks and Fulcher is a feeling of disillusionment with our current assessment practices**

Both authors are creating new systems. However, their systems look quite different. In Fulcher's vision, assessment practice seems to remain largely the same (presumably, there are still student learning outcomes, curriculum mapping, and aligned measurement components). Assessment practice, however, becomes subordinate to educational development in the new learning system. The vision behind Fulcher's learning system is clear - achieve evidence of improved student learning. I love that educational development is front and center. Indeed, it is clear that faculty and administrators need assistance in improvement efforts. However, I wonder – would Fulcher's future still mandate the "dreaded" assessment reporting requirements? And I wonder what evidence of learning improvement would achieve? It sounds nice! But senior administrators are not bothered with this metric now, why would they be in the future? The public is certainly not demanding it (though the COVID-19 pandemic may change this).

Dr. Eubanks's future creates an entirely new system. It seems assessment is replaced with a more robust analytic system using as much data as possible (notably including grades) to understand student learning in real-time. The analytic system would yield findings for faculty, administrators, and educational developers to improve. Students could be helped when they need it. In such a system, it seems the workload could shift from the faculty to the analysts, creating space that might be used for improvement conversations. And, although I was trained that grades are not assessment measures, I recognize that using them as Eubanks described could be powerful and create new faculty partners. Finally, I appreciate that Eubanks has acknowledged educational development's role in the new system (though perhaps to a lesser degree than Fulcher). I have one primary question/concern for this future – how do we know that the new robust system doesn't yield reports that sit on a shelf like current assessment reports? In the new system, why would faculty and administrators be any more motivated to apply these new data to drive change?

Perhaps there could be a hybrid vision that combines Eubanks's and Fulcher's perspectives. I imagine Eubanks's analytic system could replace assessment and could exist within Fulcher's learning system. In this way, the data would be different, but the primary focus would still be on improvement with educational developers central to the system's success. Regardless of the future, these visionary pieces have left me excited for the possibilities to change higher education for the better.

**Regardless of the future, these visionary pieces have left me excited for the possibilities to change higher education for the better.**

Assessment practice has certainly been a beneficial practice in higher education over the last thirty years. Much good has come from it, notably faculty working together within a program as opposed to merely independent contractors. Now, it's time to level-up and build on the progress made. I leave you with three hopes for the future. First, I hope accrediting agencies will recognize the need to create space to engage in more interesting work. It's hard to imagine innovation while turning the crank to ensure assessment reports are neatly filed ahead of an accreditation visit. Second, I hope assessment practitioners continue to create partnerships with educational developers. Regardless of vision, the next assessment movement must rely more heavily on the expertise of educational developers. Last, I hope we can create a strategy to inspire higher education to truly value learning and improvement. Learning is often taken for granted (of course it's happening!) and senior administrators are generally not obsessed with ensuring we're getting this right and making it better. The pandemic has opened this conversation – many parents and students have started asking tough questions about learning as modalities shifted. Can we harness this attention to support our next vision? We can and we must.

## References


Beatty, A. S., Walmsley, P. T., Sackett, P. R., Kuncel, N. R., & Koch, A. J. (2015). The reliability of college grades. *Educational Measurement: Issues and Practice*, *34*, 31–40. https://doi.org/10.1111/emip.12096

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). http://www.bu.edu/provost/files/2015/09/From-Gathering-to-Using-Assessment-Results_Lessons-from-the-Wabash-Study-C.-Blaich-K.-Wise1.pdf

Council for Higher Education Accreditation. (2019). *Accreditation and student success outcomes: Perspectives from accrediting organizations*. CHEA/CIQG Publication Series. https://www.chea.org/sites/default/files/pdf/Accreditation%20and%20Student%20Learning%20Outcomes%20-%20Final.pdf

Eubanks, D. (2017). A guide for the perplexed. *Intersection*, 4-14. https://www.aalhe.org/assets/docs/AAHLE_Fall_2017_Intersection.pdf

Eubanks, D. A., Vanovac, S. (2020) Divergent Development of Writers in College, *Journal of Writing Assessment*, *4*.

Lederman, D. (2019, April 17). Harsh take on assessment … from assessment pros. *Inside Higher Ed*. https://www.insidehighered.com/news/2019/04/17/advocates-student-learning-assessment-say-its-time-different-approach

Millet, I. (2010) Improving grading consistency through grade lift reporting. *Practical Assessment Research and Evaluation*,*15*(4), 1–8. https://doi.org/10.7275/zp5s-xe72

Millet, I. (2018) The relationship between grading leniency and grading reliability. *Studies in Higher Education*, *43*(9), 1524-1535. https://doi.org/10.1080/03075079.2016.1265497

United States Department of Education. (2020). *Predictive analytics to improve student outcomes* [Webinar]. https://mediasite.ed.gov/webcast/Play/44bab8fe1b54425792e62f452837a5641d

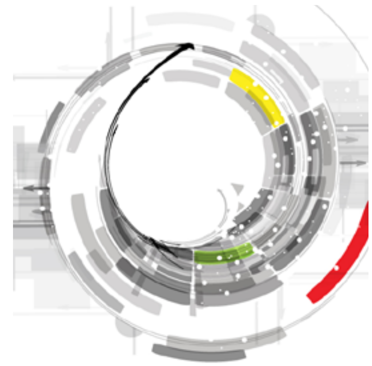


**Podcast: Leading Improvements in Higher Education**

**Stay connected to hear the latest research-based topics featured on IUPUI Assessment Institute!**
**https://assessmentinstitute.iupui.edu/overview/podcast-episodes.html.**

## *Abstract*

In an effort to create a meaningful but reduced set of institutional core competencies, the Oregon Health & Science University (OHSU) Core Competency Project was developed. This paper reviews the importance of core competencies as a unifying institutional tool to examine equitable outcomes for all learners across schools and programs and to meet the expectations of external accreditors. Researchers utilized textual analysis to collect data from 60+ accreditors' guidelines and constant comparative analysis to interpret the data. The results of the study highlight a data-informed approach to competency development that engages stakeholders and provides an approach for other institutions to consider. This research study occurred in the midst of community calls for social justice and during a global pandemic, and these social contexts impacted the study in significant ways. It is from a process of rigorous debate paired with passionate calls for change that meaningful core competency definitions emerged. Researchers conclude by reflecting on the lessons learned from core competency development in times of crisis.


*AUTHORS*

Constance R. Tucker,
MA., Ph.D.
*Educational Improvement & Innovation, Oregon Health & Science University*

Kirstin Moreno,
MSEd, PhD.
*Educational Improvement & Innovation, OHSU*

Masooma Jafari,
PharmD, MPH, CPH
*Educational Improvement & Innovation, OHSU*


# Shifting From Alignment to Transformation: Crosswalk to Graduation Core Competency Development

The Office of Educational Improvement and Innovation (EII) at Oregon Health & Science University (OHSU) started the OHSU Core Competency Project in July 2019 to revise the institutional core competencies at OHSU as part of the Northwest Commission on Colleges and Universities Mission Fulfillment Fellowship project (Northwest Commission on Colleges and Universities [NWCCU], 2020). The revision project sought both to reduce the number of competencies and to update to more meaningful competencies. The following sections describe the process used to revise institutional core competencies: 1) a brief explanation of the importance of core competencies; 2) a statement of purpose for the core competency project; 3) a summary of the core competency project methods, analyses, and results; and 4) reflections on the lessons learned from the study, including future actions.

## The Importance of Core Competencies

Competency-based education (CBE) has been gaining more attention as a practical approach for training a more knowledgeable and skilled healthcare workforce. In competency-based education, observable and measurable performance metrics and core competencies are established. Students in health professions must achieve the metrics and competencies to be considered proficient (Epstein & Hundert, 2002; Frank et. al, 2010).


*CORRESPONDENCE*

***Email***
tuckeco@ohsu.edu

Core competencies articulate a robust set of explicit expectations for student learning outcomes and can shape the culture of equity within higher education (Brower et al., 2017; Cleary & Breathnach, 2017). Core competencies are essential as they represent the desired learning outcomes of a program or across programs in an observable and measurable way to specify the expectations of the program(s). Approaches to developing institution-wide core competencies are limited within the literature. After review of the literature, the most common approaches to compiling institutional core competencies are expert panels, crowdsourcing, and backward design (Council on Linkages between Academic and Public Health Practice, 2014; Interprofessional Education Collaborative's Core Competency, 2016; Kerchner et al., 2012).

**Core competencies are essential as they represent the desired learning outcomes of a program or across programs in an observable and measurable way to specify the expectations of the program(s).**

Even more limited are reports and studies on the development of institutional core competencies at health science centers. One such report is the Interprofessional Education Collaborative's Core Competencies for Interprofessional Collaborative Practice (2016) in which four competencies were developed by an expert panel to move beyond profession-specific educational efforts and engage learners from different health professions to learn with, from, and about each other. This previous work informed and grounded the institutional core competency development process at OHSU.

## Purpose of Core Competency Refinement

Oregon Health & Science University's main campus is in Portland, Oregon. OHSU is Oregon's only academic health center and is nationally distinguished as a research university dedicated solely to advancing health sciences. This allows us to focus on discoveries to prevent and cure disease, on education that prepares the health care and health science professionals of the future, and on patient care that incorporates the latest advances. OHSU has five schools and colleges including the School of Dentistry, School of Medicine, School of Nursing (which has students around the state), School of Public Health (a collaboration with Portland State University), and the College of Pharmacy (a collaboration with Oregon State University). As of April 2020, OHSU is a comprehensive university with 102 programs, the majority of which are graduate and professional programs. Successful assessment at the institution level requires a multidimensional and highly collaborative process that recognizes the diversity of programs, degrees, and unique contributions to the health sciences. In 2013, the first set of 10 OHSU graduation core competencies were developed as part of a university-wide interprofessional education initiative. As diverse academic programs learned from, with, and about each other, administrators came to a consensus around 10 graduation core competencies in which all OHSU graduates were required to demonstrate proficiency (Figure 1).

To ensure achievement, all academic programs at OHSU aligned their curriculum to the 10 core competencies. The alignment of curriculum to the institutional core competencies is reviewed annually by the OHSU Assessment Council. During the 2019 academic year, the Assessment Council data indicated that some of the core competencies such as teamwork, patient/client-centered care, lifelong learning, (patient) safety and quality improvement, and systems were not adequately represented in all OHSU programs. Also, because some competencies such as patient safety and quality improvement and patient/client-centered care were difficult to measure, a core competency project was developed to provide recommendations to academic leadership about developing fewer and more meaningful competencies.

The project's objectives were two-fold. First, the project reviewed and identified the core competencies that were recommended or required for all or most of the dozens of specialized accreditors in health professions. While the practice of reviewing key documents and position descriptions to develop professional competencies is common in undergraduate education, it is less common in the health professions literature (Interprofessional Education Collaborative, 2016; Rhodes, 2010). Second, this project analyzed the existing institutional data to make recommendations to the Assessment Council and other committees responsible for the ultimate approval of the revised set of the core competencies.

Figure 1: OHSU Graduation Core Competencies (2012-2020)


## Methodology

The epistemological approach for this project is constructionism. Social constructionism views knowledge as constructed as opposed to created (Charmaz, 2006). The methodology used was constructivist grounded theory and the methods of analysis included comparative analysis, textual analysis, and conversation analysis to explore, explain, and predict future actions (Charmaz, 2006). The study's qualitative approach follows a systematic but flexible process to collect data, code data, compare data, and generate results (Thomas, 2006).

**To ensure achievement, all academic programs at OHSU aligned their curriculum to the 10 core competencies.**

The authors conducted a preliminary literature review to understand the context of the project. The methods of textual and conversation analysis were used by answering question about information in the texts, including: 1) What and whose facts are represented? 2) What does the document leave out? 3) Who is the intended audience for the document? 4) How does the information impact behavior? In addition to placing the data in context, authors' reflections on the content of the text set the stage for in-depth analysis of the data.

The data were collected through textual analysis. Initial readings of textual data were followed by identification and labeling of segments into categories. Subsequently, redundancies and overlapping categories were eliminated to produce a model incorporating the most important competencies and representing all OHSU programs.

### Textual Analysis

Textual analysis of specialized accreditors' descriptions of competencies formed the primary source of data for this project. These documents were extant and not elicited (Charmaz, 2006). All of the gathered documents were obtained from the process of manually combing official websites of the professional accreditation bodies, committee meeting minutes from national and international health organizations, the OHSU website, and published research literature. The authors reviewed and organized 79 documents which fell into three categories of data. The authors also presented analyses to the Assessment Council and their input constituted a fourth type of data.

1. *The Specialized Accreditation Standards (SAS)*: 69 documents were collected from the specialized accreditors for different existing programs at OHSU. These documents were accessed through the official websites of the specialized accrediting bodies (i.e., Comission on Dental Accreditation and Commission on Collegiate Nursing Education).

2. *The Other Accrediting Standards (OAS)*: eight documents were collected from other accreditation bodies. These other accreditation bodies were selected based on providing more comprehensive (interprofessional) and PhD-level accreditation standards and also the specialized accreditation standards that were not applied by OHSU programs but will be used in the future (Health Professions Networks Nursing & Midwifery Human Resources for Health, 2020).
3. *OHSU (Institutional) Core Competencies (OCC)*: One document with a description of OHSU's current institutional core competencies was accessed through the OHSU website.
4. *OHSU Assessment Council's Recommended Core Competencies (ACR)*: As mentioned before, the Assessment Council (AC) oversees assessment processes at OHSU and ensures that student learning outcomes (SLOs) are connected to the curriculum and the OHSU Core Competencies. The AC had some recommendations regarding updating the core competencies informed by their understanding of assessment data and results. The authors accessed the Assessment Council's Recommendations through the Assessment Council discussions and meeting minutes.

### Constant Comparative Analysis

**The significance of these two emerging competencies is highlighted by the social context in which they were explored…**

Constant comparative analysis was used to understand the collected data (Charmaz, 2006). In this approach, the analysis process is a continuous coding and categorizing process, which involves constant comparison between the coded and categorized data. First, the authors compared core competency names for both their similarities and differences. Both initial and focused coding occurred in this phase of analysis. Second, authors utilized axial and theoretical coding to explore the nuances of core competency definitions. The summary of the constant comparative analysis process is provided in Figure 2.

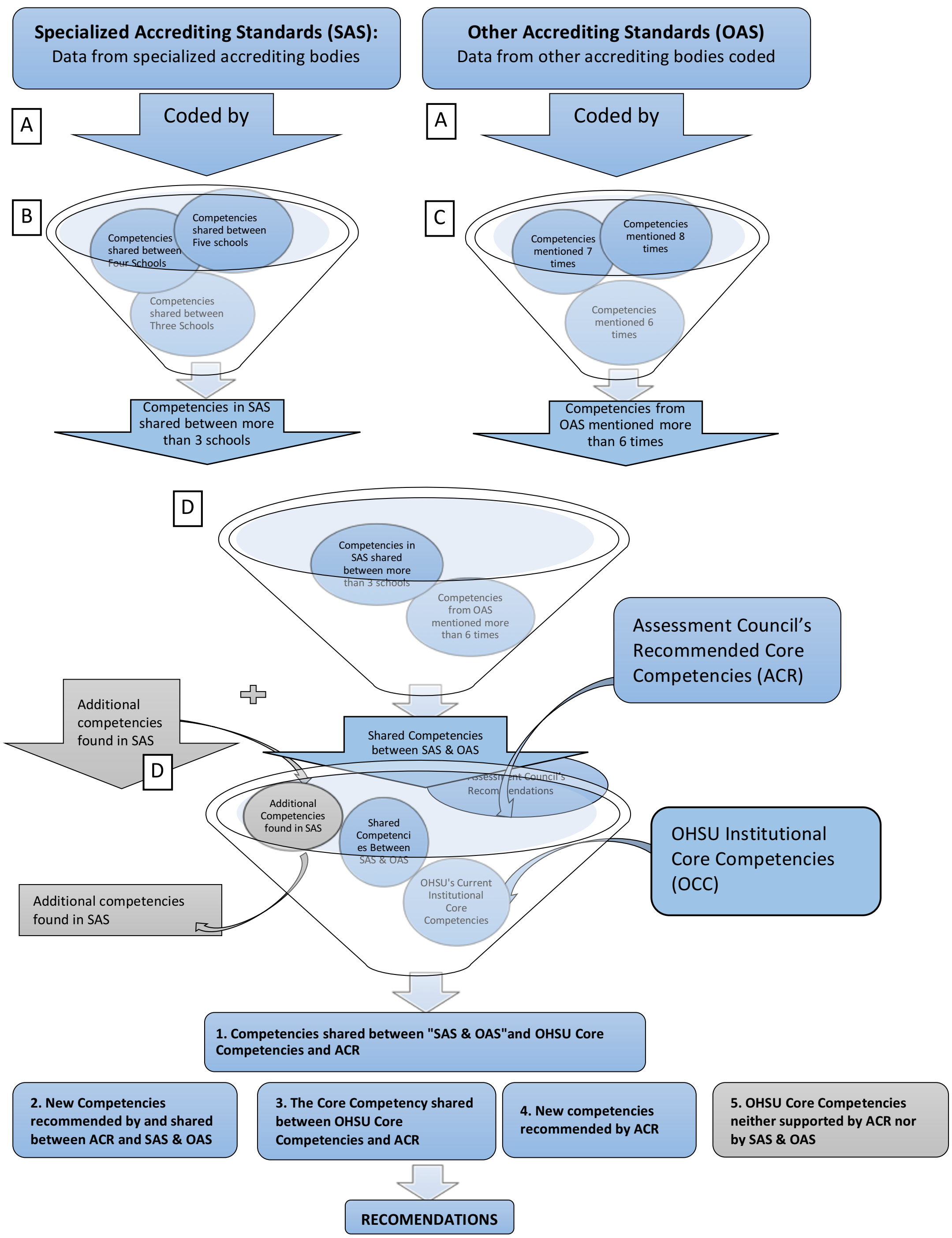
At the conclusion of thorough textual analysis, the preliminary findings were provided to numerous stakeholders for feedback and refinement. Stakeholders reviewed the findings using adapted charrette activity, a faculty-driven collaborative peer review process, to corroborate or refine the authors' findings (National Institute of Learning Outcomes Assessment, 2018). This iterative process resulted in a final set of core competencies and their associated definitions.

## Results

The results of the textual analysis and constant comparative analysis were illustrated in the form of a crosswalk that can be found in Table 1. It is in this chart that one can see the evolution of coding from line by line, initial coding, focused coding, and ultimately to theoretical coding.

The coding practice resulted in a modified list of core competencies (Appendix A). The list of original core competencies was reduced from ten to five and two new core competencies were added for a total of seven core competencies. The two new competencies included *Information Literacy and Community Engagement, Social Justice, and Equity* (Navarre Cleary & Breathnach, 2017). The significance of these two emerging competencies is highlighted by the social context in which they were explored. The authors conducted this study in Portland, OR which is a "left-leaning" culture during a period of pandemic and violent protest in both physical and virtual space. As the authors were exploring core competencies, numerous community members, students, faculty, and staff demanded policy and procedural changes that ensured that #BlackLivesMatter. In contrast, other faculty, staff, students, and community members counter-protested with #BlueLivesMatter and #ProudBoys. Protest and pandemic influenced the prioritization of Information Literacy in which one is challenged to think critically, as well as Community Engagement, Social Justice, and Equity. The project stakeholders intentionally stepped back from the data and thought carefully about what was missing from the literature. Ultimately, the project stakeholders decided, in light of a nationwide call for individuals and institutions to be allies for social

Figure 2: The Analysis Process




justice and the urgency we felt about developing information literacy and critical thinking in all students, that we would include those two as new core competencies. These two were added to the five which had emerged from the extensive data analysis. Once we had finalized the final set of seven graduation core competencies, we spent the following eight months in small groups iterating new definitions for these competencies. The Assessment Council members were the primary participants in this process. Small groups of Assessment Council members worked on definitions and then sent drafts of the definitions to the whole Council for anonymous voting. After several rounds of voting, the Assessment Council settled on a set of core competency definitions and solicited feedback from various other stakeholder groups on campus. This resulted in additional rounds of iterating definitions by a different

Table 1:
Crosswalk of Institutional and Health Professions Competencies

| 2016 OHSU Core Competencies | ACR* | SAS & OAS* | SAS |
|---|---|---|---|
| Professional Knowledge & Skills | Professional Knowledge And Skills | Application Of Knowledge Into Practice | |
| Patient/Client Centered Care | Patient/Client Centered Care | | Comprehensive, Patient-Centered Care<br>Population Health |
| Communication | Communication | Communication | |
| Teamwork/ Collaboration | Teamwork/ Collaboration | Interprofessional Collaborative Skills | |
| Professional Practice & Ethical | Professional Practice And Ethical | Interpersonal Relations And Teamwork | |
| Reasoning And Judgement | | Leadership | Decision Making<br>Problem-Solving |
| Lifelong Learning | | Ethics | Continuing Education<br>Assessment<br>Teaching And Mentoring/Educating |
| Evidence-Based Practice & Research | | Legal/Regulatory Standards | Research<br>Evidence-Based Care/ Practice<br>Prevention |
| Safety & Quality Improvement | | Professionalism | |
| Systems | | | Systems Thinking<br>Health Policy<br>Organizational Dynamics |
| | Information Literacy | | Data Management |
| | Community Engagement/Social Justice & Equity | Cultural Competence | |
| | Critical Thinking | | Critical Thinking |
| | Professional Identity | Professional Values | |

* ACR: Assessment Council Recommendations
* SAS & OAS: Overlapping between Specialized Accrediting Standards & Other Accreditation Standards
* OAS: Other Accreditation Standards

set of Assessment Council members broken into small groups. Members took the drafts of definitions back to their program faculty, staff, and students to get input several times during the eight-month process. At the end of the eight-month period, Assessment Council members were set to take a final vote and seek the Board of Trustees' approval (Appendix B).

## Conclusion

We are pleased that we were able to accomplish the goal of the NWCCU fellowshipproject which prompted this work: to create a shorter, more meaningful and well-aligned list of graduation core competencies. This accomplishment is significant as it allows the institution to document competency-based learning and growth of the whole student across all academic programs and student services. In addition, when the graduation core competencies are meaningful, the strong alignment of instruction, assessment, and faculty development are achievable and reinforced. It is the authors' intent to stimulate discussion and actions at other institutions that build on this work to develop institutional graduation core competencies that are meaningful, measurable, data-based, and accurate.

**When the graduation core competencies are meaningful, the strong alignment of instruction, assessment, and faculty development are achievable and reinforced.**

## Epilogue

While measurable and aligned core competencies are not transformative in and of themselves, the context in which we engaged in this process was transformational. This core competency revision project was wrapping up in late spring of 2020. At the conclusion of the research project, many of our learners, staff, and faculty started fighting the COVID-19 pandemic on the frontlines and speaking up and protesting in support of Black lives, while abruptly shifting to online learning. This sudden new context shifted the way we felt about "what students were supposed to achieve upon graduation." It became clear that the work ofrevising our graduation core competencies was not done until we examined ways to integrate anti-racism and equity into each of the new core competency definitions. We unexpectedly spent the summer of 2020 creating definitions for the competencies which resonated with the call for OHSU to become an anti-racist institution. The process of core competency revision in this context created unique opportunities to think differently about how we engage stakeholders, advocate for social justice, and reinforce the humanity of all.

## References

Brower, A. M., Humphreys, D., Karoff, R., & Kallio, S. (2017). Designing quality into direct-assessment competency-based education. *Competency-Based Education*, (2)2. https://doi.org/10.1002/cbe2.1043

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage Publications.

Council on Linkages Between Academia and Public Health Practice. (2014). *Core competencies for public health professionals*. http://www.phf.org/resourcestools/pages/core_public_health_competencies.aspx

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *Journal of the American Medical Association*, *287*(2), 226-235. https://doi.org/10.1001/jama.287.2.226

Frank, J. R., Mungroo, R., Ahmad, Y., Wang, M., De Rossi, S., & Horsley, T. (2010). Toward a definition of competency-based education in medicine: A systematic review of published definitions. *Medical Teacher*, *32*(8), 631-637. https://doi.org/10.3109/0142159x.2010.500898

Health Professions Networks Nursing & Midwifery Human Resources for Health. (2010). *Framework for Action on Interprofessional Education & Collaborative Practice*. World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/70185/WHO_HRH_HPN_10.3_eng pdf;jsessionid=8E1C5B6B6BB299C695E3F30E93D9B034?sequence=1

Interprofessional Education Collaborative. (2016). *Core Competencies fo Interprofessional Collaborative Practice: 2016 Update*. https://nebula.wsimg.com/2f68a39520b03336b41038c370497473?AccessKeyId=DC06780E69ED19E2B3A5&disposition=0&alloworigin=1

Kerchner, M., Hardwick, J. C., & Thornton, J. E. (2012). Identifying and using 'core competencies' to help design and assess undergraduate neuroscience curricula. *Journal of Undergraduate Neuroscience Education*, *11*(1), A27–A37. https://www.funjournal.org/wp-content/uploads/2015/09/kerchner_etal_11_1_a27_a37.pdf?x89760

Navarre Cleary, M., & Breathnach, C. (2017). Competency-based education as a force for equity. *Competency-Based Education*, (*2*)1. https://doi.org/10.1002/cbe2.1040

Northwest Commission on Colleges and Universities. (2020, December). *NWCCU Mission Fullfillment Fellowship*. https://www.nwccu.org/programs_and_events/fellowship/

Rhodes, T. (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics*. Washington, DC: Association of American Colleges and Universities. https://www.aacu.org/publications-research/publications/assessing-outcomes-and-improving-achievement-tips-and-tools-using

Thomas, D.R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, *27*(2), 237-246. https://doi.org/10.1177%2F1098214005283748

## Authors Note

Masooma Jafari, PharmD, MPH, CPH is now the Presidential Health Advisory Office in Kabul, Afghanistan. We have no conflicts of interest to disclose. Acknowledgements: The authors acknowledge the efforts of the OHSU Assessment Council and the NWCCU Mission Fulfillment Fellowship and are grateful for their partnership. We acknowledge the hours of thought, care, and work our Assessment Council members and program faculty put into assessment and are grateful to partner with them in improving student learning at OHSU.

*Correspondence concerning this article should be addressed to Constance Tucker, 3181 SW Sam Jackson Park Rd. MC: L349, Portland, OR 97239 tuckeco@ohsu.edu*

Constance R. Tucker, MA., PhD.
https://orcid.org/0000-0002-6507-8832

Kirstin Moreno, MSEd, PhD.
https://orcid.org/0000-0002-1725-5494

## Appendix A

Proposed OHSU Graduation Core Competencies

### OHSU Graduation Core Competencies

2012-2020

- Professional Knowledge and Skills
- Teamwork
- Communication
- Patient/client-centered Care
- Evidence-based Practice and Research
- Lifelong Learning
- Reasoning and Judgement
- Professionalism and Ethics
- Safety and Quality Improvement
- Systems

### OHSU Graduation Core Competencies

*Proposed* starting Sept 2020

- Professional Knowledge and Skills
- Teamwork
- Communication
- Patient Centered Care
- Information Literacy
- Professional Identity
- Community Engagement, Social Justice, and Equity

## Appendix B

# OHSU Graduation Core Competencies

In the summer of 2020, OHSU affirmed its commitment to the health and wellbeing of all Oregonians and asked everyone to work together to shatter structural racism. The new core competency definitions align the Education Mission with OHSU's anti-racism work. The revision was undertaken with the following principles in mind:

- *Power, privilege, and positionality impact how people function as professionals and interact in the world.*
- *Seeking and listening to diverse voices results in better outcomes.*
- *Knowledge and authority are constructed and contextual.*
- *Information has power and existing systems privilege some perspectives and present barriers to others.*
- *Systemic racism causes undue burden and may not impact everyone in the same way.*
- *Open-mindedness and compassion are core OHSU values that enhance our effectiveness.*
- *Our audience should inform how we communicate.*
- *We are a professional community, dedicated to improving the human condition.*

**Professional knowledge and skills**
Demonstrate core knowledge, skills, and practices as defined by the discipline, professional licensing, or accreditation organization while being open to new perspectives, additional voices, and changes in schools of thought that impact the core knowledge, skills, and practices in the discipline.

**Professional Identity and Ethical Behavior**
Demonstrate discipline-specific behaviors, norms, and ethics while also recognizing and challenging racist professional expectations which can cause undue burden and/or deny the full humanity of ourselves, our peers, and our patients.

**Information Literacy**
Recognize the power of information in educating, influencing, and understanding the world, while seeking and amplifying missing perspectives. With this lens, locate, critically evaluate, and effectively use information to participate in decision-making, quality improvement, and broader scholarly discourse.

**Communication**
Communicate effectively and equitably with diverse individuals, organizations, and communities to support stakeholder decision-making and promote culturally responsive exchanges of information.

**Teamwork**
Work effectively within collaborative, team- or teaming-based interprofessional environments while acknowledging positionality and intentionally making space for diverse perspectives.

**Community Engagement, Social Justice and Equity**
Apply principles of social justice, equity, and/or anti-racism through community-engaged practice, service, or scholarship.

**Patient Centered Care**
Clinical degree program graduates will collaborate with diverse individuals, families, and communities to provide quality trauma-informed care that is anti-racist and respectful of and responsive to preferences, needs, attitudes, beliefs, and values.

OHSU

*Approved by OHSU Board September 2020*

## Abstract

Kolb's experiential learning cycle theorizes the process of learning through a hands-on experience. Although the Association of American Colleges and Universities (AAC&U) Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics provide a direct measure of the qualities of this learning cycle, few indirect measures have been developed to accompany the rubrics and the learning cycle. This paper aims to demonstrate construct validity and measurement invariance of pre-experience and post-experience surveys intended to measure undergraduate students' perception of learning in an experiential learning context. Construct validity and longitudinal measurement invariance were examined through a confirmatory factor analysis. Findings suggest the instruments provide an adequate measure of students' perceptions of learning. In addition, partial scalar measurement invariance was achieved supporting the ability to compare growth between surveys. The survey instruments serve as strong indirect measures of Experience Learning Student Learning Outcomes. These indirect measures, coupled with direct measures, provide evidence of learning through hands-on experiences, however evidence of growth is less robust.



### AUTHORS

John P. Walker, MPS
*University of Tennessee*

Louis M. Rocconi, Ph.D.
*University of Tennessee*


# Experiential Learning Student Surveys: Indirect Measures of Student Growth

A steady increase of scholarship in experiential learning has demonstrated a need for effective measures of these experiences (Seaman et al., 2017). Boyatzis et al. (1995) proposed that institutions conduct longitudinal studies to determine the value added to learning and continuously assess the learning process. Although there are several methods to collect student data to measure learning in an experiential context, indirect measures, such as surveys, allow students to reflect on the learning experience from their perspective (Banta & Palomba, 2015). Surveys used in a within-subjects design provide vital information on growth; however, researchers and practitioners need to demonstrate that these surveys measure what they intend to measure and do so consistently. This paper assesses the construct validity and measurement invariance of two surveys, pre-experience and post-experience, used to measure student attainment of the Experience Learning Student Learning Outcomes (SLOs), as established by a southeastern U.S., four-year, research university.


### CORRESPONDENCE

**Email**
jwalk141@utk.edu


## Kolb's Experiential Learning Cycle

The notion of learning by experience is not a new concept. Notable educational analysts, John Dewey and David Kolb, each laid the groundwork for the importance of experiential learning. Dewey (1938) contended that students' potential is hindered by the traditional classroom approach to learning which focused on delivering knowledge and little emphasis on the application of the knowledge. Kolb (1984) formalized the fundamental theory of experiential education which summarized the steps needed for learning to happen in a hands-on experience. Kolb posited, "learning is the process whereby knowledge is created through the transformation of experience" (Kolb, 1984, p. 38). He theorized that experience, coupled with structured reflection, allows students to grow through hands-on learning-a cyclical process that outlines the pathway to learning by doing, reviewing, concluding, and planning. Reflection and thoughtful planning follow each experience which leads to improvements for the next experience.

Kolb's experiential learning model is based on his identification of two ways of *acquiring knowledge and skills* through an experience (concrete experience and abstract conceptualization) and two ways of *transforming* through an experience (reflective observation and active experimentation). The first step of Kolb's experiential learning cycle states that learning begins by doing. Students will only begin to construct the skills and knowledge needed through observation and application in an experience. The first step serves as the cornerstone to the learning process. Once students have completed the initial experience, reflective behavior pushes learning forward. Reflection, the second step, allows students to strategize improvements and brainstorm new ideas for the next experience. The third step, abstract conceptualization, occurs when students formulate solutions for improvement to apply to the next experience based on their reflection. Students can then implement the solution as the final step of the learning process through experimentation. The iterative process continues through further reflection, planning, and testing that builds on one experience after the next.

Kolb's experiential learning theory has been adapted and applied in several higher education contexts. The theory has been applied to many course-based curricula (Abdulwahed & Nagy, 2009; Healey & Jenkins, 2000; Petkus, 2000; Russell-Bowie, 2013) and aligned to other educational models in higher education (Baker et al., 2012; Poore et al., 2014). For instance, Reshmad'sa and Vijayakumari (2017) investigated the pedagogical skills of pre-service teachers ($n$ = 40) in student teaching roles. The authors measured students' teaching aptitude and their use of active learning strategies in a classroom setting. They found the use of Kolb's experiential learning strategy was substantially more effective than conventional teaching strategies. That is, student teachers that used Kolb's experiential learning strategies were more reflective and demonstrated better development of pedagogical skills compared to student teachers using conventional teaching strategies.

**....experience, coupled with structured reflection, allows students to grow through hands-on learning**

Some researchers have attempted to directly measure experiential learning opportunities based on Kolb's experiential learning theory model. For example, Smith and Rayfield (2017) used Kolb's Learning Style Inventory (KLSI) to examine the preferred learning styles of individuals in an experiential setting. The KLSI is a direct measure designed to categorize students into nine learning styles related to Kolb's learning model. However, the inventory has received criticism for its "pigeonhole" approach to evaluate learner style (Manolis et al., 2013). The KLSI is focused on identifying an individual's learning style with little emphasis on the measure of actual learning. Additionally, the inventory does not take learner perception of self-efficacy into account. An indirect measure of learner perceptions could provide a new perspective of self-assessment and self-awareness through the learning process.

Some studies have noted that Kolb's experiential learning cycle falls short of capturing the learning process through hands-on experiences (Bergsteiner et al., 2010; Miettinen, 2010). Bergsteiner and Avery (2014) suggested that the model insufficiently attends to the numerous facets of real-world learning including cultural and emotional contexts that can be captured through student perceptions of their learning environment. Warren et al. (1995) argued, "experiential methodology is not linear, cyclical, or even patterned. It is a series of working principles, all of which are equally important or must be present to varying degrees at some time during experiential learning" (p. 243). Through an examination of the literature, Chapman et al. (1992) found experiential education is grounded on, among other principles, meaningful relationships built from collaborative endeavors, structured reflection, and emotional investment and engagement in the experience. Although Kolb and colleagues have examined the role of teamwork and collaboration in experiential learning in later work (Kayes et al., 2005), the cycle fails to emphasize the importance of collaboration, and the principles of lifelong learning and engaged scholarship. Thus, we sought to develop a valid and reliable instrument that incorporated these missing learning characteristics into the survey instruments.

## AAC&U's VALUE Rubrics

Two experiential learning survey instruments, based on research and instruments from the Association of American Colleges and Universities (AAC&U), were developed to measure undergraduate students' perception of their learning in an experiential learning context. The AAC&U created 16 rubrics to measure student learning on skills essential to employers and faculty known as the Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics (Rhodes, 2010). Several VALUE rubrics measure skills essential to experiential learning theory (e.g., critical thinking, foundations and skills for lifelong learning, teamwork, integrative learning) and have been widely used across many institutions to measure these latent traits objectively and reliably (Finley, 2011; Rhodes, 2010; Rhodes & Finley, 2013). They are designed to be a direct measure of student achievement, but modifying the language allows the rubrics to be used as an indirect measure to quantify learning from the learners' perspective. The rubrics served as an important framework to develop the Experiential Learning Student Surveys. The surveys include measures on teamwork (collaboration), lifelong learning, and engagement, elements not explicitly mentioned in Kolb's experiential learning cycle. Table 1 shows how each item in the surveys is aligned with the VALUE rubrics.

**The purpose of this study was to assess the construct validity and longitudinal measurement invariance of two surveys designed to measure students' perceptions of learning in an experiential learning context.**

## Experience Learning SLOs

The primary goal of experiential learning is to enhance students' development and educational experiences by providing more opportunities for real-world learning. Experiential learning is most effective when it is a dynamic approach in which students engage, apply, collaborate, and reflect on course content and lessons learned (Kolb & Kolb, 2011). The Experience Learning SLOs were designed to incorporate Kolb's four-stage learning cycle with the addition of collaboration and lifelong learning. Since learning occurs at all of these stages, it is important to measure students' learning and growth throughout the process. These stages of experiential learning therefore formed the foundation for defining our desired SLOs:

SLO 1: Students will value the importance of engaged scholarship and lifelong learning.

SLO 2: Students will apply knowledge, values, and skills in solving real-world problems.

SLO 3: Students will work collaboratively with others.

SLO 4: Students will engage in structured reflection as part of the inquiry process.

The Experience Learning SLOs represent a holistic approach to learning that emphasizes learning through experiences. The four interrelated SLOs are assessed using two indirect measures (pre-experience and post-experience surveys, described later) to understand student perceptions of self-efficacy in an experiential learning context.

## Purpose

The purpose of this study was to assess the construct validity and longitudinal measurement invariance of two surveys designed to measure students' perceptions of learning in an experiential learning context. Specifically, we were interested in measuring students' attainment of the above-mentioned Experience Learning SLOs. We utilized confirmatory factor analysis to assess the extent to which the pre-experience and post-experience surveys measure the SLOs.

Table 1.
*Alignment of survey items with Experience Learning SLOs and AAC&U VALUE rubric*

| Item No. | Pre-Experience Survey Items | Post-Experience Survey Items | SLO | VALUE Rubric Item | VALUE Rubric |
|---|---|---|---|---|---|
| 1 | I often participate in activities that serve the needs of others | I am interested in exploring the problems of society (i.e. the needs of others) | 1 | Show evidence of interest in the problems of society (needs of others) | Foundations and Skills for Lifelong Learning |
| 2 | I think it is important for the university to use its resources for the benefit of society | I think it is important for academia to use their resources for the benefit of society | 1 | Value (i.e., offer a positive attitude toward) the use of engaged scholarship to address societal problems | Foundations and Skills for Lifelong Learning |
| 3 | I often participate in academic activities/events that aim to help others | I am interested in using the skills and knowledge that I have acquired from this course to contribute to the public good | 1 | Demonstrate a desire to utilize engaged scholarship | Civic Engagement |
| 4 | I typically like to explore more than usual when I am learning something new that interests me | I want to continue to develop relevant skills that are related to this experience | 1 | Demonstrate a commitment to lifelong learning | Foundations and Skills for Lifelong Learning |
| 5 | I can clearly describe a real world problem related to this course to someone that knows little about the problem | I can clearly describe a real-world problem related to this course to someone that knows little about the problem | 2 | Clearly describe a real-world problem amenable to engaged scholarship | Critical Thinking |
| 6 | I have been introduced to more than one way to address real-world problem(s) related to this course | I have been introduced to more than one way to address real world problem(s) that my faculty member/professor brought up in this course. | 2 | Analyze literature (content/research methods) related to the problem | Critical Thinking |
| 7 | I feel confident in my ability to develop a logical, consistent approach to address a real world problem related to this course | I feel confident in my ability to develop a logical, consistent approach to address a real world problem related to this course | 2 | Formulate an inquiry approach driven by questions relevant to the problem | Critical Thinking |
| 8 | I can list many potential ethical issues for real world problems related to this course | I can list many potential ethical issues for real world problems related to this course | 2 | Recognize potential ethical issues related to addressing the problem | Ethical Reasoning |
| 9 | I can draw conclusions from data that has been collected | I can draw conclusions from data collected through this experience | 2 | Employ the selected inquiry approach<br>• Collect and analyze data<br>• Draw conclusions/ inferences (interpret) | Inquiry and Analysis |
| 10 | I am able to identify and apply information from this course to address and potentially improve real-world problem(s) | I am able to identify and apply information from this course to address and potentially improve real world problem(s) | 2 | Apply findings toward addressing the problem | Global Learning |

| Item No. | Pre-Experience Survey Items | Post-Experience Survey Items | SLO | VALUE Rubrics Items | VALUE Rubric |
|---|---|---|---|---|---|
| 11 | I am often told I listen to and respect the ideas of others | My classmates would say that I often listened to and respected the ideas of others | 3 | Participate in collaborative interactions; Support group Processes; Be attentive to the ideas of others | Teamwork |
| 12 | I am often told I offer relevant questions and comments within a group setting | My classmates would say that I was able to offer relevant question and comments within a group setting | 3 | Participate in collaborative interactions; Support group processes; Offer relevant questions and comment | Teamwork and Civic Engagement |
| 13 | I meet obligations for group assignments on a timely basis | I met obligations for group assignments on a timely basis | 3 | Support group processes; Meet obligations for group assignments on a timely basis | Teamwork |
| 14 | In the past, I have purposefully reflected on what I learned from problems I encountered during a learning experience | I purposefully reflected on what I learned from problems I encountered during this experience | 4 | Use structured reflection in assessing an engaged inquiry experience; Use reflection on the inquiry process to guide lifelong learning | Integrative Learning |
| 15 | In the past, I often reflected on what I have learned about myself from learning experiences | During this experience, I reflected on what I have learned about myself from this experience | 4 | Assess what they have learned about themselves as an individual (self-awareness) from experiences; Use reflection on the inquiry process to guide lifelong learning | Integrative Learning |
| 16 | I have thought about what it means to be a member of the broader community | During this experience, I thought about what it means to be a member of the broader community | 4 | Assess what they have learned about themselves as members of the broader community | Integrative Learning |

We also examined the extent to which longitudinal measurement invariance holds between the two surveys to examine growth in the Experience Learning SLOs over time. A key aspect of measuring students' growth over time hinges on the assumption that the instruments represent the same construct in the same metric over time (i.e., longitudinal measurement invariance). Findings from the analysis and the subsequent discussion will provide insight into the quality of the surveys as measures of the Experience Learning SLOs. Moreover, the results will describe the relationship between the Experience Learning SLOs and the caliber in which the surveys can indirectly measure student growth.

## Methods

### Study Sample

Two surveys were completed by different cohorts of students enrolled in courses that were redesigned to incorporate experiential learning as the main pedagogy and dispersed across five semesters (Fall 2017 through Fall 2019). The first survey, the pre-experience survey, was administered at the beginning of each semester, while the second survey, the post-experience survey, was administered at the end of each semester. All students were exposed to lifelong learning, application of knowledge and skills, collaboration with others, and structured reflection, regardless of the experiential learning course platform (e.g., internship, service learning, simulation/gaming/role-playing, study abroad, and undergraduate research); therefore, all items were deemed relevant to all survey respondents. Of the 990 students who completed at least one survey (78.7% response rate), 858 students completed the pre-experience student survey (68.2% response rate), 683 students completed the post-experience student survey (54.3% response rate), and 551 students completed both surveys (43.8% response rate).

### Measurement Instruments

**A key aspect of measuring students' growth over time hinges on the assumption that the instruments represent the same construct in the same metric over time (i.e., longitudinal measurement invariance).**

The pre-experience and post-experience surveys measure student perceptions of achievement of the SLOs as a form of indirect assessment. The surveys were developed to provide supportive evidence for the institutional continuous improvement initiative, Experience Learning, as required by the regional accreditation agency, the Southern Association of Colleges and Schools Commission on Colleges (SACSCOC). The pre-experience survey serves as a baseline measure, whereas the post-experience survey measures perceived learning after the experience and is compared to the pre-experience survey with the intention to measure growth. The SLOs represent the culmination of Experience Learning, a program that seeks to enhance student learning in four particular areas: lifelong learning, application of knowledge and skills, collaboration, and structured reflection. The four interrelated Experience Learning SLOs are derived from the Experience Learning mission statement, which calls for "enhancing opportunities for students to learn through actual involvement with problems and needs in the larger community," and Kolb's experiential learning cycle.

The Experience Learning SLOs highlight each stage of Kolb's experiential learning cycle. SLO 2 "Students will apply knowledge, values, and skills in solving real-world problems" aligns with the "concrete experience" and "active experimentation" stages in Kolb's cycle, while SLO 4 "Students will engage in structured reflection as part of the inquiry process" is aligned with the "reflective observation" and "abstract conceptualization" stages in Kolb's cycle. SLO 1, "Students will value the importance of engaged scholarship and lifelong learning" and SLO 3 "Students will work collaboratively with others" address the criticisms of Kolb's theory (e.g., Bergsteiner & Avery, 2014) to include collaboration and lifelong learning.

In addition to the connection with Kolb's experiential learning cycle, each SLO is accompanied by a set of benchmarks that are modified from the AAC&U VALUE rubrics. The benchmarks are used to operationalize the SLOs and guide the assessment measures. The surveys serve as an indirect measure of these learning outcomes such that each item is theoretically aligned with a benchmark from each SLO, as shown in Table 1. For example, the benchmark from SLO 1, related to lifelong learning, is constructed from language found in the "Foundations and Skills for Lifelong Learning" and "Civic Engagement" rubrics. Survey items are then reconstructed from the benchmarks to use simpler language, more suitable for undergraduate students to comprehend and answer. A similar method was used to reconstruct benchmark language for survey use with SLOs 2, 3, and 4.

After the instruments were constructed, a panel of 25 experiential learning campus experts examined the surveys for content validity. Three items from each survey were suggested to be dropped or consolidated because they were considered to be redundant and to limit survey fatigue. The surveys were pilot tested with a cohort of 80 students from five experiential learning courses. Initial results through an exploratory factor analysis revealed strong evidence that items in each survey factored onto the anticipated latent trait. Only one change occurred as a result of the pilot test; the rating scale was expanded from a 5-point Likert scale to a 7-point Likert scale (i.e., strongly disagree to strongly agree) to better examine variability between responses and to mitigate a ceiling effect. Table 2 provides descriptive statistics for the final items.

**This study examined construct validity and measurement invariance between the pre-experience and post-experience surveys.**

Table 2

| | Mean | SD | Mean | SD |
|---|---|---|---|---|
| SLO1: Lifelong learning | | | | |
| Item 1 | 5.561 | 1.12 | 6.088 | 1.14 |
| Item 2 | 6.424 | 0.80 | 6.441 | 0.87 |
| Item 3 | 5.166 | 1.27 | 6.201 | 1.09 |
| Item 4 | 6.128 | 0.94 | 6.167 | 1.13 |
| SLO 2: Solving real-world problems | | | | |
| Item 5 | 5.411 | 1.24 | 6.183 | 0.96 |
| Item 6 | 5.279 | 1.28 | 6.116 | 1.08 |
| Item 7 | 5.389 | 1.27 | 6.183 | 0.97 |
| Item 8 | 5.388 | 1.29 | 6.110 | 1.06 |
| Item 9 | 5.971 | 0.98 | 6.199 | 1.00 |
| Item 10 | 5.515 | 1.21 | 6.221 | 1.00 |
| SLO 3: Collaboration | | | | |
| Item 11 | 6.045 | 1.02 | 6.517 | 0.71 |
| Item 12 | 5.897 | 0.95 | 6.328 | 0.91 |
| Item 13 | 6.443 | 0.74 | 6.505 | 0.83 |
| SLO 4: Structured reflection | | | | |
| Item 14 | 5.930 | 0.94 | 6.136 | 1.02 |
| Item 15 | 5.980 | 0.97 | 5.990 | 1.22 |
| Item 16 | 5.892 | 1.10 | 6.044 | 1.18 |

## Analysis

This study examined construct validity and measurement invariance between the pre-experience and post-experience surveys. Both surveys were designed to measure the same SLOs, and therefore, the same latent factors; however, items between the pre-experience and post-experience surveys were worded differently to better articulate the students' experiences relevant to the timing of the administered surveys. Empirical differences between corresponding items will need to be examined to justify that both surveys are measuring the same latent factors (i.e., measurement invariance). While rare, measurement invariance has been examined across surveys with altered, but theoretically aligned, items to better capture experiences across time (e.g., Vianello et al., 2018; Wang et al., 2017). Confirmation of the relationship between the survey items and the Experience Learning SLOs (i.e., the latent factors) will first be examined through a confirmatory factor analysis, as will the relationship between each of the Experience Learning SLOs. A confirmatory factor analysis provides a psychometric evaluation of the latent structure of the measurement model (Brown & Moore, 2012). A variety of fit indices (i.e., scaled-MLR chi-squared test, CFI, TLI, RMSEA, and SRMR) were used to evaluate the measurement model. Specifically, we assessed model fit using the following guidelines (Brown, 2015; Gana & Broc, 2019): comparative fit index (CFI) and Tucker-Lewis index (TLI) ≥ .90 for adequate fit and ≥ .95 for good fit; root mean square error of approximation (RMSEA) ≤ .08 for adequate fit and ≤ .05 for good fit; and standardized root mean square residual (SRMR) ≤ .08 for adequate fit and ≤ .06 for good fit. Examination of correlation between latent factors within and between occasions provided insight into relationships between the SLOs.

Next, we examined longitudinal measurement invariance between corresponding indicators and latent constructs in the pre-experience and post-experience surveys. Longitudinal measurement invariance demonstrates that the same indicators consistently measure the same construct over multiple occasions (Meredith, 1993; Millsap & Olivera-Aguilar, 2012). Model fit statistics were used to analyze invariance between models. Measurement invariance between nested models is typically assessed through a chi-squared difference test; however, research has shown that the chi-squared test is sample size dependent (Brannick, 1995; Cheung & Rensvold 2002; Kelloway, 1995). Chen (2007) argued that changes in fit indices that are independent of sample size (e.g., CFI and RMSEA) between nested models should be used to test measurement invariance rather than the chi-squared difference test. Specifically, $\Delta$CFI ≥ -.010 and $\Delta$RMSEA ≥ .015 indicate non-invariance (Chen, 2007); however, values of $\Delta$CFI should take precedent as model complexity affects $\Delta$RMSEA.

## Results

A confirmatory factor model measuring SLO 1 "lifelong learning" (four manifest variables), SLO 2 "application of knowledge and skills" (six manifest variables), SLO 3 "collaboration" (three manifest variables), and SLO 4 "structured reflection" (three manifest variables) was examined using Mplus v. 8.1 (Muthén & Muthén, 1998-2017). Linearity was assessed by bivariate scatterplots and correlations among indicators. Violations of univariate and multivariate normality were prevalent in the sample. The violation of normality was addressed using a robust estimation method, robust maximum likelihood (MLR), as recommended by Lei and Shiverdecker (2019).

First, a configural invariance model was specified to examine the factor structure (i.e., whether the item measure the experience learning SLOs) separately for the pre- and post-experience surveys. This model included four correlated factors across two occasions (i.e., pre-experience survey and post-experience survey), such that eight latent factors were allowed to correlate without constraint. Correlated factors were estimated simultaneously with factor means fixed to 0 and factor variances fixed to 1 for identification. Factor loadings and intercepts were allowed to freely estimate. Residual covariances between the corresponding indicators across occasions were also freely estimated. Figure 1 outlines the standardized parameters and identification decisions for the model. The model displayed good fit, $\chi^2_{MLR}(420) = 843.65$, $p < .001$, CFI = .957, TLI = .949, RMSEA = .032, 90% $CI_{RMSEA}$ (.029, .035), SRMR = .048. Statistically significant correlations at each occasion were found between all latent factors, as shown in Figure 1 and Table 3. Specifically, correlations between the "lifelong learning" latent factors were .354, "application of knowledge and skills" factors were .272 "collaboration" factors were .293, and "structured reflection" factors were .339. Parameter constraints were then applied to subsequent models (i.e., metric and scalar invariance models) to examine measurement invariance between two occasions.

**Examination of correlation between latent factors within and between occasions provided insight into relationships between the SLOs.**

Unstandardized factor loadings between corresponding indicators were constrained to be equal across both occasions to examine metric invariance. Metric measurement invariance is achieved when factor loadings are equal between occasions indicating that the factors have the same meaning across occasions. Factor variance was fixed to 1 for the pre-experience factors but allowed to estimate freely for the post-experience factors. Intercepts and residual variances were allowed to vary across occasions. Factor and residual covariances continued to be estimated freely between corresponding factors and indicators, respectively. Although, the metric model fit worse than the configural model[1] , $\Delta\chi^2_{MLR}(12) = 124.67$, $p < .001$, $\Delta$CFI = -0.015, $\Delta$RMSEA = 0.005, $\Delta$RMSEA was acceptable; moreover, the model fit indices for the full metric model were within acceptable ranges, CFI = .942, TLI = .933, RMSEA = .037, 90% $CI_{RMSEA}$ (.034, .040), SRMR = .072. The Lagrange Multiplier Test suggested item 9 as a source

[1] It should be noted that the metric model will inherently fit worse than the configural model; instead, we measured the extent to which the metric model fits worse than the configural model with the desire that the two models do not statistically differ from one another.

Figure 1.
*Configural invariance model with fixed factor variance and standardized parameters*

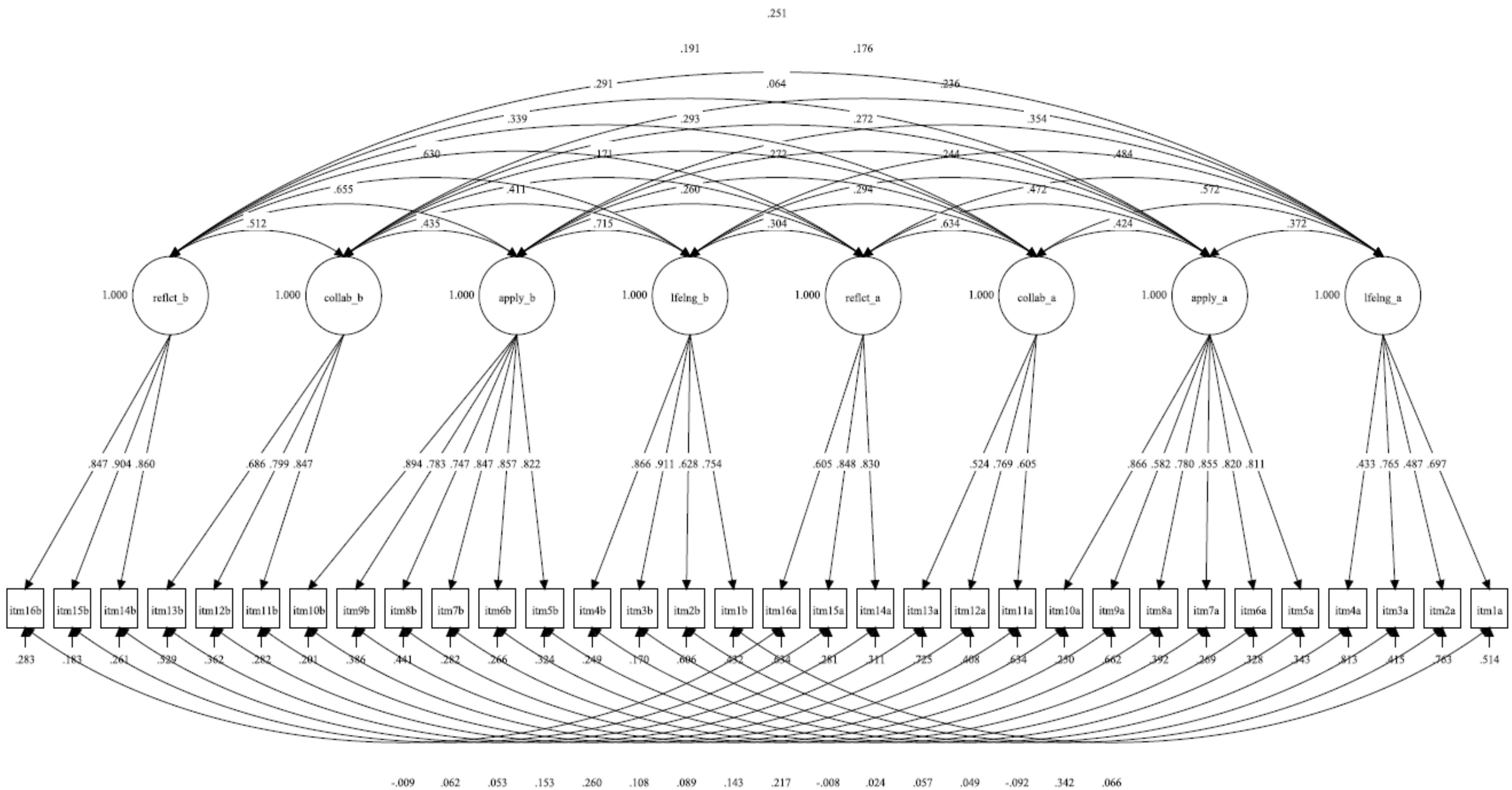


Table 3
*Standardized factor correlations in configural model*

| | Pre-Experience Survey | | | | Post-Experience Survey | | | |
|---|---|---|---|---|---|---|---|---|
| | SLO 1 | SLO 2 | SLO 3 | SLO 4 | SLO 1 | SLO 2 | SLO 3 | SLO 4 |
| Pre-Experience Survey | | | | | | | | |
| SLO 1: Lifelong learning | 1.00*** | | | | | | | |
| SLO 2: Application of knowledge & skills | .372*** | 1.00*** | | | | | | |
| SLO 3: Collaboration | .572*** | .424*** | 1.00*** | | | | | |
| SLO 4: Structured reflection | .484*** | .472*** | .634*** | 1.00*** | | | | |
| Post-Experience Survey | | | | | | | | |
| SLO 1: Lifelong learning | .354*** | .244*** | .294*** | .304*** | 1.00*** | | | |
| SLO 2: Application of knowledge & skills | .236*** | .272*** | .272*** | .260*** | .715*** | 1.00*** | | |
| SLO 3: Collaboration | .176** | .064 | .293*** | .171** | .411*** | .435*** | 1.00*** | |
| SLO 4: Structured reflection | .251*** | .191*** | .291*** | .339*** | .630*** | .655*** | .512*** | 1.00*** |

***$p < .001$, **$p < .01$

**Differences between corresponding latent means between the pre-experience and post-experience surveys indicate significant student growth when post-experience survey latent means were allowed to vary, except in SLO 1.**

of misfit. Modification to the model allowed factor loadings for item 9 to freely vary across occasions to produce a partial metric model[2]. The partial metric model was compared to the configural model and found to be invariant, $\Delta\chi^2_{MLR}(11) = 78.61, p < .001, \Delta CFI = -0.009, \Delta RMSEA = 0.003$. Table 4 outlines the model comparisons between the full and partial metric model to the configural model. Partial metric invariance held, which shows that the same latent factors were being measured at both occasions when item 9 was allowed to vary. Next, intercepts and factor loadings between corresponding indicators were constrained equal across both occasions to examine scalar invariance, except for item 9 which was allowed to vary freely. Factor variance and means were fixed to 1 and 0, respectively, for the pre-experience factors to allow for identification. Factor variance and means for the post-experience factors were permitted to freely estimate, while residual variances were still allowed to vary across occasions. Factor and residual covariances continued to be estimated freely between corresponding factors and indicators, respectively. The full scalar invariance model fit worse than the partial metric invariance model , $\Delta\chi^2_{MLR}(11) = 435.23, p < .001, \Delta CFI = -0.036, \Delta RMSEA = 0.009$, particularly in regards to the ΔCFI. The Lagrange Multiplier Test suggested that the intercept of item 3 to be a source of misfit. Modification to the model allowed the intercepts between occasions to freely vary. The partial scalar model was compared to the partial metric model and found to still be a worse fit, $\Delta\chi^2_{MLR}$ (10) = 186.97, $p < .001$, ΔCFI = -0.015, ΔRMSEA = 0.004. The Lagrange Multiplier Test suggested that the intercept of item 1 to be the next largest remaining source of misfit and was allowed to freely vary. After doing so, the new partial scalar model was found to fit the model similarly to the partial metric model, $\Delta\chi^2_{MLR}$ (9) = 96.57, $p < .001$, ΔCFI = -0.008, ΔRMSEA = 0.002. Table 4 outlines the model comparisons between the full scalar model and each modified scalar model to the partial metric model.

Scalar invariance holds across 13 of the 16 items which indicates that the observed differences in these indicator means between the pre-experience and post-experience surveys are due to factor mean differences only; however, items 1 and 3 had a lower expected indicator response at the same absolute level of the "lifelong learning" factor in the pre-experience survey than the post-experience survey. Item 9 had higher expected responses in the pre-experience survey than the post-experience survey in the "application of knowledge and skills" latent factor. Differences in indicator intercepts between occasions suggests that precautions should be considered when comparing factor mean differences across occasions for the "lifelong learning" latent factor. Differences between corresponding latent means between the pre-experience and post-experience surveys indicate significant student growth when post-experience survey latent means were allowed to vary, except in SLO 1, $\Delta\gamma = .087$, S.E. = .066,

[2] It should be noted that the scalar model will inherently fit worse than the metric model; instead, we measured the extent to which the scalar model fits worse than the metric model with the desire that the two models do not statistically differ from one another.

Table 4
*Model comparisons for validation of longitudinal measurement invariance*

| Model | $\chi^2_{MLR}$ | *df* | Scaling | CFI | TLI | RMSEA | SRMR | BIC | $\Delta\chi^2_{MLR}$ | $\Delta df$ | ΔCFI | ΔRMSEA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Configural Model | 843.65 | 420 | 1.272 | .957 | .949 | .032 | .048 | 59054.5 | | | | |
| Metric Models[a] | | | | | | | | | | | | |
| Full metric | 1006.92 | 432 | 1.286 | .942 | .933 | .037 | .072 | 59193.4 | 124.67*** | 12 | -.015 | .005 |
| Partial metric (item 9) | 944.05 | 431 | 1.285 | .948 | .940 | .035 | .059 | 59118.5 | 78.61*** | 11 | -.009 | .003 |
| Scalar Models[b] | | | | | | | | | | | | |
| Full scalar (item 9) | 1303.71 | 442 | 1.279 | .912 | .902 | .044 | .079 | 59603.8 | 435.23*** | 11 | -.036 | .009 |
| Partial scalar (items 9, 3) | 1103.23 | 441 | 1.280 | .933 | .924 | .039 | .069 | 59249.3 | 186.97*** | 10 | -.015 | .004 |
| Partial scalar (items 9, 3, 1) | 1029.13 | 440 | 1.281 | .940 | .933 | .037 | .064 | 59161.7 | 96.57*** | 9 | -.008 | .002 |

*Note*. $\chi^2_{MLR}$, scaled robust maximum likelihood chi-square test; Scaling, scaling correction factor for chi-squared test for MLR estimator; CFI, comparative fit index; TLI, Tucker-Lewis index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual; BIC, Bayesian information criterion; $\Delta\chi^2_{MLR}$, scaled chi-squared difference test for MLR estimator; ΔCFI, change in CFI; ΔRMESA, change in RMESA.
[a]factor loadings in items listed in parenthesis are free to vary between occasions.
[b]intercepts in items listed in parenthesis are free to vary between occasions.

$p$ = .188. There exists a significant difference between the latent means of pre-post SLO 2, $\Delta\gamma$ = .742, S.E. = .041, $p$ < .001; pre-post SLO 3, $\Delta\gamma$ = .611, S.E. = .054, $p$ < .001; and pre-post SLO 4, $\Delta\gamma$ = .163, S.E. = .058, $p$ = .005.

In addition to assessing the construct validity and longitudinal measurement invariance, we also calculated the average variance extracted (AVE) from each factor and the internal consistency reliability for each factor. The AVE is useful in assessing the presence of convergent and divergent validity in the model. Convergent and divergent validity help supplement the claim for construct validity by determining the degree to which the factors are related and unrelated to one another, respectively. Cheung and Wang (2017) recommended that convergent validity be established provided that AVE and standardized factor loadings of all items are not considerably less than .500. The AVE was greater than .500 for six out of the eight factors. Two factors in the pre-experience survey were slightly lower than this threshold: "lifelong learning" AVE = .430 and "collaboration" factor was .428. The AVE values for each factor are presented in Table 4. Additionally, standardized factor loadings for all but two items were above the .500 threshold: item 2 ($\lambda$ = .433) and item 4 ($\lambda$ = .487) in the pre-experience survey. These results suggest strong convergent validity for the post-experience survey. While the instruments overall exhibited strong factor loadings and AVE, the "lifelong learning" and "collaboration" factors in the pre-experience survey revealed values that were slightly below the threshold to conclude convergent validity for the pre-experience survey. Divergent validity is concluded if the correlation between any two factors is not considerably greater than .700 (Cheung & Wang, 2017). Table 3 shows correlations between factors to be below the threshold with the exception of the "lifelong learning" and "application of knowledge and skills" factors in the post-experience survey, which has a correlation of .715. Thus, divergent validity is evident within the model.

Internal consistency was measured using McDonald's (1999) $\omega_t$ coefficient. The ωt coefficient, along with a confidence interval, provides a better reflection of variability within the point estimation process than the τ-equivalent model (i.e., Cronbach's alpha); that is, $\omega_t$ reflects a more accurate degree of confidence in the internal consistency of the factor (Dunn et al., 2013). Confidence intervals (95%) were obtained through the bias-corrected and accelerated bootstrapping technique in the MBESS package 4.6.0 in R 3.5.1. Table 5 outlines the $\omega_t$ internal consistency coefficient for the model factors. Factors display strong internal consistency with all but one factor above $\omega_t$ = .700. The "collaboration" factor in the pre-experience survey showed moderate levels of internal consistency, $\omega_t$ = .682, 95% CI (.632, .722).

**These results suggest strong convergent validity for the post-experience survey.**

Table 5
*Average variance extracted (AVE); composite reliability (CR); and ωt internal consistency estimates, standard errors, and confidence intervals for model factor*

| | | | | | $\omega_t$ CI 95% | |
|---|---|---|---|---|---|---|
| | AVE | CR | $\omega_t$ | S.E. | Lower | Upper |
| Pre-Experience Survey | | | | | | |
| SLO 1: Lifelong learning | .430 | .694 | .730 | .018 | .691 | .761 |
| SLO 2: Solving real-world problems | .644 | .908 | .913 | .005 | .903 | .922 |
| SLO 3: Collaboration | .428 | .671 | .682 | .023 | .632 | .722 |
| SLO 4: Structured reflection | .571 | .810 | .794 | .015 | .759 | .821 |
| Post-Experience Survey | | | | | | |
| SLO 1: Lifelong learning | .644 | .873 | .883 | .013 | .856 | .908 |
| SLO 2: Solving real-world problems | .681 | .928 | .927 | .008 | .912 | .944 |
| SLO 3: Collaboration | .602 | .823 | .815 | .027 | .758 | .866 |
| SLO 4: Structured reflection | .761 | .904 | .905 | .010 | .883 | .921 |

*Note*: Bias-corrected and accelerated bootstrap confidence interval used with 1000 bootstrap iterations.

## Discussion

**Our findings provide support for the instruments to measure the desired Experience Learning SLOs, and importantly that this structure is mostly invariant over time (i.e., equality of factor structure, factor loadings, and intercepts).**

Kolb's experiential learning model focuses on the active (application of knowledge and skills) and transformative (structured reflection) aspects of learning. As part of the survey construction, items related to these facets of the model were carefully constructed based on the language used in the critical thinking, ethical reasoning, inquiry and analysis, global learning, and integrative learning AAC&U VALUE rubrics and the principles laid out in Kolb's experiential learning theory (Kolb, 1984). We set out to develop pre-experience and post-experience surveys to measure students' attainment of specific Experience Learning SLOs. A key aspect of measuring students' growth over time hinges on the assumption of longitudinal measurement invariance (i.e., the instruments need to represent the same construct in the same metric over time). Thus, the purpose of the present study was to test the construct validity and longitudinal measurement invariance of the pre-experience and post-experience surveys.

Our findings provide support for the instruments to measure the desired Experience Learning SLOs, and importantly that this structure is mostly invariant over time (i.e., equality of factor structure, factor loadings, and intercepts). Our results overall indicate good psychometric properties of the instruments to measure SLOs in experiential learning courses and demonstrate that meaningful comparisons can be made to assess students' growth on these learning outcomes. Although we did not achieve full metric or scalar invariance, partial scalar invariance was achieved for the vast majority of items (i.e., 13 out of 16 items). According to Sass and Schmitt (2013), "if the number non-invariant items is small compared to total number of items, or the overall amounts of non-invariance is small, the latent factor means used for group comparisons should not be drastically impacted" (p. 324). Knowing which specific items are non-invariant, as evidenced in our sample, researchers and assessment professional applying these instruments can modify the problematic items in the future with the aim of achieving full metric and scalar invariance.

Our results demonstrate at least partial scalar longitudinal measurement invariance between the two instruments over the course of a semester suggesting that the surveys measured the same SLOs at different occasions. This also implies that the mean differences (or student growth) can be interpreted as true changes in students' attainment of the SLOs. These findings have significant implications for assessment professionals examining learning outcomes in experiential learning settings. In particular, attainment and growth in learning outcomes commonly found in experiential learning contexts can be measured, at least from the students' perspective. These results, in conjunction with valid direct measures, provide strong evidence of student attainment over the course of an experiential course or other learning experience (Banta & Palomba, 2015).

Convergent validity occurs when the degree to which items under the same construct are related to one another. Measures of AVE and standardized factor loadings greater than

.500 were used to support the claim for convergent validity (Cheung & Wang, 2017). While two factors in the pre-experience survey (i.e., "lifelong learning" and "collaboration") failed to meet the AVE threshold and two items in the pre-experience "lifelong learning" factor failed to meet the standardized factor loading threshold to support convergent validity, Fornell and Larcker (1981) argue that AVE values greater than .400 are adequate to support convergent validity, provided that composite reliability (CR) values are at least .600. The two factors whose AVE < .500 were still above the .400 threshold and had CR values exceeding .600 (see Table 4); thus, evidence for convergent validity is arguable. The internal consistency coefficient ($\omega_t$) also supports that the items measure their respective factors. The degree to which the surveys discriminate between the constructs within the model (i.e., divergent validity) is supported through examination of the correlations between factors. Cheung and Wang (2017) posited that divergent validity is evident when any correlation between two factors is not significantly greater than .700. The factor correlation between post-experience "lifelong learning" and "applications of knowledge and skills" did exceed .700, but not substantially so (i.e., $r$ = .715); moreover, Kline (2010) suggested a more liberal cut-off of .850, which may be more reasonable considering the theorized relationship between all factors (Kolb & Kolb, 2011).

**When the surveys are used jointly, they can quantify the practical impact of the experiential learning environment from the learners' perspective.**

Correlations between corresponding factors (e.g., pre-experience survey "lifelong learning" and post-experience survey "lifelong learning") exhibited the highest coefficients compared to non-corresponding factors between occasions; that is, the correlation between the pre-experience survey "lifelong learning" factor was strongest with the post-experience "lifelong learning" factor than any other post-experience factor. The same results were found among all corresponding factors between surveys. These results indicate that the differences between occasions is mostly to do with the change across the same SLO. Additionally, strong internal consistency reliability within each factor was evident and suggests that items were likely reliably measuring their latent factors.

The pattern of correlations among the factors both within and between occasions indicates that our SLOs exhibit non-cyclical relationships between one another. Warren et al. (1995) suggested that experiential learning objectives are not achieved in any particular order or pattern. Kolb's (1984) model shows that learning in an experiential context is patterned such that lifelong learning would follow reflection, reflection would follow applications, and so forth. Although this relationship still holds in our model, one SLO does not necessarily precede a specific SLO but rather, any factor could precede or succeed another and provide substantive evidence for student growth across all SLOs.

The pre-experience and post-experience surveys are intended to be used to understand student attainment of learning outcomes commonly found in an experiential learning opportunity, such as study abroad, service-learning, internships, undergraduate research, and others. Intended users should implement the surveys before and immediately after an experiential activity to better measure student growth as a result of the real-world learning experience. Survey results can be examined by individual items and many can be compared across multiple occasions with exceptions to a few items. In particular, items in SLO 3 "collaboration" and SLO 4 "structured reflection" are invariant over time. Users are encouraged to replicate the study with different samples and adjusted language to validate and improve our findings. When the surveys are used jointly, they can quantify the practical impact of the experiential learning environment from the learners' perspective. The learners' perspective is key when they are the most valuable and reliable source to understand the role of the experiential learning such as during internships, externships, practicums, and study abroad experiences. In particular, the surveys serve as strong indirect measures of the Experience Learning SLOs and Kolb's experiential learning cycle.

## References

Abdulwahed, M., & Nagy, Z.K. (2009). Applying Kolb's experiential learning cycle for laboratory education. *Journal of Engineering Education*, *98*(3), 283-294. https://doi.org/10.1002/j.2168-9830.2009.tb01025.x

Baker, M.A., Robinson, J.S., & Kolb, D.A. (2012). Aligning Kolb's experiential learning theory with comprehensive agricultural education model. *Journal of Agricultural Education*, *53*(4), 1-16. https://doi.org/10.5032/jae.2012.04001

Banta, T.W., & Palomba, C.A. (2015). *Assessment Essentials: Planning, implementing, and improving assessment in higher education* (2nd ed.). Jossey-Boss.

Bergsteiner, H., & Avery, G.C. (2014). The twin-cycle experiential learning model: Reconceptualising Kolb's theory. *Studies in Continuing Education*, *36*(3), 257-274. https://doi.org/10.1080/0158037X.2014.904782

Bergsteiner, H., Avery, G.C., & Neumann, R. (2010). Kolb's experiential learning model: critique from a modelling perspective. *Studies in Continuing Education*, *32*(1), 29-46. https://doi.org/10.1080/01580370903534355

Boyatzis, R.E., Cowen, S.S., & Kolb, D.A. (1995). *Innovation in professional education: Steps on a journey from teaching to learning*. Jossey-Boss.

Brannick, M.T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, *16*(3), 201-213. https://doi.org/10.1002/job.4030160303

Brown, T.A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Brown, T.A., & Moore, M.T. (2012). Confirmatory factor analysis. In R.H. Hoyle. (Ed.), *Handbook of structural equation modeling*, (pp. 361-379). Guilford Press.

Chapman, S., McPhee, P., & Proudman, B. (1992). What is experiential education? *Journal of Experiential Education*, *15*(2), 16-23. https://doi.org/10.1177/105382599201500203

Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Cheung, G.W., & Wang, C. (2017). Current approaches for assessing convergent and discriminant validity with SEM: Issues and solutions. [Abstract]. *Academy of Management*, *1*. https://doi.org/10.5465/ AMBPP.2017.12706abstract

Dewey, J. (1938). *Experience and education*. MacMillan.

Dunn, T.J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *The British Journal of Psychology*, *105*(3), 399-412. https://doi.org/10.1111/bjop.12046

Finley, A. (2011). How reliable are the VALUE rubrics? *Peer Review: Emerging trends and key debates in undergraduate education*, *14*(4), 31-34. https://www.aacu.org/sites/default/files/files/peerreview/PRfa11wi12.pdf

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement errors. *Journal of Marketing Research*, *18*(1), 39-80. https://doi.org/10.2307/3151312

Gana, K., & Broc, G. (2019). *Structural equation modeling with Lavaan*. Wiley.

Healey, M., & Jenkins, A. (2000). Kolb's experiential learning theory and its application in geography in higher education. *Journal of Geography*, *99*(5), 185-195. https://doi.org/10.1080/00221340008978967.

Kayes, A.B., Kayes, C., & Kolb, D.A. (2005). Experiential learning in teams. *Simulation & Gaming*, *36*(3), 330-354. https://doi.org/10.1177/1046878105279012

Kelloway, E.K. (1995). Structural equation modelling in perspective. *Journal of Organizational Behavior*, *16*(3), 215-224. https://doi.org/10.1002/job.4030160304

Kline, R.B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). Guilford Publications.

Kolb, D.A. (1984). *Experiential learning: Experience as a course of learning and development*. Prentice-Hall.

Kolb, Y.A., & Kolb, D.A. (2011). Experiential learning theory: A dynamic, holistic approach to management learning, education and development. In S.J. Armstrong, & C.V. Fukami (Eds.), *The SAGE handbook of management learning, education and development*, (pp. 42-68). SAGE Publications. https://doi.org/10.4135/9780857021038.n3

Lei, P., & Shiverdecker, L.K. (2019). Performance of estimators for confirmatory factor analysis of ordinal variables with missing data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(4), 584-601. https://doi.org/10.1080/10705511.2019.1680292

Manolis, C., Burns, D.J., Assudani, R., & Chinta, R. (2013). Assessing experiential learning styles: A methodological reconstruction and validation of the Kolb learning style inventory. *Learning and Individual Differences*, *23*(1), 44-52. https://doi.org/10.1016/j.lindif.2012.10.009

McDonald, R.P. (1999). Test theory: A unified treatment. Lawrence Erlbaum Associates. https://doi org/10.1111/j.2044-8317.1981.tb00621.x

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543. https://doi.org/10.1007/BF02294825

Miettinen, R., (2010). The concept of experiential learning and John Dewey's theory of reflective thought and action. *International Journal of Lifelong Education*, *19*(1), 54-72. https://doi.org/10.1080/026013700293458

Millsap, R.E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R.H. Hoyle. (Ed.), *Handbook of structural equation modeling*, (pp. 380-392). Guilford Press.

Muthén, L.K., & Muthén, B.O. (1998-2017). *Mplus user's guide* (8th ed.). Author.

Poore, J.A., Cullen, D.L., & Schaar, G.L. (2014). Simulation-based interprofessional education guided by Kolb's experiential learning theory. *Clinical Simulation in Nursing*, *10*(5), 241-247. https://doi.org/10.1016/j.ecns.2014.01.004

Petkus, E. (2000). A theoretical and practical framework for service-learning in marketing: Kolb's experiential learning cycle. *Journal of Marketing Education*, *22*(1), 64-70. https://doi.org/10.1177/0273475300221008

Reshmad'sa, L., & Vijayakumari, S.N. (2017). Effect of Kolb's experiential learning strategy on enhancing pedagogical skills of pre-service teachers in secondary school level. *Journal of School Educational Technology*, *13*(2), 1-6. https://doi.org/10.26634/jsch.13.2.13825

Rhodes, T., & Finley, A. (2013). *Using the VALUE rubrics for improvement of learning and authentic assessment*. Association of American Colleges and Universities. https://www.aacu.org/publications/using-value-rubrics

Rhodes, T. (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics*. Association of American Colleges and Universities. https://www.aacu.org/publications-research/publications/assessing-outcomes-and-improving-achievement-tips-and-tools-using

Russell-Bowie, D. (2013). Mission impossible or possible mission? Changing confidence and attitudes of primary preservice music education students using Kolb's experiential learning theory. *Australian Journal of Music Education*, *2*, 46-63. https://files.eric.ed.gov/fulltext/EJ1061833.pdf

Sass, D.A., & Schmitt, T.A. (2013). Testing measurement and structural invariance: Implications for practice. In T. Teo (Ed.), *Handbook of quantitative methods for educational research*, (pp. 315-345). Sense Publishers.

Seaman, J., Brown, M., & Quay, J. (2017). The evolution of experiential learning theory: Tracing lines of research in the JEE. *Journal of Experiential Education*, *40*(4), https://doi.org/10.1177/1053825916689268

Smith, K.L., & Rayfield, J. (2017). Student teaching changed me: A look at Kolb's learning style inventory scores before and after the student teaching experience. *Journal of Agricultural Education*, *58*(1), 102-117. https://doi.org/10.5032/jae.2017.01102

Vianello, M., Dalla Rosa, A., Anselmi, P., & Galliani, E.M. (2018). Validity and measurement invariance of the unified multidimensional calling scale (UMCS): A three-wave survey study. *PLoS ONE*, *13*(12). https://doi.org/10.1371/journal.pone.0209348

Wang, M., Chen, R.C., Usinger, D.S., & Reeve, B.B. (2017). Evaluating measurement invariance across assessment modes of phone interview and computer self-administrated survey for the PROMIS measures in a population-based cohort of localized prostate cancer survivors. *Quality of Life Research*, *26*, 2973-2985. https://doi.org/10.1007/s11136-017-1640-3

Warren, K., Sakofs, M., & Hunt, J.S. (1995). *The theory of experiential education*. Kendall Hunt.

Widaman, K.F., Ferrer, E., & Conger, R.D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10-18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

## *Abstract*

Identifying evidence-informed programming (e.g., strategies, activities, pedagogies) facilitates both the intentional offering of programming that should "work" and the use of the outcomes assessment process to evaluate program effectiveness. Evidence-informed programming is more efficient than unsupported programming because the programming is more likely to improve learning and development. Thus, faculty and student affairs professionals require fewer iterations of the assessment cycle to inform programming changes in order to achieve desired outcomes. To help locate evidence-informed programming, we describe systematic review repositories (e.g., *Campbell Collaboration*, *What Works Clearinghouse*) that synthesize high-quality research to identify "what works". We share a tool we created that organizes relevant systematic review repositories and other collections of evidence of effectiveness, providing numerous examples of evidence-informed programming pertinent to higher education. These resources aid faculty and student affairs professionals in achieving their ethical obligation to engage students in effective learning and development experiences.



***AUTHORS***

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

Holly A. Buchanan,
DNP, ANP-BC
*James Madison University*


# A More Efficient Path to Learning Improvement: Using Repositories of Effectiveness Studies to Guide Evidence-Informed Programming

Institutions of higher education are expected to gather and use outcomes data to improve student learning and development (Jankowski et al., 2018; U.S. Department of Education, 2006). It is hoped that learning improvement will be evidenced by employing an iterative process of building educational programming, implementing programming, assessing outcomes, and using results to make changes to programming (Fulcher et al., 2014). Changes to pedagogy, activities, or educational content are common strategies employed in the hope of creating more effective programming and in turn improving student learning and development (Jankowski et al., 2018).


***CORRESPONDENCE***

***Email***
finneysj@jmu.edu


We endorse this improvement science (Bryk et al., 2015; Lewis, 2015) approach promoted in higher education (Fulcher et al., 2014). However, echoing others, we call for a process of improvement that begins with programming that *should* be effective based on research (Kerr et al., 2020; Pope et al., 2019; Pope et al., in press; Slavin, 2020; Smith & Finney, 2020; Wight et al., 2016). Our recommendation is informed by concerns of inefficiency, engagement, and ethics.

There is great *inefficiency* in the outcomes assessment process when programming is either built from "scratch" based on good intentions, assumptions, and hunches, or programming is based on ineffective strategies. Depending on the initial quality of programming, major changes may be required for programming to be effective. Although outcomes data can indicate that students did not achieve expectations, outcomes data do not suggest changes to programming. Moreover, faculty and student affairs professionals may not know what programming is necessary to achieve intended outcomes (e.g., Brownell & Tanner, 2012; Hutchings, 2010; Jones, 2008). Thus, changes to programming may be exploratory in nature (e.g., "Let's try this approach"), based on tradition (e.g., "This is what I experienced as a student"), or avoided. Moreover, changes may be minor. Thus, it may take years to implement effective programming that results in intended outcomes. An analogy offered by Eubanks (2017) makes this point clearly: "Imagine if each town and village were required to research and produce its own drugs, and ignore large-scale science-based medical research. That is our current situation with respect to assessment" (p. 11). Instead, we recommend offering evidence-informed programming that is supported by research and instills greater confidence that students' knowledge, attitudes, and skills will be impacted in desired ways. Subsequent outcomes assessment is still needed to formally examine the effectiveness of the programming. In this context, the outcomes assessment process is used in a confirmatory way to assess if the research-informed (thus, should-be-effective) programming is actually effective in the specific institutional context. This confirmatory approach should be much more efficient than the exploratory approach. Less time and resources are needed to improve the programming because it is evidence informed and more likely to be effective. In turn, fewer iterations of the assessment cycle are required to inform changes to programming to obtain the desired impact on student learning and development.

Implementing programming with no prior effectiveness information requires consistent *engagement* by faculty and staff to assess outcomes and use results for improvement. Yet, many student affairs professionals and faculty do not consistently engage in outcomes assessment (e.g., Bresciani, 2010; Hutchings, 2010). If assessment data are gathered, there are few examples of iterative, continued improvement efforts until intended outcomes are achieved (Jankowski et al., 2018). Adopting new or unsupported programming requires a great deal of active, thoughtful engagement in assessment and improvement activities that may be perceived as demanding, unrealistic, and unsustainable by faculty and staff. There are tremendous challenges to building and improving new interventions, programming, or pedagogy (Gitlin, 2013) and faculty and staff may not be interested in these innovation activities or able to assume the trajectory of this work (e.g., Brownwell & Tanner, 2012). Engaging in empirical study of the effectiveness of new programming versus the use of pre-existing research to inform programming is much like the distinction between the Scholarship of Teaching and Learning (SoTL) and Scholarly Teaching in higher education. SoTL is the systematic study (i.e., intentional, planned, occurring over time) of teaching and learning using an established scholarly process to understand what maximizes learning, resulting in findings publicly shared for use (Potter & Kustra, 2011). Scholarly Teaching is consuming evidence on what effectively fosters learning (often drawn from SoTL literature) and using that evidence to inform practice. We recommend an approach similar to Scholarly Teaching. We recommend faculty and staff identify and offer existing evidence-informed programming to reduce the burden associated with building novel programming or continuously improving less effective approaches to learning and development.

**We recommend faculty and staff identify and offer existing evidence-informed programming to reduce the burden associated with building novel programming or continuously improving less effective approaches to learning and development.**

Student affairs professionals and faculty have an *ethical* responsibility to offer effective learning and development experiences to students (Finney & Horst, 2019b; Svinicki & McKeachie, 2013). The implementation of unassessed programming that is ineffective or harmful to students is unacceptable. Although we hope our colleagues in higher education are continuously assessing and improving their programming, we are realistic that many are not. Thus, the implementation of unsupported programming under the assumption that it will be improved until effective is unethical if programming is rarely or never assessed and improved. Fortunately, three sets of professional standards in higher education (Assessment Skills and Knowledge Standards, ACPA-NASPA Professional Competencies, CAS Standards) call for programming to be intentionally built using current research that indicates what effectively impacts particular outcomes (Finney & Horst, 2019a).[1] Moreover, Horst and Prendergast's (2020) *Assessment Skills Framework* outlined the knowledge, skills, and attitudes necessary for assessment work. The ability to articulate program theory, create a logic model, and identify literature domains to inform program development were considered necessary for high-quality assessment practice.

These standards and frameworks echo previous statements regarding the ethics and expectations of implementing research-informed programming in higher education.

> Any student affairs professional not reading the literature, not becoming knowledgeable of research and theory, is not acting ethically. Students have a right to expect that student affairs professionals are knowledgeable of appropriate theories, current research, and proven best practices. (Carpenter, 2001, p. 311)

**There is an expectation that student affairs professionals and faculty can answer a basic question: "What evidence suggests the intended programming should be effective?**

Likewise, for faculty the "ethical principle of competence" emphasizes that "both departments and instructors have the obligation to place competent teachers in classrooms and hold them accountable for providing undergraduates with a quality educational experience" (Komarraju & Handelsman, 2012, p. 192). Meeting the ethical obligation to provide high-quality, impactful opportunities to learn requires understanding what is effective. "An effective curriculum uses research-informed strategies to help students learn and succeed" (Suskie, 2018, p. 69). Thus, when Banta and Blaich (2011) described the outcomes assessment process, they noted the importance of understanding what programming should be effective when trying to "close the loop" (bold emphasis added):

> An internally driven, formative approach to assessment is based on the belief that a key factor inhibiting improvements in student learning or allowing students to graduate without learning enough is that faculty and staff who deal with students **lack high-quality information about the experiences and conditions that help students learn**. If they had information about how much their students were or were not learning and **the practices and conditions that helped them learn**, practitioners would put this knowledge to work, and improvement would naturally follow. (p. 27)

In short, there is an expectation that student affairs professionals and faculty can answer a basic question: "What evidence suggests the intended programming should be effective?" (Finney, et.al., 2021). However, many student affairs professionals and faculty have not been trained in cognition, learning, or pedagogy (e.g., Bresciani, 2010; Brownell & Tanner, 2012; Jones, 2008). Given the lack of training, Kerr and colleagues (2020) noted the need to build this knowledge base:

> If the learning goals focus on identity development, scholarship in this area will require significant exploration and expert consultation. If the learning goals are specified in self-advocacy or self-efficacy, the relevant literature must be mined to identify the right content and develop effective techniques intended to stimulate student learning. This is true for any learning goal selected. Those trained as generalists will need to connect with topic and discipline experts and literature to move beyond surface-level understandings of student learning concepts and practices to achieve the learning. (p. 27)

Where can faculty and student affairs professionals find "high-quality information about the experiences and conditions that help students learn", as Banta and Blaich (2011) noted? How can faculty and staff determine what scholarship is providing credible evidence

[1] Expectations are found at other levels of education. For K-12, the primary source of federal aid is the Elementary and Secondary Education Act, as amended by the Every Student Succeeds Act, which calls on states, districts, and schools to use evidence-based programming. Section 8101(21) defines "evidence-based." For a strategy, intervention, or activity, the definition establishes three tiers of evidence that demonstrate a significant effect on improving student outcomes: (1) strong, (2) moderate, and (3) promising. The definition also includes an activity, strategy, or intervention supported by a rationale based on high-quality research or positive evaluation that such activity, strategy, or intervention has a high likelihood of improving student outcomes (Skinner, 2019; Slavin, 2020).

of effectiveness versus (mis)information that should be ignored? How should faculty and staff summarize the existing credible evidence to inform their programming decisions? We describe the use of systematic review repositories to support the selection, implementation, and assessment of "should-be-effective" programming. Implementing evidence-informed programming is ethical and should result in more efficient engagement in learning improvement efforts.

## Systematic Review Repositories

Faculty and student affairs professionals hope to impact a wide variety of student learning and development outcomes. Often faculty target what we consider academic outcomes, such as written communication, critical thinking, quantitative reasoning, oral communication, among other outcomes. Thus, faculty search for programming, pedagogy, and strategies shown to facilitate students achieving these outcomes. Student affairs professionals are often tasked with targeting these same outcomes in addition to outcomes related to health, civic engagement, diversity, leadership, among other outcomes. *The CAS Standards* (2019) provide the breadth of outcomes that student affairs professionals hope to impact via effective programming. Of course, faculty and student affairs professionals often work together to offer effective co-curricular programming that impacts a variety of desired and shared outcomes.

Knowing where and how to find credible evidence regarding program effectiveness can empower faculty and student affairs professionals to make evidence-based programming decisions. Conducting a search for research on a particular topic (e.g., effective leadership development programming) can be daunting if one is not trained to conduct such a search. An internet search using Google Scholar (which we often observe in practice) often yields an immense number of articles and chapters. The articles providing empirical study of programming need to be read to evaluate the type and quality of evidence, which impact the credibility of effectiveness statements. Faculty and student affairs professionals may not have the time or skill to sort studies into evidence categories, rank them based on a set of evidence standards, and then synthesize the evidence in a meaningful way (Bambra, 2009). Thus, the search for and synthesis of credible evidence of effectiveness may be characterized as time consuming, tedious, demanding, and, for some, overwhelming.

Partly as a consequence of this overwhelming challenge, but also in response to a call for evidence-based programming, organizations have developed evidence grading schemes and repositories of systematic reviews (Boruch & Rui, 2008). These grading schemes and systematic review repositories are forward-facing so the public can easily access already conducted reviews of credible evidence to guide decision making.

The *systematic review*, which has been associated with healthcare evidence and evidence-based medicine for over two decades (Bearman et al., 2012), is becoming an established research method in public health and education, as well as the social sciences (Methods Group of the Campbell Collaboration, 2017). The goal of a systematic review is to describe the effectiveness of programming based on the most credible research evidence available. This goal is accomplished by applying transparent, standardized, and reproducible methods to find and evaluate the quality of evidence from effectiveness studies.

**Knowing where and how to find credible evidence regarding program effectiveness can empower faculty and student affairs professionals to make evidence-based programming decisions.**

A high-quality systematic review follows a formal procedure that begins with the formulation of a precise question, including the definition of the population, the intervention, any comparison group, and outcomes to be measured. A question relevant for faculty overseeing general education may be: Do first-year experience courses for college students positively impact credit accumulation, degree attainment, and academic achievement relative to no first-year course? (for answer, see review by *What Works Clearinghouse*). A question relevant for university health center professionals may be: Does mindfulness-based stress reduction programming improve health, quality of life, and social functioning for students relative to no programming? (for answer, see systematic review by *Campbell Collaboration*).

After the question is delineated, the search for studies to include in the systematic review can begin. After the search has been conducted, the evidence produced by each study is appraised for quality. Increasingly, decision-makers recognize the importance of standards when finding and sorting evidence (Boruch & Rui, 2008). Evidence from individual studies can be equivocal or biased, even if the authors claim otherwise. At their best, systematic reviews produced using evidence grading schemes can reduce the possibility of bias and screen out studies producing ambiguous results. Evidence grading schemes take the design of the study into consideration when screening individual studies of effectiveness. Evidence from each study is sorted within a hierarchy of evidence, with randomized controlled trials (RCTs) at the top of the pyramid and professional opinion articles at the bottom (see Figure 1).

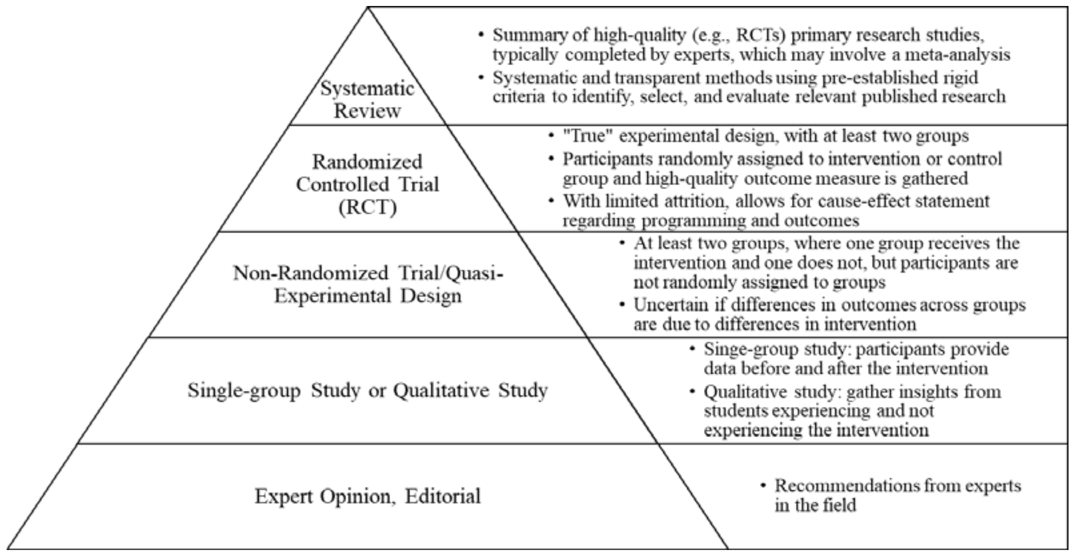



*Figure 1*. Pyramid of Evidence for Program Effectiveness Inferences. The research design producing the evidence should be appropriate to the question being asked (e.g., Slavin, 2020). In the case of effectiveness studies, the question being asked is whether students receiving programming will be more likely to achieve outcomes than students not receiving programming. Answering this question necessitates at least two groups of students who differ in the experience of receiving programming. Other data (e.g., implementation fidelity) and approaches (e.g., qualititaitve methods) provide other valuable insights; yet, when the question centers on "what works", experimental designs are optimal.

**Evidence grading schemes and repositories are incredibly helpful to identify the best available evidence given the intended inference (e.g., program effectiveness) and the variety of designs employed in a domain.**

Although RCTs remain the "gold standard" for effectiveness claims, numerous studies in education and the social sciences may not employ this design. Thus, evidence grading schemes are incredibly helpful to identify the best available evidence given the intended inference (e.g., program effectiveness) and the variety of designs employed in a domain. After high-quality evidence is identified, it is concatenated across studies (meta-analysis may be used) to understand the size of the effect of programming on an outcome.

Since the early 1990's, a number of organizations created evidence grading schemes and repositories for systematic reviews. These organizations differ in the specific discipline(s) they target (see Table 1). Yet, all apply strict grading schemes when reviewing each study's methods and results, all follow a rigorous peer-review process, and all reviews are conducted by qualified researchers. Three well-known organizations are the *Campbell Collaboration* (education, crime, welfare), *Cochrane Collaboration* (health), and *What Works Clearinghouse* (WWC) of the U.S. Department of Education. Below we provide an example of a systematic review provided by The *Campbell Collaboration* and *WWC* to showcase the utility of these repositories. We recommend exploring each repository, as there are many reviews relevant to outcomes in higher education.

## The Campbell Collaboration

*The Campbell Collaboration* "promotes positive social and economic change through the production and use of systematic reviews and other evidence synthesis for evidence-based policy and practice" (Campbell Collaboration, n.d.). The repository provides a user-friendly keyword search by program or outcome. Moreover, each detailed systematic review (i.e., "full report") is coupled with a short (i.e., one to two page) plain language summary.

Table 1
*Description and examples from systematic review repositories*

| Repository | Description | Examples Relevant to Higher Education |
|---|---|---|
| *Campbell Collaboration* | Exists to help people make well-informed decisions about **social & behavioral interventions**. Provides systematic reviews of programs or interventions using rigorous review & synthesis processes of high-quality (RCTs or quasi-experimental designs) primary research. Some research designs have such weak internal validity that they are unacceptable in reviews to inform effective claims (e.g., simple before-after programming studies without comparison groups). | • Bystander Intervention<br>• Mindfulness-based Stress Reduction<br>• Motivational Interviewing for Substance Abuse<br>• Exercise to Improve Self-Esteem in Young People<br>• Advocacy Interventions to Reduce Violence & Promote Well-Being of Women who Experience Partner Abuse |
| *What Works Clearinghouse* | A trusted source of scientific evidence on **education programs, practices, & policies**. WWC reviews research, determines which studies meet rigorous standards (RCTs, quasi-experimental designs), summarizes findings, and provides practice guides. | • Using Technology To Support Postsecondary Learning<br>• Linked Learning Communities<br>• Organizing Instruction & Study to Improve Learning<br>• First Year Experience Courses<br>• Strategies for Postsecondary Students in Developmental Education |
| *Cochrane Library* | Provides short plain language summaries of their longer systematic reviews of empirical research that focus on interventions for **health outcomes** (e.g., alcohol, STIs). Indicates the quality of the studies that informed their conclusions. | • Social norms interventions are not effective enough on their own to reduce alcohol misuse among college students<br>• Self-help & Guided Self-help for Eating Disorders<br>• Prevention of Suicide in University Settings |

*Note*. RCTs = Randomized Controlled Trials.

The systematic review *Effects of Bystander Programs on the Prevention of Sexual Assault among Adolescents and College Students: A Systematic Review* (Kettrey et al., 2019) is (unfortunately) quite relevant to higher education. The full report of the program's effectiveness begins with a description of the purpose for the review, including background information on the problem, research question of interest, and current state-of-the-evidence. In this example, the review "examines the effects bystander programs have on knowledge and attitudes concerning sexual assault and bystander behavior, bystander intervention when witnessing sexual assault or its warning signs, and participants' rates of perpetration of sexual assault" (p. 1).

**This information is particularly helpful for faculty and student affairs professionals seeking to align their desired outcomes with evidence-informed programming.**

Next, the review includes a description of the studies included in the review. Of note are details regarding the types of interventions and various outcomes. This information is particularly helpful for faculty and student affairs professionals seeking to align their desired outcomes with evidence-informed programming. For example, this review summarizes research on the effects of bystander programs on the following outcomes: knowledge concerning sexual assault and intervening, attitudes concerning sexual assault and intervening, behavior when witnessing a sexual assault or its warning signs, and perpetration of sexual assault. Thus, if professionals were interested in influencing these outcomes, this review would provide insight into what programming was and was not effective for which outcome. Evidence from 27 high-quality studies was summarized, including 21 RCTs. Inclusion criteria required that eligible studies have an experimental or controlled quasi-experimental research design, comparing an intervention group (i.e., students assigned to a bystander program) to a comparison group (i.e., students not assigned to a bystander program). Reviewers limited the types of studies included to RCTs and quasi-experimental designs because these typically have lower risk of bias relative to other research designs (e.g., single-group designs).

**Reading this review replaces years of creating novel programming, collecting assessment data produced by rigorous designs, and using results to improve programming in order to uncover effective programming.**

Lastly, the authors include an interpretation of the findings (including what outcomes were and were not impacted) with implications for real-life application and acknowledgment of any remaining gaps in the literature. For example, bystander programs were found to have an effect on some but not all outcomes reflecting knowledge and attitudes concerning sexual assault and intervening. Bystander programs had the most pronounced beneficial effect on rape myth acceptance. The effect on bystander efficacy (i.e., respondents' confidence in their ability to intervene) was also fairly pronounced. There were significant delayed effects (i.e., 1 to 4 months after the intervention) on taking responsibility for intervening/acting, knowing strategies for intervening, and intentions to intervene. Additionally, the effects of bystander programs on intervention behavior outcomes diminished 6-months post-intervention; thus, reviewers concluded that booster sessions may be needed to yield sustained intervention effects. Little or no evidence of effects were found for gender attitudes, victim empathy, date rape attitudes, or noticing sexual assault.

This review provides credible evidence to faculty and student affairs professionals seeking to engage students in effective programming to impact the following outcomes: rape myth acceptance, bystander self-efficacy, increased knowledge and attitudes toward taking responsibility for intervening/acting, knowing strategies for intervening, and intentions to intervene. This review would also suggest to faculty and staff that this programming may not be effective for changing the behavior of potential perpetrators. Reading this review replaces years of creating novel programming, collecting assessment data produced by rigorous designs, and using results to improve programming in order to uncover effective programming.

## What Works Clearinghouse

The *WWC* of the U.S. Department of Education reviews the existing research on different programs, products, practices, and policies in education (*WWC*, n.d.). The *WWC* offers a number of resources for researchers, practitioners, and policymakers, including the following:

- systematic reviews, which provide a synthesis and analysis of all available research on a particular program or intervention in order to assess its effectiveness;
- intervention reports, which provide a brief summary or snapshot of the evidence on a practice, program, or curriculum;
- practice guides for educators, which are based on reviews of research, experiences of practitioners, and expert opinions;
- resources for researchers, which include methodological guidelines and training to further the field of education research.

For example, the *WWC* educator's practice guide *Using Technology to Support Postsecondary Student Learning* provides five evidence-based recommendations on the effective uses of technologies associated with improving postsecondary student learning outcomes (Dabbagh et al., 2019). Each recommendation has a summary of the evidence for that specific recommendation, along with a level of evidence rating (i.e., minimal, moderate, or strong). This rating is informed by the number of studies supporting the recommended practice, the types of study designs included (e.g., RCT, quasi-experimental), and whether the study was conducted in different contexts and with different populations. Due to these strict criteria, it is common for a recommendation to get a minimum level of evidence rating.

Also included in each recommendation are the outcome measure domains impacted. Once again, this information is particularly useful for faculty and student affairs professionals seeking to align their intended outcomes with evidence-based strategies. In this particular practice guide, three of the recommendations (e.g., the use of varied, personalized and readily available digital content; the incorporation of technology that models self-regulated learning; the use of technology to provide targeted feedback) received a moderate level of evidence rating.

The outcome domains associated with the recommendation for the use of varied, personalized, and readily available digital content include student achievement and credit accumulation. The same outcomes were associated with the use of technology to provide targeted feedback. The outcome of student achievement was associated with the recommendation of incorporating technology that models self-regulated learning. Two strategies (e.g., the use of communication and collaboration tools to increase interaction; the use of simulation technologies that help students engage in problem-solving) received a minimal level of evidence rating because only one study met the WWC design standards without reservations. This systematic review provides an efficient mechanism to inform faculty and staff's pedagogy and programming decisions, which can be assessed for effectiveness in their specific context with their students.

## Additional Resources

Clearly, systematic reviews and their derivative products (e.g., practice guides) have utility. They serve as an efficient way to find effective programming, particularly if individuals do not feel qualified to rate the quality of evidence, keep pace with new studies, or wade through large amounts of research (Hempenstall, 2006). Because they bring together a whole body of evidence, systematic reviews can also reduce confusion stemming from individual studies having conflicting results (Cochrane Training, n.d.). They also identify contexts or individuals' characteristics that moderate effectiveness (e.g., program is more effective for one group of students than another). Finally, systematic reviews spotlight areas where there is insufficient evidence to guide programming decisions. That is, for some programs or outcomes, there is no research using adequate methodology to make trustworthy effectiveness claims. A review may be undertaken to formally demonstrate the absence of evidence for common programming.

With that said, there are limitations of systematic reviews. As with any research, a systematic review is time sensitive in that new studies are continually produced and not included in the review. A systematic review is time-consuming and effortful and may take years before available to inform practice. Moreover, a high-quality systematic review requires particular research designs for studies to be included. Thus, systematic reviews may not exist for many programs. We created a resource that provides numerous examples of systematic reviews relevant for higher education. Notice, there are approximately half-a-dozen to a dozen reviews for each repository, not hundreds of reviews within each repository due to the limitations noted above. Given these limitations, we offer three additional resources that may be useful to create evidence-informed programming when a systematic review is not available: collections of research on a topic, the pyramid of evidence, and the Wise Interventions database.

### Collections of Research on a Topic

Not all collections of empirical research on a topic meet the criteria of systematic review repositories (e.g., *WWC*). Yet, these other collections can provide useful information. In the resource we created, we included these other collections of research and information. For each collection, we provide a description, a brief summary of how research is identified, selected, and synthesized, and numerous examples relevant to higher education.

**Clearly, systematic reviews and their derivative products (e.g., practice guides) have utility.**

For example, *Culture of Respect*, a NASPA initiative, is a curated list of theory-driven and evidence-informed sexual violence prevention programs. Programming included on the list may be deemed "supported by evidence" (one peer-reviewed publication using a RCT or quasi-experimental design), "promising" (report or peer-reviewed publication using non-experimental design) or "emerging" (program based on theory but no empirical evidence). There are no systematic reviews and evidence of effectiveness may be weak. Likewise, *CollegeAIM*, a resource developed to address harmful drinking on campuses, does not engage in systematic reviews of effectiveness studies. Instead, this collection lists potential interventions and rates their cost, implementation, and amount of research evaluating their effectiveness. Collections of this sort may be helpful to guide program creation if formal systematic reviews do not exist.

## Pyramid of Credible Evidence

If no systematic reviews or collections of evidence exist, faculty and staff should begin their search for evidence of program effectiveness by identifying individual studies employing RCTs. Recall, high-quality systematic reviews synthesize findings from RCTs to make credible claims regarding program effectiveness. A few well-conducted RCTs may provide necessary evidence to support claims of program effectiveness. If RCTs do not exist, quasi-experimental studies that involve the intervention group and a comparison group should be located. If RCTs or quasi-experimental studies do not exist, studies involving a single group assessed before and after experiencing the programming may be available. Likewise, qualitative studies of data gathered from students who did and did not experience the programming may exist. As one moves down the pyramid, the effectiveness statement becomes less credible. Beginning the search for credible evidence at the top of the pyramid supports efficiency in that less credible evidence may not need to be gathered or evaluated. Expert opinion and testimonials are seductive and we have found they often distract from finding credible evidence of effectiveness.

**Using pre-existing evidence of effectiveness to inform programming forces a focus on student outcomes because evidence-informed strategies are intentionally selected to achieve these outcomes.**

The pyramid of evidence serves as a guide and reminder of how to find the most credible evidence and how to adjust effectiveness claims given the level of evidence. Working with a librarian to identify RCTs before wading through the other types of studies can save a tremendous amount of time. In fact, we have found this pyramid coupled with a few consultations with a librarian not only results in efficient searches for the most credible evidence of effectiveness but also efficacy in future searches.

## "Wise Interventions"

In addition to using the common search engines (e.g., ERIC, PsycNet, PubMed) to find primary research, we recommend a curated resource by Greg Walton and colleagues. This useful website summarizes short yet powerful interventions to impact behavior, self-control, health, belonging, achievement, among other outcomes. Although this database of "wise" interventions is not a concatenation of several RCTs (as found in systematic reviews), there tends to be a body of research regarding intervention impact (Walton & Wilson, 2018).

These Wise Interventions showcase that interventions do not need to be long, complex, or difficult to implement (Walton, 2014). They can be short activities that are not marketed or perceived as interventions by the students engaged in the activity. They are "wise" because they target the underlying psychological process influencing the outcome of interest. For example, a one-hour intervention where students learned and then explained to others that social adversities are normal to college buffered the impact of negative experiences on sense of belonging, which resulted in improvement in grades and health outcomes for minoritized students (Walton & Cohen, 2011). The database includes several interventions relevant for Offices of Student Success that focus on academic achievement outcomes, Offices of Health and Wellness that focus on wellbeing and physical health outcomes, Offices of Civic Engagement that focus on voting and other civic behavior outcomes, Offices of Orientation and Residence Life that focus on sense of belonging outcomes, STEM degree programs with the intended outcomes of retaining and supporting underserved populations, among other short interventions to impact outcomes relevant to higher education (Walton & Wilson, 2018).

## Conclusion

Improving student learning and development involves answering "What works?" Answering this question involves two fundamental steps: 1) identifying proven effective evidence-informed strategies; 2) assessing if the strategies are effective in the current setting (Bryk et al., 2015). We focused on the first step given it addresses our concerns regarding efficiency and engagement in outcomes assessment and ethical practice in higher education. More specifically, using pre-existing evidence of effectiveness to inform programming forces a focus on student outcomes because evidence-informed strategies are intentionally selected to achieve these outcomes. Faculty and staff need to evaluate the credibility of the pre-existing evidence, with the most credible evidence for effectiveness coming from RCTs. Each RCT provides insight into program effectiveness under narrow conditions with a specific population. Accumulation of several RCTs across different contexts and populations provides insight into context or student characteristics that may moderate program effectiveness. This body of research has potential to create more equitable programming given demonstrated impact on outcomes across diverse student populations.

Yet, not all higher education professionals feel comfortable reading studies of RCTs. Not all campuses have a Center for Teaching and Learning that offers training in evidence-informed practices. Moreover, colleagues who engage in evidence-informed programming should not be relied upon to support the development of other colleagues, as research shows those engaging with evidence-informed, innovative practices tend to talk to each other (Lane et al., 2020). Thus, we introduced systematic review repositories to counter the deluge of misinformation and encourage the use of evidence-informed programming. Our goal was to support faculty and student affairs professionals who yearn for resources to help them do their jobs well.

We embrace outcomes assessment as a mechanism for assessing should-be-effective programming. However, we urge the higher education community to acknowledge its inefficiency and dependency on consistent engagement by faculty and staff, which hinders its impact on learning improvement when applied to unsubstantiated programming. We recognize that some decision-makers lack a scientific framework and are inclined to accept programming proposals based on opinions, testimonials, intuition, and good intentions, not empirically linked to intended outcomes (Hempenstall, 2006). Thus, to ensure students have the opportunity to learn and develop as promised by higher education institutions, we call on colleagues, administrators, and students to consistently ask those creating and improving programming to share their process of using credible evidence to inform decision making.

## References


Bambra, C. (2009). Real world reviews: a beginner's guide to undertaking systematic reviews of public health policy interventions. *Journal of Epidemiology and Community Health*, 65, 14-19. 10.1136/jech.2009.088740

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

Bearman, M., Smith, C., Carbone, A., Slade, S., Baik, C., Hughes-Warrington, M., & Neumann, D. (2012). Systematic review methodology in higher education. *Higher Education Research & Development*, *31*, 625-640. https://doi.org/10.1080/07294360.2012.702735

Boruch, R., & Rui, N. (2008). From randomized controlled trials to evidence grading schemes: current state of evidence-based practice in social sciences. *Journal of Evidence-Based Medicine*, *1*, 41-49. 10.1111/j.1756-5391.2008.00004.x

Bresciani, M.J. (2010). Understanding barriers to student affairs professionals' engagement in outcomes-based assessment of student learning and development. *Journal of Student Affairs*, *14*, 81–90.

Brownell, S., & Tanner, K. (2012). Barriers to faculty pedagogical change: Lack of training, time, incentives, and… tensions with professional identity? *CBE—Life Sciences Education*, *11*, 339 – 346. 10.1187/cbe.12-09-0163

Bryk, A., Gomez, L., Grunow, A. & LeMahieu, P. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Press. https://eric.ed.gov/?id=ED568744

Campbell Collaboration (n.d.). *Campbell's vision, mission and key principles*. https://campbellcollaboration.org/about-campbell/vision-mission-and-principle.html

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

Cochrane Training (n.d.). *Module 3*: *Introduction to systematic reviews*. Cochrane Evidence Essentials. https://training.cochrane.org/essentials

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

Dabbagh, N., Bass, R., Bishop, M., Costelloe, S., Cummings, K., Freeman, B., Frye, M., Picciano, A. G., Porowski, A., Sparrow, J., & Wilson, S. J. (2019). *Using technology to support postsecondary student learning: A practice guide for college and university administrators, advisors, and faculty*. Institute of Education Sciences, What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/wwc-using-tech-postsecondary.pdf

Eubanks, D. (2017). A guide for the perplexed. *Intersection: A Publication of the Association for the Assessment of Learning in Higher Education*, 4-13. Retrieved from https://www.aalhe.org/assets/docs/AAHLE_Fall_2017_Intersection.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). Charles Thomas Publisher. https://eric.ed.gov/?id=ED593398

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. 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.H., Good, M.R., Coleman, C.M., & Smith, K.L. (2014). *A simple model for learning improvement: Weigh pig, feed pig, weigh pig*. (Occasional Paper No. 23). University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Hempenstall, K. (2006). What does evidence-based practice in education mean? *Australian Journal of Learning Disabilities*, *11*, 83 – 92. https://doi.org/10.1080/19404150609546811

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

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. https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/OccasionalPaper4.pdf

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

Jones, A. (2008). Preparing new faculty members for their teaching role. *New Directions For Higher Education*, *143*, 93-100. https://doi.org/10.1002/he.317

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

Kettrey, H. H., Marx, R. A., & Tanner-Smith, E. E. (2019). Effects of bystander programs on the prevention of sexual assault among adolescents and college students: A systematic review. *Campbell Collaboration*, *15*, e1013. https://doi.org/10.4073/csr.2019.1

Komarraju, M., & Handelsman, M. (2012). *Preparing to teach*: *Becoming part of an ethical culture*. In R. Landrum & M. McCarthy (Eds.), Teaching ethically: Challenges and opportunities (p. 191–201). American Psychological Association. https://doi.org/10.1037/13496-017

Lane, A., McAlpin, J., Earl, B., Feola, S., Lewis, J., Mertens, K., Shadle, S., Skvoretz, J., Ziker, J., Couch, B., Prevost, L., & Stains, M. (2020). Innovative teaching knowledge stays with users. *Proceedings of the National Academy of Sciences*, *117*, 22665-22667. https://doi.org/10.1073/pnas.2012372117

Lewis, C. (2015). What is improvement science? Do we need it in education? *Educational Researcher*, *44*, 54-61. https://doi.org/10.3102/0013189X15570388

Methods Group of the Campbell Collaboration (2017). *Methodological expectations of Campbell Collaboration intervention reviews*: *Conduct standards* (Campbell Policies and Guidelines Series No. 3). Campbell Collaboration. 10.4073/cpg.2016.3

Pope, A., 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., 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*.

Potter, M. & Kustra, E. (2011). The relationship between scholarly teaching and SoTL: Models, distinctions, and clarifications. *International Journal for the Scholarship of Teaching and Learning*, *5*(1). https://doi.org/10.20429/ijsotl.2011.050123

Skinner, R.R. (2019). *The Elementary and Secondary Education Act (ESEA), as amended by the Every Student Succeeds Act (ESSA)*: *A primer*. Congressional Research Service. R45977. https://fas.org/sgp/crs/misc/R45977.pdf

Slavin, R.E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist*, *55*, 21-31. https://doi.org/10.1080/00461520.2019.1611432

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*, 1-13. https://www.rpajournal.com/dev/wp-content/uploads/2020/09/Elevating-Program-Theory-and-Implementation-Fidelity-in-Higher-Education.pdf

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

Svinicki, M.D., & McKeachie, W.J. (2013) *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (14th ed.). Wadsworth Cengage Learning.

U.S. Department of Education (2006). *A test of leadership*: *Charting the future of US higher Education*. U.S. Department of Education. https://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf

Walton, G.M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science*, 23, 73–82. https://doi.org/10.1177/0963721413512856

Walton, G.M., & Cohen, G.L. (2011). A brief social-belonging intervention improves academic and health outcomes among minority students. *Science*, *331*, 1447–1451. 10.1126/science.1198364

Walton, G.M., & Wilson, T.D. (2018). Wise interventions: Psychological remedies for social and personal problems. *Psychological Review*, *125*, 617 – 655. https://doi.org/10.1037/rev0000115

What Works Clearinghouse (n.d.). *What we do*. https://ies.ed.gov/ncee/wwc/

Wight, D., Winbush, E., Jepson, R., & Doi, L. (2016). Six steps in quality intervention development (6SQuID). *Journal of Epidemiology and Community Health*, *70*, 520-525. 10.1136/jech-2015-205952

## Abstract

This paper examines students' patterns of success in classes with high DFW rates at a research-intensive university. We investigated whether certain assignment types were associated with inequitable grade distributions for underrepresented minority (URM) and transfer students and whether assignment grade patterns were similar to final grade patterns. Across eight classes, 745 students' grades were analyzed from 27 assignments including tests, papers, projects, homework, and oral reports. In every class, URM students received lower final grades than non-URM students, and transfer students received lower final grades than non-transfer students. In five classes, different patterns of equity emerged across different assignment types and different groups of students. These findings support the importance of going beyond the disaggregation of final grades by disaggregating grades on individual assignments, and the need to develop institutional practices that examine the presence of equity gaps in the classroom.



### AUTHORS

Harriet T. Hobbs, MPA
*University of North Carolina at Charlotte*

Karen E. Singer-Freeman, PhD
*University of North Carolina at Charlotte*

Christine Robinson, EdD
*University of North Carolina at Charlotte*


# Considering the Effects of Assignment Choices on Equity Gaps

The Aspen Education and Society Program and the Council of Chief State School Officers (2017) defined equitable institutions as those in which "every student has access to the resources and educational rigor they need at the right moment in their education, despite race, gender, ethnicity, language, disability, family background, or family income" (p. 3). However, as a nation, we are failing to create equitable institutions of higher education. Many colleges and universities still require standardized scores from the SAT or ACT for entry, despite evidence that historically underserved students receive lower scores than other students (College Board, 2018; National Center for Education Statistics, 2019). Lower scores may reduce financial aid awards and discourage students from applying to or being admitted by competitive institutions (Zwick, 2019). Once students gain admission to a college, over a fifth leave without obtaining a credential (Rosenbaum et al., 2015), and over a third of students who matriculate at four-year public universities fail to graduate (Shapiro et al., 2018). A disproportionately high number of students leaving college without degrees are from underrepresented ethnic minority populations (URM) or low-income families. For URM students who transfer between institutions, the completion gaps are larger (Shapiro et al., 2018).

Transferring between institutions creates stress for students and is followed by a period of adaptation called "transfer shock" (Diaz, 1992; Fauria & Fuller, 2015). For


### CORRESPONDENCE

**Email**
hhobbs2@uncc.edu

example, transfer students in Texas were four times less likely to be retained after one year than non-transfer students (Fauria & Slate, 2014). Many transfer students experience a dip in grade point average (GPA) during the first or second semester at a new institution (Jacobson et al., 2017). Low grades can contribute to students' doubts about their ability to succeed. Ishitani (2008) found that transfer students with higher first semester GPAs were more likely to persist than students with lower first semester GPAs.

**For many students, the first step toward leaving college can be a low or failing grade in a class.**

Students who leave college without credentials have invested substantial amounts of time and money in the pursuit of higher education without any tangible benefit. Rosenbaum et al. (2015) called these students the "new forgotten half." With rapid demographic, economic, and cultural transitions, even more students will transfer between institutions of higher education and be first-generation, low-income, and students of color (McGee, 2015). Consequently, it is essential that institutions of higher education initiate practices to increase completion rates of underserved students (Association of American Colleges and Universities, 2018; Harper & Harris, 2012; Olson, 2020).

For many students, the first step toward leaving college can be a low or failing grade in a class. At our institution, we found that among students experiencing financial distress, every unit increase in GPA increased the odds the student would be retained by a factor of 1.68. Thus, closing gaps in class grades is an important element of closing gaps in college completion. Differences in college GPA are only partially explained by differences in income and prior academic preparation (Fletcher & Tienda, 2010; Lorah & Ndum, 2013). Spenner et al. (2004) found that only 40% of the variance between White and Black students' first semester grades could be explained by differences in socioeconomic background and academic preparation, leaving 60% of the gap unexplained.

Even when low assignment grades do not impact a student's final grade, low assignment grades can negatively impact retention by reducing a student's sense of academic self-efficacy (Montenegro et al., 2020). Academic self-efficacy describes students' beliefs about their ability to execute a course of action to successfully complete an academic task (Bandura, 1997). When students lack a sense of academic self-efficacy, they are less likely to persist to overcome academic challenges (Chemers et al., 2001; Han et al., 2017; Shen et al., 2016). Thus, even in instances in which low assignment grades do not translate directly to low course grades, when low assignment grades reduce students' sense of academic self-efficacy, there could be long-term reductions in academic success.

Because educational equity gaps represent institutional failure, improving equity requires organizational change and faculty engagement (Bensimon, 2005). To engage faculty, institutions must create cultures of inquiry in which the examination of data informs faculty-driven responses to inequities (Bensimon, 2005). Disaggregation of student learning data reveals educational equity gaps and supports the establishment of institutional cultures of inquiry (Maki, 2017). Currently, most colleges and universities only report aggregated student outcomes data, which obscures evidence of privilege-based stratification (Bauman et al., 2005; Singer-Freeman et al., 2021). To date, little research examines equity gaps within assignments. Campuses that disaggregate grades do so based on course grades. When faculty learn of equity gaps in their classes, it can be difficult for them to determine the source of the inequity. An examination of disaggregated data across different assignments in a course can provide faculty with actionable information. Identifying assignments that result in inequitable patterns of performance can lead to evidence-based assignment modifications. Demonstrating that different patterns of equity exist across different assignment types can be the first step toward engaging faculty in disaggregating assignment grades in their classes.

In the current work, we examined disaggregated grades across different assignments in classes with 50 or more enrolled students and with high numbers of D, F, or W (withdrawal) grades at a research-intensive university. This work did not involve direct contact with either students or faculty. Our goal was to determine whether grading distributions differed for URM and transfer students compared to non-URM and non-transfer students across different assignments and final grades within classes. We focused our exploratory work on large classes in which many students received grades of D, F, or W (DFW rates) because success or failure in these classes has consequences for retention in the major and at the university. Because faculty and administrators are currently examining the role these classes play in student success, evidence of different grading distributions on assignments in these classes will help to establish the importance of disaggregating assignment grades.

## Methods

### Procedure

We obtained a list of 88 classes enrolling 50 or more students that had DFW rates of 30% or higher during the fall and spring semesters in 2017 and 2018. The courses that were listed included multiple sections taught by different instructors. We reviewed assignments from all sections of each course that recorded grades in the university's learning management system. The review of assignments revealed eight classes that stored grades in the learning management system and included graded assignments other than quizzes, tests, exams, or completion-based grades (such as attendance grades or assignments in which students received full credit for completion). Because we wished to examine patterns of performance across different assignment types, we excluded the 80 classes that did not offer forms of assignments other than quizzes, tests, exams, or completion-based assignments. Of the classes that did not include different forms of assignments, 42 (53%) were introductory-level classes and 57 (71%) were science, technology, engineering, and mathematics (STEM) classes. The eight remaining classes included in analyses were four introductory classes: Pre-Calculus (MATH), Introduction to Communication Theory (COMM), Network Theory II (ENGR), and Principles of Accounting (ACCT) and four advanced classes: Organic Chemistry Lab (CHEM), Design & Implementation – Object-Oriented Systems (INFO), Physiological Psychology (PSYC), and Sociology of Health and Illness (SOCY).

**Even in instances in which low assignment grades do not translate directly to low course grades, when low assignment grades reduce students' sense of academic self-efficacy, there could be long-term reductions in academic success.**

When a class was taught by the same instructor using the same assignments for more than one semester, we included data from all offerings between 2016 and 2018. The classes are listed in Table 1, along with the number of class offerings, percentage of students receiving final grades of D or F (DF rates), and special features of the class. We do not report withdrawal rates because this information was not available in the learning management system. As seen in Table 1, DF rates varied widely between classes ranging from 3% in ENGR to 25% in SOCY. Most of the classes were offered in the College of Liberal Arts. Several classes required completion of prerequisite courses (with a final grade of C or above) prior to enrollment.

Table 1
*Classes Included in the Study*

| Class | Sections | % DF Rates | College | Special Features |
|---|---|---|---|---|
| MATH | 2 | 19% | Liberal Arts | Prerequisite for Engineering Calculus |
| COMM | 1 | 16% | Liberal Arts | ——— |
| ENGR | 1 | 3% | Engineering | 3 prerequisites required for enrollment |
| ACCT | 1 | 13% | Business | Flipped Delivery – students viewed lectures online at home and spent class time working on problems |
| CHEM | 9 | 14% | Liberal Arts | Lab, 1 prerequisite required for enrollment |
| INFO | 1 | 5% | Computing | ——— |
| PSYC | 1 | 11% | Liberal Arts | 4 prerequisites required for enrollment, online delivery |
| SOCY | 3 | 25% | Liberal Arts | 1 prerequisite required for enrollment |

Every class included at least two different forms of graded assignments. The types of assignments included exams (cumulative finals and mid-terms), tests (covering several weeks of work), quizzes (low-stakes frequent assessments covering a single week or day of work), homework (frequent low stakes work to check for understanding and allow practice), writing (scientific lab reports, formal essays, and reading responses), group projects, in-class activities, and oral reports. The proportion of the final class grade determined by each assignment type is reported in Table 2. Tests were the most common form of assignment, followed by homework and writing. Generally, introductory classes (the first four in the table) relied more heavily on tests and homework than advanced courses which were more likely to include writing assignments, projects, activities, or an oral report.

Table 2
*Proportion of Final Class Grade Determined by each Assignment*

| Class | Exams, Quizzes or Tests | Homework | Writing | Group Project | Class Activity | Oral Report |
|---|---|---|---|---|---|---|
| MATH | 80% | 20% | | | | |
| COMM | 83% | | 8% | | | |
| ENGR | 85% | 15% | | | | |
| ACCT | 72% | 7% | | 14% | | |
| CHEM | 5% | | 95% | | | |
| INFO | 50% | 40% | | | 10% | |
| PSYC | 75% | | | | | 15% |
| SOCY | 30% | 20% | 30% | | | |

*Note*. Rows may not total to 100% because completion-based grades were excluded.

## Participants

We report the number of participants and demographic information in Table 3. We had a total sample size of 745 students which included 53% female, 47% transfer, 51% White, 23% African American, 14% Hispanic, 8% Asian, 3% two or more races, and .01% Native American. Four percent of the sample did not provide information about their race or ethnicity.

Table 3
*Demographic Information*

| Class | Total | Female | Transfer | White | African American | Hispanic | Asian | 2 or more Races | Native American | No report |
|---|---|---|---|---|---|---|---|---|---|---|
| MATH | 109 | 36 | 20 | 56 | 23 | 10 | 13 | 6 | 1 | 1 |
| COMM | 146 | 81 | 93 | 79 | 35 | 19 | 3 | 4 | 0 | 6 |
| ENGR | 41 | 2 | 17 | 22 | 4 | 4 | 7 | 1 | 0 | 3 |
| ACCT | 53 | 13 | 31 | 30 | 7 | 11 | 2 | 1 | 0 | 2 |
| CHEM | 150 | 101 | 57 | 74 | 24 | 19 | 24 | 1 | 0 | 9 |
| INFO | 61 | 12 | 16 | 30 | 9 | 2 | 15 | 4 | 0 | 1 |
| PSYC | 54 | 43 | 37 | 33 | 15 | 3 | 1 | 1 | 0 | 3 |
| SOCY | 131 | 107 | 81 | 50 | 51 | 13 | 5 | 6 | 0 | 6 |
| Total | 745 | 395 | 352 | 379 | 169 | 103 | 62 | 24 | 1 | 31 |
| % | | 53% | 47% | 51% | 23% | 14% | 8% | 3% | .01% | 4% |

Because many classes had limited enrollment of students from certain underserved groups, we compared URM students, which included African American, Hispanic, and Native American students (37% of total sample), to non-URM students which included White and Asian students (59% of total sample). We chose to classify both White and Asian students as non-URM because students from these groups are either well-represented or over-represented at four-year institutions of higher education in the United States when compared to their representation in the population of the United States (Monarrez & Washington, 2020). We excluded participants who did not report race or ethnicity or reported two or more races. We compared students who transferred to the university (transfer students) to students who began their studies at the university (non-transfer students).

## Coding

To compare patterns of performance on different assignment types without influence of assignment weighting, we converted scores into percentages and created a single average score for each assignment type for each student. We included scores of 0 for missing assignments in average scores. For example, a single average homework score was created by totaling the number of homework points received and dividing it by the total number of possible homework points. Independent samples t-tests were conducted using SPSS to evaluate differences between URM and non-URM students and differences between transfer and non-transfer students on individual assignments and in final grades. Cohen's *d* was calculated by hand.

## Results

Final course grades are reported as a function of URM and transfer status in Tables 4 and 5. An inspection of scores prior to data analysis revealed that in every class, URM students received lower final grades than non-URM students, and transfer students received lower final grades than non-transfer students. To determine if these differences were statistically significant, we calculated independent samples t-tests comparing final grades of URM students to non-URM students and transfer students to non-transfer students. We observed significant differences with moderate effect sizes in SOCY in which URM students received lower average grades (70%) than non-URM students (77%), $t(102) = 2.75$, $p = .01$, $d = .57$ and transfer students received lower average grades (71%) than non-transfer students (76%), $t(129) = 2.29$, $p = .02$, $d = .39$. A significant difference was observed for transfer students in ACCT $t(51) = 2.18$, $p = .04$, $d = .54$ such that transfer students received lower average grades (74%) than non-transfer students (79%).

**Demonstrating that different patterns of equity exist across different assignment types can be the first step toward engaging faculty in disaggregating assignment grades in their classes.**

Table 4
*Non-URM and URM Student Final Grades Reported as Percentages with Corresponding t-Tests*

| Class | Non-URM | URM | t-test | *p* | Cohen's *d* |
|---|---|---|---|---|---|
| MATH | 76% (17) | 75% (14) | $t(99) = .28$ | .78 | .06 |
| COMM | 72% (11) | 70% (12) | $t(134) = .91$ | .37 | .17 |
| ENGR | 55% (16) | 54% (5) | $t(40) = .17$ | .87 | .08 |
| ACCT | 77% (9) | 75% (9) | $t(48) = .74$ | .46 | .22 |
| CHEM | 81% (18) | 76% (20) | $t(139) = 1.37$ | .18 | .26 |
| INFO | 90% (12) | 84% (12) | $t(58) = 1.66$ | .10 | .50 |
| PSYC | 82% (16) | 81% (14) | $t(49) = .32$ | .75 | .07 |
| SOCY | 77% (9) | 70% (15) | $t(102) = 2.75$ | .01 | .57 |

*Note*. Standard deviations are reported in parentheses.

Table 5
*Non-Transfer and Transfer Student Final Grades Reported as Percentages with Corresponding t-Tests*

| Class | Non-Transfer | Transfer | t-test | *p* | Cohen's *d* |
|---|---|---|---|---|---|
| MATH | 76% (17) | 75% (6) | $t(105) = .14$ | .78 | .06 |
| COMM | 73% (13) | 69% (15) | $t(144) = 1.72$ | .09 | .29 |
| ENGR | 57% (14) | 54% (15) | $t(43) = .66$ | .51 | .21 |
| ACCT | 79% (7) | 74% (11) | $t(51) = 2.18$ | .04 | .54 |
| CHEM | 81% (17) | 78% (17) | $t(137) = .94$ | .35 | .18 |
| INFO | 89% (13) | 87% (10) | $t(59) = .65$ | .52 | .17 |
| PSYC | 82% (19) | 80% (15) | $t(53) = .52$ | .60 | .12 |
| SOCY | 76% (13) | 71% (13) | $t(129) = 2.29$ | .02 | .39 |

*Note*. Standard deviations are reported in parentheses.

Table 6
*Non-URM and URM Student Quiz, Test, and Exam Grades Reported as Percentages with Corresponding t-Tests*

| Class | Test Type | Non-URM | URM | t-test | *p* | Cohen's *d* |
|---|---|---|---|---|---|---|
| MATH | FR Test | 78% (15) | 78% (12) | $t(99)=.12$ | .81 | 0 |
| | FR Exam | 72% (21) | 70% (21) | $t(99) =.44$ | .78 | .10 |
| ACCT | MC Quiz* | 83% (13) | 81% (22) | $t(48) =.25$ | .80 | .11 |
| | MC Exam | 74% (10) | 69% (14) | $t(48)=1.31$ | .20 | .41 |
| COMM | MC Exam | 73% (15) | 71% (15) | $t(134)=.65$ | .52 | .13 |
| ENGR | MC Quiz | 67% (21) | 62% (17) | $t(40)= .61$ | .54 | .26 |
| | FR Test | 53% (13) | 53% (6) | $t(40)=.004$ | .99 | 0 |
| | FR Exam | 72% (15) | 65% (10) | $t(40)=1.28$ | .21 | .55 |
| CHEM | MC Quiz* | 78% (23) | 70% (31) | $t(139)= 1.62$ | .10 | .29 |
| INFO | MC Exam* | 91% (9) | 90% (7) | $t(58) =.39$ | .70 | .12 |
| PSYC | MC Quiz* | 87% (8) | 83% (8) | $t(49)=1.59$ | .12 | .50 |
| | MC Exam* | 80% (11) | 79% (12) | $t(49)= .53$ | .60 | .09 |
| SOCY | MC Exam* | 77% (12) | 71% (11) | $t(116)= 2.71$ | .01 | .52 |

*Note*. Online assessments are marked with an asterisk. Standard deviations are reported in parentheses.

To investigate the extent to which different assignment types resulted in different grading distributions, we conducted independent samples *t*-tests comparing assignment grades of URM students to non-URM students and transfer students to non-transfer students. Every class included quizzes, tests, or exams. Quizzes included frequent low-stakes assessments that covered a small amount of material, tests included non-cumulative assessments that were given to cover several weeks of material, and exams included cumulative mid-terms or finals. Each assessment included either multiple-choice question formats (MC) or free response question formats (FR). As seen in Tables 6 and 7, across the eight classes, three had significant grade differences, with moderate to large effect sizes. In SOCY, non-URM students received higher online multiple-choice exam grades (77%) than URM students (71%), $t(116) = 2.71$, $p = .01$, $d = .52$ and non-transfer students received higher online multiple-choice exam grades (77%) than transfer students (71%), $t(128) = 2.50$, $p = .01$, $d = .50$. In ACCT, non-transfer students received higher multiple-choice exam grades (77%) than transfer students (69%), $t(51) = 2.62$, $p = .01$, $d = .72$. In PSYC non-transfer students received higher online multiple-choice exam grades (84%) than transfer students (77%), $t(53) = 2.02$, $p = .05$, $d = .63$ and non-transfer students received higher online multiple-choice quiz grades (88%) than transfer students (84%), $t(52) = 2.58$, $p = .05$, $d = .53$

Table 7
*Non-Transfer and Transfer Student Quiz, Test, and Exam Grades Reported as Percentages with Corresponding t-Tests*

| Class | Test Type | Non-transfer | Transfer | t-test | *p* | Cohen's *d* |
|---|---|---|---|---|---|---|
| MATH | FR Test | 78% (15) | 76% (15) | $t(105)=.51$ | .78 | .13 |
| | FR Exam | 72% (21) | 72% (20) | $t(105)=.03$ | .78 | 0 |
| ACCT | MC Quiz* | 87% (11) | 78% (22) | $t(47)=1.90$ | .06 | .82 |
| | MC Exam | 77% (9) | 69% (13) | $t(51) = 2.62$ | .01 | .72 |
| COMM | MC Exam | 73% (15) | 71% (15) | $t(144) =.57$ | .57 | .13 |
| ENGR | MC Quiz | 70% (18) | 63% (23) | $t(43) = 1.23$ | .23 | .34 |
| | FR Test | 53% (13) | 53% (8) | $t(42) =.14$ | .89 | 0 |
| | FR Exam | 72% (15) | 71% (12) | $t(41) =.16$ | .87 | .07 |
| CHEM | MC Quiz* | 77% (27) | 74% (24) | $t(137) =.63$ | .53 | .12 |
| INFO | MC Exam* | 91% (9) | 89% (6) | $t(59) =.75$ | .46 | .26 |
| PSYC | MC Quiz* | 88% (7) | 84% (8) | $t(52) = 2.58$ | .05 | .53 |
| | MC Exam* | 84% (10) | 77% (12) | $t(53) = 2.02$ | .05 | .63 |
| SOCY | MC Exam* | 77% (11) | 71% (13) | $t(128) = 2.50$ | .01 | .50 |

*Note*. Online assessments are marked with an asterisk. Standard deviations are reported in parentheses.

Five classes included homework assignments. Average homework grades are reported as a function of URM and Transfer status in Tables 8 and 9. Significant differences with moderate effect sizes were observed in SOCY in which non-URM students received higher homework (reading response) grades (78%) than URM students (72%), $t(103) = 2.24$, $p = .03$, $d = .37$ and non-transfer students received higher homework (reading response) grades (80%) than transfer students (72%), $t(129) = 2.87$, $p = .01$, $d = .52$.

Three classes included writing assignments. Average writing grades are reported as a function of URM and Transfer status in Tables 10 and 11. Significant differences with moderate to large effect sizes were observed. In COMM non-URM students received higher in-class writing grades (88%) than URM students (80%), $t(134) = 2.79$, $p = .01$, $d =.43$. In SOCY

Table 8
*Non-URM and URM Student Homework Grades Reported as Percentages with Corresponding t-Tests*

| Class | Assignment | Non-URM | URM | t-test | *p* | Cohen's *d* |
|---|---|---|---|---|---|---|
| MATH | Problem Sets | 77% (23) | 76% (23) | $t(99)=.18$ | .78 | .04 |
| ACCT | Problem Sets | 71% (22) | 69% (27) | $t(48)=.21$ | .84 | .20 |
| ENGR | Problem Sets | 70% (21) | 60% (17) | $t(40)=1.24$ | .22 | .52 |
| INFO | Programming | 91% (16) | 81% (23) | $t(58) = 1.87$ | .07 | .51 |
| SOCY | Reading Responses | 78% (11) | 72% (20) | $t(103) = 2.24$ | .03 | .37 |

*Note*. Standard deviations are reported in parentheses.

Table 9
*Non-Transfer and Transfer Student Homework Grades Reported as Percentages with Corresponding t-Tests*

| Class | Assignment | Non-Transfer | Transfer | t-test | *p* | Cohen's *d* |
|---|---|---|---|---|---|---|
| MATH | Problem Sets | 76% (24) | 76% (21) | $t(105)=.04$ | .78 | 0 |
| ACCT | Problem Sets | 74% (16) | 67% (30) | $t(51)=1.11$ | .27 | .29 |
| ENGR | Problem Sets | 69% (18) | 67% (24) | $t(43) = .40$ | .69 | .09 |
| INFO | Programming | 90% (20) | 86% (15) | $t(59) =.59$ | .56 | .23 |
| SOCY | Reading Responses | 80% (15) | 72% (16) | $t(129) = 2.87$ | .01 | .52 |

*Note*. Standard deviations are reported in parentheses.

Table 10
*Non-URM and URM Student Writing Grades Reported as Percentages with Corresponding t-Tests*

*Non-URM and URM Student Writing Grades in Percentages with Corresponding t-Tests*

| Class | Assignment | Non-URM | URM | t-test | *p* | Cohen's *d* |
|---|---|---|---|---|---|---|
| COMM | In Class | 88% (13) | 80% (23) | $t(134) = 2.79$ | .01 | .43 |
| CHEM | Lab Report | 80% (19) | 75% (21) | $t(139)=1.35$ | .18 | .25 |
| SOCY | Essay | 82% (7) | 78% (13) | $t(93) = 2.07$ | .04 | .38 |

*Note*. Standard deviations are reported in parentheses.

non-URM students received higher essay grades (82%) than URM students (78%), $t(93) = 2.07$, $p = .04$, $d = .38$, and non-transfer students received higher essay grades (84%) than transfer students (77%), $t(125) = 3.95$, $p = .00$, $d = .74$.

Three classes included other forms of assignments: a group project, in-class activities, and an oral report. Average assignment grades are reported as a function of URM and Transfer status in Tables 12 and 13. A significant difference with a moderate effect size was observed in INFO in which non-URM students received higher in-class activity grades (83%) than URM students (70%), $t(58) = 2.16$, $p = .04$, $d = .60$.

Table 11
*Non-URM and URM Student Writing Grades Reported as Percentages with Corresponding t-Tests*

| Class | Assignment | Non-transfer | Transfer | t-test | *p* | Cohen's *d* |
|---|---|---|---|---|---|---|
| COMM | In Class | 85% (15) | 85% (16) | $t(144)=.22$ | .82 | 0 |
| CHEM | Lab Report | 80% (18) | 77% (18) | $t(137)=1.03$ | .31 | .17 |
| SOCY | Essay | 84% (6) | 77% (12) | $t(125) = 3.95$ | .00 | .74 |

*Note*. Standard deviations are reported in parentheses.

Table 12
*Non-URM and URM Assignment Grades Reported as Percentages with Corresponding t-Tests*

| Class | Assignment | Non-URM | URM | t-test | *p* | Cohen's *d* |
|---|---|---|---|---|---|---|
| ACCT | Group Project | 86% (9) | 91% (9) | $t(48) = 1.90$ | .06 | .56 |
| INFO | Class Activities | 83% (20) | 70% (23) | $t(58) = 2.16$ | .04 | .60 |
| PSYC | Oral Report | 97% (12) | 94% (11) | $t(49) = .73$ | .47 | .26 |

*Note*. Standard deviations are reported in parentheses.

Table 13
*Non-transfer and Transfer Student Assignment Grades Reported as Percentages with Corresponding t-Tests*

| Class | Assignment | Non-transfer | Transfer | t-test | *p* | Cohen's *d* |
|---|---|---|---|---|---|---|
| ACCT | Group Project | 86% (12) | 89% (8) | $t(51)=1.05$ | .30 | .29 |
| INFO | Class Activities | 81% (23) | 80% (15) | $t(59) =.12$ | .90 | .05 |
| PSYC | Oral Report | 93% (16) | 96% (10) | $t(53) =.82$ | .93 | .23 |

*Note*. Standard deviations are reported in parentheses.

## Discussion

We began this work with the goal of demonstrating the importance of disaggregating assignment and final grades as a first step towards identifying patterns of performance in different student populations. We investigated whether certain assignments were associated with grade distributions in which URM or transfer students received lower grades than non-URM or non-transfer students. Both URM students and transfer students have been shown to be underserved by institutions of higher education (Bensimon, 2005; Nuñez & Yoshimi, 2017). We hypothesized that differing grade distributions in which students from underserved groups receive lower grades than those from other groups are evidence of educational equity gaps. Further, we hypothesized that examining assignments with uneven distributions of grades will engage faculty in a culture of equity in which changes to assignment design might be a route to improving equity in educational attainment.

We found a great deal of variability in the patterns of performance that emerged from final grades and individual assignment grades. In four of the eight classes, different patterns of performance emerged across individual assignments and final grades. These results support the importance of considering patterns of performance on assignments to clarify and address educational equity gaps. In every class, we found URM students received lower final grades than non-URM students and transfer students received lower final grades than non-transfer students. There were several instances in which these differences had moderate effect sizes despite not reaching conventional levels of significance. Strikingly, of the 27 assignments analyzed across eight classes, non-URM students received higher average grades than URM students in 23 assignments (85%), and non-transfer students received higher average grades than transfer students in 21 assignments (78%). For both URM and transfer students, significant differences were observed in six assignments (22%).

**Structuring assignments so that content is equally familiar to all students reduces educational equity gaps by limiting the effects of prior knowledge and privilege.**

Given the prevalence of assignment grade distributions that favored students from well-served groups over students from underserved groups, it is likely that small, non-significant, grade differences across several assignments did contribute to significant differences in final grades. Accordingly, we believe that even non-significant grade differences should be considered by faculty who are interested in improving equity in their classes. Additionally, we posit that inequitable patterns of assignment grades matter even in instances in which these grades do not contribute to low final grades. Low assignment grades matter because assignment grades provide students with information about how they are viewed by faculty in a discipline (Singer-Freeman & Bastone, 2019b). Low grades communicate a lack of success, which may become part of the student's academic sense of self, reducing feelings of academic self-efficacy and the student's sense of belonging. A diminishment in any of these areas can reduce persistence within a major or within an institution (Chemers et al., 2001; Han et al., 2017; Shen et al., 2016; Singer-Freeman & Bastone, 2019a, 2019b; Singer-Freeman et al., 2019).

There are several methods for creating equitable assessments. One is to accept that the transmission of knowledge is not a neutral activity (Montenegro & Jankowski, 2020) and consider positionality and agency at each phase of the assessment cycle (Heiser et al., 2017). Life experiences, privilege, and biases can influence the types of questions that are asked, what is viewed as a correct response, and the types of assessment methodologies that are selected. Each of these factors can contribute to educational equity gaps (Cumming & Dickson, 2007; Stowell, 2004). Montenegro and Jankowski (2017; 2020) suggest that when instructors dictate how students will demonstrate learning, it privileges certain types of learning over others. They encourage adopting differentiated assignments to allow students to select assignment structures that best demonstrate their mastery. Although providing students with a choice of assignments may be an effective way to increase equity, it can be impractical and make uniform grading difficult (Singer-Freeman et al., 2019, 2021).

Other approaches to increasing equity in assignments have examined ways specific forms of assessment might misrepresent the abilities of certain student groups (Sleeter, 2004) or be culturally inappropriate to underserved students (Cahill et al., 2004). We and others have begun to explore whether specific features of assignments might increase or reduce equity gaps (Harackiewicz et al., 2015; Singer-Freeman & Bastone, 2018, 2019a, 2019b, 2021;

Singer-Freeman et al., 2019, 2021; Steele & Aronson, 1995; Stiggins & Chappuis, 2005). In our work, we found that assignments often vary along two dimensions: utility value and inclusive content (Singer-Freeman et al., 2019). Utility value describes the extent to which students perceive work to have value (Eccles et al., 1983). Assignments can be professionally, academically, or personally useful. Experimental and applied work have established that increasing the utility value of assignments reduces educational equity gaps (Harackiewicz et al., 2015; Singer-Freeman & Bastone, 2019a, 2021; Singer-Freeman et al., 2019; 2021). Inclusive content describes material that is equally accessible to all students (Gay, 2010). If examples are drawn from the dominant culture, they are less accessible to students from other cultures. Structuring assignments so that content is equally familiar to all students reduces educational equity gaps by limiting the effects of prior knowledge and privilege. Providing clear and detailed instructions and grading rubrics makes content more inclusive by eliminating the benefits of prior preparation from other classes (Gay, 2010; Singer-Freeman et al., 2019, 2021). We hypothesize that increasing assignments' perceived utility value and inclusive content has the potential and power to mitigate equity gaps.

Improving equity requires faculty engagement in a culture of inquiry in which the examination of data informs responses to inequities (Bensimon, 2005; Maki, 2017). We believe the data presented in this paper are an example of the kinds of data that can be shared with faculty and students as a starting place for conversations about increasing equity in classes. As faculty review patterns of equity and inequity at the assignment level and discuss their assignments with students, they will be able to make informed changes to assignments that will increase equity. In some instances, assignments that evoke equity gaps may examine similar competencies as alternative assignments that do not evoke inequity. In these cases, faculty might consider replacing assignments that result in equity gaps with more equitable methods of assessment. In other instances, assignments that result in equity gaps may be revealing incomplete mastery of an essential learning outcome. In these cases, it might be important to consider whether all students have equal access to educational resources and prior learning. For example, if transfer students are struggling to demonstrate mastery in an area, it might be worth considering whether the course is assuming levels of prior preparation that transfer students may lack.

## Limitations and Future Directions

There were some limitations of the current work. Because this work was exploratory, we did not discuss the assignments with either students or faculty. Having relied on class syllabi and materials available in the learning management system to classify assessments, we cannot know the extent to which students viewed the assignments as being high in inclusive content or utility value and how those perceptions might have impacted student performance. Having established the importance of disaggregating assignment grades in this work, we are currently working directly with students to examine whether their views of assessments predict equity gaps. Because we did not partner with faculty, we cannot establish if the assessments with equity gaps were evaluating the same learning as assessments without equity gaps.

**As faculty review patterns of equity and inequity at the assignment level and discuss their assignments with students, they will be able to make informed changes to assignments that will increase equity.**

Finally, we did not evaluate the long-term effects of equity gaps on students. There is evidence of completion gaps in higher education (Shapiro et al., 2018). In future work, it will be important to examine how academic self-efficacy, identity, and sense of belonging are impacted by low assignment grades and whether low course and assignment grades increase the likelihood students will leave a major or fail to complete a degree.

## Conclusion

The current work found frequent equity gaps for both URM and transfer students. Importantly, equity gaps appeared to be more common in multiple-choice tests and formal writing than in other assignment types. Because patterns of equity gaps differed between final course grades and individual assignment grades, faculty should consider disaggregating grades on individual assignments. Because patterns of equity gaps varied within assignment types, future research should investigate whether specific features of assignments such as utility value and inclusive content influence the size of equity gaps. We believe that assessment professionals play a critical role in this work. Encouraging the disaggregation of student outcomes data can be the first step toward establishing a culture of inquiry in which faculty, students, and assessment professionals explore how assignments are contributing to inequities in higher education. These considerations can direct learning improvements that are sensitive to the needs of every student rather than the needs of the average student.

## References


Association of American Colleges and Universities. (2018). A Vision for Equity, results from AAC&U'S project committing to equity and inclusive excellence: Campus-based strategies for student success. https://www.aacu.org/committing-to-equity/publications

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.

Bauman, G.L., Bustillos, L. T., Bensimon, E. M., Brown, M. C., & Bartee, R. D. (2005). Achieving equitable educational outcomes with all students: The institution's roles and responsibilities. https://www.aacu.org/sites/default/files/files/mei/bauman_et_al.pdf

Bensimon, E. M. (2005). Closing the achievement gap in higher education: An organizational learning perspective. *New Directions in Higher Education*, *131*, 99-111. https://doi.org/10.1002/he.190

Cahill, M., Philpott, D., Nesbit, W., & Jeffery, G. (2004). Educational assessments of first nations students: A Review of the Literature. In W. Nesbit (ed.), *Cultural Diversity and Education: Interface Issues*. Memorial University, 77–102

Chemers, M., Hu, L., & Garcia, B. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology*, *93*(1), 55–64. https://doi.org/10.1037//0022-0663.93.1.55

College Board. (2018). More than 2 million students in the class of 2018 took the SAT, highest ever. https://www.collegeboard.org/releases/2018/more-than-2-million-students-in-class-of-2018-took-sat-highest-ever#:~:text=More%20Than%202%20Million%20Students,Highest%20Ever%20%7C%20The%20College%20Board

Cumming, J.J., & Dickson, E.A. (2007). Equity in assessment: Discrimination and disability issues from an Australian legal perspective. *Education and the Law*, *19*, 201-220. https://doi:10.1080/09539960701762854

Diaz, P. (1992). Effects of transfer on academic performance of community college students at the four-year institution. *Community Junior College Quarterly of Research and Practice*, *16*(3), 279-291. https://doi.org/10.1080/0361697920160307

Eccles J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and Achievement Motivation* (pp. 75–146). W. H. Freeman.

Fauria, R. M. & Fuller, M. B. (2015). Transfer student success: Educationally purposeful activities predictive of undergraduate GPA. *Research and Practice in Assessment*, *10*, 39-52. https://files.eric.ed.gov/fulltext/EJ1064764.pdf

Fauria, R. M., & Slate, J. R. (2014). Persistence rate differences of university students by race: A within groups comparison. *International Journal of University Teaching and Faculty Development*, *4*(1), 1-10. https://uncc.primo.exlibrisgroup.com/permalink/01UNCC_INST/ 1rqb8fi/cdi_proquest_journ als_1626178943

Fletcher, J. M. & Tienda, M. (2010). Race and ethnic differences in college achievement: Does high school attended matter? *The Annals of the American Academy of Political and Social Science*, *627*(1), 144-166. https://doi.org/10.1177/0002716209348749

Gay, G. (2010). *Culturally Responsive Teaching: Theory, Research and Practice*. Teachers College Press.

Han, C., Farruggia, S.P., & Moss, T.P. (2017). Effects of academic mindsets on college students' achievement and retention. *Journal of College Student Development 58*(8), 1119-1134. https://doi.org/10.1353/csd.2017.0089

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2015). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of personality and social psychology*, *111*(5), 745-765. https://doi.org/10.1037/pspp0000075

Harper, S. R., & Harris, F, I. (2012). Men of color: A role for policymakers in improving the status of black male students in U.S. higher education. *Institute for Higher Education Policy*, 18.

Heiser, C.A., Prince, K., & Levy, J.D. (2017). Examining critical theory as a framework to advance equity through student affairs assessment. *The Journal of Student Affairs Inquiry*, *3*(1). https://jsai.scholasticahq.com/article/1621-examining-critical-theory-as-a-framework-to-advance-equity-through-student-affairs-assessment

Ishitani, T. T. (2008). How do transfers survive after 'transfer shock'? A longitudinal study of transfer student departure at a four-year institution, *Research in Higher Education*, *49*(5), 403-419. https://doi.org/10.1007/s11162-008-9091-x

Jacobson, T., Delano, J., Krzykowski, L., Garafola, L., Nyman, M., & Barker-Flynn, H. (2017). Transfer student analysis and retention: a collaborative endeavor. *Reference Services Review*, *45*(3), 421–439. https://doi.org/10.1108/RSR-10-2016-0069

Lorah, J. & Ndum, E. (2013). Trends in achievement gaps in first-year college courses for racial/ethnic, income, and gender subgroups: A 12-year studwy. ACT. https://files.eric.ed.gov/fulltext/ED546852.pdf

Maki, P. (2017). *Real-time student assessment: meeting the imperative for improved time to degree*, *closing the opportunity gap*, *and assuring student competencies for 21st century needs*. Stylus Publishing.

McGee, J. (2015). *Breakpoint*: *The changing marketplace for higher education*. Johns Hopkins University Press.

Monarrez, T. & Washington, K. (2020). Racial and ethnic representation in postsecondary education. Urban Institute. https://www.urban.org/research/publication/racial-and-ethnic-representation-postsecondary-education/view/full_report

Montenegro, E., & Jankowski, N. A. (2017). *Equity and assessment*: *Moving towards culturally responsive assessment* (Occasional Paper No. 29). Urbana: University of Illinois and Indiana University, NILOA.

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

National Center for Education Statistics. (2019). *Digest of Education Statistics*, 2018. https://nces.ed.gov/fastfacts/display.asp?id=171

Nuñez, A. M., & Yoshimi, J. A. (2017). Phenomenology of transfer: Students' experiences at a receiving institution. *Innovative Higher Education 42*, 173–187 https://doi.org/10.1007/s10755-016-9374-7

Olson, A. (2020). *Equity audits should be commonplace*. Inside Higher Education. https://www.insidehighered.com/views/2020/03/25/more-colleges-should-use-equity-audits-address-inequalities-their-institutions.

Rosenbaum, J., Ahearn, C. & Becker, K. (2015). *The new forgotten half and research directions to support them*. William T. Grant Foundation. https://wtgrantfoundation.org/library/uploads/2015/09/The-New-Forgotten-Half-and-Research-Directions-to-Support-Them.pdf

Shapiro, D., Dundar, A., Huie, F., Wakhungu, P.K., Bhimdiwala, A. & Wilson, S. E. (2018). Completing college: A national view of student completion rates – Fall 2012 Cohort (Signature Report No. 16). Herndon, VA: National Student Clearinghouse Research Center. https://nscresearchcenter.org/signaturereport16/

Shen, C., Miele, D. B., & Vasilyeva, M. (2016). The relation between college students' academic mindsets and their persistence during math problem solving. *Psychology in Russia*: *State of the Art*, *9*(3), 38–56. https://doi.org/10.11621/pir.2016.0303

Singer-Freeman, K. E., & Bastone, L. (2018). ePortfolio and declarations of academic self: A tale of two contexts. In Eynon and Gambino (Eds.), *Catalyst in Action*: *Case Studies of High-Impact ePortfolio Practice*. Stylus Publishing, LLC.

Singer-Freeman, K. E., & Bastone, L. (2019a). Increasing equity in general education using self-relevant writing. *Intersections*: *A Journal at the Intersection of Assessment and Learning*. 24-27. https://www.aalhe.org/assets/docs/Intersection/AAHLE_fall_2019_Intersection_final.pdf

Singer-Freeman, K. E. & Bastone, L. (2019b) Developmental science concepts guide effective support of underrepresented STEM students. *Journal of Biochemistry and Molecular Biology Education*. *47*(5) 506-512. https://doi.org/10.1002/bmb.21292

Singer-Freeman, K. E. & Bastone, L. (2021). Social sciences general education: Incorporating self-relevant academic writing using ePortfolios. In D. Kelly-Riley & E. Norbert (Eds.), *Improving Outcomes: Disciplinary Writing, Local Assessment, and the Aim of Fairness*. Modern Language Association: New York.

Singer-Freeman, K. E., Hobbs, H., & Robinson, C. (2019). Theoretical matrix of culturally relevant assessment. *Assessment Update*, *31*(4). https://doi.org/10.1002/au.30176

Singer-Freeman, K. E., Robinson, C., & Bastone, L. (2021). Balancing the freedom to teach with the freedom to learn: The critical role for assessment and accreditation. In P. Blessinger and E. Sengupta (Eds.). *Innovations in Higher Education Teaching and Learning: Academic Freedom in Teaching and Learning Practices*. Emerald Group Publishing. 39-51. https://doi.org/10.1108/S2055-364120200000034005

Sleeter, C. (2004) Critical multicultural curriculum and the standards movement. *English Teaching: Practice and Critique*, *3*(2), 122–138. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.539.4348&rep=rep1&type=pdf

Spenner, K., Buchmann, C., & Landerman, L. (2004). The Black-White achievement gap in the first college year: Evidence from a new longitudinal case study. *Research in Social Stratification and Mobility*, *22*, 187-216. https://doi.org/10.1016/S0276-5624(04)22007-8

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797-811. https://doi.org/10.1037/0022-3514.69.5.797

Stiggins, R. & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory into Practice*, *44*(1), 11-18. https://doi.org/10.1207/s15430421tip4401_3

Stowell, M. (2004). Equity, justice and standards: Assessment decision making in higher education. *Assessment & Evaluation in Higher Education*, *29*(4), 495-510. https://doi.org/10.1080/02602930310001689055

The Aspen Education & Society Program and the Council of Chief State School Officers. (2017). Leading for equity: Opportunities for state education chiefs. https://ccsso.org/sites/default/files/2018-01/Leading%20for%20Equity_011618.pdf

Zwick, R. (2019). Assessment in American higher education: The role of admissions tests. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 130-148. https://doi.org/10.1177/0002716219843469

## Abstract

A Faculty Encouragement Scale (FES) was created to measure students' perception of faculty encouragement (challenge-focused and potential-focused encouragement) in college. This paper reports the psychometric properties of scores from the FES using a sample of 237 first-year engineering undergraduate students in a suburban public university. Analyses were conducted in both confirmatory factor analysis (CFA) and exploratory structural equation modeling (ESEM) frameworks, where CFA constrains all cross-loadings to be zero, but ESEM estimates all cross-loadings. Both CFA and ESEM analyses suggested two-factor models had better goodness-of-fit than one-factor models. However, we discovered a high factor correlation in CFA model that could result from forced-zero cross-loadings. We chose the ESEM model over CFA model because the estimate of factor correlation in CFA model might be inflated. Moreover, we found one item was closely loaded on both encouragement factors. Considering a high communality of this item, we did not suggest a further revision.



### AUTHORS

Hsien-Yuan Hsu, Ph.D.
*University of Massachusetts Lowell*

Yanfen Li, Ph.D.
*University of Massachusetts Lowell*

Sandra Acosta, Ph.D.
*Texas A&M University*


# Psychometric Properties of the Faculty Encouragement Scale with First-year Engineering Undergraduate Students

The culture of highly competitive classrooms can create a harsh learning environment that discourages first-year engineering students from pursuing of an engineering degree (National Academies of Sciences Engineering and Medicine, 2016). Not surprisingly, one of the most pressing concerns in engineering education is that the percentage of engineering students who persist beyond the first year has remained stagnant (American Society for Engineering Education, 2016). To address this issue, extensive research has been conducted to identify factors relating to student persistence. The Social Cognitive Career Theory (SCCT) choice model (Lent et al., 1994) built on the foundations of Bandura's (1986, 1991, 1997) social cognitive theory is a theoretical framework developed to reveal the persistence of racial/ethnic minorities in STEM (science, technology, engineering, and mathematics; Lent & Brown, 2019; Lent et al., 2018). As summarized by Lent and Brown (2019), previous meta-analyses of the SCCT choice model has consistently found that both self-efficacy and outcome expectations can promote students' interest (e.g., interest in performing various engineering-related activities), choice goal (e.g., intent to declare engineering degree), and action (e.g., persistence to the second year). Therefore, a better understanding of the sources of self-efficacy and outcome expectations may hold valuable implications for persistence among engineering majors.


### CORRESPONDENCE

**Email**
hsienyaun_hsu@uml.edu

In the SCCT choice model, self-efficacy refers to beliefs about one's ability to successfully perform particular behaviors or courses of action, while outcome expectations are beliefs about the consequences of given actions (Lent et al., 2008). Both self-efficacy and outcome expectations are informed by four types of sources: (a) previous personal performance accomplishments; (b) vicarious learning (or modeling); (c) verbal persuasion (e.g., supportive messages from significant others); and (d) physiological and affective states (Bandura, 1997; Lent & Brown, 2006; Sheu et al., 2018). Prior meta-analyses (Byars-Winston et al., 2017; Sheu et al., 2018) have found that performance accomplishments — strongly influential, vicarious learning and social persuasion — were moderately compelling, while affective arousal was only weakly related to self-efficacy and outcome expectations. Among these strongly or moderately impactful sources, verbal persuasion has received increasing attention in engineering education because it is a manipulable factor that can be facilitated by people who are significant to students (Bandura, 1991, 1993). Particularly, in college context, faculty are usually perceived as authority figures by students. Verbal persuasion from faculty is considered a critical source of students' self-efficacy beliefs (Garriott et al., 2021; Wong, 2015; Wong et al., 2019).

At present, current instruments do not allow researchers to measure verbal persuasion from faculty as an independent construct. In most existing instruments, verbal persuasion is considered a single-faceted construct, measured by asking students whether they received encouraging messages about their academic capabilities from multiple significant others, such as parents, teachers, peers, and other adults (Usher & Pajares, 2008). For example, the verbal persuasion scale proposed by Lent et al.'s (1991), one of the most popular instruments for measuring verbal persuasion, seeks college students' level of agreement with 10 statements on a 5-point scale (1 = strongly disagree; 5 = strongly agree). Two statements in the scale are related to friends (e.g., "My friends have discouraged me from taking math classes"), two are related to parents (e.g., "My parents have encouraged me to be proud of my math ability"), two are related to high school teachers (e.g., "Teachers have discouraged me from pursuing occupations that require a strong math background"), three are related to unspecified significant others (e.g., "Other people generally see me as being poor at math."), and only one statement is related to faculty (e.g., My adviser has singled me out as having good math skills and has encouraged me to take college math courses."). Because verbal persuasion is conceptually defined as a single-faceted construct, a composite score is computed to indicate students' perception of verbal persuasion. Other instruments, developed to capture college students' verbal persuasion in the SCCT framework, also adopted a similar measuring approach [e.g., Schaub's (2004) *Learning Experiences Questionnaire* and Garriott et al.'s (2021) *Engineering Learning Experiences Scale*].

**An instrument for measuring verbal persuasion from faculty in college environments has not been well developed. This study attempts to close this gap by creating a scale that permits researchers to measure verbal persuasion from faculty.**

Yet, an instrument for measuring verbal persuasion from faculty in college environments has not been well developed. This study attempts to close this gap by creating a scale that permits researchers to measure verbal persuasion from faculty. Specifically, this study used faculty encouragement as an indicator of verbal persuasion. The rationale underlying the use of faculty encouragement as an indicator for measuring faculty verbal persuasion is presented below.

## Verbal Persuasion from Faculty can be Measured as an Independent Construct

Prior studies using samples of middle school students have shown verbal persuasions from specific informants (e.g., parents, teachers, peers) can be measured and distinguished. Specifically, Falco and Summers (2021) modified Usher and Pajares' (2009) measure of middle school students' verbal persuasions by clearly specifying the informants (e.g., teachers, peers, and adult family members) rather than using confusing language (e.g., other or adult). Falco and Summers found verbal persuasions from teachers, peers, and family members were distinguishable and that three verbal persuasions exert distinct and unique impact on the math self-efficacy of middle school students. Similarly, Gebauer et al. (2020) reworked items in Lent et al.'s (1991) verbal persuasion scale by asking 7th graders to explicitly answer items with reference to their parents, teachers and classmates, and friends who are not classmates, respectively. Results from Gebauer et al. (2020) established that the multifaceted structure of verbal persuasions from different informants showed unique positive effects on student academic self-efficacy. These findings support the notion that students can differentiate between informants of verbal persuasion, and verbal persuasion from faculty can be measured as an independent construct.

## Use Encouragement as an Indicator of Verbal Persuasion

As in previous studies (e.g., Anderson & Betz, 2001; Lent et al., 1991; Loo & Choy, 2013), we expected the measure of verbal persuasion from faculty would focus on the positive side of verbal persuasion even though the verbal persuasion could be positive or negative (e.g., doubt in an individual's capabilities). Thus, we proposed to use *encouragement* as an indicator of verbal persuasion. Encouragement refers to messages of affirmation and motivation enhancement (Wong, 2015). The construct of encouragement from faculty has been used and studied in previous engineering education research. However, it was ambiguously defined, and the measures were not necessarily about positive verbal persuasion. For example, Branch et al. (2015) viewed perceived faculty encouragement as a form of social support that provides students with positive feedback regarding their belonging and performance, while some statements measuring encouragement from faculty were actually asking students to rate the resources provided by faculty (e.g., "I've been provided with opportunities to pursue research").

**The construct of encouragement from faculty has been used and studied in previous engineering education research. However, it was ambiguously defined, and the measures were not necessarily about positive verbal persuasion.**

## The Definition of Encouragement and Wong et al.'s (2019) Academic Encouragement Scale

Unlike previous studies, we adopted the definition of encouragement provided in Wong's (2015) Tripartite Encouragement Model (TEM). The conceptual basis of TEM is drawn in part from the psychology of character strengths and virtues, Bandura's (1997) concept of verbal persuasion, and some Adlerian conceptual insights on encouragement (Wong, 2015). Wong's (2015) TEM describes three facets of encouragement: (1) foci of encouragement, (2) features of effective encouragement, and (3) levels of encouragement. The first facet posits two foci of encouragement—challenge-focused encouragement and potential-focused encouragement—providing a theorical framework for the two-factor structure of encouragement in the academic context. The second facet describes the features influencing the extent to which encouragement produces positive outcomes for recipients (e.g., encouragement is more effective in fostering self-efficacy when it commutes recipients' effort or strategy). The third facet of TEM distinguishes three levels of encouragement: interpersonal communication, character strength (e.g., some people are more effective encouragers than others), and group norms (some groups/settings are more encouraging than others). Note the second and third facets are not directly related to the factor structure of encouragement nor to the definition of the encouragement, but they could potentially inform the design of future investigations and/or the design of faculty development programs focusing on providing effective encouragement. In TEM, Wong (2015) defined encouragement as "the expression of affirmation through language or other symbolic representations to instill courage, perseverance, confidence, inspiration, or hope in a person(s) within the context of addressing a challenging situation or realizing a potential" (p.180).

Based on foci of encouragement, Wong et al. (2019) further developed the *Academic Encouragement Scale* (AES) to measure college students' perception of encouragement. AES proposes five statements to measure challenge-focused encouragement (e.g., instilling hope in students when they feel like giving up on an academic task) and five to measure potential-focused encouragement (e.g., noticing that students are doing well in school and encouraging them to dream bigger and to aim higher). Using a sample of 714 undergraduate students, Wong et al. (2019) found both exploratory factor analysis and confirmatory factor analysis (CFA) favored the two-factor structure of academic encouragement. The Cronbach's alpha coefficients of the challenge-focused encouragement and potential-focused encouragement were .93 and .90, respectively. The correlation between two encouragement factors was extremely high (.94). Wong et al. (2019) also regressed college students' academic self-efficacy (i.e., student's degree of confidence on successfully completing a college-related task such as taking notes or asking a question in class) on two encouragement factors. Wong et al. (2019)

found academic self-efficacy was positively and significantly predicted by both challenge- and potential-focused encouragement after controlling for each other's effects. It should be noted that AES was designed to measure students' perception of encouragement from their significant others in a generic academic setting. In other words, AES did not specify the informants of encouragement and, thus, cannot be used to measure encouragement from faculty.

## The Creation of Faculty Encouragement Scale

**The intent of the present study is to report the psychometric properties of scores from the FES using a sample of first-year engineering undergraduate students in a suburban public university.**

Because we specifically focused on encouragement from faculty rather than a broad interest in academic encouragement, we created a modified version of AES, the *Faculty Encouragement Scale* (FES). More specifically, items of AES were drafted to measure encouragement from significant others in a generic academic setting (e.g., *Someone I respect encouraged me to believe in myself when I doubted my academic abilities*). We created the FES by specifying the informant of encouragement as faculty members in each item of AES (e.g., *An engineering professor I respect, or I am familiar with encouraged me to believe in myself when I doubted my academic abilities*). Although the intent of the current version of FES specified "engineering professor" as the informant of encouragement, the FES could be used to measure encouragement from faculty in different disciplines (e.g., changing "engineering professor" to "chemistry professor"). The FES is provided in Figure 1.

Figure 1
*Faculty Encouragement Scale*

Instructions/Items:
*Please recall your experiences of interacting with engineering professors at [Name of University]. For each statement, please decide how accurately it describes your situation by checking the box that precedes it. An engineering professor I respect, or I am familiar with* ______________.

FE_C1. Encouraged me to believe in myself when I doubted my academic abilities.

FE_C2. Instilled hope in me when I felt like giving up on an academic task.

FE_C3. Reminded me of my strengths when I was discouraged about a challenging academic task.

FE_C4. Assured me that I was competent in dealing with my academic difficulties.

FE_C5. Expressed confidence in me and told me to keep trying in school even though it was hard.

FE_P1. Pointed out my strengths when she/he suggested I pursue a new academic opportunity.

FE_P2. Noticed I was doing well in school and encouraged me to dream bigger and aim higher.

FE_P3. Insisted that I should strive for higher academic standards because I was capable.

FE_P4. Explained why I had the skills to succeed in school at an advanced level.

FE_P5. Said something positive to motivate me to consider a new academic goal.

*Note*. FE_C = Challenge-focused faculty encouragement. FE_P = Potential-focused faculty encouragement.

## Purpose of This Study

The intent of this study is to report the psychometric properties of scores from the FES using a sample of first-year engineering undergraduate students in a suburban public university. Researchers hypothesized that there would be two underlying factors (challenge-focused and potential-focused encouragement) as stated in TEM (Wong, 2015). Regarding criterion validity, it was hypothesized that both factors would demonstrate significantly positive relationships with students' self-efficacy, as posited by the SCCT (Lent & Brown, 2019). Since the FES was the first attempt to measure encouragement from faculty based on TEM, we did not have any expectation that one of the encouragement from faculty would exhibit a stronger relationship with self-efficacy than the other encouragement; however, we did expect challenge-focused encouragement and potential-focused encouragement would differently correlate with self-efficacy if the two encouragements were distinct.

## Contribution of The Study

Our study not only contributes to the SCCT literature but also has meaningful implications in engineering education. Because of the absence of instruments for measuring verbal persuasion from faculty, the existing SCCT literature still lacks rigorous evidence illustrating the specific role of verbal persuasion from faculty members on students' self-efficacy. The availability of the instrument could extend our understanding of the determinants of engineering students' self-efficacy beyond current SCCT studies. On the other hand, some engineering faculty members consider student attrition in the first year of an engineering program to be the result of weeding out under-prepared or unmotivated students. Consequently, these faculty members continue to overlook or denigrate their influence on student persistence. Clearly, identifying the relationship between verbal persuasion from faculty and students' self-efficacy beliefs would provide an empirical foundation for explaining why and how verbal persuasion from faculty matters. In turn, these understandings could change faculty attitudes towards their role in student persistence and, ultimately, inform faculty actions accordingly.

# Method

## Participants and Procedures

The present study was conducted at a midsized public four-year university in Massachusetts, USA. Institutional review board approval has been obtained from the institution research team. Data were collected using the Qualtrics online survey tool. First-year engineering students enrolled in the fall 2019 semester were eligible for the study. The data collection period was from October 31, 2019 to December 6, 2019. The FES was completed by 237 students. Diversity breakdown of the sample was 34.05% female, with 63.98% of students identifying as White, 11.44% Asian, 8.90% Hispanic/Latinx, 4.66% Black/African American, and 11.02% multiracial.

## Measures

*Faculty Encouragement Scale* (FES). The FES was designed to measure students' perception of faculty encouragement in college contexts. The FES comprises 10 items: five items devoted to challenge encouragement (e.g., *Instilled hope in me when I felt like giving up on an academic task*), and five describing potential encouragement (e.g., *Said something positive to motivate me to consider a new academic goal*). In this study, the intent of the FES is to measure engineering students' perceived encouragement from engineering faculty. The FES asks students to recall their interactions with an engineering professor whom they respect or are familiar with and to indicate how accurately the 10 items in the FES describe their situations on a 6-point scale (1 = very untrue of me; 6 = very true of me).

**Identifying the relationship between verbal persuasion from faculty and students' self-efficacy beliefs would provide an empirical foundation for explaining why and how verbal persuasion from faculty matters. In turn, these understandings could change faculty attitudes towards their role in student persistence and, ultimately, inform faculty actions accordingly.**

**Self-efficacy Scales**. In this study, two types of self-efficacy beliefs (Lent et al., 2008) were measured. The academic milestone self-efficacy scale, measuring students' confidence in their ability to complete academic requirements, comprises four items (e.g., How much confidence do you have in your ability to excel in your engineering major over the next semester). On the other hand, coping self-efficacy, which assesses students' confidence in their ability to cope with a variety of barriers that engineering students might experience, was measured on a 7-item scale (e.g., How confident are you that you could find ways to overcome communication problems with professors or teaching assistants in engineering courses?). Both self-efficacy scales were measured on a 9-point scale, from no confidence (1) to complete confidence (9).

In this study, the mean and standard deviation of the academic milestones self-efficacy composite scores were 6.551 and 1.587, respectively, and those of coping efficacy composite scores were 6.448 and 1.391, respectively. Cronbach's α estimates for academic milestones efficacy and coping efficacy were .91 and .88, respectively. In addition, we also found that the two-factor structure of self-efficacy was supported by a two-factor confirmatory factor analysis model, where academic milestone self-efficacy and coping efficacy loaded on corresponding items [$\chi^2(df)$= 63.074(42), $p$ <.05, *CFI*=.979, *TLI*=.973, *RMSEA*=.046, *SRMR*=.043]. The correlation between academic milestones efficacy and coping self-efficacy in CFA was 0.732 (SE = 0.045, $p$ < .05).

## Data Analysis

The FES was predicted to have a two-factor structure (challenge-focused and potential-focused encouragement). To examine the factor structure of FES, data were fitted to both a one-factor model (a competitive model) and two-factor model. The model fit was evaluated by using chi-square ($\chi^2$) statistics and fit indices-comparative fit index (CFI), Tucker-Lewis Index (TLI), root-mean-square-error-of-approximation (RMSEA), and standardized root mean squared residual (SRMR) with cutoff values (CFI, TLI ≥ .95; RMSEA ≤ .06; SRMR ≤ .08; Hu & Bentler, 1999). Additionally, $\chi^2$ difference ($\Delta\chi^2$) tests were conducted to compare relative fit of a one-factor model versus a two-factor model (Satorra & Bentler, 2010). Further, Akaike's Information Criterion (AIC; Akaike, 1973), Bayesian IC (BIC; Schwarz, 1978), and sample-size adjusted BIC (SABIC; Sclove, 1987) were applied to assist model selection. A model with relatively smaller values of IC indices is preferred.

**The FES was predicted to have a two-factor structure (challenge-focused and potential-focused encouragement). To examine the factor structure of FES, data were fitted to both a one-factor model (a competitive model) and two-factor model.**

Analyses were conducted in both CFA and exploratory structural equation modeling (ESEM) frameworks, where CFA constrains all cross-loadings to be zero, but ESEM estimates all cross-loadings. If the ESEM solution is not clearly superior, the CFA solution, which is more parsimonious (fewer free parameters), is adopted to determine the factor structure of FES (Morin et al., 2013). However, previous studies have shown that forcing cross-loadings to zero in CFA might result in inflated factor correlations (Hsu et al., 2014; Liang et al., 2020). Therefore, when the CFA model fit is good and approaches that of the ESEM, "the sizes of the factor correlations are a primary justification for choosing ESEM over CFA" (Marsh et al., 2020, p. 114). As presented in the results section, we used Marsh et al.'s (2020) recommendation because we experienced a similar model fit in two-factor CFA and ESEM, but the factor correlation in CFA was unusually high, which persuaded us to choose a two-factor ESEM. Detailed information about model selection is provided in the next section. Table 1 presents the means, standard deviations (SDs), and correlations of 10 FES items that were used for CFA and ESEM analyses. Note the target rotation method was applied in ESEM analysis (Marsh et al., 2020).

We evaluated the criterion validity by regressing academic milestone self-efficacy and coping self-efficacy on challenge-focused encouragement and potential-focused encouragement, controlling for students' demographic information (gender, age, English is the primary language at home, first-generation students, and transfer students) in one regression model using SEM. Criterion validity was supported when the slopes of challenge-focused encouragement and potential-focused encouragement were positive and statistically significant. All analyses were conducted using the structural equational modeling method in *Mplus* 8. Score reliability was indicated by Cronbach's $\alpha$.

## Results

Table 2 presents values of model fit indicators for one-factor and two-factor models, as well as model comparison results in CFA and ESEM frameworks. Both CFA and ESEM analyses suggested two-factor models had better goodness-of-fit than one-factor models (i.e., greater *CFI* and *TLI*, smaller *RMSEA* and *SRMR*, and smaller AIC, BIC, and saBIC). Results of model comparison ($\Delta\chi^2$) showed that the one-factor model demonstrated a model fit significantly worse than that of the two-factor model in CFA ($\Delta\chi^2$ ($df$)=23.589(1), $p$<.05) and in ESEM (($\Delta\chi^2(df)$=91.958(1), $p$<.05). The two-factor factor structure of FES was supported in this study.

Results showed the two-factor CFA model [$\chi^2$(df)= 84.669(34), $p$ <.05, *CFI*=.960, *TLI*=.948, *RMSEA*=.079, *SRMR*=.027] and the two-factor ESEM model [$\chi^2(df)$= 71.082(26), $p$ <.05, *CFI*=.965, *TLI*=.939, *RMSEA*=.086, *SRMR*=.015] had an adequate and similar model fit. Both two-factor CFA model and ESEM model had an *RMSEA* value slightly greater than .06, which is still considered an indication of fair fit (MacCallum et al., 1996). Table 3

Table 1
*Descriptive Statistics and Correlations for Items of Faculty Encouragement Scale*

| Item | Mean (*SD*) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) FE_C1 | 3.597(1.465) | - | | | | | | | | | |
| (2) FE_C2 | 3.534(1.471) | .878 | - | | | | | | | | |
| (3) FE_C3 | 3.415(1.463) | .851 | .827 | - | | | | | | | |
| (4) FE_C4 | 3.540(1.442) | .808 | .809 | .794 | - | | | | | | |
| (5) FE_C5 | 3.591(1.466) | .805 | .832 | .827 | .833 | - | | | | | |
| (6) FE_P1 | 3.264(1.487) | .746 | .741 | .772 | .703 | .790 | - | | | | |
| (7) FE_P2 | 3.284(1.532) | .731 | .716 | .757 | .646 | .760 | .782 | - | | | |
| (8) FE_P3 | 3.366(1.522) | .758 | .776 | .775 | .760 | .774 | .761 | .859 | - | | |
| (9) FE_P4 | 3.302(1.472) | .796 | .788 | .821 | .748 | .773 | .765 | .817 | .863 | - | |
| (10) FE_P5 | 3.708(1.572) | .731 | .765 | .792 | .743 | .818 | .695 | .714 | .777 | .780 | - |

*Note*. n = 235. All correlation coefficients were statistically significant ($p < .05$). FE_C = Challenge-focused faculty encouragement. FE_P = Potential-focused faculty encouragement.

Table 2
*Model Fit of One-factor and Two-factor Model in CFA and ESEM and Model Comparison Results*

| Model | $\chi^2$(*df*) | *CFI* | *TLI* | *RMSEA* | *SRMR* | AIC | BIC | saBIC | $\Delta\chi^2$ (*df*): One-factor versus two-factor |
|---|---|---|---|---|---|---|---|---|---|
| *CFA* | | | | | | | | | |
| One-factor | 125.827(35), $p$ <.05 | .929 | .909 | .105 | .030 | 5879.879 | 5983.920 | 5888.831 | 23.589(1), $p$ <.05 |
| Two-factor | 84.669(34), $p$ <.05 | .960 | .948 | .079 | .027 | 5792.729 | 5900.239 | 5801.980 | |
| *ESEM* | | | | | | | | | |
| One-factor | 125.827(35), $p$ <.05 | .929 | .909 | .105 | .030 | 5879.879 | 5983.920 | 5888.831 | 91.958(9), $p$ <.05 |
| Two-factor | 71.082(26), $p$ <.05 | .965 | .939 | .086 | .015 | 5754.335 | 5889.590 | 5765.973 | |

presents two-factor CFA model and ESEM model solutions. In the CFA model, standardized factor loadings of challenge-focused encouragement ranged from 0.881 to 0921, while those of potential-focused encouragement ranged from 0.847 to 0.925. The communalities of indicators (i.e., proportion of indicator's variance that can be explained by the factor) ranged from 0.717 to 0.856. The factor correlation in CFA ($\gamma$= 0.935, $\gamma^2$= 87.42%) was exceptionally high, suggesting the two factors were not distinguishable. The size of factor correlation in CFA could be overestimated due to fix- to-zero cross loadings.

Table 3
*Two-factor CFA Model and ESEM Model Solutions*

| | CFA Solution Standardized Factor Loadings (and *SE*) | | | ESEM Solution Standardized Factor Loadings (and *SE*) | | |
|---|---|---|---|---|---|---|
| | Challenge-focused Encouragement | Potential-focused Encouragement | Communalities | Challenge-focused Encouragement | Potential-focused Encouragement | Communalities |
| FE_C1 | 0.920(0.015)* | - | 0.846 | 0.710(0.100)* | 0.280(0.104)* | 0.844 |
| FE_C2 | 0.921(0.015)* | - | 0.848 | 0.729(0.066)* | 0.264(0.066)* | 0.855 |
| FE_C3 | 0.915(0.022)* | - | 0.837 | 0.611(0.052)* | 0.387(0.057)* | 0.834 |
| FE_C4 | 0.881(0.024)* | - | 0.776 | 0.734(0.098)* | 0.214(0.096)* | 0.792 |
| FE_C5 | 0.908(0.018)* | - | 0.825 | 0.618(0.079)* | 0.375(0.077)* | 0.827 |
| FE_P1 | - | 0.847(0.031)* | 0.717 | 0.357(0.094)* | 0.573(0.095)* | 0.724 |
| FE_P2 | - | 0.889(0.025)* | 0.790 | 0.055(0.131) | 0.895(0.139)* | 0.869 |
| FE_P3 | - | 0.925(0.017)* | 0.856 | 0.215(0.113) | 0.773(0.116)* | 0.862 |
| FE_P4 | - | 0.923(0.016)* | 0.852 | 0.328(0.104)* | 0.663(0.103)* | 0.834 |
| FE_P5 | - | 0.848(0.033)* | 0.718 | 0.488(0.105)* | 0.449(0.106)* | 0.728 |
| | Factor correlation = 0.935 (*SE* = 0.016, $p$<.05, $\gamma^2$ = 87.42%) | | | Factor correlation = 0.657 (*SE* = 0.030, $p$<.05, $\gamma^2$ = 43.16%) | | |

*Note*. $^*p$<.05. FE_C = Challenge-focused faculty encouragement. FE_P = Potential-focused faculty encouragement.

**Based on Marsh et al.'s (2020) recommendation, we chose the ESEM model over CFA model because the estimate of factor correlation in CFA model might be inflated. As a result, the two-factor ESEM model was selected as a final model.**

Alternatively, in the ESEM model, indicators FE_C1 to FE_C5 were mainly loaded on challenge-focused encouragement with standardized factor loadings ranging from 0.611 to 0.724, and were cross-loaded on potential-focused encouragement with relatively smaller and statistically significant standardized factor loadings (ranging from 0.214 to 0.387). On the other hand, FE_P1 to FE_P4 were mainly loaded on potential-focused encouragement with standardized factor loadings ranging from 0.573 to 0.895, whereas only FE_P1 and FE_P4 were statistically significantly cross-loaded on challenge-focused encouragement (standardized factor loading was 0.357 to 0.328, respectively). Note FE_P5 was closely loaded on challenge-focused encouragement (loading = 0.488) and potential-focused encouragement (loading = 0.449). The communalities of indicators ranged from 0.724 to 0.862. The factor correlation in ESEM was moderate ($\gamma$= 0.657, $\gamma^2$= 43.16%).

Although the two-factor CFA model was more parsimonious (i.e., cross-loading were constrained to zero) than two-factor ESEM model, we found the factor correlation in CFA was close to 1, which could result from fixed-to-zero cross loadings. Based on Marsh et al.'s (2020) recommendation, we chose the ESEM model over CFA model because the estimate of factor correlation in CFA model might be inflated. As a result, the two-factor ESEM model was selected as a final model. Cronbach's α results for challenge-focused and potential-focused encouragement were 0.959 and 0.946, respectively. In order to examine the criterion validity of FES, we further regressed academic milestones self-efficacy and coping self-efficacy on challenge-focused faculty encouragement and potential-focused faculty encouragement, controlling for students' demographic information.

**Particularly, challenge-focused faculty encouragement demonstrated statistically significant predictive power to academic milestones self-efficacy and coping self efficacy. In contrast, potential-focused faculty encouragement could not statistically significantly predict academic milestones self-efficacy.**

Note that in the regression model, self-efficacy factors were specified as CFA model and encouragement factors were specified as ESEM model. Table 4 presents the results of regression analysis and descriptive statistics of two self-efficacy variables and two encouragement variables. Results suggested that the regression model had a satisfactory model fit [$\chi^2(df)$= 421.199 (287), $p$ <.05, $CFI$=.962, $TLI$=.955, $RMSEA$=.046, $SRMR$=.054]. $R^2$ academic milestones self-efficacy (0.227) was similar to that of coping self-efficacy (0.223). Two encouragement factors predicted self-efficacy variables in different patterns. Particularly, challenge-focused faculty encouragement demonstrated statistically significant predictive power to academic milestones self-efficacy (standardized slope = 0.392) and coping self-efficacy (standardized slope = 0.216). In contrast, potential-focused faculty encouragement could not statistically significantly predict academic milestones self-efficacy (standardized slope = 0.083) but was able to predict coping self-efficacy (standardized slope = 0.219).

Table 4

*Regression Results and Descriptive Statistics*

| *Regression Results* | Outcome = academic milestones self-efficacy | Outcome = coping self-efficacy |
|---|---|---|
| *Predictor* | Standardized slope (and *SE*) | Standardized slope (and *SE*) |
| FE_C | 0.392(0.090)* | 0.216(0.109)* |
| FE_P | 0.083(0.094) | 0.219(0.110)* |
| $R^2$= | 0.227 | 0.223 |

*Descriptive Statistics*

| | Mean (*SD*) | (1) | (2) | (3) | (4) |
|---|---|---|---|---|---|
| (1) FE_C | 3.521(1.349) | - | | | |
| (2) FE_P | 3.385(1.375) | .658* | - | | |
| (3) Academic milestones efficacy | 6.551(1.587) | .429* | .324* | - | |
| (4) Coping self-efficacy | 6.448(1.391) | .337* | .329* | .733* | - |

*Note*. *$p$<.05. FE_C = Challenge-focused faculty encouragement.
FE_P = Potential-focused faculty encouragement.

## Discussion

In this study, the Faculty Encouragement Scale was created for measuring encouragement received by students from a specific informant (i.e., faculty). The design of FES was built on Wong's (2015) Tripartite Encouragement Model that articulates two foci of encouragement – challenge-focused and potential-focused encouragement. According to SCCT choice mode, encouragement from faculty captured by the FES is a positive side of verbal persuasion received from faculty and thus was hypothesized to be correlated to students' self-efficacy beliefs.

Factor analysis results (Table 2) suggested that one-factor model demonstrated a model fit significantly worse than two-factor. That is, the two-factor structure in FES was better supported by the data. This finding echoes Wong's (2015) TEM that proposes encouragement in the academic context can be either challenge-focused encouragement or potential-focused encouragement. When determining the final two-factor model, we found the correlation between challenge-focused encouragement and potential-focused encouragement in CFA was extremely high ($\gamma$= 0.935, $\gamma^2$= 87.42%), while that in ESEM was moderate ($\gamma$= 0.657, $\gamma^2$= 43.16%), although both two-factor CFA and ESEM models had similar model fit. In this case, as Marsh et al. (2020) suggested, it was very likely that the factor correlation in CFA model was inflated due to fixed-to-zero cross-loadings. Therefore, we choose a two-factor ESEM model as our final model. This finding suggested that researchers should not blindly ignore cross-loadings. In fact, as pointed out by Hsu et al. (2014), constraining cross-loadings to zero might become a model misspecification when cross-loadings were not ignorable. Our two-factor ESEM model results (Table 3) suggested only two indicators had non-significant cross-loadings and forced zero cross-loadings. Future studies are needed to investigate whether ESEM model is preferred using data collected from other samples.

Our findings could be used to explain the reason why Wong et al. (2019) derived a high correlation between challenge-focused encouragement and potential-focused encouragement ($\gamma$= 0.94) when using the Academic Encouragement Scale to measure two encouragement factors in a generic academic setting. High factor correlation in Wong et al.'s (2019) work was derived by using CFA model, which could lead to inflated factor correlation, thus, raising a concern that the two factors might be redundant.

In general, the two-factor ESEM model solution presented a simple factor structure, where all indictors were mainly loaded on one factor except for FE_P5 ("*Said something positive to motivate me to consider a new academic goal*"). Specifically, FE_P5 was statistically significantly loaded on both challenge-focused encouragement (loading = 0.488) and potential- focused encouragement (loading = 0.449) with comparable loadings. The communality of FE_P5 was 0.728, meaning 72.8% of variance in FE_P5 can be explain by two encouragement factors jointly. This result suggested FE_P5 was not a unidimensional indicator, however; two encouragement factors explained most variance. This finding made sense because "considering a new academic goal" can mean either adjusting the goal when students encounter a challenge or setting up a higher goal as a recognition of a student's potential. Considering a high communality of FE_P5, we did not suggest revising this item. Instead, we recommend ESEM be utilized for FES data so that cross-loadings could be appropriately modeled.

**These findings not only supported the criterion validity of FES, but also suggested challenge-focused faculty encouragement and potential-focused faculty encouragement were distinguishable factors.**

Furthermore, we found challenge-focused faculty encouragement statistically significantly predicted both academic milestones self-efficacy (standardized slope = 0.392) and coping self-efficacy (standardized slope = 0.216). Potential-focused faculty encouragement only statistically significantly predicted coping self-efficacy (standardized slope = 0.219). Those findings not only supported the criterion validity of FES, but also suggested challenge-focused faculty encouragement and potential-focused faculty encouragement were distinguishable factors. Future studies are needed to investigate this issue further.

A few limitations of the current study provide a window into other future research needs. First, the findings of this study were derived from a small sample size (n = 237). Future studies are encouraged to validate our findings using a larger sample size. Increasing the sample size not only enhances the quality of parameter estimates in data analysis but also permits more extensive analysis (e.g., testing the measurement invariance of FES among gender groups and racial/ethnic groups). Second, suggested by SCCT choice model, two students' self-efficacy beliefs were collected to test the criterion validity of FES. Future studies are needed to better validate evidence to foster understanding of the discrimination and validity of two types of encouragement in FES.

Third, each of the measures, including FES and two self-efficacy scales, used a self-reported Likert-scale. Although the data were collected through an anonymous online survey, it is possible students hid true feelings when replying to the survey. Future studies need to ensure a safe and secure space for students when measuring these psychological factors. Fourth, nesting in data could occur due to groups of students taught by the same faculty members. This study did not ask students to identify the names of faculty giving the encouragement. Future studies could take into account the nesting in data by applying multilevel analytical approaches (Stapleton et al., 2016). Finally, the generalizability of the findings was limited to engineering first-year students studying in universities similar to our research site. Future studies are encouraged to replicate our study with samples in other STEM fields or other institutes of higher education (e.g., two-year colleges, private universities) and compare the findings with ours.

## Conclusion

The results based on a sample of engineering students suggested the two-factor structure in FES was better supported by the data, which was aligned with Wong's (2015) Tripartite Encouragement Model. Both two-factor CFA model and ESEM model had similar model fit. However, we discovered a high factor correlation in CFA model which could result from forced- zero cross-loadings. Following Marsh et al.'s (2020) recommendation, the two-factor ESEM model was selected as the final model. In general, FES had good psychometric properties. The indicators of FES were reasonably loaded on theoretically corresponding factors except for item FE_P5, which was loaded on both encouragement factors. Notwithstanding, considering the high communality of FE_P5 (0.728), we did not recommend revising this item.

**Future studies are encouraged to replicate our study with samples in other STEM fields or other institutes of higher education (e.g., two-year colleges, private universities) and compare the findings with ours.**

The criterion validity of FES was supported by the results that encouragement factors can predict students' self-efficacy beliefs. Nevertheless, the prediction patterns of two encouragement factors were different – challenge-focused faculty encouragement statistically significantly predicted both academic milestones self-efficacy (standardized slope = 0.392) and coping self-efficacy (standardized slope = 0.216), while potential-focused faculty encouragement could only predict coping self-efficacy (standardized slope = 0.219). Prediction patterns suggested two faculty encouragement factors were distinguishable. Future studies are encouraged to verify our findings using participants in other STEM-related fields.

### AUTHOR NOTE

Hsien-Yuan Hsu is an assistant professor of Research and Evaluation in Education at the University of Massachusetts Lowell. His research interests include the development and retention of underrepresented groups in post-secondary institutions, multilevel modeling, model evaluation, and large-scale data analysis.

Yanfen Li is an Assistant Teaching Professor at the University of Massachusetts Lowell. Her current research is in engineering education with a focus on curriculum development and formation of engineering identity in undergraduate students in the Department of Educational Psychology at Texas A&M University.

Sandra T. Acosta is an Associate Professor at Texas A&M University in the Department of Educational Psychology. Her research interests include bilingualism, oral language development in English learners, and action research in PreK-12 general education settings.

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Second international symposium on information theory, Budapest, Hungary.

American Society for Engineering Education. (2016). *Engineering by the Numbers: ASEE retention and time-to-graduation benchmarks for undergraduate engineering schools, departments and programs*. https://ira.asee.org/wp-content/uploads/2017/07/2017- Engineering-by-the-Numbers-3.pdf

Anderson, S. L., & Betz, N. E. (2001). Sources of social self-efficacy expectations: Their measurement and relation to career development. *Journal of Vocational Behavior*, *58*(1), 98-117. https://doi.org/https://doi.org/10.1006/jvbe.2000.1753

Bandura, A. (1986). *Social foundations of thought and action*: *A social cognitive theory*. Englewood Cliffs.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, *50*(2), 248-287. https://doi.org/ 10.1016/0749-5978(91)90022-L

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist*, *28*(2), 117–148. https://doi.org/10.1207/s15326985ep2802_3

Bandura, A. (1997). *Self-efficacy*: *The exercise of control*. W.H. Freeman and Company.

Branch, S. E., Woodcock, A., & Graziano, W. G. (2015). Person orientation and encouragement:

Predicting interest in engineering research. *Journal of Engineering Education*, *104*(2), 119-138. https://doi.org/10.1002/jee.20068\

Byars-Winston, A., Diestelmann, J., Savoy, J. N., & Hoyt, W. T. (2017). Unique effects and moderators of effects of sources on self-efficacy: A model-based meta-analysis. *Journal of Counseling Psychology*, *64*(6), 645-658. https://doi.org/10.1037/cou0000219

Falco, L. D., & Summers, J. J. (2021). Social persuasions in math and their prediction of STEM courses self-efficacy in middle school. *The Journal of Experimental Education*, *89*(2), 326-343. https://doi.org/10.1080/00220973.2019.1681350

Garriott, P. O., Hunt, H. K., Navarro, R. L., Flores, L. Y., Lee, B. H., Suh, H. N., Brionez, J., Slivensky, D., & Lee, H.-S. (2021). Development and initial validation of the Engineering Learning Experiences Scale. *Journal of Vocational Behavior*, *124*, 103516. https://doi.org/https://doi.org/10.1016/j.jvb.2020.103516

Gebauer, M. M., McElvany, N., Bos, W., Köller, O., & Schöber, C. (2020). Determinants of academic self-efficacy in different socialization contexts: investigating the relationship between students' academic self-efficacy and its sources in different contexts. *Social Psychology of Education*, *23*(2), 339-358. https://doi.org/10.1007/s11218-019-09535-0

Hsu, H.-Y., Skidmore, S. T., Li, Y., & Thompson, B. (2014). Forced zero cross-loading misspecifications in measurement component of structural equation models: Beware of even "small" misspecifications. Methodology: *European Journal of Research Methods for the Behavioral and Social Sciences*, *10*(4), 138-152. https://doi.org/10.1027/1614-2241/a000084

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

Lent, R. W., & Brown, S. D. (2006). On conceptualizing and assessing social cognitive constructs in career research: A measurement guide. *Journal of Career Assessment*, *14*(1), 12-35. https://doi.org/10.1177/1069072705281364

Lent, R. W., & Brown, S. D. (2019). Social cognitive career theory at 25: Empirical status of the interest, choice, and performance models. *Journal of Vocational Behavior*, *115*, 103316. https://doi.org/10.1016/j.jvb.2019.06.004

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior*, *45*(1), 79-122. https://doi.org/https://doi.org/10.1006/jvbe.1994.1027

Lent, R. W., Lopez, F. G., & Bieschke, K. J. (1991). Mathematics self-efficacy: Sources and relation to science-based career choice. *Journal of Counseling Psychology*, *38*(4), 424-430. https://doi.org/10.1037/0022-0167.38.4.424

Lent, R. W., Sheu, H.-B., Miller, M. J., Cusick, M. E., Penn, L. T., & Truong, N. N. (2018). Predictors of science, technology, engineering, and mathematics choice options: A meta- analytic path analysis of the social–cognitive choice model by gender and race/ethnicity. *Journal of Counseling Psychology*, *65*(1), 17-35. https://doi.org/10.1037/cou0000243

Lent, R. W., Sheu, H.-B., Singley, D., Schmidt, J. A., Schmidt, L. C., & Gloster, C. S. (2008). Longitudinal relations of self-efficacy to outcome expectations, interests, and major choice goals in engineering students. *Journal of Vocational Behavior*, *73*(2), 328-335. https://doi.org/10.1016/j.jvb.2008.07.005

Liang, X., Yang, Y., & Cao, C. (2020). The Performance of ESEM and BSEM in Structural Equation Models with Ordinal Indicators. *Structural equation modeling*, *27*(6), 874-887. https://doi.org/10.1080/10705511.2020.1716770

Loo, C. W., & Choy, J. L. F. (2013). Sources of self-efficacy influencing academic performance of engineering students. *American Journal of Educational Research*, *1*(1), 86-92. https:// doi.org/10.12691/education-1-3-4

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130-149. https://doi.org/10.1037/1082-989X.1.2.130

Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2020). Confirmatory factor analysis (CFA), exploratory structural equation Modeling (ESEM), and set-ESEM: Optimal balance between goodness of fit and parsimony. *Multivariate Behavioral Research*, *55*(1), 102-119. https://doi.org/10.1080/00273171.2019.1602503

Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling*: *A second course*. Information Age Publishing, Inc.

National Academies of Sciences Engineering and Medicine. (2016). *Barriers and opportunities for 2-Year and 4-Year STEM degrees: Systemic change to support students' diverse pathways*. https://www.nap.edu/catalog/21739/barriers-and-opportunities-for-2-year- and-4-year-stem-degrees

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243-248. https://doi.org/10.1007/s11336-009-9135-y

Schaub, M. (2004). *Social cognitive career theory*: *Examining the mediating role of sociocognitive variables in the relation of personality to vocational interests* (Publication Number 3098455) University of Akron].

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464. https://doi.org/10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333-343. https://doi.org/10.1007/BF02294360

Sheu, H.-B., Lent, R. W., Miller, M. J., Penn, L. T., Cusick, M. E., & Truong, N. N. (2018). Sources of self-efficacy and outcome expectations in science, technology, engineering, and mathematics domains: A meta-analysis. *Journal of Vocational Behavior*, *109*, 118-136. https://doi.org/https://doi.org/10.1016/j.jvb.2018.10.003

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481-520. https://doi.org/ 10.3102/1076998616646200

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research*, *78*(4), 751-796. https:// doi.org/10.3102/0034654308321456

Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology*, *34*(1), 89-101. https://doi.org/10.1016/j.cedpsych.2008.09.002

Wong, Y. J. (2015). The psychology of encouragement: theory, research, and applications. *The Counseling Psychologist*, *43*(2), 178-216. https://doi.org/10.1177/0011000014545091

Wong, Y. J., Cheng, H.-L., McDermott, R. C., Deng, K., & McCullough, K. M. (2019). I believe in you! Measuring the experience of encouragement using the academic encouragement scale. *The Journal of Positive Psychology*, *14*(6), 820-828. https://doi.org/10.1080/17439760.2019.1579357