RESEARCH & PRACTICE IN ASSESSMENT

VOLUME TWENTY | 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 © 2025

TABLE OF CONTENTS

FROM THE EDITOR

4 **From the Editor**

- Nicholas A. Curtis

ARTICLES

5 **A Systematic Review of the International Assessment Literacy Measures in Higher Education (2013–2023)**

- Beyza Aksu Dunya, Mehmet Can Demir & Stefanie Wind

21 **Leveling Up Outcome-Based Assessment: Using Propensity Score Matching and Cost Analysis to Meet Contemporary Assessment Needs**

- Kendra Thompson-Dyck, Michelle Sogge, Lucas Schalewski & Alexandra Robie

34 **The Influence of Student Disengagement on a Non-Cognitive Measure: Practical Solutions for Assessment Practitioners**

- Katarina E. Schaefer & Sara J. Finney

49 **Multimodality as an Equitable Approach to Summative Assessment in Higher Education**

- Alison Cook-Sather, Daniela Moreira, Piper Rolfes & Jess Smith

62 **A Review Of Practices For Adjusting Exam Scores And A Proposed Nonlinear Scaling Method**

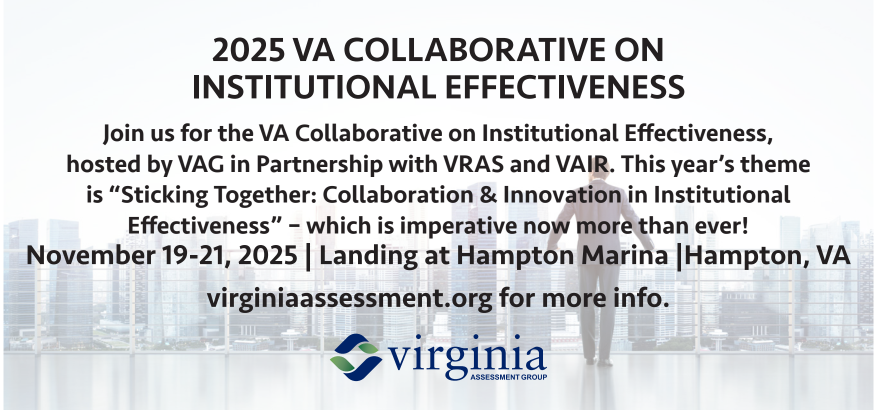
- Ryan K. Orchard

76 **Increasing Expended Effort on Low-Stakes Accountability Tests via Priming: Effectiveness with Graduating University Students**

- Sara J. Finney, Stuart A. Miller & Kendall M. McGoey

92 **Predictors of Student Technology Competencies in Assurance of Learning Assessment**

- Melodie Philhours & Kelly E. Fish



FROM THE EDITOR

"Change is the process by which the future invades our lives."

— *Alvin Toffler*

"As we begin Volume 20, Issue 1 of *Research & Practice in Assessment*, we are reminded that assessment continues to evolve. This issue reflects that ongoing change, with articles that challenge long-standing practices, introduce new methodologies, and offer fresh perspectives on equity, engagement, and technology.

In A Systematic Review of the International Assessment Literacy Measures in Higher Education (2013–2023), Dunya, Demir, and Wind synthesize a decade of measures, finding strong but incomplete psychometric evidence and urging the development of updated tools. In Leveling Up Outcome-Based Assessment: Using Propensity Score Matching and Cost Analysis to Meet Contemporary Assessment Needs, Thompson-Dyck, Sogge, Schalewski, and Robie provide a model that pairs rigorous statistical methods with cost analysis to inform student success initiatives. In The Influence of Student Disengagement on a Non-Cognitive Measure: Practical Solutions for Assessment Practitioners, Schaefer and Finney show how disengagement threatens validity and offer strategies to ensure fairer interpretations.

Equity and fairness also take center stage in Multimodality as an Equitable Approach to Summative Assessment in Higher Education, where Cook-Sather, Moreira, Rolfes, and Smith illustrate how student-choice and multimodal portfolios expand opportunities to demonstrate learning. In A Review of Practices for Adjusting Exam Scores and a Proposed Nonlinear Scaling Method, Orchard critiques conventional adjustment methods and proposes a nonlinear approach to better account for variation in performance.

Two final articles examine student motivation and competencies. In Increasing Expended Effort on Low-Stakes Accountability Tests via Priming: Effectiveness with Graduating University Students, Finney, Miller, and McGoey show that priming can increase effort and time spent, while narrowing some performance gaps, though without raising overall scores. In Predictors of Student Technology Competencies in Assurance of Learning Assessment, Philhours and Fish identify practice exam performance and introductory computing coursework as the strongest predictors of success, highlighting clear levers for curriculum design.

Together, these contributions remind us that assessment is never static. Each article adds to our collective effort to learn, unlearn, and relearn in pursuit of better practices for higher education.

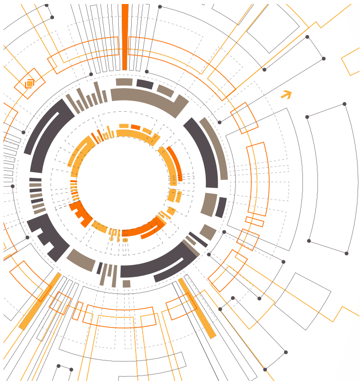
Regards,

Nicholas Curtis

Editor-in-Chief,
Research & Practice in Assessment

Abstract

This paper aims to synthesize measures of assessment literacy in higher education by forging a connection between two research domains: educational assessment and psychometrics. It begins with a systematic review of assessment literacy measures within the context of higher education published within the last ten years. AL measures, including tests of assessment literacy, self-report measures such as inventories, surveys, and rubrics in assessment literacy studies, were reviewed. Psychometric properties of the measures were evaluated against standards related to validity and reliability. Across a number of 11 measures reviewed, we found that while the reviewed studies demonstrated strong adherence to rigorous validation processes, the psychometric evidence presented for the available measures is neither complete nor up to date, concerning researchers' and educators' needs in terms of assessment. Nearly all measures were grounded in substantial literature reviews and expert evaluations, with five measures providing detailed content validity evidence and several studies reporting good fit indices for internal structure analyses. Reliability evidence was generally robust, with most Cronbach's Alpha coefficients ranging between .79 and .94, and high reliability indices reported in Rasch measurement theory studies. Despite these strengths, the identified gaps highlight the need for establishing psychometrically sound, comprehensive, and up-to-date assessment literacy measures. The paper concludes by discussing the development of enhanced assessment literacy measures that are adaptive to the changing landscape of assessment, and their implications for policy and practice in higher education.



AUTHORS

Beyza Aksu Dunya, Ph.D.
University of Alabama

Mehmet Can Demir, Ph.D.
Bartin University

Stefanie Wind, Ph.D.
University of Alabama

A Systematic Review of the International Assessment Literacy Measures in Higher Education (2013–2023)

Assessment literacy (AL) refers to an individual's understanding and use of assessment concepts and procedures (Coombs & DeLuca, 2022; DeLuca et al., 2019; Popham, 2011). As a construct, AL has a strong theoretical background rooted in well-established standards (American Federation of Teachers [AFT], National Council on Measurement in Education [NCME], & National Education Association [NEA], 1990) and research (e.g., Stiggins, 1991). The definition of AL has evolved over time to extend beyond the knowledge of the Standards for Educational and Psychological Testing (AERA et al., 2014), with a recent emphasis on socio-cultural and historical context that shapes assessment practices and beliefs. Even as the specific definition of AL has evolved over time, researchers have consistently recognized this construct as a central component of effective teaching across contexts, grade levels, and disciplinary areas (e.g., Danielson, 2013; Gotch & French, 2014).

In previous studies related to AL, researchers have mostly focused on the development, measurement, and perceptions related to this construct among in-service and pre-service primary and secondary education teachers (e.g., Alkharusi, 2011; Gotch & French, 2011; Plake et al., 1993; Zhang & Burry-Stock, 2003). Despite this emphasis, it is important to recognize that AL has critical implications in higher education, where assessment tasks are often complex (Friesen, 2022). As is true in primary and secondary educational contexts, the effectiveness of assessment practices within and beyond the classroom in higher education settings depends heavily on faculty involvement and proficiency in assessment efforts (Ray

CORRESPONDENCE
Email
baksu@ua.edu

et al., 2012). Despite the central role of assessment in higher education, faculty members often have little formal training related to assessment (Knapper, 2010), which may result in minimal integration of formative feedback and learner-driven assessments in their instruction (Massey et al., 2020).

Reflecting this lack of training, higher education in general has been repeatedly critiqued for emphasizing exam-based summative measures of student achievement while neglecting other potentially useful assessment techniques (Yorke, 2003). In response, many higher education institutions have launched faculty development units that offer training programs to support faculty's curriculum planning, instruction, and assessment practices (Taylor & Colet, 2010). Those training programs include efforts to enhance faculty members' conceptions and practices in student assessment using approaches that integrate summative measures, formative feedback, and dialogical assessment structures to advance learning (Nicol & Macfarlane-Dick, 2006). Today, higher education institutions incorporate a variety of assessment types, focusing not just on outcomes but also on processes and experiences. They are increasingly embracing technological innovations such as artificial intelligence (AI) to create a more detailed and accurate picture of how institutions are meeting their objectives (Watermark, 2023). Organizations like the Association for the Assessment of Learning in Higher Education (AALHE) offer events and webinars to support ongoing faculty development in AL. Despite these advancements, there is still no widely recognized, commonly used measure of AL specifically designed for faculty members. Establishing what AL means for a faculty member in the contemporary higher education sector and how we measure it is necessary to continue improving educational outcomes. With this in mind, the primary objective is to review and revise the psychometric properties of existing AL tools, drawing on a comprehensive analysis of recent literature in the field. This review not only evaluates these tools against a set of predefined criteria but also identifies gaps and areas for improvement. The findings will inform the development of enhanced AL measures that are adaptive to the changing landscape of assessment. In this context, we propose to explore the concept of *Artificial Intelligence Assessment Literacy*, which integrates AI technologies to refine and optimize the assessment process. This exploration is intended to pave the way for future research and practical implementations that respond effectively to the dynamic nature of educational assessment.

Despite the central role of assessment in higher education, faculty members often have little formal training related to assessment...

Theoretical Framework

Our framework draws extensively from the fundamental conceptual foundations of AL, prominently incorporating Stiggins' (1991) model. This model defines AL as the foundational comprehension of educational assessment combined with the skills required to apply this knowledge effectively across various measures of student achievement. Specifically, Stiggins identifies five standards of effective assessment that assessment-literate individuals should achieve (Stiggins, 1991). We selected Stiggins' model because it provides a comprehensive and widely recognized framework that emphasizes both conceptual understanding and practical application, aligning well with the goals of our study:

1. Formulating Precise Assessment Objectives: Focusing on defining clear intentions for what each assessment aims to evaluate. This involves recognizing the various functions assessments serve at the instructional level, such as discerning individual student needs, evaluating group dynamics within the classroom, organizing students into appropriate learning groups, and assigning grades.

2. Focusing on Achievement Targets: Individuals who are literate in the core principles of effective assessment recognize that students must achieve various interrelated objectives. These include mastering content knowledge, acquiring performance skills, and creating high-quality products.

3. Selecting Proper Assessment Methods: Assessment-literate educators understand the appropriate times and methods to employ various assessment techniques within these categories: selected response, essay, performance assessment, and personal communication (i.e. discussions, interviews).

4. Sampling student achievement: Any assessment is essentially a sample from a broader set of potential questions that could be asked if the assessment had no length constraints. It should include a sufficient number of questions to ensure that the assessment accurately represents this wider range of possibilities.

5. Avoiding bias and distortion: An assessment-literate educator must consistntly be alert to various specific technical and practical issues that could lead to inaccuracies and bias in measuring student achievement.

In summary, as introduced by Stiggins (1991), the concept of AL outlines that sound assessments rest upon five essential standards: they originate from and reinforce clearly articulated purposes, emerge from and conform to well-defined achievement targets, rely on the appropriate employment of assessment methods, efficiently sample student performance, and proactively prevent and eliminate bias and distortion. Over time, the idea and meaning of AL has evolved from merely being a collection of skills and knowledge to being viewed as a social practice. Drawing on Bernstein's sociocultural theory, Willis et al. (2013) described AL as a social practice characterized by interactions between teachers and students, encompassing collaborative engagement and shared understanding. More recently, Pastore and Andrade (2019) provided a theoretical definition of AL as a context-dependent concept with multiple components that integrate social, cultural, policy, and professional factors.

The significance of AL within higher education is underscored by its capacity to empower educators with the competence needed to create, administer, and interpret assessments effectively. Our proposed framework's content and structure are shaped by a synthesis of prior studies and our experiences as psychometricians and educational measurement professionals. The overarching research questions for this review are: "*How is AL measured in the higher education context*?" and "*What is the psychometric evidence presented for the AL measures used in higher education settings*?" This review is unique given the specific population that is elaborated with an updated definition of AL. The following specific research questions guided our study:

1. What evidence of *reliability*, if any, was provided for the available AL measures in the higher education context?

2. What evidence of *validity*, if any, was provided for the available AL measures in the higher education context?

Method

We used the principles outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to design and carry out our systematic review. As illustrated in Figure 1, these principles involve four major phases: (1) identification, (2) screening, (3) eligibility, and (4) inclusion (Moher et al., 2009). We discuss our methods specific to each of these phases below.

Database Search

Our search was conducted in databases focused on educational research, namely ERIC (Educational Resources Information Center), PsycINFO, Web of Science (WoS) and Education Full Text. These databases have a wealth of peer-reviewed journals and publications related to higher education assessment. We also searched Google Scholar for additional published work citing Stiggins (1991).

Over time, the idea and meaning of AL has evolved from merely being a collection of skills and knowledge to being viewed as a social practice.

Search Terms

To locate relevant studies, we used and modified keywords according to the specific formats and requirements for each database. Specifically, text databases were systematically queried using a comprehensive set of search terms, which included '*assessment literacy*' and the combined usage of '*assessment literacy*' with '*higher education*' to ensure a thorough exploration of the literature in this domain.

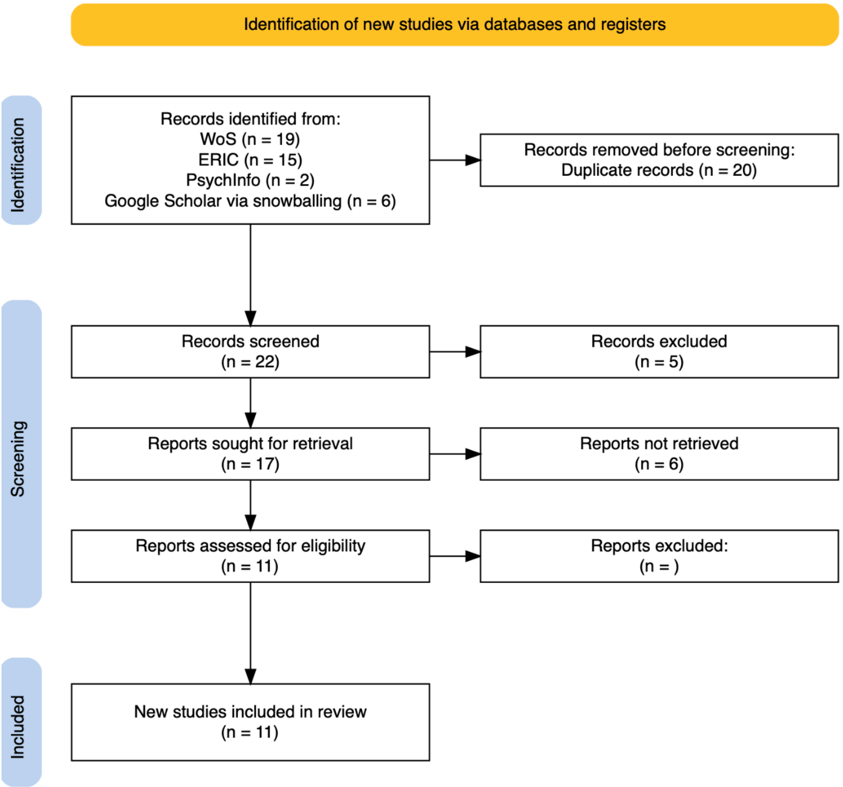

Identification of new studies via databases and registers
Identification
Records identified from:
WoS (n = 19)
ERIC (n = 15)
PsychInfo (n = 2)
Google Scholar via snowballing (n = 6)
Records removed before screening:
Duplicate records (n = 20)
Screening
Records screened
(n = 22)
Records excluded
(n = 5)
Reports sought for retrieval
(n = 17)
Reports not retrieved
(n = 6)
Reports assessed for eligibility
(n = 11)
Reports excluded:
(n =)
Included
New studies included in review
(n = 11)

Figure 1
PRISMA flow chart study inclusion process

Inclusion and Exclusion Criteria

We identified a set of inclusion and exclusion criteria that reflected our guiding research questions and objectives. We selected articles that met the following criteria: 1) published in a peer reviewed journal, 2) full text in English was available to the researchers, and 3) the use of the construct directly related to higher education context. Exclusion criteria included: 1) articles that were not published in a peer-reviewed publication (e.g., non-peer reviewed article, dissertation, technical report), and 2) articles not directly related to higher education institutions but to other stakeholders (e.g., in-service teachers, school administrators).

We have diligently reviewed studies conducted over the past decade, recognizing the significance of updates in the AL construct.

The Reviewing Process

In our investigation, we have diligently reviewed studies conducted over the past decade, recognizing the significance of updates in the AL construct. The screening process included the following stages: 1) eliminating duplicate articles, 2) filtering out articles that did not align with the inclusion criteria based on their titles and abstracts, 3) reviewing full texts and excluding articles that did not meet the criteria, 4) employing a snowballing approach to identify additional relevant articles in Google Scholar, and 5) extracting results from the selected articles. This process returned 11 peer-reviewed articles. Any work employing an AL instrument in the United States higher education setting or an international equivalent was retained for review.

Evaluation of Psychometric Properties

We evaluated the psychometric properties of the identified measures by identifying and interpreting evidence related to validity and reliability. We organized our synthesis and interpretation of validity and reliability-related results using the criteria summarized in Table 1.

Table 1
Psychometric Property Evaluation Criteria

Type	Criteria	Citation
Concurrent Validity Evidence	Correlation > .25	Litwin (1995, p. 45)
Content Validity Evidence	Scale level: S-CVI > .90 or Item level: I-CVI = 1.00 for 3-5 experts; CVI > .78 for 6-10 experts	Polit & Beck (2006, p. 496)
Construct Validity Evidence	CFI > .95 SRMR < .08 RMSE < .05	Hu & Bentler (1999)
Reliability Estimate	Cronbach alpha > .70 for attitude scales or KR-20 or KR-21 > .80 for competency tests	Field (2009) Nunnally (1978) Salvucci et al. (1997)

Note. This table displays evaluation criteria for the validity and reliability aspects of the measures. The details about definition and calculation are presented in the following section. *KR-20/KR-21*: Kuder-Richardson formulas which provide an estimate of proportion of the total variance in the test scores that is attributable to the construct being measured; *CVI*: Content validity index showing the degree of agreement among content experts on each item's relevancy to the intended construct; *SRMR*: Standardized Root Mean-Square Residual which evaluated the discrepancy between observed covariance matrix and model-implied covariance matrix; *RMSE*: Root mean square error which is the average difference between predicted and observed values is used to assess precision of estimates.

Evidence Related to Validity

The current *Standards for Educational and Psychological Testing* (AERA et al., 2014) define validity as the accumulation of evidence to support the interpretation and use of test scores for a particular purpose. This consensus view of validity includes five primary sources of evidence that may be relevant for developing a validity argument to support a particular interpretation and use of test scores, including evidence related to test content, response processes, internal structure, relations to other variables, and consequences of testing.

Researchers operationalize construct validity differently depending on their selected measurement framework…

As we will discuss in our results section, the articles included in our review reported validity-related evidence that reflects qualitative approaches to gathering content-related validity evidence or quantitative techniques to examine the internal structure of the instrument for construct-related validity evidence.

Construct validity is defined as "the degree to which inferences based on test scores are justified in relation to the conceptualization and theoretical framework of the construct that the test is intended to measure" (Messick, 1980, p. 8). Because researchers operationalize construct validity differently depending on their selected measurement framework, we started our analysis related to construct validity evidence by first considering whether researchers used a test-score approach (e.g., Classical Test Theory (CTT) or factor analysis) or scaling model approach (e.g., Item Response Theory (IRT) or Rasch measurement theory); please see Engelhard (2013) for a detailed discussion of these two measurement frameworks.

Among the articles included in our review, those that reflected the test score tradition tended to use a factor-analytic approach to collecting construct-related validity evidence; these analyses also provide validity evidence related to the internal structure of an assessment instrument. We evaluated this evidence using well-established critical values for model fit indices as presented by Hu and Bentler (1999): CFI > .95; SRMR < .08; RMSEA < .05. For the scaling model framework, researchers may use several techniques to evaluate the internal structure of an instrument using various fit statistics. For example, researchers who use Rasch models often use residual-based analyses to examine adherence to model requirements such as unidimensionality or invariance.

Most of the identified measures relied on self-reports where the participants were asked to evaluate their own knowledge, practices, confidence or needs in terms of assessment…

Evidence Related to Reliability

According to the Standards (AERA et al., 2014), reliability refers to the consistency of results over replications of a measurement procedure. For example, replications may include repeated observations of test-takers over multiple items within a form (e.g., internal consistency reliability) or between administrations of a test (e.g., repeated measures reliability). In our research synthesis, we established criteria for interpreting reliability results based on the type of instrument and selected method for evaluating reliability. For example, reliability evidence included in the articles in our review used reliability analyses based on test-score methods, where the authors reported internal consistency statistics such as Cronbach's alpha, KR-20, or KR-21. We interpreted these results using critical values from the literature. Specifically, we used a value of 0.70 for Cronbach's alpha (Nunnally, 1978) and a value of 0.80 for KR-20 or KR-21 (Salvucci et al., 1997). Researchers also reported reliability evidence based on scaling models from Rasch measurement theory. In these analyses, *reliability of separation statistics* describe the precision of location estimates for items, persons, or other facets. As a general guideline, a value exceeding 0.8 typically suggests reproducibility of measures within a Rasch context (Linacre, 2002).

Results

After applying our inclusion and exclusion criteria, we identified a total of eleven instruments. The instruments took three forms as presented in Table 1:

1) Scales, tests or inventories of assessment knowledge or AL (*n*=5),

2) Surveys on attitudes, perceptions or needs on AL (*n*=5), and

3) Rubrics evaluating other faculty members' work on assessment (*n*=1).

Most of the identified measures relied on self-reports where the participants were asked to evaluate their own knowledge, practices, confidence or needs in terms of assessment. Details about psychometric properties of the measures are presented in Table 2 and discussed below.

Evaluation of Psychometric Properties

In this section, we summarize the reported psychometric properties of the identified measures as they relate to validity and reliability.

Validity Evidence. Among the identified measures for our review, researchers reported validity evidence related to content validity and construct validity using internal structure analyses. Content validity evidence was provided in detail for a total of five measures. No detail regarding development of item content was provided for two measures (DiLoreto et al., 2017 and McGrath et al., 2020). Almost all the measures that were reviewed were grounded on a substantial literature review process followed by expert reviews. The process for identification of content and development of items involved both inductive (i.e., focus groups) and deductive evidence in one study (Alonzo et al., 2019). In two studies (Mokshein et al., 2015, 2019), content was grounded on AFT et al. (2009), but no information was provided related to expert review in these studies. Overall, no quantitative index was presented as evidence of content validity.

Five studies reported statistical evidence related to internal structure in the form of either EFA or PCA results. Among them, Alonzo et al. (2019) conducted a CFA as well as

Table 2
Psychometric Property Evaluation Criteria

Measure Name	Construct of Interest	Instrument & Item Characteristics	Respondents	Test Content or Content Validity	Internal Consistency /Reliability	Construct Validity
Adapted version of the Classroom Assessment Literacy Test (CAL) (Mertler, 2003)	Classroom assessment literacy skills	35 dichotomously scored multiple choice items	Regular teaching faculty in public and private higher education institutions in Pakistan	Expert opinion was collected qualitatively. No quantitative evidence was provided.	Cronbach alpha = .89	Not reported
Confidence in assessment survey for faculty members (Massey et al., 2020)	Conceptions and confidence in assessment	23 scale items in confidence in assessment and 2 open-ended items in conceptions of assessment	27 instructors at a two-year college in TX, USA	The instrument adapted from DeLuca et al., (2013) but no quantitative content validity evidence was provided	Cronbach alpha for the confidence in assessment scale items was found .84 and .68 for two subscales, namely assessment approaches and assessment praxis	Principal Component Analysis (PCA) resulted in a minimum subscale loading equals to .40.
Modified Conceptions of Assessment III (CoA-III) survey (DiLoreto et al., 2017)	Attitude about assessment & knowledge about assessment types	2 open-ended items were added to the modified CoA-III scale (Q1. What does the term assessment mean to you? Q2. What types of activities come to mind when you think of the term assessment?)	156 faculty members from 10 higher education institutions from the Southeast U.S.	Not reported	Not reported	The original CoA-III scale was validated through second-order CFA model based on a different population (teachers) but no evidence was provided for the modified survey
Assessment practice inventory for teacher educators – Section B (MAPITE) (Mokshein et al., 2015)	Practice with respect to assessment literacy standards	The inventory had five sections. Section B directly related to AL practices so was involved in this review. It includes 10 items on a 5-point scale.	254 faculty members from a teacher education university in Malaysia	Content was grounded on AFT, NCME & NEA (1990). No information related to expert review was provided.	Section B included two factors. Cronbach alpha for factor 1= .863 and for factor 2= .786	EFA results yielded all factor loadings > .60 Fit values were not reported.
Rasch-validated version of MAPITE Section B (Mokshein et al., 2019)	Practice with respect to assessment literacy standards	One item was dropped as the results of the Rasch validation	763 faculty members from various teacher education institutions in Malaysia	Content was grounded on AFT, NCME & NEA (1990).	Person reliability index = .84 Item reliability index = .91	Variance explained by Rasch measures = %53.1 (more than 20% variance explained is needed for accurate estimation, Reckase, 1979)
Assessment literacy survey for medical education faculty (McGrath et al., 2020)	Assessment literacy and practices	A survey with 4 open-ended items on assessment literacy (e.g., What do formative and summative assessment mean to them), 7 open-ended items particularly on peer assessment	35 faculty members from the departments of Medicine and Biomedical	No detail was provided regarding how the survey items were developed and survey content were determined	Not Applicable	Not Applicable
Questionnaire on formative and summative assessment (type 1) (Davies & Taras, 2018)	Assessment literacy on summative and formative assessment	A 44-item questionnaire with Yes-No response format. Some items required written comment (e.g., give a definition of formative assessment)	100 faculty members in The U.K.	Items were pilot tested with 5 faculty members when they were initially developed (Taras, 2008)	Not Applicable	Not Applicable

Table 2
Psychometric Property Evaluation Criteria, continued

Measure Name	Construct of Interest	Instrument & Item Characteristics	Respondents	Test Content or Content Validity	Internal Consistency /Reliability	Construct Validity
SALRubric-summative assessment literacy (Edwards, 2017)	Summative assessment literacy and knowledge	A rubric with descriptors on ten dimensions (i.e., knowledge on purpose of summative assessment, interpretation of results, fairness) developed for five levels of expertise	Academic and teaching staff in New Zealand	Inductive evidence (questionnaire, interviews, summative assessment tasks used by teachers) was used to ensure content validity	The evaluation criteria was reviewed by assessment experts to ensure the evaluation criteria measures what it purports to measure	Two senior academics in assessment independently scored data using SALRubric. No statistics were provided regarding the degree of agreement.
Academic SBA Practices Tool – ASBAPT (Alonzo et al. 2019)	Standard based assessment literacy of academics	21 items with 6 theoretical dimensions	410 academics in public universities in Philippines	8 experts reviewed the tool. Items and content were determined using inductive (focus groups) and deductive (literature review) evidence	Cronbach alpha values for the six subscales as follows: .92, .88, .89, .90, .92, .88, .94	EFA results for the 6-factor model: RMSEA= 0.02, SRMR= 0.03, CFI= 0.95, TLI= 0.96 CFA Results: RMSEA= 0.03, CFI= 0.98, TLI= 0.97
Language assessment literacy survey (Kremmel, 2020)	Perceived need on language assessment literacy	71 self-report items on a 5-point scale from No knowledgeable to extremely knowledgeable. Item format example: "How knowledgeable do people in your chosen group/profession need to be about…"	138 language assessment researchers, 198 language assessment developers and 645 language instructors	Content was grounded on a well-known theory (Taylor, 2013). 6 expert reviews on the initial instrument and 2 additional expert reviews on the piloted instrument were taken.	Cronbach alpha values for the identified nine subscales were ranged from .85 to .96	EFA results only included eigenvalues on a large sample. No fit values were reported.
Questionnaire on Language Assessment Literacy and Assessment Training Needs (Sayyadi, 2022)	Perceived need for language assessment literacy and received training on it	22 items on a 3-point scale from not at all to advanced on received training (if you received training on…) and perceived training needs (if you need training on…) on LAL	68 university instructors on English in Iran	The questionnaire was adapted from a well-known instrument (Vogt & Tsagari, 2014) developed for teachers and piloted with four instructors	Cronbach alpha value of .92 was reported	Not reported

an EFA and reported that the results from both analyses showed good fit according to our criteria. In other studies that employed factor analytic methods to collect construct validity evidence, minimum factor loadings were reported as .40 for Massey et al. (2020) and .60 for Mokshein et al. (2015). The dimensional structure of the measures was analyzed and reported in several studies. For example, Mokshein et. al. (2019) investigated dimensionality for the MAPITE using a Rasch measurement approach and found that Rasch measures based on data collected through explained 53.1% of the variance, supporting a unidimensional structure. Kremmel (2020) investigated dimensionality using EFA, where the number of dimensions was determined based on eigenvalues. However, fit statistics were not reported in either study. No statistical evidence was provided for a total of four studies, including the adapted version of Mertler's (2003) assessment literacy inventory. In one study, no further statistical analysis was conducted with the new sample, but fit statistics for the original scale were reported (DiLoreto et. al., 2017).

Reliability Evidence. Among the identified measures, researchers reported reliability using either internal consistency statistics based on CTT or reliability of separation statistics based on Rasch measurement theory models. Cronbach Alpha coefficients were reported in a total of six studies. Among the Cronbach alpha coefficient values that were reported, only the assessment praxis dimension in Massey et al. (2020) had a relatively low reliability index value against our criteria, while the other values were reported between .79 and .94. The validation study of MAPITE (Mokshein et. al., 2019) based on Rasch measurement theory suggested that the person reliability index and item reliability index values were high, .84 and .91, respectively. Lastly, a total of three measures were not accompanied with any reliability evidence.

Synthesis of the Reviewed Studies

Our systematic review of the literature on AL measures in higher education has revealed several critical insights, highlighting both strengths and gaps in the existing instruments. We identified eleven instruments classified into three categories: scales, tests, or inventories of assessment knowledge or literacy (n=5); surveys on attitudes, perceptions, or needs related to AL (n=5); and rubrics evaluating faculty members' assessment work (n=1). The predominant reliance on self-report measures indicates a strong focus on subjective evaluations of knowledge, practices, confidence, and needs in assessment.

Validity Evidence. The psychometric evaluation of these instruments revealed mixed results. While five measures provided detailed content validity evidence, two lacked specific information regarding item content development. Most measures were grounded in substantial literature reviews and expert evaluations, indicating a rigorous approach to establishing content validity. However, the absence of a quantitative index for content validity in many studies highlights a gap that needs to be addressed. Construct validity evidence, primarily through factor analytic methods, was reported in five studies. While some studies, like Alonzo et al. (2019), reported good fit indices, others did not provide sufficient statistical evidence.

Reliability Evidence. The reliability analysis revealed that six studies reported Cronbach Alpha coefficients, with most values falling between .79 and .94, indicating acceptable internal consistency. The MAPITE study by Mokshein et al. (2019) reported high person and item reliability indices using Rasch measurement theory. However, three measures lacked any reported reliability evidence.

General Analysis. The current instruments, while grounded in solid methodological foundations, often fall short of providing up-to-date content. This gap is particularly significant given the evolving educational landscape and the increasing integration of innovative technologies such as artificial intelligence (AI). The advancement of AI presents new opportunities to refine and optimize assessment processes, making it imperative to explore the concept of *Artificial Intelligence Assessment Literacy*. Educators need to be literate in integrating, using, and tracking the effects of AI in assessment to fully leverage these technologies. In conclusion, our review highlights the critical need for ongoing research and development in the field of AL. Establishing robust and adaptive AL measures will contribute significantly to improving educational outcomes and aligning assessment practices with the rapidly changing landscape of higher education.

Redefining Assessment Literacy and Measuring It

In light of recent advancements in technology, particularly with the rise of generative AI (GAI), there is a pressing need to redefine AL. Scholars have suggested that we need to reconsider the kinds of assessment tasks instructors assign to make them 'AI-resistant' by reducing the likelihood that GAI can complete the entire assignment task (Moorhouse et al., 2023). While there is an increasing amount of advice available to instructors on how to modify their assignment tasks in the GAI world (e.g., blogs, newsletters), many instructors need to look to their institutions for guidance and direction regarding GAI (Moorhouse et al., 2023).

The advancement of AI presents new opportunities to refine and optimize assessment processes, making it imperative to explore the concept of Artificial Intelligence Assessment Literacy.

For this reason, a comprehensive framework on Artificial Intelligence Assessment Literacy (AI AL) must be developed, and resources must be provided to faculty. This framework should include guidelines on creating AI-resistant assignments, training on the use of AI tools in assessment, and strategies for leveraging AI to enhance learning outcomes. By equipping faculty with these resources, institutions can help ensure that assessment practices remain effective and relevant in the face of rapidly evolving technological capabilities.

Discussion

In this study we examined the reported psychometric evidence of existing AL measures developed within the international higher education context over the past decade (2013-2023).

Based on psychometric evaluation of existing measures against a set of criteria, we concluded that the available psychometric evidence supporting these measures is not strong. Despite the importance of understanding AL levels, perceptions, and practices of faculty members in higher education being widely recognized, our findings raise doubts about the preparedness of these measures to meet such demands.

First, we examined the literature on measures of AL for evidence related to validity. In terms of content validity, the majority of the reviewed measures were grounded on a substantial literature review process followed by expert reviews, although these procedures were not accompanied by statistical indices such as the Content Validity Index (Lawshe, 1975). However, a noteworthy critique of the existing measures pertains to their content. These measures lack coverage of current topics such as digital AL, fairness in the era of Artificial Intelligence (AI), and fairness related to Assessment for Learning. This deficiency can be attributed to their content development and validation processes being based on outdated standards that no longer adequately encompass these areas. One significant practical implication of this study regarding this finding is the need to update and expand the content of AL measures, while ensuring the collection of robust content validity evidence that aligns with current standards. This entails seeking consensus from content experts, which can be quantified using appropriate indices like CVI.

Many of the reviewed measures primarily concentrate on individuals' *perceived* knowledge or skills, while there is a notable scarcity of research examining individuals' actual skills, knowledge, or competency. On the other hand, self-report measures are limited in terms of their susceptibility to response biases and social desirability effects (Fisher, 1993). Among the reviewed measures, five of them included direct statements (i.e., "I can"). Self-reports are used most often since they are easiest to administer and shown to correlate with actual assessment practices (Kelly, 2020) despite their well-researched limitations. In future work, we suggest that researchers consider using objective assessments of individuals' knowledge, literacy, and practices in assessment rather than their subjective perceptions in future studies.

Several studies acknowledged the limitation of sample size and sample characteristics as potential constraints on the validity of their findings (Massey et. al., 2020; McGrath et. al., 2020). It is crucial to consider both the size and characteristics of the sample when evaluating the validity and reliability of scores obtained from measures. However, it is worth noting that the Rasch model, which is a member of IRT-based models, can address limitations regarding small sample size (Linacre, 1994), since these models are known for their robustness to handle small sample sizes and can be effectively utilized for validation purposes in measurement processes with limited samples. Of the reviewed studies, only one measure employed the Rasch framework for validation purposes. Future studies on developing and adapting measures in higher education assessment can utilize a scaling model approach more often to address sampling limitations.

Many of the reviewed measures primarily concentrate on individuals' perceived knowledge or skills, while there is a notable scarcity of research examining individuals' actual skills, knowledge, or competency.

Another concern about the reviewed measures was related to the presented reliability evidence. We observed that none of the reviewed studies took into account the dimensional structure when calculating reliability indices. However, it is crucial to consider the dimensionality of the measure to determine the appropriate type of reliability coefficient to report. For multidimensional scales, reporting composite reliability provides a smaller margin of error compared to reporting separate Cronbach alpha values for individual sub-scales (Cronbach et al., 1965); in other words, reliability may be overestimated. Therefore, future research studies aiming to develop AL scales, including those for assessing attitudes, should carefully consider the dimensionality of the measure when calculating and presenting reliability evidence. By doing so, researchers can enhance the accuracy and precision of the reliability estimates, providing more robust evidence of the measures' internal consistency and stability over time.

The results of the review also suggested that there is a lack of evidence regarding validity and reliability evidence for the modified instruments, which were adapted from existing measures through the addition, omission, or revision of items. However, any adaptation may raise concerns about the trustworthiness of the modified instruments if they lack the necessary validation and reliability evidence to support their use. It is crucial for researchers to provide comprehensive evidence of the validity and reliability of modified instruments to ensure the robustness of their results (AERA et al., 2014).

None of the reviewed measures presented evidence related to relationships with external variables (e.g., concurrent validity; Lin & Yao, 2014). However, the newly developed measures could report their association with relevant measures such as student outcomes to provide evidence of concurrent validity (Murphy & Davidshofer, 1988).

Lastly, providing institutional support and resources for ongoing research into AL measures is essential for their continuous development. This support can include funding for validation studies, access to advanced technologies, and dedicated time for faculty to engage in research activities. By investing in these resources, institutions can foster innovation in assessment practices and ensure that AL measures are grounded in the latest educational research and methodologies.

By adopting these strategies, institutions can ensure that AL measures are robust, comprehensive, and well-aligned with contemporary educational needs. This proactive approach will help institutions stay ahead of emerging trends and challenges in education, ultimately leading to improved teaching and learning outcomes.

Higher education institutions and faculty development centers can utilize well-established measures to inform the design and implementation of targeted training programs aimed at enhancing faculty's understanding, attitude, and skills in assessment.

Limitations

This review has several limitations that must be acknowledged. First, our review focused on published studies that met specific inclusion criteria, which may have resulted in the exclusion of relevant unpublished or non-peer-reviewed works, such as dissertations, that could offer additional insights into the psychometric properties of AL measures. This selection bias may limit the generalizability of our findings. Second, the rapid advancement of AI technologies is reshaping assessment requirements and expectations. While our study discusses the definition of *AI assessment literacy* and the development of enhanced AL measures, it does not fully encompass the dynamic and rapidly evolving landscape of educational assessment.

Lastly, while we have highlighted both the strengths and gaps in the existing measures, the review itself is limited by the quality and depth of the original studies.

Despite these limitations, this review aims to provide insights into the current state of AL measures and discuss avenues for future research aimed at developing psychometrically sound, comprehensive, and up-to-date tools that are adaptive to the changing landscape of assessment.

Future Work

Overall, the findings of this review study highlight the need for further research on refining existing AL measures in diverse higher education contexts. This can also be considered as a new research avenue for developing psychometrically sound measures of AL in higher education. In practical terms, the study emphasizes the significance of integrating psychometrically-sound AL measures when planning training and professional development initiatives in higher education. Higher education institutions and faculty development centers can utilize well-established measures to inform the design and implementation of targeted training programs aimed at enhancing faculty's understanding, attitude, and skills in assessment. This may entail offering customized resources, workshops, or ongoing support to assist faculty in effectively designing and implementing assessments that align with learning objectives, promote student engagement, and provide meaningful feedback. By incorporating these measures and implementing comprehensive training programs, institutions can foster a culture of AL among faculty, thereby enhancing the overall quality of assessment practices in higher education.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for Educational and Psychological Testing*. AERA.

Alkharusi, H. (2011). An Analysis of the internal and external structure of the teacher assessment literacy questionnaire. *International Journal of Learning*, *18*(1), 515-528. https://doi.org/10.18848/1447-9494/CGP/v18i01/47461

Alonzo, D., Mirriahi, N., & Davison, C. (2019). The standards for academics' standards-based assessment practices. *Assessment & Evaluation in Higher Education*, *44*(4), 636-652. https://doi.org/10.1080/02602938.2018.1521373

American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), & National Education Association (NEA) (1990). *Standards for Teacher Competence in the Educational Assessment of Students*. Retrieved from https://eric.ed.gov/?id=ED323186

American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), & National Education Association (NEA) (2009). *Assessment Literacy Standard*.

Coombs, A., & DeLuca, C. (2022). Mapping the constellation of assessment discourses: a scoping review study on assessment competence, literacy, capability, and identity. *Educational Assessment, Evaluation and Accountability*, *34*(3), 279-301. https://doi.org/10.1007/s11092-022-09389-9

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. J. , Schoneman, P., & McKie, D. (1965). Alpha coefficient for stratified parallel tests. *Educational & Psychological Measurement*, *25*(2), 291-312. https://doi.org/10.1177/001316446502500201

Danielson, C. (2013). *The framework for teaching evaluation instrument*. The Danielson Group.

Davies, M.S. & Taras, M. (2018). Coherence and disparity in assessment literacies among higher education staff. *London Review of Education*, *16*(3), 474–490. https://doi.org/10.18546/LRE.16.3.09

DeLuca, C., Chavez, T. & Cao, C. (2013). Establishing a foundation for valid teacher judgement on student learning: The role of pre-service assessment education. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 107–126. https://doi.org/10.1080/0969594X.2012.668870

DeLuca, C., Coombs, A., MacGregor, S., & Rasooli, A. (2019). Toward a differential and situated view of assessment literacy: Studying teachers' responses to classroom assessment scenarios. *Frontiers in Education*, *4*(94), 1-10. https://doi.org/10.3389/feduc.2019.00094

DiLoreto, M. A., Pellow, C., & Stout, D. L. (2017). Exploration of conceptions of assessment within high-stakes US culture. *International Journal of Learning, Teaching and Educational Research*, *16*(7), 1-9.

Edwards, F. (2017). A rubric to track the development of secondary pre-service and novice teachers' summative assessment literacy. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 205-227. https://doi.org/10.1080/0969594X.2016.1245651

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Field, A. (2009). Discovering statistics using SPSS (3rd edition). *Sage*.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, *20*(2), 303-315. https://doi.org/10.1086/209351

F Flaugher, R. (2000). Item pools. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, & R. J. Mislevy (Eds.), *Computerized adaptive testing* (pp. 37–59). Routledge. https://doi.org/10.4324/9781410605931

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., & de Beurs, E. (2017). Development of a computer adaptive test for depression based on the Dutch-Flemish version of the PROMIS item bank. *Evaluation & the Health Professions*, *40*(1), 79-105. https://doi.org/10.1177/0163278716684168

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of life Research*, *14*, 2277-2291. https://doi.org/10.1007/s11136-005-6651-9

Fonseca-Pedrero, E., Menéndez, L. F., Paino, M., Lemos-Giráldez, S., & Muñiz, J. (2013). Development of a computerized adaptive test for schizotypy assessment. *PLoS One*, *8*(9), e73201. https://doi.org/10.1371/journal.pone.0073201

Forbey, J. D., & Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: a review and illustration with the MMPI-2 Computerized Adaptive Version. *Psychological Assessment*, *19*(1), 14-24. https://doi.org/10.1037/1040-3590.19.1.14

Friesen, D. W. (2022). *Towards a situated view of assessment literacy for higher education* [Unpublished master's thesis]. [University of Saskatchewan]. *ProQuest Dissertations & Theses Global*.

Gorgun, G., & Bulut, O. (2023). Incorporating test-taking engagement into the item selection algorithm in low-stakes computerized adaptive tests. *Large-scale Assessments in Education*, *11*(1), 1-21. https://doi.org/10.1186/s40536-023-00177-5

Gotch, C. M., & French, B. F. (2011). *Development and validity evidence for the teacher educational measurement literacy scale* [Conference presentation]. National Council on Measurement in Education, New Orleans, LA, United States.

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice*, *33*(2), 14–18. https://doi.org/10.1111/emip.12030

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

Kelly, M. P., Feistman, R., Dodge, E., St. Rose, A., & Littenberg-Tobias, J. (2020). Exploring the dimensionality of self-perceived performance assessment literacy (PAL). *Educational Assessment, Evaluation and Accountability*, *32*, 499-517. https://doi.org/10.1007/s11092-020-09343-7

Knapper, C. (2010). Plus ça change… educational development past and future. *New directions for teaching and learning*, *122*, 1-5. https://doi.org/10.1002/tl.392

Kozierkiewicz-Hetmańska, A., & Nguyen, N. T. (2010, September). A computer adaptive testing method for intelligent tutoring systems. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 281-289). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kraska, J., Bell, K., & Costello, S. (2023). Graded response model analysis and computer adaptive test simulation of the depression anxiety stress scale 21: Evaluation and validation study. *Journal of Medical Internet Research*, *25*, e45334. https://doi.org/10.2196/45334

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly*, *17(*1), 100-120. https://doi.org/10.1080/15434303.2019.1674855

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology*, **28**(4), 563-575. https://psycnet.apa.org/doi/10.1111/j.1744-6570.1975.tb01393.x

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Lin, WL. & Yao, G. (2014). Concurrent Validity. In *Encyclopedia of Quality of Life and Well-Being Research*. https://doi.org/10.1007/978-94-007-0753-5_516

Litwin, M. S., & Fink, A. (1995). *How to measure survey reliability and validity*. Sage.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382-386.

Massey, K. D., DeLuca, C., & LaPointe-McEwan, D. (2020). Assessment literacy in college teaching: Empirical evidence on the role and effectiveness of a faculty training course. *To Improve the Academy: A Journal of Educational Development*, *39*(1). http://dx.doi.org/10.3998/tia.17063888.0039.109

McGrath, M. F., Scott, L., & Logue, P. (2020). Peer assessment in Irish medical science education: Exploring staff assessment literacy and assessment practice. *Practitioner Research in Higher Education*, *13*(1), 37-56.

Mertler, C. A. (2003). *Pre-service versus in-service teachers' assessment literacy: Does classroom experience make a difference?* [Conference presentation]. Mid-Western Educational Research Association, Columbus, OH, United States.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., ... & Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement (Chinese edition). *Journal of Chinese Integrative Medicine*, *7*(9), 889-896.

Mokshein, S. E., Lebar, O., Yunus, J. Y., Rahmat, A., Dollah, M. U., Muhammad, A., Mansor, N. A., Mahmood, A., & Noor, N. M. (2015). Development and validation of assessment practice inventory for teacher educators. *Asian Journal of Assessment in Teaching and Learning*, *5*, 25-43. Retrieved from https://ojs.upsi.edu.my/index.php/AJATeL/article/view/2036

Mokshein, S. E., Ahmad, H., Lebar, O., Dollah, M. U., Yunus, J., Rahmat, A., & Ahmed, H. H. (2019). Validation of the Malaysian-Based Assessment Practice Inventory for Teacher Educators (MAPITE) using Rasch model. *Journal of Engineering and Applied Sciences*, *14*(9), 2783-2798.

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, *5*, 100151. https://doi.org/10.1016/j.caeo.2023.100151

Murphy, K. R., & Davidshofer, C. O. (1988). *Psychological testing: Principles & applications* (4th ed.). Prentice-Hall.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199-218. https://doi.org/10.1080/03075070600572090

Nunnally, J. C. (1978) *Psychometric Theory* (2nd edition). McGraw-Hill.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*(4), 679. https://psycnet.apa.org/doi/10.1037/0021-9010.78.4.679

Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education*, *84*, 128-138. https://doi.org/10.1016/j.tate.2019.05.003

Petersen, M. A., Aaronson, N. K., Arraras, J. I., Chie, W. C., Conroy, T., Costantini, A., ... & European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group. (2018). The EORTC CAT Core—The computer daptive version of the EORTC QLQ-C30 questionnaire. *European Journal of Cancer*, *100*, 8-16. https://doi.org/10.1016/j.ejca.2018.04.016

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice*, *12*(4), 10-12. https://doi.org/10.1111/j.1745-3992.1993.tb00548.x

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489-497. https://doi.org/10.1002/nur.20147

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, *46*(4), 265-273. https://doi.org/10.1080/08878730.2011.605048

Ray, C. M., Peterson, C. M., & Montgomery, D. M. (2012). Perceptions of college faculty concerning the purpose of assessment in higher education. *Journal of Human Subjectivity*, *10*(1), 77-102.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207-230. https://doi.org/10.3102/10769986004003207

Salvucci, S., Walter, E., Conley, V., Fink, S., & Mehrdad, S. (1997). *Measurement error studies at the National Center for Education Statistics* (NCES 97-464). U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Sayyadi, A. (2022). In-service university-level EFL instructors' language assessment literacy and training needs. *Profile Issues in Teachers Professional Development*, *24*(1), 77-95. https://doi.org/10.15446/profile.v24n1.93676

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: Introducing the CAT–PD project. *Journal of Personality Assessment*, *93*(4), 380-389. https://doi.org/10.1080/00223891.2011.577475

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, *72*(7), 534-539.

Suskie, L. (2004). *Assessing student learning: A common sense guide*. Anker Publishing Company, Inc.

Taras, M. (2008). Summative and formative assessment: Perceptions and realities. *Active Learning in Higher Education*, *9*(2), 172-192. https://doi.org/10.1177/1469787408091655

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing*, *30*(3), 403-412. https://doi.org/10.1177/0265532213480338

Taylor, K. L. & Colet, N. (2010). Making the shift from faculty development to educational development. In A. Saroyan & M. Frenay (Eds.), *Building teaching capacities in higher education* (pp. 139-167). Stylus.

Thompson, N. A. (2011). *Advantages of computerized adaptive testing* (CAT). https://assess.com/docs/Advantages-of-CAT-Testing.pdf

Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, *16*(1), 1.

Van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly*, *11*(4), 374-402. https://doi.org/10.1080/15434303.2014.960046

Watermark (2023). *The importance of assessment in higher education*. https://www.watermarkinsights.com/resources/blog/importance-of-assessment-in-higher-education

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Willis, J., Adie, L., & Klenowski, V. (2013). Conceptualising teachers' assessment literacies in an era of curriculum and assessment reform. *The Australian Educational Researcher*, *40*, 241-256. https://doi.org/10.1007/s13384-013-0089-9

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education*, *45*, 477-501. https://doi.org/10.1023/A:1023967026413

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education*, *16*(4), 323-342. https://doi.org/10.1207/S15324818AME1604_4

Appendix A.
Reading List for CAT

1. What is CAT? http://www.iacat.org/what-is-cat

2. Elements of Adaptive Testing https://doi.org/10.1007/978-0-387-85461-8

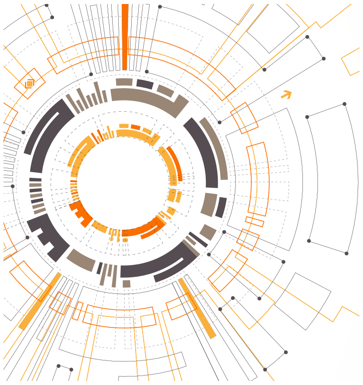
3. First Adaptive Test http://www.iacat.org/first-adaptive-test

4. Some Current Issues in CAT http://www.iacat.org/some-current-issues-cat

5. Computer-Adaptive Testing: A Methodology whose Time Has Come https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d41985bb9c543b94c60fc7dc6ab5e0ca31b8362f

6. The Impacts of Computer Adaptive Testing from a Variety of Perspectives https://doi.org/10.3352/jeehp.2017.14.12

Abstract
This article provides a model for robust student affairs program assessment using diverse data sources, multiple outcomes, propensity score matching, and cost analysis. Contemporary outcome-based assessment in student affairs requires equity-minded approaches paired with methods that support causal inference and actionable results. As a guide for practitioners, we summarize the approach we used to assess a student success program at a large, research-intensive, Hispanic Serving Institution. We describe the team structure, data sources, and analytical framework for assessing the student journey from admissions through graduation. Essential project management templates such as a sample logic model, data tables, and a project calendar are also included. These tools and strategies can be successfully adapted to meet contemporary assessment needs in student affairs at institutions of all types.



AUTHORS
Kendra Thompson-Dyck, Ph.D.
The University of Arizona

Michelle Sogge, MPA
The University of Arizona

Lucas Schalewski, Ph.D.
Columbia College Chicago

Alexandra Robie, EdD
The University of Arizona

Leveling Up Outcome-Based Assessment: Using Propensity Score Matching and Cost Analysis to Meet Contemporary Assessment Needs

Outcome-based assessment is at the forefront of student affairs practice (Biddix, 2018; Henning & Roberts, 2016; Schuh et al., 2016). Robust assessment is increasingly an expectation for student affairs professionals. There remains an ongoing need for high-quality reviews that support causal inferences between student participation and impact to ensure educational advancement for all (Henning et al. 2023, Horst et al., 2022; Montenegro & Jankowski, 2020). Outcome-based assessments are also critical evidence for university accreditation requirements (Gordon et al., 2019; Levy et al., 2018). Further, fiscal challenges in higher education make it pertinent to identify cost-effective strategies that support student success.

Unfortunately, methodological approaches have not kept pace with the changing landscape of assessment needs. Horst and colleagues' (2022) review of student affairs journal articles found program effectiveness claims were often not sufficiently supported in the corresponding methods. They suggest more rigorous assessment training for student affairs professionals. New approaches are needed to demonstrate credible evidence of success program effectiveness and produce useful insights for continuous improvement (Henning & Roberts, 2016).

To meet this need, we offer a recent success program study as an illustrative example. Our analytical approach combines descriptive, inferential, and cost analysis methods to examine program metrics across different domains. We describe our assessment

CORRESPONDENCE
Email
kthompd@arizona.edu

process and methodological choices using concrete examples and templates so that student affairs professionals can apply these tools to their work. We highlight the use of propensity score matching and cost analysis as advanced statistical techniques for assessing impact and cost.

Background

Outcome-based Assessment

Outcome-based assessment within student affairs is a systematic approach to gathering, analyzing, and interpreting data to evaluate a program's effectiveness in meeting its goals and to use those findings to make improvements (Henning & Roberts 2016; Schuh et al., 2016). Outcome-based assessment is essential to ensure that research-based practices used to design new high-impact programs are effective for the specific context where they are implemented (Finney & Buchanan, 2021).

Bresciani Ludvik (2019) used the analogy of a mechanic to illustrate the purpose of outcomes-based assessment. Mechanics run diagnostics on a vehicle's performance indicators to discern if optimal performance is met. If this optimal state is not achieved, performance indicators may point towards further diagnostics or analysis. As assessment professionals may act as mechanics to student learning and success, outcomes-based assessment provides core data points that inform how students can perform optimally (e.g., student learning and success). Bresciani Ludvik (2019) also emphasized the importance of studying individual students' experiences and outcomes to improve overall university performance. Further, efforts to disaggregate performance indicators by student identity (i.e., race, gender, first-generation) support equity-driven practices.

Financial assistance is a critical component. This recognizes that socio-economic status is strongly correlated with college enrollment and completion and that need-based aid boosts student retention, while financial stress increases the likelihood of discontinuing college.

Equity-minded Assessment

Historically, many assessment methods have failed to examine different outcomes, needs, and experiences between groups. Leaders in the field encourage assessment that is rooted in theoretical frameworks such as cultural competency and deploys meaningful data disaggregation (McNair et al., 2020; Montenegro & Jankowski, 2020). Bourke (2017) encourages using data to support action for social justice. Framing data and results in an equity-minded manner can produce novel critical questions to shape campus dialogues. Such frameworks are particularly important for minority serving institutions, including Hispanic Serving Institutions (HSIs). While the federal designation for HSIs is based on thresholds of Hispanic and low-income student enrollment, the concept of what it means for institutions to serve Hispanic students, referred to as "Servingness," embraces a holistic approach (Garcia 2020). A Servingness framework centers culturally affirming, transformative educational experiences for students that lead to positive academic and non-academic outcomes (Garcia 2020). Assessment is key to measuring and improving institutional capacity to serve Hispanic students equitably (Franco & Hernandez, 2018; Garcia et al., 2019). Equity-mindedness should be woven throughout the entire assessment process to the greatest extent possible (Montenegro & Jankowski, 2020).

Example Student Success Program

Program History

In 2008, the university president established the student success program (SSP) to improve educational attainment and upward mobility for low-income residents. Table 1 shows the program logic model which includes high-impact practices such as financial, academic, social, and emotional support associated with student retention (Collings et al., 2014; Nora & Crisp, 2007; Snowden & Hardy, 2012).

Financial assistance is a critical component. This recognizes that socioeconomic status is strongly correlated with college enrollment and completion (Engberg & Wolniak, 2014; Palardy, 2013; Wilbur & Roscigno, 2016) and that need-based aid boosts student retention (Bettinger, 2004; Bettinger, 2015; Millea et al., 2018), while financial stress increases the likelihood of discontinuing college (Britt et al., 2017). At its inception, funding was expansive.

Table 1
Simplified Student Success Program Logic Model

Resources and Inputs	Activities	Outputs	Outcomes
Operational Staff	Invite eligible students	Distributed funds to participants	**Short and Mid-Term:** Increased retention
Funding	Award financial support	Program participation	Increased financial, social, and academic support
Program Staff	Engage program participants in affiliated programs	Advisor meetings	Greater sense of belonging on campus and engagement
Program Partners	Renew participation based on program, financial, and academic requirements	Academic outcomes (GPA 2.0 or greater)	**Long Term:** Increased graduation
		FAFSAs completed	Less debt at graduation
			Better job and graduate school placement
			Upward mobility
			Invest in state and communities

A combination of federal and institutional grant aid covered tuition and fees, housing, food, and books. Gift aid was awarded up to the cost of attendance, resulting in a significant reduction in student loan borrowing for participants. Funding at the time of the program review shifted to a flat amount of $10,000 per student per year, combined with other federal, institutional, or private aid a student received.

Tailored support services are required for students to receive continued funding. First-year students receive one-on-one peer mentoring and workshops on adjusting to college, building academic skills, and making connections. Identity-specific mentoring is offered to students who seek community based on their racial, sexual, or first-generation college student identity, which is a successful strategy supporting historically minoritized groups (Queener & Ford, 2019). Second-year programs emphasize campus involvement, leadership development, and experiential learning while third- and fourth-year options focus on career development, graduate/professional school planning, and preparing for life after college. This comprehensive suite of services was designed with a culturally responsive, asset-based philosophy that recognizes and honors the vast knowledge and skills students bring with them to college (González et al., 2006).

Prior Assessment Efforts

Beginning in 2013, success program staff collected and stored extensive student-level data annually including participant sex, race/ethnicity, high school, major, and a detailed record of support programs usage. These data were regularly used to analyze retention and graduation annually, but a holistic review of the program efficacy or cost-effectiveness was needed.

Analytical and Methodological Approach to the Comprehensive Review

Program Assessment Team Structure

Cochran et al. (2018) recommend an assessment team with internal and external perspectives. This composition can reduce bias while also retaining integral program context and buy-in from those who are close to the program. A multi-person review team can bring diverse personal and professional experiences and identities to the assessment process which supports equity-mindedness.

Our program review team included four on-campus staff members: two internally situated and two externally situated in relation to the program. The internal team provided an in-depth understanding of the program's history, context, and data. The other two team members were from a centralized assessment office for university student affairs and student services. This composition offered nuanced and balanced viewpoints, which we recommend as standard practice for comprehensive assessment whenever possible.

Assessment that Reflects the Student Journey: Beyond Single Outcomes

Henning and Roberts (2016) emphasize that student success programs with broad goals should not be narrowly assessed. Since the cohort-based student success program (SSP) was designed to support access, retention, graduation, and post-graduation outcomes, we designed an assessment to examine program participants' experiences and outcomes throughout the undergraduate journey, starting with admissions and culminating with post-graduation metrics. Since the program provided direct financial aid, campus leaders were interested in understanding the impact of substantial aid and the associated costs and benefits.

A multi-person review team can bring diverse personal and professional experiences and identities to the assessment process which supports equity-mindedness.

We collected and analyzed data from across the institution and student journey (e.g., admissions yield, post-graduation outcomes) as shown in Table 2. We began with traditional descriptive and comparative analyses common in student affairs assessments before adding predictive models and cost-effectiveness analyses to level up our approach (Schuh et al., 2016). This framework could be adapted for teams with varying capacities and time constraints.

Data Collection and Integration

Rather than collect new data, we maximized existing data and leveraged local data steward expertise. This strategy reduces campus survey fatigue that can lead to low response rates (Porter et al., 2004). Montenegro and Jankowski (2020) view this practice as equity-minded since it reduces the data collection burden on students, particularly historically minoritized groups.

However, historical program-level participation datasets are not always clean and readily available. Having designated data stewards who manage the responsible acquisition, cleaning, storing, and use of data and metadata is essential for a comprehensive program review (Plotkin, 2014; Rosenbaum, 2010). Our strategy was to collate siloed departmental data into an expansive dataset across domains and functional areas (e.g., institutional student data records, admissions, financial aid, summer transition programs, fraternity and sorority programs, leadership programs, academic support, campus recreation, student engagement, and career development). This was possible because our team had trusted relationships with other student data stewards across the university through a student data coalition that meets monthly. These colleagues provided essential context for data analysis (e.g., historical trends in data coding/collection) and situated our findings in relation to departmental and university policies and practices.

First, we made an explicit request to each functional area partner and discussed the larger purpose of the project. We scheduled 'Collegial Check-In' meetings after receiving the data to ensure that our use and interpretations were correct. For example, we worked closely with the financial aid data team to clarify our understanding of financial aid award codes during cleaning and coding, as well as during the reporting phase when we translated data insights into recommendations. These steps were crucial to ensure validity and improve trust in information-sharing among partners before sharing findings with campus leaders. Setting this expectation with partners for collaboration at the beginning of an assessment can increase buy-in and dispel trepidation about assessment findings.

Having designated data stewards who manage the responsible acquisition, cleaning, storing, and use of data and metadata is essential for a comprehensive program review

Propensity Score Matching: Determining Differential Impact

Experimental or quasi-experimental approaches can improve the credibility of educational research. These methods account for the counterfactual, the control condition where the program experience or 'treatment' is not administered, to ascertain the average treatment effect of the program among participants (Murnane & Willet, 2010; Horst et al., 2022). It is often not feasible or desirable to utilize a true randomized control trial in higher

Table 2
Assessment Domains and Methods Used

Domain	Analysis	Methodology	Inferences
Admissions	Yield rate of students selected for SSP and comparison groups (e.g., eligible, applied not selected, eligible, did not apply)	Descriptive statistics	Descriptive
Demographic, Academic and Financial Aid	Student background / financial aid comparison participants vs. eligible non-participant peers	Pearson's chi-square or t-test measures of association by subgroup	Descriptive; Comparative
First Year Program *within the affiliated support programs* for SSP Participants	Participation rates, disaggregated by student characteristics (race/eth., first gen, gender). Retention by program choice	Pearson's chi-square measures of association by subgroup	Descriptive; Comparative
Participation in Other Cocurricular Activities	Rates of engagement in other programs, SSP participants vs. eligible non-SSP peers	Pearson's chi-square measures of association by subgroup	Descriptive; Comparative
Retention and Graduation – Descriptive	YR1 Retention, YR2, YR3 Persistence, YR4, YR5, YR6 Graduation rates of SSP participants and non-SSP peers	Rates among pooled cohorts (multiple years)	Descriptive
Retention and Graduation – Propensity Score Matching	YR1 Retention, YR4, YR5, YR6 Grad rates between SSP and statistically matched comparison group within non-SSP peers	PSM to produce rates among pooled cohorts, disaggregated by subgroups (e.g., first gen Latinx female)	Predictive
Drivers of First Year Retention	Factors that predict retention among SSP participants (demographic, financial, academic, participation)	Logistic regression driver analysis	Predictive
Cost-Effectiveness Analysis	Dollars spent per additional student retained and graduated due to program participation	Cost-effectiveness ratio analysis	Cost Analysis
Loan Debt among Graduates	Average loan accumulation of SSP vs. eligible peers	Descriptive statistics	Descriptive; Comparative
Post-Graduation Career and Graduate School	Rates of employment, continuing education of SSP vs. peers	Descriptive statistics	Descriptive; Comparative
University Foundation Fund-Development and Endowments	Foundation funds to support endowments and scholarships	Descriptive statistics	Descriptive

education programs. One method we increasingly recommend for program assessment is Propensity Score Matching, a quasi-experimental approach.

Propensity score matching (PSM) is an alternative to randomized control trials that creates a control condition using statistics to compare outcomes between groups of participants and non-participants (see Harris and Horst's 2016 article for a step-by-step guide). Specifically, PSM generates a balanced comparison group by matching students on variables or covariates that are predictors of self-selecting into the program and the outcome(s) of interest. Propensity score values are generated by a logistic regression model predicting participation and reflect the probability of students participating in a program. These scores are then used to match participants with non-participants. The program's effect is evaluated by comparing the average treatment effects for the participant group in relation to the statistically similar non-participant group. This method has been used to demonstrate the impact of fraternity and sorority membership (Holmes & Bowman, 2017), an engineering grading program (Novak et al., 2016), and honors program participation (Keller & Lacy, 2013) on student success. Propensity score matching can demonstrate the differential impact of the program overall and by subgroup (e.g., first-generation, Black, male students).

Specifically, PSM generates a balanced comparison group by matching students on variables or covariates that are predictors of self-selecting into the program and the outcome(s) of interest.

The first and most critical step within PSM is to identify the appropriate covariates. A core assumption is that all potential cofounding variables that are related to both the selection into treatment and our intended outcome have been included. In higher education, common covariates include student demographics, academic preparation and achievement, and campus engagement indicators. Critically, these must have a hypothesized relationship with either program participation, the outcome of interest, or both (Harris & Horst, 2016).

Program evaluations with a cost component are relatively uncommon in higher education but provide value to decision-makers for contextualizing the return on institutional investments.

The PSM model first conducts a logistic regression analysis predicting participation in treatment from the covariates. These scores are used to generate balanced treatment and comparison groups, which the analyst verifies using post-estimation commands. Once balanced groups are generated, the outcome analysis is performed to demonstrate the difference in average treatment effects between the treated and untreated groups.

In most cases, a one-to-one nearest neighbor matching method is used so that each program participant is matched with a statistically similar non-participant peer. A caliper threshold (e.g., 0.2) can be set by the analyst to limit the absolute distance between propensity scores suitable to be matched to ensure a high-quality comparison group (Austin, 2009; Stuart, 2010; Harris & Horst, 2016). A one-to-one nearest neighbor matching approach with no replacement where non-participants are matched only once has been shown to reduce bias between various PSM techniques (Austin, 2014; Caliendo & Kopeinig, 2008).

In our assessment, we first limited the analytical dataset to program participants as well as non-participants who met the initial eligibility criteria and could have participated but did not. Then, we identified covariates associated with program participation and key outcomes of retention and graduation including student demographics, academic background, financial need, and financial aid award package indicators. A model with 10 covariates using one-to-one nearest neighbor matching, a caliper width of 0.2, and no replacement created an appropriate, balanced comparison group. We then compared the average treatment effects of participants to non-participants overall for retention and graduation rates, reported as a percentage point difference. In applying an equity lens to our assessment, we provided supplemental breakouts by subgroups, such as for first-generation, Hispanic, female students, which indicated even greater returns from the program for students from typically marginalized communities or identities. Appendix A shows sample results as an example using synthetic figures.

Cost Analysis

We then conducted a cost analysis to identify and monetize the various inputs required to support the program. Program evaluations with a cost component are relatively uncommon in higher education but provide value to decision-makers for contextualizing the return on institutional investments (Henning & Roberts, 2016). The scope can vary from costs and benefits borne within an organization, across a national program, or for society at large. For example, Levin and Garcia (2018) considered the cost-savings to the taxpayer for investment in community college programs in the state of New York, whereas Bowden and Belfield (2015) examined the Talent Search TRIO programs nationwide, and Walcott et al. (2018) framed undergraduate research initiatives in terms of students' earnings potential in relation to the university's costs.

Cost-effectiveness identifies and monetizes program implementation costs in cases where the benefits are difficult to monetize or where analysts want to compare alternate interventions seeking to impact a similar outcome (Cellini & Kee, 2015). This method distills costs and outcomes into a cost-effectiveness ratio with costs as the numerator and the unit of effectiveness as the denominator such as "dollars per dropout prevented," (Cellini & Kee, 2015, p. 637).

We used a cost-effectiveness framework to identify the costs of retaining and graduating participants using the procedural approach described by Cellini and Kee (2015). In our case, we applied an organizational lens to consider costs and benefits to the university, acknowledging that this does not account for alternative costs and benefits to other stakeholders (e.g., students, program staff, and state).

Conceptually, costs were divided into direct financial aid award costs from program-specific scholarships and grants, and program implementation costs (e.g., staff salaries, employee-related expenses, and materials/training). In consultation with the budget office and campus partners, we determined that indirect costs such as facility space and services provided by ancillary units should be excluded since those would not be reappropriated or realize any cost savings if the program were to be discontinued. Direct aid and program cost totals across years were combined to provide a total cost of program administration for seven cohorts in their first year of program participation.

Since returning students increase tuition revenue, we provided a conservative adjustment to the costs based on Federal Pell Grant monies. Nearly all participants received a Pell Grant, so continued enrollment among this population translates into tuition dollars from federal grant funding. In consultation with the budget and financial aid offices, we reduced the total costs for the university by the average of the maximum Pell Grant amount across the years of study, multiplied by the number of additionally retained students. Although additional revenue streams through increases in student persistence do occur (i.e., Housing & Residential Life, Bookstore, Athletics), these amounts were inestimable and were not included in the revenue calculation.

An example table in Appendix A shows the high-level reported cost calculations, though more detailed year-by-year breakdowns were included in the full report. The key figure divides the total cost by the number of additionally retained students, determined by the PSM modeling, to obtain a calculated cost per student. Given the sensitivity of aligning a dollar amount with a specific program and intervention for the first time without any comparison program figures, our evaluation team cross-checked our cost model at each stage of the process with the program director, budget office, scholarship and financial aid, and other leadership members before reporting our findings internally.

Audiences will read public or widely shared reports with varying degrees of background and data literacy, necessitating clear deliverables that communicate insights, not just data.

While the cost-effectiveness analysis produced useful and actionable insights, it also risked causing inadvertent sticker shock for those unfamiliar with the level of institutional investment required to administer a success program of this size. The program review team was also apprehensive of conducting this novel assessment on a program that primarily serves low-income, first-generation, students of color. We acknowledge that programs and services designed to attract and retain high-socioeconomic students exist and likely carry similar, or even greater, price tags. Therefore, we recommend that program cost reporting should always be contextualized with other similar programs to reduce sticker shock and prevent the unintentional targeting of cuts to programs that support historically underserved populations.

Communication and Use of Findings

We used a multifaceted, targeted communication plan to ensure insights were used for program improvement (Bourke, 2017). It is critically important to bridge insight-to-action, especially with lengthy data reports (Henning & Roberts, 2016). High-level decision-makers often have limited program-specific knowledge and time to digest nuanced information. Audiences will read public or widely shared reports with varying degrees of background and data literacy, necessitating clear deliverables that communicate insights, not just data. This step reduces the chances of misinterpretation or misuse.

The review team created one public and one confidential report. The public report was widely available to provide transparency (Montenegro & Jankowski, 2020). The confidential version was for university-affiliated members only and included sensitive details on the budget (e.g., salary, fundraising). The report purposefully integrated context-specific divisional and institutional language informed by the strategic plan, mission statements, and metrics that leadership has top of mind (Henning & Roberts, 2016). Importantly, we authored substantive recommendations based on the results (e.g., optimize award allocation with larger awards vs. larger cohort). Data visualization components were added to enhance communication (Evergreen, 2020). We sent tailored email memos to concisely share the most relevant and useful findings for decision-makers with the full reports attached.

To maximize the use of insights, we invited senior leaders, student affairs administrators, budget office representatives, and program staff members to a virtual presentation and debrief. The goal was to talk through the assessment process, highlight key recommendations, answer questions, and discuss the recommendations with the added context and perspectives provided by the audience. We distributed the confidential report two weeks in advance.

These communication strategies facilitated productive discussion by decision-makers on the report's evidence and aligned recommendations. Following those conversations, a debrief meeting with the program review team occurred where nearly all our recommendations were adopted in practice. Table 3 is the model that was used during this meeting, which ties each recommendation to a policy and practice plan for action.

Leveraging key partnerships and engaging in thoughtful, strategic communication of assessment findings were key to maximizing the use of data-informed insights.

Table 3
Template Model for Demonstrating Use of Assessment Findings

Recommendation	Additional Information	Proposed Plan
Restate the evidence-based recommendation here.	Space to provide additional institutional context and information to broaden understandings.	Identify the actions taken based on the recommendation moving forward.

Project Timetable

Considerable time and staff resources are required to execute a comprehensive review successfully. In our case, we had the advantage of clean historical data and an established collegial network of data stewards. In this best-case scenario, the project unfolded over 12 months. A detailed timetable located in Appendix B provides the month-by-month breakdown of activities which may be useful for replication or modification based on institutional needs and staff capacity.

Limitations

As with all assessment projects, there are limitations. In most cases, mixed-methods approaches are ideal to support equity-mindedness and integrate student voices (Henning et al. 2023). Due to time, resources, and decision-maker priorities, we used exclusively quantitative methods. Future comprehensive program reviews are expected to incorporate learning assessment as well as qualitative data. We recommend aligning methodological choices with institutional priorities, integrating equity-centered strategies in data collection and reporting, and using mixed methods when time and resources allow.

Conclusion

In this article, we reviewed the process used to execute a comprehensive outcome-based assessment project on a long-standing student success program at a large, Research I, Hispanic Serving Institution. We detailed various project management and methodological considerations to provide fellow assessment professionals with a valuable roadmap to replicate similar work on their campuses. Leveraging key partnerships and engaging in thoughtful, strategic communication of assessment findings were key to maximizing the use of data-informed insights. We support Henning and Robert's (2016) claim that assessment professionals are not passive evaluators, but agents charged to work with stakeholders to facilitate planning and action. Student affairs and higher education stakeholders will continue to require data driven causal inferences to determine the impact of student success program participation. Assessment methodological approaches should continue to evolve to meet the profession's current and future needs.

References

Austin, P. C. (2009). Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal*, *51*(1), 171-184. https://doi.org/10.1002/bimj.200810488

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, *33*(6), 1057-1069. https://doi.org/10.1002/sim.6004

Bettinger, E. (2004). How financial aid affects persistence. In C. Hoxby (Ed.), College choices: The economics of where to go, when to go, and how to pay for it (pp. 207-238). University of Chicago Press. https://www.nber.org/papers/w10242

Bettinger, E. (2015). Need-based aid and college persistence: The effects of the Ohio College Opportunity Grant. *Educational Evaluation and Policy Analysis*, *37*(1), 102S-119S. https://doi.org/10.3102/0162373715576072

Biddix, J. P. (2018). Research methods and applications for student affairs. Jossey-Bass.

Bourke, B. (2017). Advancing towards social justice via student affairs inquiry. *Journal of Student Affairs Inquiry, Improvement, and Impact*, *3*(1), 1-18. https://doi.org/10.18060/27837

Bowden, A. B., & Belfield, C. (2015). Evaluating the Talent Search TRIO program: A benefit-cost analysis and cost-effectiveness analysis. *Journal of Benefit-Cost Analysis*, *6*(3), 572-602. https://doi.org/10.1017/bca.2015.48

Bresciani Ludvik, M. J. (2019). What makes a performance indicator an equity-driven, high performance indicator? *Assessment Update*, *31*(2), 1-2, 15-16. https://doi.org/10.1002/au.30163

Britt, S. L., Ammerman, D. A., Barrett, S. F., & Jones, S. (2017). Student loans, financial stress, and college student retention. *Journal of Student Financial Aid*, *47*(1), 3. https://eric.ed.gov/?id=EJ1141137

Caliendo, M. & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31-72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Cellini, S. R. & Kee, J. E. (2015). Cost-effectiveness and cost-benefit analysis. In: Newcomer, K. E, Hatry, H. P., & Wholey, J. S. (Eds.), *Handbook of practical program evaluation* (4th ed., pp. 636-672). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119171386.ch24

Cochran, M. F., Shefman, P. K., & Hettiarachchi, M. M. (2018). Assessing the assessors: Views from the inside and outside. *Journal of Student Affairs Inquiry, Improvement, and Impact*, *4*(1), 1-18. https://doi.org/10.18060/27909

Collings, R., Swanson, V., & Watkins, R. (2014). The impact of peer mentoring on levels of student wellbeing, integration and retention: A controlled comparative evaluation of residential students in UK higher education. *Higher Education*, *68*, 927–942. https://doi.org/10.1007/s10734-014-9752-y

Engberg, M. E. & Wolniak, G. C. (2014). An examination of the moderating effects of the high school socioeconomic context on college enrollment. *The High School Journal*, *97*(4), 240–263. https://doi.org/10.1353/hsj.2014.0004

Evergreen, S. (2020). *Effective data visualization: The right chart for the right data* (2nd edition). Sage Publications.

Finney, S. J., & Buchanan, H. A. (2021). A more efficient path to learning improvement: Using repositories of effectiveness studies to guide evidence-informed programming. *Research & Practice in Assessment*, *16*(1), 36-48. https://eric.ed.gov/?id=EJ1307022

Franco, M. A., & Hernández, S. (2018). Assessing the capacity of Hispanic serving institutions to serve Latinx students: Moving beyond compositional diversity. *New Directions for Institutional Research*, *2018*(177), 57-71. https://doi.org/10.1002/ir.20256

Garcia G. A. (Ed.). (2020). *Hispanic serving institutions (HSIs) in practice: Defining "Servingness" at HSIs*. Information Age Publishing.

Garcia, G. A., Núñez, A. M., & Sansone, V.A. (2019). Toward a multidimensional conceptual framework for understanding "servingness" in Hispanic-serving institutions: A synthesis of the research. *Review of Educational Research*, *89*(5), 745-784. https://doi.org/10.3102/0034654319864591

González, N., Moll, L. C., & Amanti, C. (Eds.). (2006). *Funds of knowledge: Theorizing practices in households, communities, and classrooms*. Routledge.

Gordon, S. R., Shefman, P., Heinrich, B., & Gage, K. (2019). The role of student affairs in regional accreditation: Why and how to be included. *Journal of Student Affairs Inquiry, Improvement, and Impact 5*(1), 1-25. https://doi.org/10.18060/27916

Harris, H. & Horst, S. J. (2016). A brief guide to decisions at each step of the propensity score matching process. *Practical Assessment, Research & Evaluation*, *21*(4), 1-11. https://doi.org/10.7275/yq7r-4820

Henning, G. W., Rice, A., Heiser, C., & Lundquist, A. E. (2023). Equity-centered assessment practices: Survey findings and recommendations. *Research & Practice in Assessment*, *18*(2). https://www.rpajournal.com/dev/wp-content/uploads/2024/03/Equity-centered-Assessment-Practices-RPA.pdf

Henning, G. W., & Roberts, D. (Eds). (2016). *Student affairs assessment: Theory to practice*. Stylus Publishing. https://doi.org/10.4324/9781003447139

Holmes, J. & Bowman, N. (2017). A quasi-experimental analysis of fraternity or sorority membership and college student success. *Journal of College Student Development*, *58*(7), 1018-1034. https://doi.org/10.1353/csd.2017.0081

Horst, J. S., Finney, S. J., Prendergast, C. O., Pope, A. M., & Crewe, M. (2022). The credibility of inferences from program effectiveness studies published in student affairs journals: Potential impact on programming and assessment. *Research & Practice in Assessment*, *16*(2), 17-32. https://eric.ed.gov/?id=EJ1348828

Keller, R. R. & Lacy, M. G. (2013). Propensity score analysis of an honors program's contribution to students' retention and graduation outcomes. *Journal of the National Honors Council*, *14*(2), 73-84. https://files.eric.ed.gov/fulltext/EJ1082022.pdf

Levin, H. M. & Garcia, E. (2018). Accelerating community college graduation rates: A benefit-cost analysis. *The Journal of Higher Education*, *89*(1): 1-27. https://doi.org/10.1080/00221546.2017.1313087

Levy, J., Hess, R., & Thomas, A. (2018). Student affairs assessment and accreditation: History, expectations, and implications. *Journal of Student Affairs Inquiry*, *4*(1), 1-19. https://doi.org/10.18060/27888

McNair, T. B., Bensimon, E. M. & Malcom-Piqueux, L. (2020). *From equity talk to equity walk: Expanding practitioner knowledge for racial justice in higher education*. Jossey Bass.

Millea, M., Wills, R., Elder, A., & Molina, D. (2018). What matters in college student success? Determinants of college retention and graduation rates. *Education*, *138*(4), 309-322. https://aalhe.scholasticahq.com/article/24575-equity-in-assessment-the-grand-challenge and-exploration-of-the-current-landscape

Montenegro, E., & Jankowski, N. A. (2020). A new decade for assessment: Embedding equity into assessment praxis. National Institute for Learning Outcomes Assessment. https://www.learningoutcomesassessment.org/wp-content/uploads/2020/01/A-NewDecade-for-Assessment.pdf

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Nora, A., & Crisp, G. (2007). Mentoring students: Conceptualizing and validating the multidimensions of a support system. *Journal of College Student Retention: Research, Theory & Practice*, *9*(3), 337–356. https://doi.org/10.2190/CS.9.3.e

Novak, H., Paguyo, C., & Siller, T. (2016). Examining the impact of the engineering successful/unsuccessful grading program on student retention: A propensity score analysis. *Journal of College Student Retention*, *18*(1), 83-108. https://doi.org/10.1177/1521025115579674

Palardy, G. J. (2013). High school socioeconomic segregation and student attainment. *American Educational Research Journal*, *50*(4), 714-754. https://doi.org/10.3102/0002831213481240

Plotkin, D. (2014). Data stewardship: An actionable guide to effective data management and data governance. Elsevier Inc. https://doi.org/10.1016/C2012-0-07057-3

Porter, S. R., Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, *2004*(121), 63-73. https://doi.org/10.1002/ir.101

Queener, J. E., & Ford, B. A. (2019). Culturally responsive mentoring programs: Impacting retention/graduation rates of African American males attending predominately White institutions. In *Overcoming challenges and creating opportunity for African American male students* (pp. 120-132). IGI Global.

Rosenbaum, S. (2010). Data governance and stewardship: Designing data stewardship entities and advancing data access. *Health Services Research*, *45*(5p2), 1442-1445. https://doi.org/10.1111/j.1475-6773.2010.01140.x

Schuh, J. H., Biddix, J. P., Dean, L. A., & Kinzie, J. (2016). *Assessment in student affairs*. (2nd edition). Jossey-Bass.

Snowden, M. & Hardy, T. (2012). Peer mentorship and positive effects on student mentor and mentee retention and academic success. *Widening Participation and Lifelong Learning*, *14*, 76–92. https://doi.org/10.5456/WPLL.14.S.76

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1-21. https://doi.org/10.1214/09-STS313

Walcott, R. L., Corso, P. S., Rodenbusch, S. E., & Dolan, E. L. (2018). Benefit–cost analysis of undergraduate education programs: An example analysis of the freshman research initiative. *CBE—Life Sciences Education*, *17*(1). https://doi.org/10.1187/cbe.17-06-0114

Wilbur, T. G., & Roscigno, V. J. (2016). First-generation disadvantage and college enrollment/completion. *Socius*, *2*. https://doi.org/10.1177/2378023116664351

Appendix A
Example Data Tables

The following data are artificial and are examples of how results may be shared. Actual results are not presented given data restrictions.

Example Table of PSM Results with Artificial Data

First Year Student Retention

Cohorts 2013-2019	Retention YR1-YR2	% Point Diff.	# of Additional Students
Participants (n=2,005)	87.8% (n=1,760)	+9.1***	+182
Matched Non-Participants (n=2,005)	78.7% (n=1,578)		

***p<.000

First Year Retention by Subgroup

Cohort 2013-2019	First Gen White Female	First Gen White Male	First Gen Hispanic Female	First Gen Hispanic Male
Participants	86.3%	77.4%	86.1%	88.1%
Matched Non-Participants	74.0%	74.8%	78.0%	74.7%
% Point Difference	12.2***	2.6	8.1***	13.4***
Participant (n=)	200	112	780	402
Non-Participant (n=)	200	112	780	402

***p<.000, **p<.01, *p<.05

Example Table of Cost-Effectiveness Ratio Calculation for Program Impact on YR1 Retention

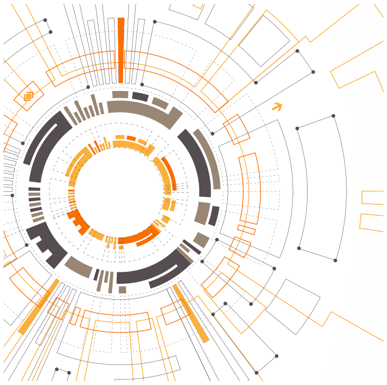
Direct Aid Financial Aid in Cohort Entry Year	$ Total A
Program Costs 2013-2019	$ Total B
Total YR1 Costs 2013-2019	$ Total C = (A+B)
Additional Students Retained due to Participation	
Participants Retained (87.8%)	1,760
Matched Peers Retained (78.7%)	1,578
Additional Participants Retained (9.1%)	182
Federal Pell Grant $ Gain from Additionally Retained Students	
Avg. Max Pell Grant 2014-2020 due to Retention ($ Avg. X 182)	$ Total D
Calculated Costs Per Student	
Total YR1 Costs Minus Federal Pell Grant $ Gained	$ Total E = (C – D)
Cost per Additional Participant Retained (182)	**$ Per Student (E / 182)**

Cost-Effectiveness Ratio = Costs / Units of Effectiveness

Dollar per additional student retained = $ Total Costs E / 182

Appendix B
Program Review Timeline

Assessment Activity Timetable for Comprehensive Review	
October	Form assessment team. Develop early framework for analyses, context, and data needs.
November	Hold information gathering meetings with campus stakeholders (Financial Aid, Enrollment Management, Program Directors, Foundation) to determine context, eligibility criteria for invitation, logic model, historical changes, cost framework.
December	Provide detailed financial aid data request and justification.
January	Conduct preliminary analyses. Collegial check-in meetings on financial aid data context, use, implications.
February	Conduct analyses. Solicit data tables from partners in enrollment, career development, and foundation.
March	Preview preliminary data analyses and insights with all contributing data stewards and stakeholder departments.
April	Develop recommendations section in consultation with Program Director.
May	Present written final report and PowerPoint presentation to leadership committee including Provost.
Summer	Attend follow-up meetings with leaders to clarify understanding of insights and provide additional data points, as requested. Final report without cost data distributed to campus.
Fall	Leadership committee subgroup presentation of program changes that were adopted based on the comprehensive review, with modifications tied specifically to recommendations in the report.



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

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

Abstract

Some college students may be disengaged when completing assessments for institutional accountability and improvement. If disengagement is not identified and the resulting data are removed, the validity of score interpretations suffers. Using data gathered from students who completed non-consequential assessments for institutional accountability, we investigated disengagement on a non-cognitive assessment. We demonstrate how we identified students who rapidly responded to items, who "streamlined" answers across items, or who self-reported low effort. We hypothesized that some students would display at least one of these disengagement behaviors and that removing their data would result in scores that better aligned with the assessment's theoretical factor structure. Half the students who self-reported low effort and half the students who streamlined also rapidly responded. The theoretical two-factor structure of the non-cognitive assessment better represented scores after removing disengaged students. We discuss the practicality of selecting a motivation filtering technique to provide more accurate outcome assessment interpretations.

The Influence of Student Disengagement on a Non-Cognitive Measure: Practical Solutions for Assessment Practitioners

Assessment practitioners generally assume that scores from assessments meaningfully represent some intended construct. That is, the goal is to gather scores with a high degree of validity whether scores are collected under high-stakes conditions (e.g., classroom exams, certification testing, admissions testing) or low-stakes conditions (e.g., institutional accountability assessment, cross-country comparisons) via cognitive assessments where items are scored correct/incorrect (e.g., quantitative reasoning, critical thinking) or non-cognitive assessments with no correct answers (e.g., civic responsibility, global perspectives, growth mindset). However, we should not simply assume that scores represent some intended construct.

CORRESPONDENCE
Email
schae2ke@jmu.edu

As assessment practitioners, we must collect validity evidence to form an argument to support our interpretations of assessment scores, whether those interpretations are shared with accreditors, board of visitors, parents, students, or other stakeholders. *The Standards for Psychological and Educational Testing* (APA, AERA, & NCME, 2014) outline five sources of validity evidence: test content, response processes, internal structure, relations to other variables, and consequences. Explanations of each source go beyond the scope of this paper.

We will focus on internal structure as a critical piece of evidence impacting score interpretations and how student disengagement on outcomes assessments impacts internal structure.

Disengagement and Low-Stakes Testing

Student disengagement is of concern in low-stakes assessment contexts (e.g., institutional accountability and improvement, cross-country comparisons) and low-stakes data collection (e.g., program evaluations, surveys, data collected for research) contexts (Finn, 2015; Wise, 2006; Wise & DeMars, 2005). In low-stakes contexts, there is no personal consequence to students based on their scores. Hence, some students expend little effort. In turn, scores on both cognitive and non-cognitive assessments may not reflect the construct of interest if a subset of students are disengaged (e.g., Rios et al., 2022; Wise & Kong, 2005).

As assessment practitioners, we must collect validity evidence to form an argument to support our interpretations of assessment scores.

Given that disengagement is an issue in low-stakes assessment contexts, some researchers have attempted to proactively reduce disengagement during testing using techniques such as offering external incentives (e.g., Rios, 2021), increasing test relevance (e.g., Liu et al., 2015), and priming (e.g., Finney et al., 2024). Unfortunately, some methods to reduce disengagement may be costly (external incentives), not possible for certain institutions (increasing test relevance), or may be less effective for certain populations (priming). Furthermore, decreasing disengagement via these proactive methods may not completely eliminate disengagement. Finally, these techniques cannot be applied to data that has already been collected. Thus, in many cases, assessment practitioners can only evaluate if test scores have been contaminated by disengagement, rather than apply a proactive strategy to increase effort. If the disengagement is not addressed, incorrect score interpretations could be made.

Consider higher educational institutions that compute gain scores from low-stakes assessments administered before and after educational programming for accountability or program improvement purposes. If student disengagement is not addressed, the low effort will bias gain scores (Finney et al., 2016; Mathers et al., 2018). These biased value-added estimates can then lead to incorrect inferences about student learning and development. This negative effect of student disengagement is usually a concern for cognitive assessments. However, non-cognitive assessments are popular in higher education to evaluate students' attitudes, perceptions, values, and behaviors (e.g., sense of social belonging, civic engagement, goal orientation, career decision self-efficacy). Surprisingly, there is less discussion surrounding the effect of disengagement on non-cognitive assessments even though these developmental constructs are often the main outcomes of first-year seminars, co-curricular experiences, and student affairs programming.

Thus, the current study focuses on student disengagement when completing non-cognitive assessments. We demonstrate three pragmatic methods that higher education assessment practitioners can use to identify disengagement (streamlining, self-reported effort, and response time) even if the data has already been collected. We then use those methods to gather more robust evidence of the internal structure of scores from the non-cognitive assessment. Finally, we provide practical recommendations for motivation filtering for assessment specialists.

Influence of Disengagement on Internal Structure

The *Standards* state that assessment developers should ensure that items align as intended with the construct(s) of interest and evaluation of this alignment is often accomplished via factor analysis. Consider the Short Grit Scale (Grit-S) used on college campuses; it has some items written to reflect consistency of interest and some items to reflect perseverance of effort (Duckworth & Quinn, 2009). The developer hypothesized that consistency and perseverance influence students' responses to particular items. Structural validity is supported if the intended two-factor structure accounts for the item covariances.

However, beyond the item's content, the factor structure of scores can be influenced by student disengagement, which is an issue for the assessment practitioner who is collecting, analyzing, and interpreting the data. Consider an extreme case of students being completely disengaged on the Grit-S (Duckworth & Quinn, 2009). If students randomly selected responses to the Likert-style items, the two-factor structure would not emerge because the relations among the item responses would not reflect differences in the two constructs. Thus, the theoretical structure of the responses (which guides the scoring of the measures) would not match the empirical structure of the scores collected by the assessment practitioner (which complicates the reporting and interpretation of scores).

Given the concern of student disengagement on non-cognitive measures, possible influences have been investigated. Barry and Finney (2009) conducted a study where data was collected from college students for accountability and improvement efforts. The assessments were low-stakes for the students (did not affect their GPA), but high-stakes for the institution (scores were reported for accreditation). Barry and Finney varied the assessment contexts to be highly controlled (i.e., proctored, small room) to uncontrolled (i.e., unproctored, remote). Barry and Finney found that for the least controlled context, their data did not align with the measure's theorized internal structure. Although Barry and Finney did not measure disengagement directly, they inferred the internal structure differences across testing contexts were due to differences in disengagement across testing contexts. Fortunately, there are empirical methods to directly identify and address disengagement that are accessible to assessment practitioners.

Methods to Identify Disengagement

Multiple methods exist to identify disengagement. Once identified, people who are disengaged can be removed from the dataset, a technique called "motivation filtering" (e.g., Wise & Kong, 2005). Studies have shown that motivation filtering resulted in higher convergent validity (e.g., Wise & DeMars, 2005; Wise et al., 2004; Wise & Kong, 2005). We discuss and then demonstrate three practical strategies to identify disengagement on non-cognitive assessments and filter non-effortful responses. We present three strategies given the constraints of assessment processes across campuses. Our hope is that at least one of these approaches will be accessible to your assessment context.

Usefulness of Negatively Worded Items to Signal Disengagement

Non-cognitive measures that include both positively and negatively worded items are useful for the identification of extreme disengagement via "streamlined" responses. Streamlined or "longstring" responses occur when people select the same response option for every item, regardless of item wording (e.g., Curran, 2016; Meade & Craig, 2012). For example, five items on a measure may be positively worded (e.g., I am confident in my communication skills), and one item may be negatively worded (e.g., I have poor communication skills). Students rate the items on a scale of 1 to 5, ranging from strongly disagree to strongly agree. It is expected that students who agree or strongly agree with the positive items should disagree or strongly disagree with the negative items. However, disengaged students may not read the items at all and may select the same response (e.g., agree) for the negatively worded items as they did the positively worded items. If some students select the same response option for every item when wording dictates a different style of response, then the factor structure will reflect this via misfit of the intended factor structure. Indeed, assessment specialists employing non-cognitive measures with negatively worded items have used streamlined responses to identify disengagement (e.g., Curran, 2016; Hong et al., 2020; Kupffer et al., 2024; Meade & Craig, 2012).

Negatively worded items are particularly useful to detect extreme disengagement via streamlined responses.

Unfortunately, negatively worded items have a complicated history. Negatively worded items were recommended for inclusion in non-cognitive assessments to help identify acquiescence or disengagement (Bandalos, 2018). However, for decades negatively worded items have been criticized because item valence has been shown to influence the factor structure of scores (Barnette, 2000; Dalal & Carter, 2015; DiStefano & Motl, 2006; Woods, 2006). In particular, factor analyses of non-cognitive assessments with both negatively and positively worded items often result in factors representing item wording (Dalal & Carter, 2015; Ponce et al., 2022). Yet, this resulting factor structure is exactly what one would expect if some students were disengaged; thus, the factor structure signals issues with the quality of responses due to low engagement. However, research also suggests that negatively phrased words can be difficult to comprehend compared to positively phrased words (e.g., Dalal & Carter, 2015; Marsh, 1986, 1996) and that negatively worded items add a level of complexity that may result in misresponse (Dalal & Carter, 2015; Swain et al., 2008). Thus, even when students are engaged, they may struggle to mentally "flip" the meaning of negatively worded items in order to respond in a way that aligns with the responses to the positively worded items, resulting in factors that represent item valence. Hence, researchers have challenged the use of negatively worded items, with some recommending against their use entirely (e.g., Lindwall et al., 2012; Quilty et al., 2006).

However, we believe that negatively worded items are particularly useful for two reasons. First, they can be used by assessment specialists to detect extreme disengagement via streamlined responses and these invalid responses can be removed. Second, and related to the first reason, negatively worded items can be used to investigate if item-wording factors represent substantively meaningful constructs, ephemeral artifacts of methods effects that are substantively irrelevant, or stable response styles (Marsh et al., 2010).

In short, if disengagement is present on the Rosenberg Self-Esteem Scale and it is not investigated and dealt with, assessment practitioners could draw inaccurate conclusions.

To showcase these two reasons, consider the following example. Assessment practitioners may wish to use the Rosenberg Self-Esteem scale (Rosenberg, 1965) to measure the self-esteem of their students on campus. The 10-item scale has 5 positively worded items and 5 negatively worded items, all intended to measure a single construct. Although the scale is intended to have one factor, numerous studies have challenged the one factor structure in favor of models that account for variance due to item valence (e.g., Lindwall et al., 2012; Marsh et al., 2010; Quilty et al., 2006). Assessment practitioners then need to investigate if the factors that reflect item wording are substantively meaningful, substantively irrelevant artifacts, or response styles. For example, when reviewing studies of the factor structure of the Rosenberg Self-Esteem scale for their meta-analysis, Gnambs et al. (2018) explained that some assessment practitioners interpreted the item-wording factors as positive self-esteem and negative self-esteem. Hence, the item-wording factors may have substantively different meanings. However, this same structure could emerge due to disengagement; thus, the structure would be substantively irrelevant. Disengagement could be manifested in streamlined responses, where some students select the same response option, regardless of the item wording. In this situation, item interrelations would not be fully explained by the one-factor model of self-esteem because responses were also influenced by level of disengagement. An EFA would likely support a two-factor solution based on item valence. If a one-factor CFA model were fitted to the data, correlated residuals between negatively worded items would emerge, necessitating an item-wording method effect factor to reproduce the data adequately. In short, if disengagement is present on the Rosenberg Self-Esteem Scale and it is not investigated and dealt with, assessment practitioners could draw inaccurate conclusions (e.g., meaningful differences between positive and negative self-esteem).

Instead, assessment practitioners could use streamlined responses to identify extreme disengagement. After filtering students who streamline from the dataset, the factor structure could be re-estimated. The factors associated with item valence would dissipate if they mainly reflected this extreme disengagement. In turn, self-esteem would be depicted as a unidimensional construct and the assessment practitioner could report a total self-esteem score for each student. If the item-wording factors remained after removing streamlined responses, then substantively meaningful factors of positive and negative self-esteem or stable response styles may be plausible. This example highlights the usefulness of negatively worded items to detect disengagement, especially for situations where item-wording factors are assumed by some assessment practitioners to be substantively relevant and by others to be substantively irrelevant.

Usefulness of Self-Reported Effort to Signal Disengagement

Not all non-cognitive assessments administered on our campuses contain negatively worded items and streamlined responses are less reliable as a method to detect disengagement without negatively worded items (Curran, 2016). Fortunately, self-report measures of expended effort can be administered by assessment practitioners after an assessment or after a series of assessments. Students who self-report that their motivation is low can be identified and their responses removed from the dataset. For example, higher education assessment practitioners have employed motivation filtering using scores from the effort subscale of the Student Opinion Scale (Thelk et al., 2009). Specifically, a cut-off of 15 has been established, where students who have scores at or below 15 on the effort subscale are identified and removed from the dataset (Swerdzewski et al., 2011).

Self-report measures are not a perfect method to evaluate disengagement. Students may lack engagement when responding to the self-report measure itself (Wise & Kong, 2005). Additionally, students may inaccurately report higher motivation in order to "look better" (Rios et al., 2014), in fear of punishment (Wise, 2020), or students may inaccurately report lower motivation to protect their self-esteem. For example, after a difficult assessment, students may falsely attribute the cause of their poor performance to low levels of effort (Myers & Finney, 2021). Moreover, a recent study found that self-reported effort is not as good of an indicator for disengagement compared to more "behavioral" (e.g., response time, number of clicks) indices (Csányi & Molnár, 2023). Hence, self-reported effort may be most useful for assessment practitioners in conjunction with additional behavioral indices of disengagement.

Usefulness of Response Time to Signal Disengagement

One of the most common behavioral measures of disengagement in higher education and K-12 contexts is response time (e.g., Rios et al., 2014; Wise & Kuhfeld, 2020). Some students answer an item so quickly that they could not have read and processed the item (e.g., Wise & Kong, 2005). Response times can be used as a method to identify and filter responses from disengaged students. For example, assessment specialists have used the mean item response time to identify rapid responses. That is, responses provided in less than an established percent of time (e.g., 20% of the mean time; Wise & Ma, 2012) are considered rapid responses. Responses from students who rapidly responded to more than a certain percentage of items (e.g., rapidly responded on 10% or more of the items) are removed from the dataset.

Unfortunately, there are also issues with using response time to identify disengaged students, particularly in low-stakes higher education testing contexts. Some filtered responses may have been effortful but removed because they were provided quickly. Likewise, some retained responses may have been non-effortful but were not removed because they were associated with long response times (Wise, 2020). Finally, some data collections do not allow for timing data to be collected at the item-level or the measure-level (e.g., some commercially available measures used in higher education for accountability do not provide the institution with timing data).

Streamlining, self-reported effort, and response time are effective means for assessment practitioners to identify student disengagement. However, the three strategies measure different manifestations of disengagement; thus, they do not identify the exact same sample of disengaged students. For example, although we know that disengagement is related to test performance (Rios et al., 2022), self-reported effort has a smaller correlation with performance (r = .33) compared to response time (r =.72), suggesting that the two reflect different manifestations of disengagement (Silm et al., 2020). Moreover, the correlation between self-reported effort and response time effort (e.g., Wise & Kong, 2005) varies between small to moderately large (r = .28, Akhtar & Firdiyanti, 2023; r = .13, Csányi & Molnár, 2023; r = .61, Rios et al., 2014; r = .25, Wise & Kong, 2005). Swerdzewski et al., (2011) found that response time effort and self-reported effort agreed (i.e., flagged the same individuals as having low motivation) for 66.01% of the disengaged students. Additionally, one study found that streamlined responses correlate in a small but positive way with self-reported effort (.16) but had a very small (-.06) correlation with response time in minutes (Kupffer et al., 2024). Thus, although it is expected that some students who self-report having low effort will also streamline or rapidly guess, there will be students who do not exhibit multiple types of disengagement. Indeed, using a "hurdle approach," where multiple methods to detect disengagement are used simultaneously, is recommended to identify different types of disengagement (e.g., Curran, 2016; Goldammer et al., 2020; Meade & Craig, 2012).

Self-reported effort may be most useful for assessment practitioners in conjunction with additional behavioral indices of disengagement.

The Current Study

Using data collected for institutional accountability and improvement purposes, we investigated the effect of different disengagement types on the internal structure of a non-cognitive assessment. Three disengagement identification methods were used. Students self-reported if they had low effort while completing assessments (e.g., Swerdzewski et al., 2011). "Rapid responders" were students who responded so quickly they could not have read the item (e.g., Wise & DeMars, 2005; Wise & Kuhfeld, 2020). "Streamliners" were students who consistently selected the same response option (e.g., Curran, 2016; Hong et al., 2020; Steedle et al., 2019). The impact of disengagement on the factor structure of scores was estimated and compared across disengagement methods with the goal that at least one of these approaches will be useful and accessible for assessment practitioners. Specifically, this study addressed two hypotheses:

1) There would be at least a moderate proportion of disengaged students identified as being disengaged by at least one method (streamlining, self-report, and response time). We did not expect that all disengaged students would be identified by all three methods, given that the three indicators of disengagement reflect different manifestations of disengagement. That is, we expected that a moderate number of students who self-reported having low-effort would also rapidly respond, given the moderate relationship between the two indicators. We expected the fewest disengaged students to be flagged using the streamline method. Moreover, of those who did streamline, only a small proportion would also self-report having low motivation and would rapidly respond, given that streamlining inconsistently aligns with other indicators of disengagement (e.g., Goldammer et al., 2020; Hong et al., 2020).

2) After removing disengaged students, the internal structure of the scores from a non-cognitive measure should be less contaminated by disengagement; thus, a CFA model reflecting the intended internal structure of the scores would fit the data better. This improved fit would be evidenced via global fit indices and reduced correlation residuals between the negatively worded items. If all disengagement methods improve model-data fit, assessment practitioners can use motivation filtering with any of these methods to improve the validity of their score interpretations.

The impact of disengagement on the factor structure of scores was estimated and compared across disengagement methods with the goal that at least one of these approaches will be useful and accessible for assessment practitioners.

Method

Participants and Procedure

Our mid-size (approximately 20,000 students) southeastern US university uses low-stakes assessments to evaluate outcomes of our general education programming (e.g., quantitative reasoning) and university-wide initiatives (e.g., civic engagement). Assessments are administered to incoming students in the fall and advanced students in the spring. Although every student was required to complete a series of assessments taking approximately two hours, scores have no personal impact on students (e.g., no impact on grades, awards, opportunities).

Data from incoming first-year students were collected in the fall of 2021 under low-stakes conditions. All students completed a series of cognitive and non-cognitive assessments. Only students who completed the non-cognitive assessment of interest, who consented to having their data used for research purposes, and who were over the age of 18 were included in the analysis, which resulted in 3,169 students. Assessments were administered online and unproctored via Qualtrics. Students had a multiweek window during which they were required to complete the assessments.

Measure

The Attitudes Towards Communication (ATC) assessment is the primary assessment of interest in the current study. The 11-item non-cognitive assessment has two subscales: willingness to communicate (6 items) and confidence in communication (5 items). These subscales are intended to measure "key communication concepts for undergraduate college students" (Williams et al., 2014). Items on the ATC were developed to assess affective components of communication and written to align with the National Communication Association standards.

We believed that some students would put forth little effort when completing the ATC assessment for two reasons. First, students tend to put forth less effort as assessments get longer (e.g., Pastor et al., 2019) or are later in the testing session (e.g., Finney & McFadden, 2023). Although the 11-item ATC assessment could be considered relatively short and less cognitively demanding than say a math assessment, we were concerned about low-effort on this assessment because it was the second assessment in a series of assessments. Thus, some students may be fatigued later in the testing session, which would result in them putting forth little effort on the ATC assessment. Second, low-stakes testing contexts are associated with lower effort than high-stakes testing contexts due to the lower perceived importance of these tests to the students (e.g., Finney et al., 2018; Satkus & Finney, 2021). Even if the ATC assessment was placed first in the series of assessments, we would still advise practitioners to investigate effort on the ATC due to the non-consequential nature of the testing context, regardless of the assessment's length or content. Finally, no methods to proactively increase test-taking motivation (e.g., incentives, increasing test relevance, priming) were used in the current study.

ATC Assessment Theorized Internal Structure

The ATC assessment was designed with two intentionally distinct subscales: willingness and confidence. Willingness items measure students' openness to communication. Confidence items measure communication self-efficacy. A two-factor structure with no cross-loadings was expected. All ATC items were responded to using a 5-point Likert scale: 1 (*Strongly Disagree*), 2 (*Disagree*), 3 (*Undecided*), 4 (*Agree*), and 5 (*Strongly Agree).* Note, one of the negatively worded items was theorized to reflect willingness (item 3) whereas the other negatively worded item was theorized to reflect confidence (item 8). Higher scores represent higher willingness or confidence to participate in speech performance. Cronbach's alpha for the willingness scores ($\alpha = 0.78$) and confidence scores ($\alpha = 0.74$) was adequate.

Indicators of Disengagement

Self-Report Measure

After approximately two hours of completing assessments, students completed a self-report measure of effort. More specifically, students completed a cognitive test, followed by the ATC scale, and then responded to the Student Opinion Scale (SOS) (Pastor et al., 2023; Thelk et al., 2009). The SOS contains a five-item subscale intended to measure expended effort for the total testing session. Thus, effort on the SOS reflects not only effort on the non-cognitive ATC scale, but also effort on the cognitive test as well. Effort scores from the SOS range from 5 (no effort) to 25 (highest effort). Filtering scores from disengaged students was accomplished using a cutoff score of 15 on the effort subscale, as was done in previous studies using this self-report measure (e.g., Swerdzewski et al., 2011). ATC scores from students who scored at or below 15 on the effort subscale were removed from the filtered datasets.

Response Time

Response times were not available for the non-cognitive ATC measure but were available for the cognitive test taken just prior to the ATC. This use of response time on a previous task was justified because rapid response behavior tends to increase throughout a series of tests (e.g., Pastor et al., 2019) and measures of motivation on one task have been used to make inferences about motivation on an accompanying task (Zamarro et al., 2019). The current study used the normative threshold setting method (NT20) to identify students who rapidly guessed on the previous cognitive test (Wise & Ma, 2012). First, the mean response time for each cognitive item was calculated. Then, if the response time for the item was lower than 20% of the mean response time for that item, the item response was flagged as a rapid response. Finally, if a student rapidly responded on more than 10% of the cognitive items, they were identified as a rapid responder (e.g., Rios et al., 2017; Wise & DeMars, 2010). ATC scores from students who rapidly responded on the cognitive test prior to the ATC were removed from the filtered datasets.

The ATC assessment was designed with two intentionally distinct subscales: willingness and confidence.

Streamlining Responses

Students were categorized as streamliners if they selected the same response option (e.g., all "strongly agree," all "disagree") to all items on the non-cognitive measure prior to reverse scoring the two negatively worded items. Students were not categorized as streamliners if they selected "undecided" for all response options, given that selecting "undecided" is a valid option for both positively and negatively worded items. Of those students who responded to the ATC assessment, only 2.2% (70 out of 3,169) selected "undecided" for all items on the confidence subscale, 3.1% (98 out of 3,169) selected "undecided" for all items on the willingness subscale, and 2.1% (66 out of 3,169) selected "undecided" for all items on the ATC (across both subscales).

These students were not categorized as streamliners. ATC scores from students who exhibited streamlining on this non-cognitive measure were removed from the filtered datasets.

Results

Number of Students Identified as Disengaged Across the Three Methods

Frequencies and proportion of students identified as disengaged by each of three methods are displayed in Table 1. Unfortunately, approximately 24% of all students (745 students out of the total sample of 3,169) were disengaged in some way (streamlined, rapidly responded, or self-reported low effort). Yet, we felt fortunate that we were able to actually identify this disengagement instead of assuming it did not exist. Of those who displayed disengagement, most rapidly responded on the previous cognitive test or self-reported low effort. Fewer students streamlined, although streamliners still accounted for a meaningful, although relatively small, number of disengaged students (141 students).

Table 1
Proportion of Total Students who were Disengaged

Disengagement Type	N	% (out of 3,169)
Streamliners	141	4.4%
Rapid Responders	488	15.4%
Low Self-Reported Effort	425	13.4%
Total Disengaged	745	23.5%

Note. Total disengaged does not equal the sum of all disengagement types in the table. Some students used more than one disengagement type.

To address our first hypothesis, we computed the proportion of students who streamlined, rapidly responded, or self-reported having low effort (Table 2). Nearly half (≈ 41%) of those who streamlined (n = 141) also rapidly responded (n = 58). Nearly half (≈ 42%) of those who self-reported having low effort (n = 425) also rapidly responded (n = 179). Very few (≈ 6%) who streamlined (n = 141) also self-reported having low effort (n = 8). Finally, very few (≈ 4%) of those who exhibited at least one of the disengagement types (n = 745) used all three disengagement types (n = 32). Figure 1 displays a proportional Venn diagram of students who exhibited different disengagement types.

Confirmatory Factor Analysis: Model Fit

We addressed our second hypothesis using confirmatory factor analysis (CFA). CFA was used to assess model-data fit for five different data sets: 1) unfiltered dataset containing all students, 2) filtered dataset without students who streamlined, 3) filtered dataset without students who rapidly responded, 4) filtered dataset without students who self-reported having low effort, and 5) filtered dataset without students who displayed any disengagement type. Due to there being only small, nuanced differences in the CFA models between the three samples in which only one type of disengagement was removed (streamlining only, rapidly responding only, and self-reporting only), we did not conduct analyses on samples in which two types of disengagement were removed (streamlining and rapidly responding, streamlining and self-reporting, and rapidly responding and self-reporting).

Approximately 24% of all students were disengaged in some way.

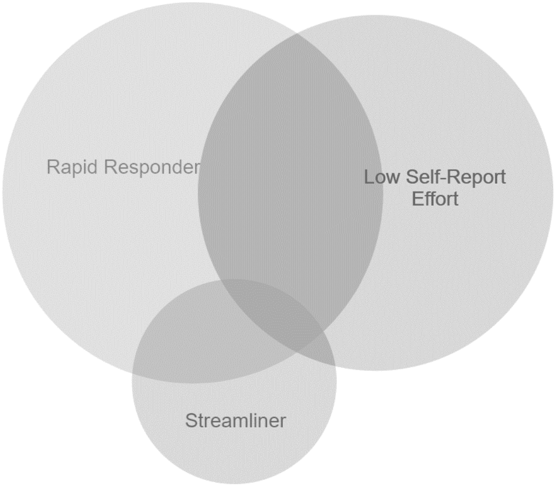
CFA analyses were conducted using Mplus version 8.6. Items were approximately normally distributed other than Item 2 having kurtosis of 5.56. We compared the results of models estimated using maximum likelihood (ML) estimation and maximum likelihood with the Satorra-Bentler adjustment (MLMV) (Finney et al., 2016). Due to the negligible difference in results and inferences when comparing ML and MLMV, we determined that the data were sufficiently normally distributed, and ML results are reported.

Statistical analyses were not used to compare the results across the different filtered samples. Instead, results across the different samples were compared via both global and local fit indices. The CFI ranges from 0 to 1 with higher values indicating better model-data fit. The

Table 2
Proportion of Disengaged Students by Disengagement Method

Disengagement Method	N	% (out of 745)
Streamliner ONLY	43	5.8%
Rapid Responder ONLY	219	29.4%
Low Self-Reported Effort ONLY	206	27.7%
Streamliner & Rapid Responder	58	7.8%
Streamliner & Low Self-Reported Effort	8	1.1%
Rapid Responder & Low Self-Reported Effort	179	24.0%
Streamliner, Rapid Responder, & Low Self-Reported Effort	32	4.3%
Total	745	100.0%

Figure 1
Overlap Between Disengagement Types



Note. Proportions are shown approximately to scale. Nearly half (≈ 41%) of those who streamlined also rapidly responded. Nearly half (≈ 42%) of those who self-reported low effort also rapidly responded. Few (≈ 6%) who streamlined also self-reported having low effort.

RMSEA and SRMR range from 0 to 1 where lower values indicate better model-data fit. As shown in Table 3, Models 2 to 5 (filtered data sets) fit the data better in a global sense than Model 1 (unfiltered data set). Sample 4 (only low self-reported effort removed) resulted in fit index values that were close in magnitude to the fit indices of Sample 1 (unfiltered). Sample 4 also resulted in the lowest CFI relative to all other samples. The largest differences in fit index values were between Sample 1 (unfiltered) and Sample 5 (all disengagements removed).

Correlation residuals, which represent the discrepancy between the corresponding observed and model-implied item-level correlations, were used to assess local model-data misfit (Bandalos & Finney, 2019). If a model fits perfectly, the correlation residuals are zero. Comparisons were made between the size of the residual correlation between the two negatively worded items, as well as the number of residual correlations over |0.10| and |0.15|. As expected, there were fewer correlation residuals over |0.10| and |0.15| in the filtered data sets compared to the unfiltered data set (see Table 4).

Table 3
Comparison of Global Fit Indices for Two-Factor Model Fit to Unfiltered and Filtered Samples

Sample Type	*n*	χ2	*df*	CFI	RMSEA	SRMR
Sample 1: Unfiltered	3169	1190.16*	43	0.886	0.092	0.058
Sample 2: Streamliners Removed	3028	949.42*	43	0.904	0.083	0.052
Sample 3: Rapid Responders Removed	2681	745.26*	43	0.895	0.078	0.049
Sample 4: Low Self-Reported Effort Removed	2744	926.52*	43	0.878	0.087	0.054
Sample 5: Any Disengagements Removed	2424	653.39*	43	0.898	0.077	0.047

The correlation residual between the two negatively worded items decreased from the unfiltered data set to the filtered data sets. That is, for the unfiltered data, even after accounting for the constructs of interest (willingness and confidence), there was still a non-negligible (residual) correlation between the negatively worded items, caused by students' disengagement. However, this construct-irrelevant relation diminished when disengagement was filtered from the dataset, as hypothesized. There are two notable changes in the correlation residuals. First, the correlation residual from the unfiltered data set dropped from 0.21 to 0.03 in the data set when all disengagement types were filtered. This difference is substantial. Second, the correlation residual dropped from 0.21 to 0.10 when only streamliners (which were only 4.4% of the sample) were removed. These results provide evidence of the impact of disengagement on the factor structure of scores and, in turn, structural validity evidence.

Table 4
Comparison of Local Fit (Correlation Residuals) for Unfiltered and Filtered Samples

Sample	Correlation Residual Between the Two Negatively Worded Items	Frequency of Correlation Residuals	
		> |.10|	> |.15|
Sample 1: Unfiltered	0.208	9	3
Sample 2: No Streamliners	0.101	6	1
Sample 3: No Rapid Responders	-0.054	6	3
Sample 4: No Low Self-Reported Effort	0.185	7	2
Sample 5: No Disengagements	0.034	4	1

Discussion

The current study demonstrated three different practical strategies that identified students who were not engaged when completing low-stakes assessments. Moreover, we used these three methods to remove invalid responses from the dataset, which in turn positively influenced the structural validity of the responses. Results of our research questions and practical implications for assessment practitioners are discussed.

The correlation residual from the unfiltered data set dropped from 0.21 to 0.03 when all disengagement types were filtered.

Disengaged Students Identified via Different Methods

We hypothesized there would be some students who were identified by at least one disengagement method and that some students would be identified by more than one disengagement method. Our results supported our first hypothesis. We found that about half of those who self-reported having low effort also rapidly responded. This finding aligns with results of studies that found moderate correlations between self-reported effort and response time (Akhtar & Firdiyanti, 2023; Csányi & Molnár, 2023; Rios et al., 2014; Wise & Kong, 2005) and aligns with results of a study that directly estimated the percent of students (66.01%) who self-reported low effort and who rapidly guessed (Swerdzewski et al., 2011). Interestingly, about half of those who streamlined also rapidly responded. This finding aligns with research that suggests both of these behavioral indicators of disengagement (rapid guessing, Akhtar & Firdiyanti, 2023; streamlining, Hong et al., 2020) are good indicators of disengagement.

Finally, few students who streamlined also self-reported having low effort. We believe that the self-reported effort measure could have been affected by streamlining. If students streamlined through the self-report effort measure, they might not be captured as having self-reported low effort. Moreover, self-reported effort is for the whole testing session, not the ATC specifically. We may expect more alignment between students who self-report low effort for the ATC and who streamline responses to the ATC items.

The Effect of Disengagement on Factor Structure

We hypothesized that removing responses from disengaged students would result in improved model-data fit (i.e., theorized internal structure would better match the empirical internal structure and thus the recommended scoring of the measure could be employed). Our results supported our hypothesis. When responses from disengaged students were removed from the dataset, model-data fit improved, regardless of the motivation filtering method chosen. In fact, the correlation residual between the negatively worded items was close to zero after removing students exhibiting any type of disengagement. In the current study, model-data fit indices did not meet recommended cut-off values. However, having good model-data fit was not the purpose of the current study. The purpose of the current study was to take steps to improve the validity of score interpretations by removing the influence of disengagement from the internal structure of scores via methods that are practical and accessible to assessment practitioners. Filtering by any disengagement method improved model-data fit, with rapid guessing and streamlining producing better results. If possible, assessment practitioners should use multiple techniques to identify disengaged students. However, practitioners can rest assured that using at least one method will still result in improved interpretations via improved internal structure.

Filtering by any disengagement method improved model-data fit, with rapid guessing and streamlining producing better results.

Some researchers note that streamlined responses are not always a good indicator of disengagement (Goldammer et al., 2020). Moreover, the detection of streamlining requires the inclusion of at least one negatively worded item, even though inclusion of negatively worded items has been admonished by some (e.g., Dalal & Carter, 2015) and may not be possible to include in all assessment contexts. However, in our study, the effect of item valence on the factor structure of scores was substantially reduced after removing a small number of students who streamlined (only 4.4% of the sample). In our sample, streamlining represented an extreme, flagrant form of disengagement used by a very small number of students. As a consequence, we believe that streamlining (which requires negatively worded items) can be used as an effective means to identify disengagement and improve the factor structure of scores without removing a large number of students and thus retain better generalizability of the scores.

Limitations

There are some limitations to the current study. First, response time was gathered using an adjacent assessment. Ideally, response time should be gathered on the assessment of interest, rather than on an adjacent assessment. With that said, we realize other assessment practitioners may encounter this same issue given response time is not always available for all measures (some commercial measures used in higher education do not report response time). Thus, our work can be used as a reference for those who find themselves unable to collect response time on their measure of interest.

Second, the current study did not vary the content nor the length of the non-cognitive measure. Thus, the generalizability of our results may only extend to measures of similar length and content. Future studies may vary the length and/or content of the non-cognitive measure to investigate the impact of low motivation on the factor structure of scores.

Third, assessment specialists caution the use of motivation filtering when disengagement is related to ability on cognitive tests (Rios et al., 2017). When disengagement is related to ability, filtering may result in a less generalizable sample because low-ability students are removed from the sample at a disproportionate rate. In short, regardless of whether a test is cognitive or non-cognitive, when motivation filtering is used, the sample characteristics may change when compared to unfiltered data (e.g., change in proportion of high-performing students, change in demographics). Thus, filtering scores from low-motivated students for any type of measure may alter the population of students to whom the test scores may generalize. Thus, higher education assessment practitioners should also focus on increasing engagement a priori. Proactive strategies such as offering external incentives (e.g., Rios, 2021), increasing test relevance (e.g., Liu et al., 2015), and priming (e.g., Finney & McFadden, 2023; Finney & Pastor, 2025) have increased test-taking motivation and reduced the percentage of responses needing to be filtered. If possible, we encourage coupling these proactive strategies to mitigate disengagement with the strategies we showcased in this study. We hope the current study provides assessment practitioners in higher education low-stakes testing with accessible tools to address disengagement, with an understanding of the limitations of the generalizability of our results.

Coupling proactive strategies to mitigate disengagement with the strategies we showcased can increase engagement.

Implications for Higher Education Assessment Practitioners

In closing, filtering invalid responses from disengaged students by any method improved the factor structure of scores and thus the validity of score interpretations. Given the popularity of non-cognitive assessments in higher education, we recommend that assessment practitioners 1) select the disengagement identification technique(s) that is most accessible to them and 2) as a matter of routine, investigate the amount of disengagement present during their collection of outcomes assessment data. Understanding and tackling the issue of student disengagement on outcomes assessments allows for more accurate interpretations of student learning and development data that can be used for accountability reporting and programmatic improvement.

References

Akhtar, H., & Firdiyanti, R. (2023). Test-taking motivation and performance: Do self-report and time-based measures of effort reflect the same aspects of test-taking motivation? *Learning and Individual Differences*, *106*. https://doi.org/10.1016/j.lindif.2023.102323

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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

Bandalos, D. L., & Finney, S. J. (2019). Factor analysis: Exploratory and confirmatory. In G. R. Hancock, L. M. Stapleton, & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (2nd ed., pp. 98–122). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315755649-8

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60*(3), 361-370. https://doi.org/10.1177/00131640021970592

Barry, C. L., & Finney, S. J. (2009). Does it matter how data are collected? A comparison of testing conditions and the implications for validity. *Research & Practice in Assessment*, *3*, 1–15. https://eric.ed.gov/?id=EJ1062735

Csányi, R., & Molnár, G. (2023). How do test-takers rate their effort? A comparative analysis of self-report and log file data. *Learning and Individual Differences*, *106*. https://doi.org/10.1016/j.lindif.2023.102340

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Dalal, D. K., & Carter, N. T. (2015). Negatively worded items negatively impact survey research. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 112–132). Routledge/Taylor & Francis Group.

DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling*, *13*(3), 440–464. https://doi.org/10.1207/s15328007sem1303_6

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment*, *91*(2), 166–174. https://doi.org/10.1080/00223890802634290

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, *2015*(2), 1–17. https://doi.org/10.1002/ets2.12067

Finney, S. J., DiStefano, C., & Kopp, J. P. (2016). Overview of estimation methods and preconditions for their application with structural equation modeling. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (1st ed., pp. 135 - 165). Hogrefe. https://psycnet.apa.org/record/2016-44627-008

Finney, S. J., & McFadden, M. E. (2023). Examining the question-behavior effect in low-stakes testing contexts: A cheap strategy to increase examinee effort. *Educational Assessment*, *28*(4), 211-228. https://doi.org/10.1080/10627197.2023.2222588

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

Finney, S. J., Schaefer, K. E., & McFadden, M. E. (2024). Priming examinees to give good effort: Differential utility across gender identity. *The Journal of Experimental Education.* https://doi.org/10.1080/00220973.2024.2310678

Finney, S. J. & Pastor, D. A. (2025). Priming non-compliant students to expend test-taking effort: How many primes are needed? *Journal of Experimental Education*. https://doi.org/10.1080/00220973.2025.2459392

Finney, S. J., Sundre, D. L., Swain, M. S., & Williams, L. M. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, *21*(1), 60–87. https://doi.org/10.1080/10627197.2015.1127753

Gnambs, T., Scharl, A., & Schroeders, U. (2018). The structure of the Rosenberg self-esteem scale: A cross-cultural meta-analysis. *Zeitschrift für Psychologie*, *226*(1), 14-29. https://doi.org/10.1027/2151-2604/a000317

Goldammer, P., Annen, H., Stockli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), https://doi.org/10.1016/j.leaqua.2020.101384

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement*, *80*(2). https://doi.org/10.1177/0013164419865316

Kupffer, R., Frick, S., & Wetzel, E. (2024). Detecting careless responding in multidimensional forced-choice questionnaires. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644231222420

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thøgersen-Ntoumani, C. (2012). Method effects: The problem with negatively versus positively keyed items. *Journal of Personality Assessment*, *94*(2), 196-204. https://doi.org/10.1080/00223891.2011.645936

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, *20*(2), 79–94. https://doi.org/10.1080/10627197.2015.1028618

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*(1), 37–49. https://doi.org/10.1037/0012-1649.22.1.37

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 810–819. https://doi.org/10.1037/0022-3514.70.4.810

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg self-esteem scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, *22*(2), 366–381. https://doi.org/10.1037/a0019225

Mathers, C., Finney, S. J., & Hathcoat, J. D. (2018). Student learning in higher education: A longitudinal analysis and faculty discussion. *Assessment & Evaluation in Higher Education*, *43*(8), 1211-1227. https://doi.org/10.1080/02602938.2018.1443202

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

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

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

Pastor, D., Patterson, C., & Finney, S. J. (2023). Development and internal validity of the Student Opinion Scale: A measure of test-taking motivation. *Journal of Psychoeducational Assessment*, *41*(2), 209-225. https://doi.org/10.1177/07342829221140957

Ponce, F. P., Irribarra, D. T., Vergés, A., & Arias, V. B. (2022). Wording effects in assessment: Missing the trees for the forest. *Multivariate Behavioral Research*, *57*(5), 718-734. https://doi.org/10.1080/00273171.2021.1925075

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg self-esteem scale method effects. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(1), 99-117. https://doi.org/10.1207/s15328007sem1301_5

Rios, J. A. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*(2), 85–106. https://doi.org/10.1080/08957347.2021.1890741

Rios, J. A., Deng, J., & Ihlenfeldt, S. D. (2022). To what degree does rapid guessing distort aggregated test scores? A meta-analytic investigation. *Educational Assessment*, *27*(4), 356-373. https://doi.org/10.1080/10627197.2022.2110465

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74-104.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, *161*, 69-82. https://doi.org/10.1002/ir.20068

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press. https://www.jstor.org/stable/j.ctt183pjjh

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

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review, *Educational Research Review*, *31*. https://doi.org/10.1016/j.edurev.2020.100335

Steedle, J., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social-emotional learning competencies. *Educational Measurement Issues and Practice*, *38*, 101-111.

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, *45*(1), 116–131. https://doi.org/10.1509/jmkr.45.1.116

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. *The Journal of General Education*, *58*(3), 129–151. https://www.jstor.org/stable/27798135

Williams, L. M., & Horst, S. J., & Sundre. D. L. (2014). Test of oral communication skills, version 2: TOCS-II test manual. Harrisonburg, VA: Center for Assessment and Research Studies and Madison Assessment. https://www.madisonassessment.com/assessment-testing/test-of-oral-communication-skills

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

Wise, S. L. (2020). Six insights regarding test-taking disengagement. *Educational Research and Evaluation*, *26*(5-6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

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

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

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program* [Paper Presentation]. National Council on Measurement in Education Annual Meeting, San Diego, CA, United States.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163 –183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Kuhfeld, M. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. Margolis & R. Feinberg (Eds.), *Integrating Timing Considerations to Improve Testing Practices* (pp. 150 – 64). Routledge. https://doi.org/10.4324/9781351064781-11

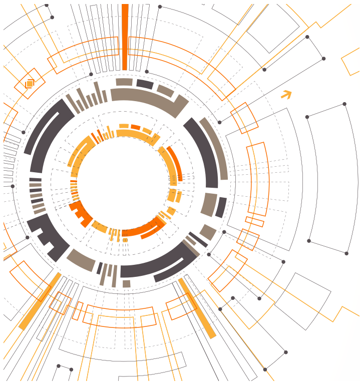
Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a cat item pool: The normative threshold method* [Paper Presentation]. National Council on Measurement in Education Annual Meeting, Vancouver, Canada. https://www.nwea.org/resources/setting-response-time-thresholds-cat-item-pool-normative-threshold-method/

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*, 189-194. https://doi.org/10.1007/s10862-005-9004-7

Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital*, *13*(4), 519-552. https://doi.org/10.1086/705799

Abstract

Through relying on limited and prescribed modes of expression, summative assessment can both create and exacerbate inequities in higher education. In this article, an instructor of an undergraduate education course and three student co-authors who completed the course discuss how the students' choice to use multimodality in their final portfolios functioned as an innovation for equity in the course's summative assessment. After introducing ourselves and the higher education context in which we have worked together, we describe the portfolio assignment from this course. Then, the three student authors present excerpts from their portfolios, each framed by some contextual information offered by the faculty author and followed by a reflection informed by the perspectives of all four co-authors. These reflections focus on how multimodality can constitute an equitable approach to summative assessment in response to specific student intentions, health needs, and preferred modes of expression.



AUTHORS

Alison Cook-Sather, Ph.D.
Bryn Mawr College

Daniela Moreira
Wake Forest University

Piper Rolfes
Bryn Mawr College

Jess Smith
Bryn Mawr College

Multimodality as an Equitable Approach to Summative Assessment in Higher Education

…The activity was to craft a shape [out of a pipe cleaner] that captures how you see yourself as an educator or educators more generally….I ultimately came up with…a spiral leading to a thought bubble….[T]he spiral represents a continuous flow of knowledge in both directions. We had talked about how education is ongoing, so the thought bubble represents that endless continuation of thinking and learning." – Jess Smith

"My third [portfolio] artifact is artwork that I engaged with while at a particularly difficult point with my concussion and overall health.…Needle felting, as a practice, is inherently slow; it asks the creator to take time in order to make the vision come to light. In the time it took for me to make this and step back into the sensations of the space, I was able to reflect on the connections…developed for me through the [course] texts and relationship building we have engaged with this semester." - Piper Rolfes

"This portfolio contains only the snapshots of some pertinent reflections on my journey seeking joy. It exists as a reaction to my discomfort with the course structure…and to my explicit desire to make and be authentic. It's more of my physical representation of what otherwise is abstract space…I hope you can give this all a complete listen and take time to see all the pieces in this portfolio in the order as presented and when you are finished, assemble it together just as you found it." - Daniela Moreira

CORRESPONDENCE
Email
acooksat@brynmawr.edu

These are excerpts from reflections that the three student co-authors of this article Jess Smith, Piper Rolfes, and Daniela Moreira, included in their final portfolios for an undergraduate education course called Community Learning Collaborative: Practicing

Partnership. The excerpts—and their fuller versions included in a subsequent section of this discussion—offer insights into the learning and growth that can result from engagement in multimodality in summative assessment as well as the movement toward equity and inclusion that an embrace of multimodality enacts.

The legitimation of diversity in forms of expression should be inextricable from affirmation of diversity in who is doing the expressing.

Our discussion of this work includes several parts. In the next section, we draw on scholarship to review trends in practice. Next, we introduce ourselves and the higher education context in which we have worked together and describe the portfolio assignment for Community Learning Collaborative. With this context established, we explain how we selected the excerpts for inclusion, and then Smith, Rolfes, and Moreira each present those excerpts. Framed briefly by some contextual information offered by faculty author, Alison Cook-Sather, these examples are original portions of the student authors' course portfolios, which the student co-authors courageously agreed to share with and collectively analyze for a wider audience. Leaning into the vulnerability of such sharing and analysis, and the continued growth and empowerment to which those contributed, the student authors add their voices to the expanding conversation about multimodality in summative assessment as an innovation for equity in higher education.

Trends in Practice

There is increasing recognition that the proliferation of modes through which information is shared—"gestures, visuals, haptics, auditory productions, text-based information, and multimedia"—both constitutes and warrants an embrace of multimodality (Bouchey et al., 2021, p. 35). Multimodality refers to the use of multiple representations (graphical, textual, auditory, textual, and gestures) in the communication of knowledge (Kress & Leeuwen, 2001). According to Kress (2010), a mode is a "socially shaped and culturally given semiotic resource for making meaning" (p. 79). As a "phenomenon of communication," multimodality focuses on combinations "of different semiotic resources, or modes, in texts and communicative events" (Adami, 2016, p. 454). Jewitt et al. (2016) emphasize that multimodality refers to how "d*ifferent means of making meaning* are not separated but almost *always appear together*: image with writing, speech with gesture, math symbolism with writing" (p. 2, emphasis in the original). The call for multimodality, then, is a call for diversification in forms of expression and for recognition of how multiple modes of expression always inform and are informed by one another.

The legitimation of diversity in forms of expression should be inextricable from affirmation of diversity in who is doing the expressing. In higher education in general and in summative assessment in particular, the person doing the expressing is the learner. Thinking about learning through focusing on "meaning making as a process of design" and "choice of representation… gives a renewed focus on the role of the learner" (Jewitt, 2008, p. 263, p. 258). By centering "design, diversity, and multiplicity" in students' meaning-making practices and interpretative work (Jewitt, 2008, p. 258), we can not only better meet the demands of the increasingly complex world, but also meet the needs and aspirations of an increasingly diverse group of learners in higher education (Bouchey et al., 2021). Among efforts to move toward greater equity and inclusion in higher education, as well as capacity to interpret and communicate in the rapidly changing world, multimodality

> approaches communication as a process in which students (as they are socially situated and constrained) make meanings by selecting from, adapting, and remaking the range of representational and communicational resources (including physical, cognitive, and social resources) available to them in the classroom. (Jewitt, 2008, p. 263).

Fiorella and Mayer (2015) describe generative learning as a process-driven motivation. Motivation requires goal-driven behavior; without it, students would be unable to begin composing and generating. Similarly, Bouchey et al. (2021) argue that multimodal learning "requires a high level of agency (self-discipline) by learners, who must have the metacognition necessary to understand how they learn and also when to challenge themselves to learn in ways that lie outside their preferred modes" (p. 36). These arguments for motivation, agency, and choice should, we suggest, be informed by an understanding of necessity, such as when students live with long-term or temporary disabilities that affect their capacity to engage with the still-dominant medium of most higher-education contexts: printed (or digital) text. The portfolio assignment we discuss in this article acknowledges and affirms students' situatedness and the range of resources through which they can demonstrate their understanding as they make choices guided by intention, necessity, and preference. As the portfolio selections we discuss here illustrate, multimodality recognizes learners' choices in how they represent understanding as forms of self-empowerment and self-authoring (Baxter Magolda, 2007), models for and educates others about diversity in expression, and contributes to a movement toward equity and inclusion in higher education.

Nowhere is this movement more important than in the realm of assessment. Inequities in assessment are an enduring concern, with most approaches "assessing students in the same way without paying attention to their differences" (Montenegro & Jankowski, 2017, p. 16). For instance, Ross at al. (2020) argue that assessment of student learning in higher education is typically through "written compositions and oral presentations, often in high-stakes exam environments" (p. 292). In contrast, equitable assessment practices for social justice and epistemic justice are linked to a radical rethinking of what is meant by the now-common term 'authentic assessment' (McArthur, 2024) and afford all learners "an equal and unbiased opportunity" to demonstrate their knowledge and achievements in different ways (Montenegro & Jankowski, 2020, p. 10). To this end, educators must "design learning opportunities that allow students to cultivate core creative dispositions, exercise agency, engage in creative processes and produce innovative artefacts, including through multimodal assessments" (Ross et al., 2020, p. 301). In doing so, they provide a "nurturing environment to kindle the creative spark, an environment where students feel rewarded, are active learners, have a sense of ownership, and can freely discuss their problems" (Ferrari et al, 2009 as cited in Ross et al, 2020, p. 22). As important, Moreira notes, is providing scaffolding—moments of storyboarding, opportunities to iterate, and occasions to get feedback from an authentic audience (say, peers in a class). (See Reyna et al., 2017, and Reyna et al., 2021, for taxonomies of <digital> multimodality and ways to promote more sophisticated multimodes through iterative assessment and feedback.)

Who We Are as Co-authors and Our Context

The first author of this article, Alison Cook-Sather, is a white, middle-aged, able-bodied, cis-gendered woman and a full professor in the Education Department in the bi-college consortium of Bryn Mawr and Haverford Colleges. Since 1994, she has taught numerous undergraduate courses, including Community Learning Collaborative (CLC), the course from which we draw portfolio excerpts for this discussion.

Second author, Daniela Moreira, is a Latina, first-generation American and college student, able-bodied, cis-gendered woman who graduated from Haverford College in 2023 with a double major in chemistry and physics. She subsequently completed an MA in Education at Wake Forest University as well as secondary science teaching certification, and her action research was in multimodal science communication. Moreira enrolled in CLC in the fall semester of 2020, when, on Bryn Mawr's and Haverford's campuses, COVID-19 and student-led strikes for racial justice inspired by the Black Lives Matter movement prompted a shift to online teaching and learning and a recasting of course assignments to be responsive to these larger contextual realities.

Third author, Piper Rolfes, is a white, queer, and able-bodied person hailing from Twin Cities, MN. They completed a double major—an independent major in dance and disability studies and a major in Education Studies—at Bryn Mawr College in 2024. Rolfes enrolled in CLC in the fall semester of 2023. Fourth author, Jess Smith, enrolled in the same section of CLC as Rolfes, in the fall semester of 2023. She is a Black, able-bodied, queer, cis-gendered woman from the Central Valley in California. She is completing her final year as an undergraduate at Bryn Mawr College as a sociology major.

Multimodality recognizes learners' choices in how they represent understanding as forms of self-empowerment and self-authoring.

CLC was developed by a group of educators in different positions in school and community contexts with the goal of supporting students in engaging in and working towards educational justice by deepening knowledge, skills, and inquiries into the practice of relationships, facilitation, and change in in-school and out-of-school educational contexts. One of four entry-point courses for the major and minor in Education Studies at Bryn Mawr College, which is open to both Bryn Mawr and Haverford College students, this course does not strive to equip students quickly with answers that resolve big challenges and questions. Rather, it aims to strengthen students' capacity to inquire, to hold tension, to build relationships and community in order to work creatively in uncertainty, and to facilitate their own and others' learning and changing with difference as a resource. The course typically enrolls 22 students; enrollment is limited because all students work with practicing educators in field-based educational settings.

The course is organized into three phases: (1) reflection on self and past educational experiences, (2) engagement with teaching and learning theory and practice; and (3) articulation of practices learned and commitments to move forward. The portfolio assignment comes at the end of the third phase and is submitted in lieu of a final paper or exam.

The Portfolio as Summative Assessment

For the portfolio, students are asked to select a thread (key theme, word, or metaphor) that weaves throughout their learning, reading, and experiences in the course that helps answer the question: What is my work/practice as an educator? The portfolio consists of a table of contents, an introduction that directly addresses the question above, 4-6 artifact-reflection pairs that document students' development of their understanding of their work as an educator, and several other components. The portfolio is not graded as a separate assignment. Rather, the final component of the portfolio is a self-assessment that describes students' labor and growth across the entirety of the course as well as the grade they believe they have earned, and why. Criteria for summative assessment listed on the course syllabus include the following, and students are also invited to specify their own criteria to add to this list:

- you are present for embodied contribution
- your work is done on time (or with explicit extension)
- your work shows power to connect (past and current) experience with a range of frameworks to generate insights and questions
- your work shows clarity and depth in unlearning and recommitting to standards—imposed and self-authorized—for education
- your work shows a deepening degree of specificity and creativity (going beyond description and narrative to analyze and integrate those with new thinking through engaging with a range of discourses, readings, and encounters)
- your work demonstrates imaginative engagement with course goals, themes, and processes
- your work contributes to others' learning (in class and in partnership with field-based educators)
- you take initiative to communicate your needs, questions, and goals in order to make the course meaningful and effective for you and in order to share responsibility in this process

The final component of the portfolio is a self-assessment that describes students' labor and growth across the entirety of the course as well as the grade they believe they have earned, and why.

Students have opportunities to build toward their culminating portfolio and final self-assessment—through drafting portfolio components as journal entries, meeting with one another in small groups and with the co-educators with whom we collaborate for the course, and meeting with the course instructors, if they wish. The journal entries are assigned in the final weeks of the semester; meetings in small groups and with co-educators take place during regularly scheduled class sessions; and meetings with course instructors are optional, based on student-initiated requests. It is in the context of these scaffolded opportunities to make choices about modality that many, but not all, students decide to engage in multimodality for this summative assessment.

As also explained on the syllabus, the artifacts that students include in their portfolios can be in any medium or mediums, and the written reflections that accompany each one are approximately 400 words focused on highlighting the shifts in student understanding—where students unlearned, rethought, chose different emphases or language. The portfolio assignment itself, then, invites combinations "of different semiotic resources" (Adami, 2016, p. 454) in which different means of making meaning appear together (Jewitt et al., 2016). Rather than being assessed "in the same way" without "attention to their differences" (Montenegro & Jankowski, 2017, p. 16), students exercise "choice of representation" (Jewitt, 2008, p. 258) "through multimodal assessments" (Ross et al., 2020, p. 301) that aim to offer them "an equal and unbiased opportunity" to demonstrate their knowledge and achievements in different ways (Montenegro & Jankowski, 2020, p. 10). The process-driven motivation students experience (Fiorella & Mayer, 2015) and the high level of agency and metacognition students exercise (Bouchey et al., 2021) affirm their diversity and deepen their learning.

Our Approach

Cook-Sather invited Moreira, Rolfes, and Smith to co-author this discussion because of the explicit way that all of them analyzed (rather than only used) multimodality in their portfolios and because they illustrated three different reasons for choosing a multimodal approach. These reasons include specific intention (to present a nuanced analysis not possible with words alone), necessity (because of a concussion), and preference (as the most effective mode of expression based on previous experience and self-knowledge).

The excerpts below reference different moments in the semester during which Smith, Rolfes, and Moreira participated in CLC. They offer insight into modes of expression these students selected and the meaning these moments have for the students and their development as educators—in their self-authoring journeys (Baxter Magolda, 2007). We offer these examples to afford these three students an opportunity to model what embrace and analysis of multimodality can look like.

Students exercise "choice of representation" through multimodal assessments that aim to offer them "an equal and unbiased opportunity" to demonstrate their knowledge and achievements in different ways.

Limitations of this selection include that it is a very small sample and illustrative of only these three students' experiences, not a wide range of students' experiences. Our goal is not to provide fully representative examples or an exhaustive set of possible equity outcomes of embracing multimodality but rather, through offering in-depth examples, to offer insights and raise questions that can be explored across contexts.

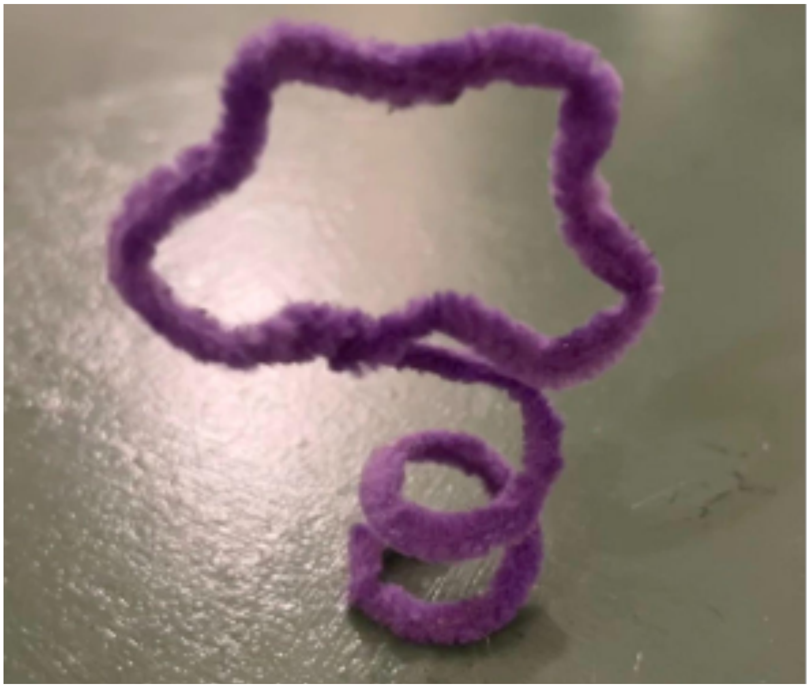
Example 1: Smith on Making Sense of and through an Unusual In-class Activity

Cook-Sather: It's 8:30 on a Monday evening in early September, and 24 enrolled students and I are halfway through the first session of CLC. We have introduced ourselves to one another and shared one true thing about ourselves or something we want others in the course to know about us. We have also discussed a text that sets the tone for the course (Mia Mingus' [2019] "Dreaming Accountability") and provides a structure for student engagement (Cook-Sather, 2023). At this point, I walk around the room holding a bag of brightly colored pipe cleaners and ask each student to select one. I then ask the students to take that single pipe cleaner and make from it a shape that captures how they see themself as an educator or educators more generally. That is all the guidance they get. Some eyes get large and glance around, and some people ask clarifying questions, but quickly the room settles into a mix of quiet chatter and focused concentration as students lean into bending and shaping the bright and flexible little wires. Fast forward 14 weeks, and I am reading Smith's portfolio. She has chosen the theme of pivoting, a term that is meaningful to her in multiple arenas of her life, including basketball as well as education. After reading Smith's explanation of this theme in her introduction to her portfolio, I come upon the following artifact-reflection pair:

This is my pipe cleaner from our very first class session on September 11th. The activity was to craft a shape that captures how you see yourself as an educator or educators more generally. My pipe cleaner has been sitting on my desk for the whole semester. I found myself glancing at it at times and reminding myself of the moment in time that I created it. For the most part, however, it sat on my desk, going unnoticed, yet I, for some reason, never got rid of it when I did my routine cleaning.

I remember when we were first explained the activity and tasked to craft a shape, I was confused about what the purpose of it was and how it would relate to what we were learning. It took me some time before I could even think of how to begin shaping my pipe cleaner. I

Figure 1
Pipe Cleaner



ultimately came up with the shape pictured: it is a spiral leading to a thought bubble. In the process of making it, I remember not really having a clue what I was creating or what I was trying to have it represent. It wasn't until I saw and heard what others in the class had to say about their pipe cleaners that I was able to solidify what my thinking behind that shape was. I came to the conclusion that the spiral represents a continuous flow of knowledge in both directions. We had talked about how education is ongoing, so the thought bubble represents that endless continuation of thinking and learning.

The pipe cleaner activity taps into unconscious and creative processes that are always unfolding within students and invites the students to make those explicit.

Thinking back to that moment in our first class, I think the reason I couldn't think of how to shape the pipe cleaner or what I wanted to depict was due to the fact that I did not yet see myself as an educator. Because of that, I was trying to think of educators more generally, but I didn't find any personal connection to that, which made it harder. Also, looking around and seeing everyone else start shaping their pipe cleaner almost immediately gave me almost a sense of imposter syndrome. I felt like I didn't resonate with the activity as much as others did. However, now that I have completed the course and looked at how that changed from week to week following our class sessions, I have become more and more aware of how I fit the role of being an educator. I realized that one doesn't have to be in a formal teaching role to be an educator. Everyone is always an educator and a student. I don't remember exactly when the shift or pivot happened to when I came to that realization, but I am proud of how far I have come in just a few months. It's funny how much can be said and learned from just a simple purple pipe cleaner.

Reflections

The pipe cleaner activity taps into unconscious and creative processes that are always unfolding within students and invites the students to make those explicit, through reflection on their own and through dialogue with others. It was in part through listening to other students' explanations of their pipe-cleaner shapes that Smith clarified the meaning and significance of her own. Linked to Ferrari et al.'s (2009) assertion regarding the importance of a "nurturing environment" in which students "can freely discuss their problems" (p. 22), Smith felt able to mention challenges or struggles—finding herself less engaged with the activity than other students, experiencing imposter syndrome. This feeling, Smith notes, was mostly due to the safe learning environment that was co-created in our class by both the students and the professor. She felt that she was allowed to admit to struggles or confusion because the environment we co-created was one where she felt unjudged and one where we all worked to help each other learn.

The final sentences of Smith's portfolio entry, and particularly words such as "proud," reflect another of Ferrari et al.'s (2009) points: "the creative spark" in an environment "where students feel rewarded, are active learners, [and] have a sense of ownership" (p. 22) as she came to see and value herself as an educator. Smith appreciated having a tangible reflection of her learning and growth that she could take away from the course and at which she could continue to look back. She illustrates in her reflection the very spiral she represented in her pipe cleaner shape: the "endless continuation of thinking and learning" and the confidence to engage in those processes, as she comes to define herself as an educator.

The multimodal activity—physically bending a pipe cleaner, talking with others about the significance of the shapes created, writing about the significance of both—allowed Smith to find her own entry-point and connection to it, especially since the instructions left room for everyone's own interpretation. This activity and the ways Smith and other students engaged with it reflect the remaking of a learning environment and the collaborative experience that is learning—fostered and deepened by the multimodality that was offered and taken up. Smith's inclusion of an image of the pipe cleaner shape and her reflection on it as part of her self-assessment through the final portfolio capture the sense of ownership (Ferrari et al., 2009) Smith experienced of her own learning process and her representation of in this form of summative assessment.

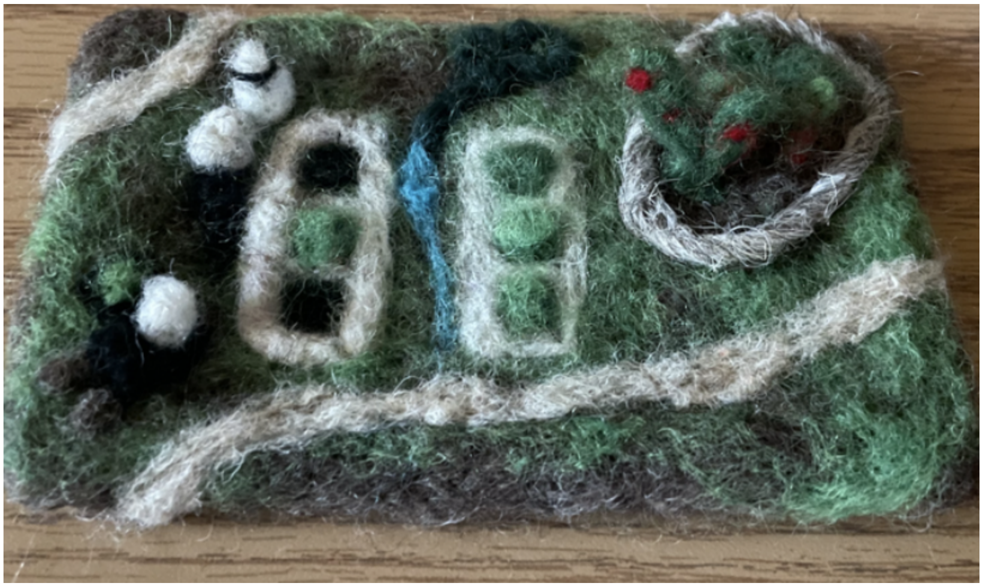
Example 2: Rolfes on Finding Ways to Complete Course Work with a Concussion

Cook-Sather: We are about ¾ of the way through the semester. Rolfes has let me know that they have suffered a concussion that precludes their spending too much time staring at a computer screen. They have been striving to find ways of engaging in the course other than only reading and writing texts. The journal entries assigned during the last several weeks in CLC invite drafts of portfolio components to give students a chance to circle back through their work and try out ways of representing their growth, as Smith's pipe-cleaner shape and reflection above describe. Rolfes has sent me an email explaining that they have made an art piece to go along with slides they have been working on that contain the bulk of their reflective and culminating work for the portfolio. In that email, they write: "I was wondering if we may be able to meet to discuss creative ways for me to still meet the assignment with integrity but also feels approachable given my current health." We have that meeting and decide on a combination of visual, written, and audio-recorded components to the portfolio, which will include reflection on Rolfes' semester-long engagement with Common Space, a local non-profit that provides "a shared space where people of all ages, races, abilities, ethnicities, and economic backgrounds make connections and cross boundaries" (https://www.commonspaceardmore.org). This non-profit organizes classes, workshops, and discussion groups, as well as provides a community space that offers fair-wage jobs for community members with many talents and abilities who are seeking part-time/flexible employment and appreciate a place to belong (https://www.commonspaceardmore.org). Several weeks later, I am reading through Rolfes' portfolio, learning from their explanation of the themes of time, structure, and slowing down that they have chosen. The third artifact-reflection pair is the art piece Rolfes had mentioned and their reflection on it:

The multimodal activity—physically bending a pipe cleaner, talking with others about the significance of the shapes created, writing about the significance of both—allowed Smith to find her own entry-point and connection to it.

My third artifact is artwork that I engaged with while at a particularly difficult point with my concussion and overall health. I wanted to still be able to reflect and spend time with the course and my partnership [with Common Space] and decided to use needle felting as a means of reflection. I decided to make a felted rendition of the garden plot behind the Haverford College which [another student] and I frequented in our partnership. I allowed myself to create the garden by my memory and recollections of the space—memories that I have been able to enhance through our mindfulness, consistent visits, and dedication. Needle felting, as a practice, is inherently slow; it asks the creator to take time in order to make the vision come to light. In the time it took for me to make this and step back into the sensations of the space, I was able to reflect on the connections this partnership has developed for me

Figure 2
Felt Garden Plot



through the CLC texts and relationship building we have engaged with this semester. The time in the Common Space garden has furthered my appreciation for varying fields of knowledge outside of the "traditional" scope. As discussed in bell hook's (1994) *Teaching to Transgress*, I was considering the connections of this work to the concept of progressive pedagogy, and the role of the student in the learning environment. Seeing as the garden requires an embodied and hands-on interaction of the student/learner with the plants and earth, there is an agreement that develops and invites the student/learner to be present in a new capacity. This particular set of knowledge and skill being steeped in the natural world also drew me to connections between the Alaska Standards for Culturally Responsive Schools. Many of these standards implore classrooms to connect the students with an understanding of the natural world and the stewardship of it, particularly, in a culturally responsive and attentive lens. One example of this is "Students who meet this cultural standard are able to: recognize and build upon the interrelationships that exist among the spiritual, natural and human realms in the world around them, as reflected in their own cultural traditions and beliefs as well as those of others." I have found being present to this experiential and nature-based education to ground me in the value of cultivating relationships with nature and people can and should occur in tandem.

As with many reconceptualizations prompted by particular needs and challenges, this kind of reconceptualization can benefit everyone.

Reflections

The necessity for Rolfes to slow down because of their concussion combined with their work in disability studies evoke the notion of "temporal re-imagining," which Levy and Young (2020) describe as "the slowing down and stretching of time—of being in the moment—as a method to enter the world of people with PMLD [profound and multiple learning disabilities]" (p. 68). As with many reconceptualizations prompted by particular needs and challenges, this kind of reconceptualization can benefit everyone. If we "re-think how we conceive of time in terms of different lives" (Levy & Young, 2020, p. 70), we can affirm and extend the ways in which all students experience and make the most of time.

Rolfes' portfolio reflection echoes Jewitt's (2008) argument for "meaning making as a process of design" (p. 263) and "choice of representation" that "gives a renewed focus on the role of the learner" (p. 258). Part of the meaning Rolfes makes is drawn from the practices learned through their internship at Common Space, such as mindfulness and presence. Rolfes also weaves into their reflection key ideas from course texts (bell hook's *Teaching to Transgress*, the Alaska Standards for Culturally Responsive Schools), demonstrating an integrated understanding in both theory and practice. Evident in Rolfes' self-assessment through their portfolio is at once process-driven motivation (Fiorella & Mayer, 2015), "a high level of agency," and the metacognition not only to challenge themself but also to respond to the necessity of finding "ways that lie outside their preferred modes" (Bouchey et al., 2021, p. 36).

Looking back on that time, Rolfes notes that not all of their professors were accommodating of their needs; some were less inclined to work with Rolfes to develop other modes of showing their cumulative knowledge. Rolfes felt in those circumstances that some professors were very attached to the way they ask students to share their learning and were not willing to push beyond that. The matter of fairness also seemed to be something of consideration. These professors seemed to feel that Rolfes' necessity to approach an assignment in a different modality or modalities from other students was unfair to those peers. In contrast, we suggest that multimodality broadens 'fairness' or equity. Instead of "assessing students in the same way without paying attention to their differences" (Montenegro & Jankowski, 2017, p. 16), an embrace of multimodality, in Rolfes' case out of necessity, afforded them "an equal and unbiased opportunity" to demonstrate their knowledge and achievements in different ways (Montenegro & Jankowski, 2020, p. 10). In addition, the possibility of multimodality and the development of project ideas, given Rolfes' situation, invited collaboration. This kind of collaboration was not just beneficial in the moment but also helped build skills in sharing knowledge in a multitude of modalities as a practice in centering accessibility and UDL in one's own learning. This development of capacity carries into future jobs, teaching/learning, relationships, partnerships, and more.

Example 3: Moreira on Her Tortured Journey toward Joy

Cook-Sather: From the first days of her participation in CLC until nearly the end, Moreira resented, resisted, and railed against the course. She is not the first to struggle with what the course invites and asks of students—to engage in new and long-standing forms of organizing (of movements and knowledge), community-creating, trauma-informed relating, and decolonization. It was not that she didn't do the work of the assignments—indeed, her weekly journal entries far exceeded the suggested word limit and included a wide range of relevant insights. Furthermore, she already had deep understanding of the premises and commitments of the course—if anything, she was too familiar with them. It was that, in her own words, she did not allow herself to accept how the course offered and asked her to engage, including through her own joy. As she explains in retrospect, at the challenging intersection of the pandemic and the Black Lives Matter movement, she was asserting autonomy, pushing against external influence, experiencing emotions of anger, frustration, resentment, engaging in active resistance, and arguing or disagreeing with the professor in the classroom. It was only in her final portfolio, for which she chose artistic modes of woodcuts and poetically structured reflections, that Moreira let the meaning she made through the course and of her own development as an educator be affirming and joyful. Indeed, education as joy was the theme Moreira chose for her portfolio. Below are two of the woodcuts from her portfolio and selections from her introduction:

These are the ideas I struggled with over the course of the semester. I feared enlightening myself to my position in this world as a learner, and reciprocally, as an educator. Because it's obvious to me. What tears me up, what riles revolutions: the need for joy for joy for joy. The search to find joy in education—the nuances of intersectionality, privilege, and spaces where academic institutions inflict violence—is rather complicated.

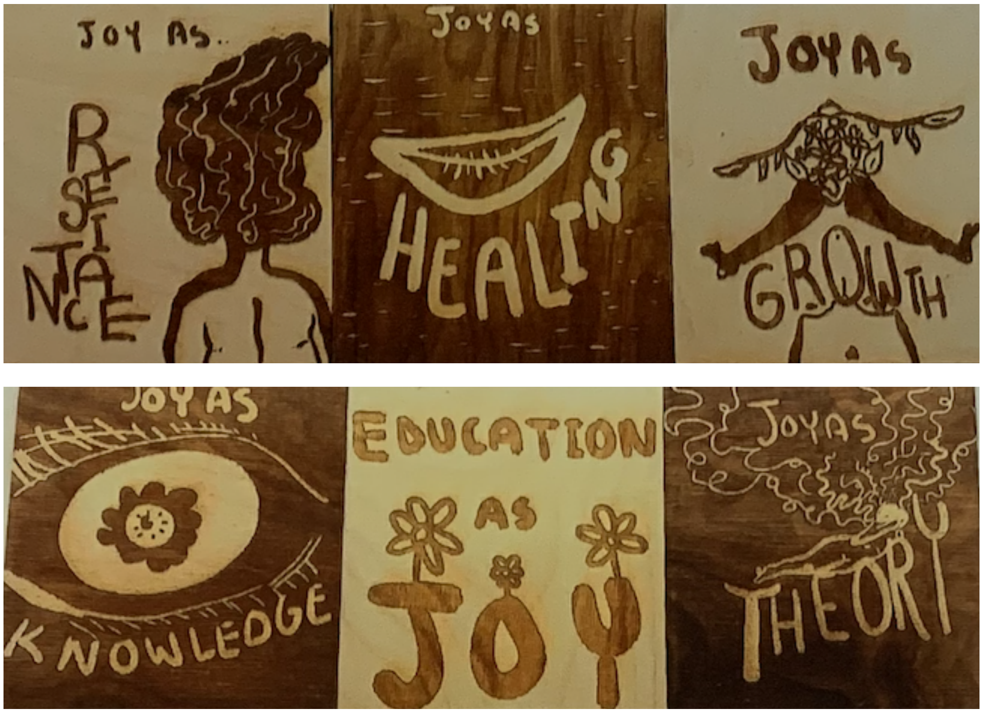
My time in CLC was an uncomfortable semester in reimagining and reevaluating my joy. It was not the place I had anticipated to feel joy but instead theorize my joy. Which, I must add, inherently sounds unjoyful.

I first needed to understand where I was as a learner. Embedded in my reflection amidst the dense, stream-of-consciousness, write-down-before-you-forget journal entries, I was stuck in a mental vortex of writing to validate instead of to learn. I wrote extensively on how I see joy in my education, the complexity to my joy, but never how joy heals me. I came to realize I am a terrible listener, that I barely listen to myself. This became a problem.

We suggest that multimodality broadens 'fairness' or equity.

I did not see reflection as a way of knowing. I squirmed and toiled. I loudly complained. I pointed fingers at moments and places in this course where I could not feel joy. I had not felt inspired or pressured to create the change I saw revolution as. I wanted to read, yet I did not

Figure 3
Woodblock Portfolio



consider the magnitude of impact of what I read. I hungered for knowledge, yet I refused to let it nourish me….

This portfolio contains only the snapshots of some pertinent reflections on my journey seeking joy. It exists as a reaction to my discomfort with the course structure, to bell hooks, and to my explicit desire to make and be authentic. It's more of my physical representation of what otherwise is abstract space…This portfolio excludes all the things that I really enjoyed about this course…

In the act of making, Moreira reflects, she slowed down to tell a story, making an interrelated set of representations of joy and what that meant for her.

And with that being said, I hope you can give this all a complete listen and take time to see all the pieces in this portfolio in the order as presented and when you are finished, assemble it together just as you found it.

Reflections

As illustrated in the woodcuts and her reflection, Moreira exercised "choice of representation" (Jewitt, 2008, p. 258), process-driven motivation (Fiorella & Mayer, 2015), and a high level of agency and metacognition (Bouchey et al., 2021), and she also moved through her frustration and resistance to create and embrace joy. (See Tolman et al., 2017, for a helpful discussion of understanding student resistance.) While she is a creative person and advocates using multiple modes to learn and convey understanding, Moreira had never used woodcuts before, and therefore challenged herself with a new mode. Although she had not originally planned to make her portfolio into a book of woodblocks, she tried (and ultimately did not succeed) to accomplish this goal. Reflecting subsequently, she noted that the effectiveness of creating was a moment to pause and practice/learn something new even though she had a "clear"-but-impossible-to-execute vision at the beginning. And while the printing project did not succeed, it catalyzed a process of finding peace (healing, one of the sections represented in the woodcut above was titled). In the act of making, Moreira reflects, she slowed down to tell a story, making an interrelated set of representations of joy and what that meant for her. Through this and other processes in which she engaged in her final work for the course, Moreira notes, "joy was being alone and realizing everything [she] touched required [her] to consider how [she] touches and how it feels to be touched by [her]."

The courage and candor Moreira demonstrated in creating the portfolio she did might have been fostered by the "nurturing environment…where students feel rewarded, are active learners, have a sense of ownership, and can freely discuss their problems" (Ferrari et al., 2009, p. 22), might have been driven entirely by her own squirming and toiling and independence, or some combination—but regardless, the multimodality affirmed a commitment to inclusion. Moreira's desire for authenticity, to ground what would otherwise have remained abstract, centers "design, diversity, and multiplicity" in her meaning-making practices and interpretative work (Jewitt, 2008, p. 258), thereby recognizing her choices as a form of self-empowerment, of modeling for and educating others about diversity in expression and representation of understanding, and of contribution to a movement toward equity and inclusion in higher education.

Moreira approached this multimodal assignment with a lot of previous experiences; she had worked in makerspaces for years teaching students how to make multimodal inventions, stories, creations, and she had a background in theater, video production, and animation. Her interest in multimodality emerged due to her identity as a bilingual English-Spanish speaker. In her educational experiences, multimodal projects served as a medium to communicate ideas with her Spanish-dominant speaking mother in ways the written English Language could not always capture. For Moreira, then, multimodality was not new, and she was a willing and eager participant. She notes that she will never reread an essay she wrote, but she will never forget when a class gave her an opportunity to create something. Moreira embraces multimodality whenever possible and notes that multimodality in science communication was her action research project Wake Forest University. Multimodal works in education are valuable because there is an audience beyond the instructor and the student (or an ephemeral class presentation); the audience is whomever you wish to share your art with, it escapes the course, it makes assignments in the academy acts of reaching out into the world.

Similar to how the collaboration necessary for Rolfes' multimodal engagement informs capacities that will be useful beyond their time in college, Moreira's master's thesis at Wake Forest University, through which she studied how her students react to creating digital multimodal science explanations as a culminating assignment, focused in part on inspiring students to share what they make ultimately with their families, their friends, their loved ones—deepening connections beyond school. Thus, the multimodality of the portfolio as summative assessment in CLC was a form of authentic assessment.

Conclusion and Call for Further Research

The invitation to embrace multimodality in the portfolio as a form of summative assessment supported Smith, Rolfes, and Moreira in forging or clarifying definitions of themselves as educators—an identity that each person has to come to in their own way, in their own time, and on their own terms as a process of self-authorship (Baxter Magolda, 2007). Since that is the goal of CLC, it is essential that every student, whether out of choice or necessity, feels empowered to create ways of sharing their knowledge and growth. Some students choose text (printed or digital) as their preferred mode of expression in their final portfolios, and others choose different modes than the ones we have highlighted here, such as images they find (rather than create) and audio or audiovisual recordings (rather than text-based analyses).

Multimodality is inherent in the activities and their processing as well as in the forms portfolios as summative assessment took. This work is at once individual and relational. For instance, Smith noted that it was not until she saw and heard what others in the class had to say about their pipe cleaners that she was able to solidify what her thinking behind her own shape was. Rolfes also highlights the individual and relational nature of their work through being present to the experiential and nature-based education in their internship, which grounded them in "the value of cultivating relationships with nature and people." And Moreira invited active engagement with what had been a very personal and individual process of creating her portfolio, linking the individual with the relational. This relationality carries over to future students: as Moreira notes, she experienced portfolio assessments in the Education Department at Bryn Mawr and Haverford Colleges as a way to shift toward equity that made her a better teacher after graduation.

The multimodality affirmed a commitment to inclusion.

With Bouchey et al. (2021), we advocate an embrace of multimodality, and we see the portfolio as a form of summative assessment that responds to the call for diversification in forms of expression and for recognition of how multiple modes of expression always inform and are informed by one another. The "renewed focus on the role of the learner" (Jewitt, 2008, p. 263) that multimodality affords allows us to move away from "assessing students in the same way without paying attention to their differences" (Montenegro & Jankowski, 2017, p. 16) and toward providing all learners "an equal and unbiased opportunity" to demonstrate their knowledge and achievements in different ways (Montenegro & Jankowski, 2020, p. 10). The ways that enrolled students like Smith, Rolfes, and Moreira have taken up the portfolio assignment demonstrate how they have cultivated creative dispositions and agency and engaged in creative processes (Ross et al., 2020) as well as "a high level of agency (self-discipline)" and a heightened awareness of when and how "to challenge themselves to learn in ways that lie outside their preferred modes" (Bouchey et al., 2021, p. 36).

We encourage others both to embrace multimodality and to pursue systematic research into those approaches and their outcomes. Instructors could start with introducing multimodal options into formative assessment or into moments of assessment prior to the end of the term, such as for oral presentations or other course projects. This strategy would afford all involved opportunities to experience this new-to-many approach at lower-stakes moments. To scale such approaches for summative assessment in larger courses and across whole departments, instructors might consider offering multimodal options for portions (not necessarily the entirety) of final exams or papers and inviting students to self-assess for what they learned from the experience.

Finally, because we share only three examples, more research is needed to understand the diversity of ways students experience multimodality. The experiences students offer here suggest possible areas of focus for such research, including: ways that multimodality can respond to long-term or temporary disabilities, build on student capacities otherwise not recognized in traditional approaches to assessment, and, encompassing both of those, move toward more equitable approaches to assessment.

References

Adami, E. (2016). Introducing multimodality. *The Oxford Handbook of Language and Society*, 451-472.

Alaska Standards for Culturally Responsive Schools. https://www.uaf.edu/ankn/publications/guides/alaska-standards-for-cult/

Baxter Magolda, M. B. (2007). Self-authorship: The foundation for twenty-first-century education. *New Directions for Teaching and Learning*, *109*, 69-83. https://doi.org/10.1002/tl.266

Bouchey, B., Castek, J., & Thygeson, J. (2021). Multimodal learning. In J. Ryoo & K. Winkelmann (Eds.), *Innovative Learning Environments in STEM Higher Education*. SpringerBriefs in Statistics. https://doi.org/10.1007/978-3-030-58948-6_3

Cook-Sather, A. (2023, February). Assigning "Accountability Partners" to support student engagement, learning, and growth. *Faculty Focus*. https://www.facultyfocus.com/articles/equality-inclusion-and-diversity/assigning-accountability-partners-to-support-student-engagement-learning-and-growth/

Ferrari A., Cachia, R., & Punie, Y. (2009). Innovation and creativity in education and training in the EU member states: Fostering creative learning and supporting innovative teaching, 1-61. http://ftp.jrc.es/EURdoc/JRC52374_TN.pdf

Fiorella, L., & Mayer, R. E. (2015). *Learning as a Generative Activity: Eight Learning Strategies that Promote Understanding* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781107707085

hooks, b. (1994). *Teaching to Transgress: Education as the Practice of Freedom*. Routledge.

Jewitt, C. (2008). Multimodality and literacy in school classrooms. *Review of Research in Education*, *32*, 241–267. DOI: 10.3102/0091732X07310586

Jewitt, C., Bezemer, J., & O'Halloran, K. (2016). *Introducing Multimodality*. Routledge.

Kress, G. R. (2010). *Multimodality: A Social Semiotic Approach to Contemporary Communication*. Routledge.

Kress, G. R., & Leeuwen, T. van. (2001). *Multimodal Discourse: The Modes and Media of Contemporary Communication*. Arnold.

Levy, S., & Young, H. (2020). Arts, disability and crip theory: Temporal re-Imagining in social care for people with profound and multiple learning disabilities. *Scandinavian Journal of Disability Research*, *22*(1), 68–79. DOI: https://doi.org/10.16993/sjdr.620

McArthur, J. (2024). Epistemic justice and authentic assessment. In M. Meredith (Ed.), *Universities and Epistemic Justice in a Plural World: Knowing Better* (pp. 121-133). (Debating Higher Education: Philosophical Perspectives; Vol. 12). Springer. https://doi.org/10.1007/978-981-99-9852-4_9

Mingus, M. (2019, May 5). "Dreaming Accountability." *Leaving Evidence*. https://leavingevidence.wordpress.com/2019/05/05/dreaming-accountability-dreaming-a-returning-to-ourselves-and-each-other/

Montenegro, E., & Jankowski, N. A. (2020, January). A New Decade for Assessment: Embedding Equity into Assessment Praxis (Occasional Paper No. 42). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA), 1-26. https://www.learningoutcomesassessment.org/wp-content/uploads/2020/01/A-New-Decade-for-Assessment.pdf

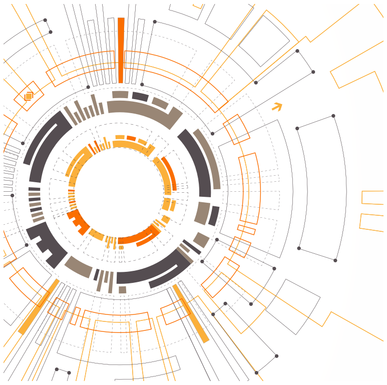
Montenegro, E., & Jankowski, N. A. (2017, January). Equity and Assessment: Moving towards Culturally Responsive Assessment. (Occasional Paper No. 29). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA), 1-23. https://learningoutcomesassessment.org/wp-content/uploads/2019/02/OccasionalPaper29.pdf

Reyna, J., Hanham, J., & Meier, P. (2017). A taxonomy of digital media types for learner-generated digital media assignments. *E-Learning and Digital Media*, *14*(6), 309–322. https://doi.org/10.1177/2042753017752973

Reyna, J., Hanham, J., Vlachopoulos, P., & Meier, P. (2021). A systematic approach to designing, implementing, and evaluating learner-generated digital media (LGDM) assignments and its effect on self-regulation in tertiary science education. *Research in Science Education*, *51*(6), 1501–1527. https://doi.org/10.1007/s11165-019-09885-x

Ross, J., Curwood, J. S., & Bell, A. (2020). A multimodal assessment framework for higher education. *E-Learning and Digital Media*, *17*(4), 290-306. https://doi.org/10.1177/2042753020927201

Tolman, A. O., Sechler, A., & Smart, S. (2017). Defining and understanding student resistance. In Tolman, A. O., & Kremling, J. (Eds). *Why Students Resist Learning: A Practical Model for Understanding and Helping*, pp. 1-20. Stylus Publishers.



Abstract

Advanced practices for summative exam development and post-exam analysis are proven to be effective but aren't always practical, and, even when these are applied to some degree, exams remain inherently imperfect measures of student ability. Instructors may thus deem it necessary to adjust overall exam scores to account for aspects of an exam that may have been ill-suited to some or all students, and often these adjustments are made in an ad hoc and/or uninformed manner. This paper reviews reasons and methods for adjusting exam scores and proposes a new method that was developed organically from observation, reflection, and literature consultation. The scaling method considers that some of the underlying reasons for adjusting exam scores may affect certain sets of students more than others and seeks to incorporate this proposition while also avoiding weaknesses of other methods.

AUTHORS
Ryan K. Orchard, MASc
Macewan University

A Review Of Practices For Adjusting Exam Scores And A Proposed Nonlinear Scaling Method

Exams are inherently imperfect measures of student ability. Practices and tools exist to improve exam validity and reliability, but, even if these were convenient and accessible to the common higher education instructor and widely-used, the exams would still be prone to some degree of measurement error. It is postulated here that it is not uncommon for higher education instructors in ordinary exam settings, such as non-standardized exams developed and graded by the instructor, to make a post-exam adjustment to scores in a manner based on their own intuition. The current paper does not condone this practice, but rather it recognizes it as a reality, reflects upon reasons and methods for adjusting exam scores, and proposes a new method that was developed organically in response to years of observation while manually grading more than ten thousand exams (and counting).

CORRESPONDENCE
Email
orchardr@macewan.ca

The following will first provide context and scope, followed by a discussion of relevant considerations and best practices for exam development, particularly as they relate to exam validity and reliability in the given context and how they imply reasons for exam score adjustment and inform the choice of method. Next, methods for adjusting exam scores will be reviewed and critiqued. A new method will then be described and compared to the other methods, and implementation advice will be offered. A conclusion and discussion of limitations will end the paper.

Context and Scope

Grade-Curving Versus Grade-Scaling

Norm-referencing exam scores refers to the practice of referencing individual scores to those in a group or to a pre-defined grade distribution and is more common in situations where students are ranked for some purpose such as admission or awards (Kibble, 2017; Kulick & Wright, 2008), or for credentialing (Ben-David, 2000). *Grading on a curve* (*curving grades*, *grade-curving*) is a form of norm-referencing and generally refers to the practice of fitting exam grades to the normal distribution (the 'curve') to achieve a pre-determined proportion of students falling into each grade category (Tan et al., 2020). Grade-curving is a common practice in higher education (Kulick & Wright 2008), but is also somewhat controversial - see, for example, Grant (2016), who also noted that the practice may be used as a foil to grade inflation, and for this reason (and/or others) may be mandated by an institution. Kulick and Wright (2008) show that grade-curving can lead to the inclusion of luck as an unwanted factor in the partitioning of students into grade categories based on relative performance, which can have significant negative consequences for some students. Overall, there are strong arguments that grading on a curve tends to have more flaws than benefits (see, for example, Close, 2009), and, except for specific applications, should not be practiced.

It is still necessary to explore the why, as a means to inform the how.

For the current paper, the term *scaling*, as applied to exam scores, will refer to the general practice of applying a mathematical operation to adjust scores (usually upward) in a direct manner that does *not* compare individual scores to the group (as grade-curving does). Although the terms *curving* and *scaling* are used interchangeably in some of the literature, the current paper will differentiate between them as described and will focus only on grade-scaling.

Application

This paper assumes a context and scope in which the exams are developed, delivered, and graded (perhaps with assistance) by individual instructors in higher education, as opposed to settings such as large-scale standardized exams. This is an important stipulation given the realities of a potential lack of best practice components for exam development such as development blueprints, peer review processes for improvement, and detailed item analysis statistics (all discussed in the next section), but also perhaps due to time constraints experienced by individual instructors. Indeed, it seems difficult to deny that at least some exam authoring and administration in higher education may be performed using imperfect processes and without extensive training. Anderson (2018) notes that single-task grades (i.e., for an individual piece of work) tend to be very unreliable (p. 9), and that there can be significant disparity between teachers' opinions about what is fair grading and what is not (p. 18). As mentioned in the introduction, the current paper does not dismiss the value of using tools and best practices for exam quality improvement, but rather it assumes that an absence of these may be the case in many settings and that instructors are adjusting overall exam scores according to their judgement, and thus aims to better inform this inevitable process.

The score-adjustment methods presented in this paper are for adjusting *total individual exam scores*, not individual items (questions), although methods for exam score adjustment could be used in combination with (after applying) measures to address any issues with individual items. The discussion of exam scores will refer to the *numerator* as the total score for the responses on an individual exam and the *denominator* as the total (maximum) possible score for the numerator.

Background - Reasons For Adjusting Exam Scores

The topic of exam score adjustment (reasons or methods for) has very little explicit coverage in the literature, and thus most of the background provided here will be based on literature concerning topics such as item and exam development, as well as observations and reflections of the current author. Although the objective of the current paper is to explore *how* exam scores can be adjusted, it is still necessary to explore the *why*, as a means to inform the *how*; indeed, the new method proposed in this paper arose organically from years of observation and was developed to specifically address the perceived *why*. Embedded within the why-how relationship is the question of whether the underlying reasons affect some students differently than others, which could justify differences in score-adjustment depending on the raw score; the current paper will argue for a scaling method that adjusts the higher exam scores of students for whom the exam may have been better suited differently than it adjusts lower exam scores.

The reasons for exam score adjustment will be organized into a few general categories and discussed in turn below: individual items (questions); overall exam composition (distribution of question difficulty, and coverage of learning objectives and cognitive skill levels); and student interpretation and other student-related factors affecting performance. The terms *validity* (whether a test measured what it is supposed to measure) and *reliability* (whether a measurement result can be reproduced) may be used during discussion of individual items and/or exams (definitions provided are based on Kibble, 2017).

Individual Items

Malau-Aduli and Zimitat (2012) found that "there is considerable evidence that multiple-choice questions (MCQs) are poorly written" (p. 920) and went on to show that a peer review process for exams comprised of MCQs can have an improving effect on the quality of an exam, in many aspects. Kibble (2017) discovered that "faculty members cannot reliably judge the difficulty of individual items they write" (p. 111) and recommended faculty development, peer review in test development, and being deliberate in matching learning objectives and exam items (the latter two topics are discussed further in the next subsection). Indeed, exam authoring can be complex and time-consuming, particularly when it comes to writing MCQs that are of high quality, and increasingly so when higher cognitive abilities (beyond memorization) are to be evaluated, as described by Stevens et al. (2022). These authors provide a thorough framework to guide the authoring of MCQs; for example, questions that are based on a vignette (scenario) should be well-edited to be "unambiguous, concise, and readable" (p. 6), so as not to affect the student's comprehension of the question [which still might not account for student-specific factors, such as language differences, or even a propensity for some to overthink]. Adhi and Aly (2018) conducted a study in which they found that MCQs of the *one-best* type performed better than those of the *one-correct* type, in terms of student scores, reliability and discrimination ability. The depth and intricacies of the principles described by these (and other) authors implies that many exams, whether employing these practices or not, likely include at least some questions that are not sound and may affect the performance of some students.

Will argue for a scaling method adjusts the higher exam scores of students for whom the exam may have been better suited differently than it adjusts lower exam scores.

Depending on how an exam is administered and graded, there may be tools available for evaluating the *individual items* of an exam. Camenares (2022) encouraged using *item analysis*, which includes measures of question difficulty (percentage success rate for each question, over all students) and discrimination index (the relative success on a given question of students who were high-performing versus low-performing on the exam). If a specific unfair question is identified, it may be necessary to remove it entirely from the exam grade calculations, such as in the case where all students answered incorrectly, or remove it only for those students who answered incorrectly (which doesn't penalize those that did answer correctly and have it counted in the numerator and denominator, unless the grades are norm-referenced). Another option is to adjust student scores on individual questions using a "proportional bonus for questions that are too difficult and/or poor discriminators" (Camenares, 2022, p. 2). Rudolph et al. (2019) suggested that, in some cases, full credit for a poorly performing item could be awarded to all students, regardless of their answer on the question, or as a bonus for those that did get it correct (although in a later discussion of exam-level grade adjustments the current author will argue against providing a grade for something that doesn't deserve a grade). Rudolph et al. also contended that 'examinations are likely to contain quite a few flaws' (p. 1502), which (in the eyes of the current author) may warrant *exam-level* score adjustments in the absence of identification of and adjustment for specific poor items on the exam. It would also seem that an overall exam score adjustment may be less necessary for the students who answered most of the questions correctly, since they may have been less affected by any poorly constructed questions; this premise is modeled by the current proposed exam score scaling method.

Exam Composition - Distribution Of Question Difficulty

Most students correctly answer questions with low difficulty indexes (i.e., easy questions), while only the top students tend to correctly answer those with high difficulty indexes (Downing, 2009). Malau-Aduli and Zimitat (2012) alluded to "the theory that the most informative test items are those of middle difficulty and they provide higher discrimination between the high-scoring and low-scoring examinees" (p. 921). Thus, even with the help of detailed item analysis, it may be that no particular question(s) can be identified as poorly constructed, but rather the *overall exam* may have had too high a proportion of difficult questions and a low average score, which would not have disadvantaged the top achievers as much as the others (and thus may not warrant as much of a grade-adjustment). Further, assuming that there were at least *some* easy questions, any very low exam scores (relative to other exam scores) may be more attributable to a lack of preparation by some students than shortcomings of the exam; a consequence of the proposed method is that very low exam scores do not receive as large of grade increases as those in the middle (as will be described with the proposed method), which may have some justification.

Exam Composition - Coverage Of Learning Objectives And Cognitive Skill Levels

Exam blueprints are outlines and plans that "guide item writers to develop sufficient items that cover important content areas and objectives at the suitable cognitive level" (Eweda et al., 2020, p. 166). Abdellatif et al. (2024) promote blueprinting as a tool to combat two major threats to exam validity - construct underrepresentation (some course content not being appropriately represented on the exam), and construct irrelevant variation (question format, questions being too easy or difficult, or the test modality being inaccurate); Kibble (2017) adds the issue of "language cueing test-wise students to the correct answer and guessing from limited option sets" (p. 115), which speaks to the premise of the current paper that any overall score adjustment should be less necessary for those whom the exam was better suited (i.e., the 'test-wise' students who may not have been harmed even by poorly constructed items).

Using a blueprinting (or other) process doesn't mean perfect exams, however. Welch et al. (2017) found that "the accuracy and reliability of [faculty member's] ability to categorize" Bloom's Taxonomy to exam questions were low (p. 103). Omar et al. (2012) also recognized the importance of balancing lower and higher cognitive level questions, and the proper classification using Bloom's, and provided preliminary results indicating that an automated system may be useful in assisting with classification. Wellberg (2023) pointed out that mathematics teachers "tend to conflate difficulty and cognitive complexity" (p. 58). In the present context, these works suggest that if a good exam requires a good balance of question cognitive levels, among other factors, then the imperfect ability for any instructor to accurately categorize questions by cognitive level may mean that an exam in which the balance is deemed to have been questionable (based on instructor judgment, post-exam) may warrant score adjustments. Again, this would seem less likely to have adversely affected those students who had high raw test scores (before adjustment), and the proposed scaling method accounts for that.

Student Interpretation

During an exam that was developed and administered by the current author, a student annotated their thoughts on many of the multiple-choice questions (by their own accord). On a few questions it was clear that the student was knowledgeable about the concept but chose what the instructor considered an incorrect answer, including one particular question in which some of their annotations were evidence that they had been distracted by a single word in the question stem, but that they clearly understood the target concept. It was difficult to give a score for the question since a grader wouldn't be able to follow the same procedure on all of the student's incorrectly-answered questions (much of the annotation was difficult to follow) or for all students (most others didn't annotate their exam at all), or to analyze annotations on correctly-answered questions to confirm understanding of the target concept. The specific question had been used for years, and generally had a high success rate, and it seemed unlikely that any formal item analysis would have flagged it, despite the clear trouble that it gave one otherwise knowledgeable student. Another example (personal): once upon returning home

A consequence of the proposed method is that very low exam scores do not receive as large of grade increases as those in the middle.

from a walk the current author asked of their teenager "which of the following people that you know did we see at the dog park today – (a) your former pre-school teacher Ms. H, (b) the father of your friend L, (c) professional athlete RNH, or (d) all of the above?" The teenager answered 'a', and upon being told that the correct answer was 'd', they replied "but I don't *know* RNH" [they only knew *of* them]. It can be surprising to a question author to find that a single and possibly extraneous word can be interpreted by one or more students in a very literal way and have unintended consequences, and this may even occur without the author ever knowing.

Noble et al. (2012) showed that a student's answers don't always reflect their knowledge; specifically, they interviewed students from different groups to evaluate target knowledge (whether the student understood the concept), which in some cases was clearly demonstrated despite the student having chosen an incorrect answer. They noted that "the difficulty [for the examinee]…was not in understanding [the target concept], but in interpreting the language of the test item and in creating a context in which the language of the item made sense to them" (p. 792). They also found that students from certain groups such as low-income households and English Language Learners were more likely to make an incorrect answer choice despite understanding the concept. They point to other studies in which answers differ between students not because of differences in knowledge or ability but because of the "interaction between students' language and life experiences and the structure and content of test items" (p. 781). According to Fencl (2019), non-academic factors such as exam anxiety and language barriers can also affect exam performance. Thelk (2008) observed that students in one group can have a greater probability of correctly answering a question than students in another group, even after controlling for ability, due to bias at the item (question) level. Crawford and Fekete (1997) found that

A measured application of exam-score adjustment, when deemed appropriate, doesn't seem out of line

> an instructor's expectations of the skills needed to answer a question were sometimes inaccurate…[and in]…some cases the most important skill that determined a student's grade on the question [was]…coping with distraction… much of the class did not even realize what key concepts were involved in the question, but instead they were distracted by irrelevant information (p. 188),

while Holmes (2021) conducted a study in which they found evidence that most students believe that some can simply be 'bad test-takers', even when they know the course material, and noted that identifying as such can have a strong association with exam anxiety (p. 297).

It seems, then, that although there may be many practices to avoid the shortcomings of exams, a measured application of exam-score adjustment, when deemed appropriate, doesn't seem out of line, nor do minor differences in how it affects raw scores at different points in the overall distribution, as will be seen in the proposed method.

Other Motivations For Exam Score Adjustment

Finally, in many settings, instructors are free to use their own intuition and judgement to guide exam authoring and grading, and may judge an overall average exam grade to be too low for reasons of their own. Additionally, career pressure on instructors may motivate them to give high grades (and thus bump low exam results), as described by Wellberg (2023) as well as Jephcote et al. (2021), who noted that "the continued emphasis on the power of student evaluations may provide instructors with an incentive to…conform to grade leniency" (p. 549). While the current paper doesn't endorse these as valid reasons, if one is going to adjust grades, they should at least be knowledgeable of the different methods.

Exam Score Adjustment Methods

The following will describe methods for exam score adjustment that are carried out *after* any adjustments for specific individual question deficiencies (i.e., methods for scaling *total* exam scores).

Three mock datasets of 1,000 exam scores were simulated using a random number generator to create desired distributions, one approximately normally-distributed, another with a bimodal distribution, and one uniformly-distributed. It was found that the distribution (normal, bimodal, or uniform) of raw exam scores had very little effect on the relative outcomes of the scaling methods, so for the sake of simplicity, the following analysis will only use the normally-distributed simulated exam scores. The mean and median of the mock exam scores were both 65%, with a standard deviation of 11.88%, a maximum grade of 100%, and a minimum grade of 28%. For the sake of a fair comparison, each of the scaling methods were calibrated so that the new (scaled) mean and median were 70% (up from 65% for the raw scores).

The simplest method for adjusting exam scores is to add a *constant value* to all scores (e.g., raise all scores by 5%, such that a 34% score becomes 39% and a 94% becomes 99%). This is obviously very simple to implement and for the exam-takers to understand, and may be considered by many to be fair since the amount of grade increase is equal across all students. Drawbacks include that *equal* and *fair* aren't necessarily the same under all circumstances (i.e., there may be reasons that some students deserve a different adjustment than others, as described previously). Further, 0% and 100% can be distorted - a blank exam would get 5% (i.e., grades are being *gifted*, despite no evidence of meeting learning objectives), and a student who did not answer all exam questions correctly could get an exam score of 100% (or higher). An example is provided in Table 1.

Another simple and common method is to *reduce the denominator* used for converting the scores to percentages, which effectively raises all exam scores (except in the case of a zero raw score). Aside from simplicity, this method has no other documented strengths while it does have some drawbacks, including that a score of 100% or higher is possible for a student that did not answer all questions correctly, and, in the words of Nelson et al. (1992), it "robs from the poor to give to the rich" (p. 463) in that already high exam scores will see a significantly larger increase (adjusted score minus raw score) than lower exam scores. For example, if raw scores are out of 40 and the denominator is adjusted to be 38, then 90% (36/40) becomes 94.7% (36/38), while 50% becomes 52.6% (20/38) and 10% becomes 10.5% (4/38). Although it seems difficult to imagine a scenario where the ultra-high scores should benefit the most from scaling, as happens with this method, it may remain common practice due to convenience, and because the quantitative implications aren't fully understood by some users. Table 1 provides scaled scores using this method for the mock exam scores, calibrated to achieve the desired mean by reducing the denominator from 100 (raw score) to approximately 92.86.

Equal and fair aren't necessarily the same under all circumstances.

Another method is *square root* scaling whereby the square root of the raw score out of 100 is multiplied by 10. 100% will still be 100% and 0% will still be 0%, but scores in between the extremes will be scaled in a way that will benefit "students who need it the most without removing incentive for higher performing students" (Page et al., 2018, p. 6). These authors propose the application of this method as a remedy for the disadvantaged underserved populations in STEM classes. The square root scaling method has a compelling property to it – it raises the scores in the middle portion of the distribution by more than those in the upper and lower sections (see Table 1), which may be justifiable (as described in a previous section of the current paper). However, this method does have a problem that may make it unusable – it raises the grades by too much: for example, 50% becomes 70.71% and 25% becomes 50%. This problem can be overcome by using a different degree (root), which was done for the mock dataset to achieve a 70% average by using a root of 2.14 instead of 2, but it required an optimization tool (Solver or Goal Seek in Excel) and it resulted in some raw scores at the higher end being adjusted negatively, which may make it ill-advised. Table 1 reports scaled values using a square root (and thus don't calibrate to the same mean as the other methods) for the sake of simplicity and to demonstrate the shape of the grade increases.

Maloy (1990) finds issue with adding a constant value to all scores and they propose a nonlinear strategy similar to the square root method in that it raises lower grades by more than it raises higher ones, except for very low grades. The equation is

$$S=100^{1-n} \bullet R^{n} \quad (1)$$

where S is the scaled score (out of 100), R the raw score (out of 100), and n a scaling parameter between zero and one which can be determined using a log-based equation and substituting a desired corresponding pair of values for S and R (for example with R being the *actual* raw score mean, median or passing score and S the *desired* mean, median or passing score). Results of this method are shown in Table 2, where n was calibrated to a value of 0.8222

Table 1
Amount of grade increase (scaled exam scores over raw scores) under three scaling methods – constant increase, denominator-reduction (provides larger score increases for higher scores) and square root (favours lower scores more than higher ones, except for very low scores)

	Constant		Reduce denominator		Square Root	
Raw Score (R)	Scaled Score (S)	Increase (S-R)	Scaled Score (S)	Increase (S-R)	Scaled Score (S)	Increase (S-R)
100	105.00	5.00	107.69	7.69	100.00	0.00
95	100.00	5.00	102.31	7.31	97.47	2.47
90	95.00	5.00	96.92	6.92	94.87	4.87
85	90.00	5.00	91.54	6.54	92.20	7.20
80	85.00	5.00	86.15	6.15	89.44	9.44
75	80.00	5.00	80.77	5.77	86.60	11.60
70	75.00	5.00	75.38	5.38	83.67	13.67
65	70.00	5.00	70.00	5.00	80.62	15.62
60	65.00	5.00	64.62	4.62	77.46	17.46
55	60.00	5.00	59.23	4.23	74.16	19.16
50	55.00	5.00	53.85	3.85	70.71	20.71
45	50.00	5.00	48.46	3.46	67.08	22.08
40	45.00	5.00	43.08	3.08	63.25	23.25
35	40.00	5.00	37.69	2.69	59.16	24.16
30	35.00	5.00	32.31	2.31	54.77	24.77
25	30.00	5.00	26.92	1.92	50.00	25.00
20	25.00	5.00	21.54	1.54	44.72	24.72
15	20.00	5.00	16.15	1.15	38.73	23.73
10	15.00	5.00	10.77	0.77	31.62	21.62
5	10.00	5.00	5.38	0.38	22.36	17.36
0	5.00	5.00	0.00	0.00	0.00	0.00

to increase the mean raw score of 65% for the mock data to a mean scaled score of 70%. (This model will be examined in more detail later when compared to the proposed method).

Becker (1991) responded to Maloy with a method that uses two scaling factors (a and b) that are determined based on the relationship between the highest raw score and desired highest scaled score, as well as the mean raw score and desired mean scaled score (i.e., *a* and *b* are found by simultaneously solving two equations); individual scores are then scaled using the equation

$$S=R \bullet a+b \quad (2)$$

where *S* and *R* are the scaled and raw scores (out of 100), respectively. The value of a will be less than one and will scale the raw grade downward before a constant b is added. This method was calibrated using the mock exam data and setting the maximum raw and scaled scores to both be 100 and the raw and scaled mean scores to be 65% and 70%, respectively, resulting in values of *a*=0.8571 and *b*=14.2857. Table 2 shows that this method benefits a zero score the most, with the difference between scaled and raw scores becoming less as raw scores increase, which avoids compensating already high scores by more than low ones, but results in the largest score increases being applied to those who may have been the least prepared for the exam.

Bailey (1992) provides a method that raises "lower grades more than higher ones but without advancing average students into the A range" (p. 221), using the equation

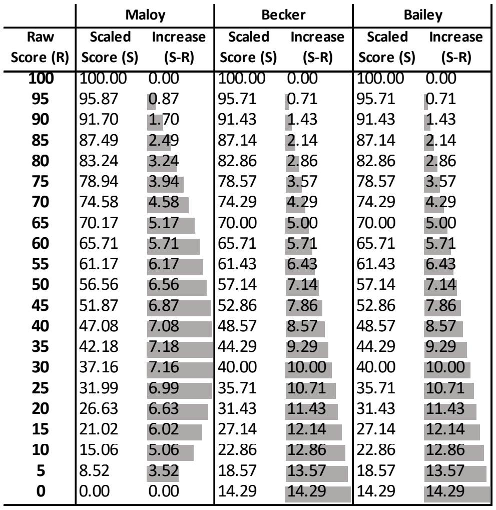
$$S=100-(W \bullet n) \quad (3)$$

where *S* is the scaled score and *W* is the total points possible (100, if a percentage) minus the raw score (*R*), which effectively makes *W* the number of lost marks. *n* is a scaling factor that can be determined by a formula and using the specific values of a desired *R*→*S* transformation (in the mock data, a 65%→70% transformation resulted in *n*=0.8571). Results from this method were identical to those of Becker (i.e., both models increase lower scores by more than higher scores and follow the same general shape, as per Table 2), which will be the case when the Becker model is calibrated using values for max *R* and max *S* of 100%.

Proposed Method

The proposed scaling method was developed in response to observations of general exam deficiencies (described previously), as well as perceived deficiencies in the basic scaling

Table 2
Comparison of amount of grade increase for Maloy, Becker, and Bailey models



	Maloy		Becker		Bailey	
Raw Score (R)	Scaled Score (S)	Increase (S-R)	Scaled Score (S)	Increase (S-R)	Scaled Score (S)	Increase (S-R)
100	100.00	0.00	100.00	0.00	100.00	0.00
95	95.87	0.87	95.71	0.71	95.71	0.71
90	91.70	1.70	91.43	1.43	91.43	1.43
85	87.49	2.49	87.14	2.14	87.14	2.14
80	83.24	3.24	82.86	2.86	82.86	2.86
75	78.94	3.94	78.57	3.57	78.57	3.57
70	74.58	4.58	74.29	4.29	74.29	4.29
65	70.17	5.17	70.00	5.00	70.00	5.00
60	65.71	5.71	65.71	5.71	65.71	5.71
55	61.17	6.17	61.43	6.43	61.43	6.43
50	56.56	6.56	57.14	7.14	57.14	7.14
45	51.87	6.87	52.86	7.86	52.86	7.86
40	47.08	7.08	48.57	8.57	48.57	8.57
35	42.18	7.18	44.29	9.29	44.29	9.29
30	37.16	7.16	40.00	10.00	40.00	10.00
25	31.99	6.99	35.71	10.71	35.71	10.71
20	26.63	6.63	31.43	11.43	31.43	11.43
15	21.02	6.02	27.14	12.14	27.14	12.14
10	15.06	5.06	22.86	12.86	22.86	12.86
5	8.52	3.52	18.57	13.57	18.57	13.57
0	0.00	0.00	14.29	14.29	14.29	14.29

methods used by some instructors, such as an adjustment that gives all students the same increase (which can result in a grade of over 100% and/or give free grades where they may not be earned or warranted), or an across-the-board reduction in the exam score denominator (which increases the grades of students that already had the highest grades by significantly more than it increases the lowest grades). The premise for the new method is that it is possible that an exam may have been *better-suited* to the students that had the highest scores (in addition to the fact that they may also have better met the learning objectives), thus a score-scaling method should scale scores in a nonlinear manner. At the same time, the scaling method should be such that the scaled grades maintain the same rank order as the raw scores, and two raw scores that have minimal difference should result in two scaled scores that also have minimal difference.

In the proposed method, the reduction in the denominator for an individual exam score is scaled according to the number of incorrect answers on the exam.

In the proposed method, the reduction in the denominator for an individual exam score is scaled according to the number of incorrect answers on the exam (i.e., a perfect score will not have a reduction in the denominator, scores with very few incorrect answers will see small reductions in the denominator, and scores at the bottom end will see the largest denominator reduction). Although the lowest raw scores receive the largest reduction in the denominator, they also have the fewest correct answers (in the numerator) and don't benefit as much from a denominator reduction, and so the largest increases in score (scaled minus raw) occurs for the scores near the middle. Described another way, the upper scores are helped less than the middle scores because the reduction in the denominator is less, which is justified by the premise that the exam may have been better suited for them, among other factors, while the lower scores are increased by less than the middle scores are increased, although they are helped by an even larger denominator reduction, because they didn't help themselves, so to speak, by having enough correct answers for the larger denominator reduction to make much difference. Grades are not bumped just to be bumped, but rather they are *scaled*, based on the number of correct and incorrect answers. The result is shown in Table 3 and has a shape similar to the square root and Maloy methods.

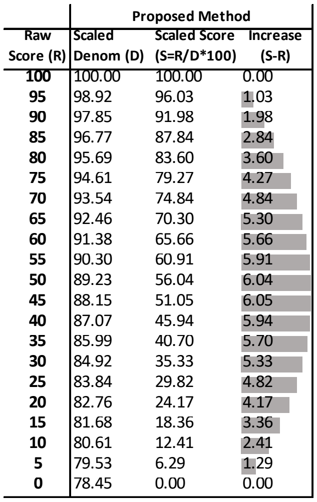
The formula is as follows, where S is a scaled score (out of 100), R is the raw score out of n total possible points, and n^* an adjusted total possible points that affects the magnitude of the scaling:

$$S = \frac{R}{n - (n - n^*)(\frac{(n-R)}{n})} \times 100 \qquad (4)$$

The denominator (which will simply equal n in the case of unscaled grades) is reduced from n by an amount (n-n^*) that is scaled by the proportion of questions that were incorrect ((n-R)/n). As described previously, although lower raw scores receive a larger denominator

Table 3

Scaled exam scores using the proposed method, which scales the amount that the denominator is reduced by a factor that incorporates the raw score



	Proposed Method		
Raw Score (R)	Scaled Denom (D)	Scaled Score (S=R/D*100)	Increase (S-R)
100	100.00	100.00	0.00
95	98.92	96.03	1.03
90	97.85	91.98	1.98
85	96.77	87.84	2.84
80	95.69	83.60	3.60
75	94.61	79.27	4.27
70	93.54	74.84	4.84
65	92.46	70.30	5.30
60	91.38	65.66	5.66
55	90.30	60.91	5.91
50	89.23	56.04	6.04
45	88.15	51.05	6.05
40	87.07	45.94	5.94
35	85.99	40.70	5.70
30	84.92	35.33	5.33
25	83.84	29.82	4.82
20	82.76	24.17	4.17
15	81.68	18.36	3.36
10	80.61	12.41	2.41
5	79.53	6.29	1.29
0	78.45	0.00	0.00

The square root method, Maloy's method, and the proposed method of the current paper are the only scaling methods that provide larger increases for scores toward the middle than for those at the high and low ends.

reduction, this won't always result in a larger adjustment to the total grade, since the scaled grade also depends on the number of correct answers (*R*, in the numerator). In other words, there must be incorrect answers for there to be a score adjustment, but there also must be correct answers to create mass for the denominator reduction to make a difference in the scaled score.

The value of *n** will be less than *n* (the maximum number of points available on the exam) and will not hold any intuitive meaning; it will be *approximately* double the difference between the raw score (R) and the scaled score (S) at the point where the difference between raw and scaled scores are largest. Lower *n** means that raw grades will receive a larger increase, and is set according to the desired scaling effect, which can be done a variety of ways, one being to specify the desired raw score (R) that should translate into a 50% scaled score, and solving the following equation (R is the raw score value out of *n* possible points or the raw percent grade where n would then be 100):

$$n^* = \frac{R \bullet n}{n - R} \quad (5)$$

For example, if n=100 and a 50% scaled score should be given to a raw score of R=45, then n*=81.82. A value for *n** can also be determined by using any numbers for *R* and *S* (in terms of a percentage out of 100) that give a desired *R*→*S* transformation and solving formula 6. Note that this method of calibrating *n** will result in a slightly different value than using equation 5, unless *S = 50* is used.

$$n^* = \frac{\frac{R}{S} \bullet 100 - R}{100 - R} \quad (6)$$

For example, using *R*=65% and *S*=70% gives a value of *n** of 79.59. (For comparison with the other scaling methods, the value for *n** was calibrated using the nonlinear solver in Excel so that the scaled mean of the mock dataset was 70%, which required using *n**=78.45). More advice for implementation is provided later in the paper.

Comparison and Discussion

As can be seen, the square root method, Maloy's method, and the proposed method of the current paper are the only scaling methods that provide larger increases for scores toward the middle than for those at the high and low ends. In looking at Maloy's method and the proposed method, the differences between the results are: (1) the magnitude of some of the adjustments (the largest score increases under the Maloy method are larger than those of the proposed method), and (2) which scores receive the most benefit for the mock dataset (scores in the 20-30% range for Maloy versus scores in the 40-50% range for the proposed method). These differences are shown graphically in Figure 1. Note that although the scaled scores of Maloy's method are quite a bit higher in some cases, the proposed method can achieve the same increase in the mean score (from 65% to 70%) with a smaller average increase in scaled versus raw scores since it has a slightly larger score increase for the higher frequency group (in other words, the Maloy method has significantly higher increases for scores at the lower end of the raw score scale, which in the mock data affected relatively few exams.

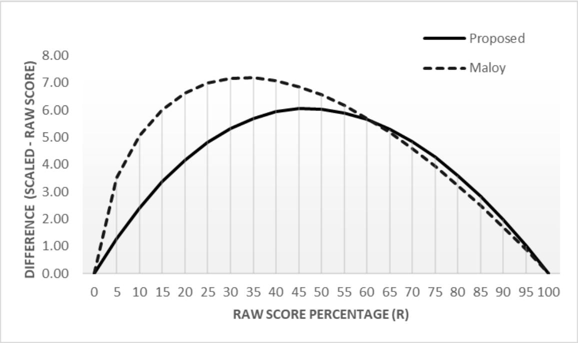
The proposed method is not being prescribed for all situations or claiming to be outright superior to all other methods.

Table 4
Comparison of scaled scores and increases under all scaling methods

	Constant		Reduce denom		Square Root		Maloy		Becker		Bailey		Proposed	
Raw Score (R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)	Scaled (S)	Increase (S-R)
100	105.00	5.00	107.69	7.69	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00
95	100.00	5.00	102.31	7.31	97.47	2.47	95.87	0.87	95.71	0.71	95.71	0.71	96.03	1.03
90	95.00	5.00	96.92	6.92	94.87	4.87	91.70	1.70	91.43	1.43	91.43	1.43	91.98	1.98
85	90.00	5.00	91.54	6.54	92.20	7.20	87.49	2.49	87.14	2.14	87.14	2.14	87.84	2.84
80	85.00	5.00	86.15	6.15	89.44	9.44	83.24	3.24	82.86	2.86	82.86	2.86	83.60	3.60
75	80.00	5.00	80.77	5.77	86.60	11.60	78.94	3.94	78.57	3.57	78.57	3.57	79.27	4.27
70	75.00	5.00	75.38	5.38	83.67	13.67	74.58	4.58	74.29	4.29	74.29	4.29	74.84	4.84
65	70.00	5.00	70.00	5.00	80.62	15.62	70.17	5.17	70.00	5.00	70.00	5.00	70.30	5.30
60	65.00	5.00	64.62	4.62	77.46	17.46	65.71	5.71	65.71	5.71	65.71	5.71	65.66	5.66
55	60.00	5.00	59.23	4.23	74.16	19.16	61.17	6.17	61.43	6.43	61.43	6.43	60.91	5.91
50	55.00	5.00	53.85	3.85	70.71	20.71	56.56	6.56	57.14	7.14	57.14	7.14	56.04	6.04
45	50.00	5.00	48.46	3.46	67.08	22.08	51.87	6.87	52.86	7.86	52.86	7.86	51.05	6.05
40	45.00	5.00	43.08	3.08	63.25	23.25	47.08	7.08	48.57	8.57	48.57	8.57	45.94	5.94
35	40.00	5.00	37.69	2.69	59.16	24.16	42.18	7.18	44.29	9.29	44.29	9.29	40.70	5.70
30	35.00	5.00	32.31	2.31	54.77	24.77	37.16	7.16	40.00	10.00	40.00	10.00	35.33	5.33
25	30.00	5.00	26.92	1.92	50.00	25.00	31.99	6.99	35.71	10.71	35.71	10.71	29.82	4.82
20	25.00	5.00	21.54	1.54	44.72	24.72	26.63	6.63	31.43	11.43	31.43	11.43	24.17	4.17
15	20.00	5.00	16.15	1.15	38.73	23.73	21.02	6.02	27.14	12.14	27.14	12.14	18.36	3.36
10	15.00	5.00	10.77	0.77	31.62	21.62	15.06	5.06	22.86	12.86	22.86	12.86	12.41	2.41
5	10.00	5.00	5.38	0.38	22.36	17.36	8.52	3.52	18.57	13.57	18.57	13.57	6.29	1.29
0	5.00	5.00	0.00	0.00	0.00	0.00	0.00	0.00	14.29	14.29	14.29	14.29	0.00	0.00

proposed method), and (2) which scores receive the most benefit for the mock dataset (scores in the 20-30% range for Maloy versus scores in the 40-50% range for the proposed method). These differences are shown graphically in Figure 1. Note that although the scaled scores of Maloy's method are quite a bit higher in some cases, the proposed method can achieve the same increase in the mean score (from 65% to 70%) with a smaller average increase in scaled versus raw scores since it has a slightly larger score increase for the higher frequency group (in other words, the Maloy method has significantly higher increases for scores at the lower end of the raw score scale, which in the mock data affected relatively few exams.

Figure 1
Increase in scaled score from raw score (S-R), for given raw score values, under the proposed method and the Maloy method



Implementation

The proposed method is not being prescribed for all situations or claiming to be outright superior to all other methods, but rather is being offered as an alternative that avoids some of the shortcomings of other methods, for consideration for use by an instructor based on their own reasons for adjusting exam scores.

Implementation requires determination of n^*. Presumably the instructor has analysed the exam results to conclude that grade-scaling is warranted in the first place. The instructor could then either decide on the raw score that should translate into a scaled score at the passing threshold (50%) and use equation 5, or determine R (raw score) and S (scaled score) values for a desired $R \rightarrow S$ transformation and use equation 6. Once the value of n^* has been determined, each grade can be individually converted using equation 4 and a calculator or spreadsheet (it may even be possible to set up a conversion formula in the LMS).

The practice of calibrating the scaling parameter (n^*) based on a raw score that is judged by the grader to be that which should convert to a passing (50%) scaled score has some similarities, albeit rudimental, to practices for *standard-setting* using cut-off points as described by Ben-David (2000) in a credentialling application. These methods, such as the Anghoff method which uses a panel of judges, consider the totality of the individual exam questions and identify a raw-score that would represent a passing grade for a "minimally competent examinee" (p. 122). Kibble (2017) also discusses best practices for standard-setting but acknowledges that these methods will not be practical in all situations (p. 117), and thus the relevant consideration for the current paper is that anchoring the scaling process with a thoughtful consideration of what raw score a student on the borderline of passing would achieve on the test (i.e., which becomes the scaled passing grade of 50%) is one straight-forward way to calibrate the proposed scaling method.

A final important aspect of implementation is transparency with students – a document or spreadsheet with the scaling formula can be posted for the students to see (and perhaps convert their own raw score into the scaled score), and/or the method can be explained at the time of review of the exam results in class. Student perceptions about the fairness (or other criteria) of this scaling method and others certainly seem a worthy area for further research.

Conclusion And Limitations

This paper began by discussing a variety of factors that can render an exam to be an imperfect measure of student ability, including some related to the construction of individual items (questions), some related to overall exam composition (including the distribution of question difficulty, and the coverage of learning objectives and cognitive skill levels), and some related to student interpretation. Although the current work does not suggest that efforts shouldn't be undertaken to improve exams and teaching, where applicable, instructors may nonetheless deem it necessary to adjust exam grades to account for the aforementioned inherent exam deficiencies and/or at the judgement of the individual instructor. It is hoped that any adjustments would be carried out in a careful and informed manner; however, some grade-adjustment methods may have drawbacks which may or may not be known to the user - for example, outright reducing the denominator of the exam by the same amount for all students will benefit higher grades more than lower grades, and can result in some grades exceeding 100%. One objective of the current paper is thus to collect and review documented processes for adjusting exam scores. Limitations include that only the methods available in the literature were reviewed (and this topic is sparsely studied), and that methods for *curving* grades according to a pre-defined distribution (e.g., the normal distribution) were not included, for reasons described in a previous section.

Based on the discussion of the underlying reasons that exams may be imperfect, which could affect some students more than others, as well as perceived deficiencies in grade-adjustment methods in the literature, the proposed method uses a reduction in the denominator that is scaled by the proportion of incorrect questions. This results in very high grades receiving small adjustments, due to having few incorrect questions and therefore minimal reduction to the denominator, very low grades receiving small adjustments despite having the largest reduction in the denominator, due to having few correct answers (causing a low numerator value), and grades towards the middle of the distribution receiving the largest adjustments, due to having a comparatively moderate reduction in the denominator as well as a large enough numerator for it to make a difference.

The proposed method is not being presented as a panacea; whether and how to use it is up to individual instructors, of course.

Limitations to this method include that it is based on the notion that high exam grades do not merit as much grade adjustment as grades towards the middle of the distribution, which reflects the intuition and experiences of the author, with implicit support by way of a discussion of reasons for exam deficiencies in the background section, but this basis may not align with the intuition and experiences of some examiners or examinees. Further, the degree of the correlation between raw grades and deserved adjustments is admittedly inexact (which is the nature of the beast in using examinations to measure student ability). However, the proposed method is not being presented as a panacea; whether and how to use it is up to individual instructors, of course.

References

Abdellatif, H., Alsemeh, A. E., Khamis, T., & Boulassel, M. R. (2024). Exam blueprinting as a tool to overcome principal validity threats: A scoping review. *Educación Médica*, *25*(3). https://doi.org/10.1016/j.edumed.2024.100906

Adhi, M. I., & Aly, S. M. (2018). Student perception and post-exam analysis of one best MCQ and one correct MCQs: A comparative study. *The Journal of the Pakistan Medical Association*, *68*(4), 570-575.

Anderson, L.W. (2018). A critique of grading: Policies, practices, and technical matters. *Education Policy Analysis Archives*. *26*(49), 1-27. https://doi.org/10.14507/epaa.26.3814

Bailey, L. C. (1992). Grade-scaling: A simplified approach. *Journal of Chemical Education*, *69*(3), 221. https://pubs.acs.org/doi/pdf/10.1021/ed069p221

Becker, C. E. (1991). Comparison of two methods for scoring examinations. *Journal of Chemical Education*, *68*(4), 309. https://pubs.acs.org/doi/pdf/10.1021/ed068p309

Ben-David, M. F. (2000). AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher*, *22*(2), 120-130. https://doi.org/10.1080/01421590078526

Camenares, D. (2022). Better Remedies For Bad Exams: correcting for difficult questions in a fair and systematic way. *International Journal for the Scholarship of Teaching and Learning*, *16*(3), 1-4. https://digitalcommons.georgiasouthern.edu/ij-sotl/vol16/iss3/4/

Close, D. (2009). Fair grades. *Teaching Philosophy*, *32*(4), 361-398. https://www.pdcnet.org/8525737F00588478/file/B08274525466752585257680005 5CC87/$FILE/teachphil_2009_0032_0004_0035_0072.pdf

Crawford, K., & Fekete, A. (1997, July). What do exam results really measure?. *In Proceedings of the 2nd Australasian conference on Computer Science Education* (pp. 185-190). https://dl.acm.org/doi/pdf/10.1145/299359.299386

Downing, S.M. (2009). Statistics of Testing. In S.M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education*, (pp. 107-109). New York: Taylor and Francis. https://doi.org/10.4324/9780203880135

Eweda, G., Bukhary, Z. A., & Hamed, O. (2020). Quality assurance of test blueprinting. *Journal of Professional Nursing*, *36*(3), 166-170. https://doi.org/10.1016/j.profnurs.2019.09.001

Fencl, H. S. (2019). Point of view: Focusing on learning as a marker of success for underrepresented students. *Journal of College Science Teaching*, *48*(5), 6-7. https://doi.org/10.1080/0047231X.2019.12290468

Grant, A. (2016, September 10). Why we should stop grading students on a curve. *The New York Times*. https://www.nytimes.com/2016/09/11/opinion/sunday/why-we-should-stop-grading-students-on-a-curve.html

Holmes, J. D. (2021). The bad test-taker identity. *Teaching of Psychology*, *48*(4), 293-299. https://doi.org/10.1177/0098628320979884

Jephcote, C., Medland, E., & Lygo-Baker, S. (2021). Grade inflation versus grade improvement: Are our students getting more intelligent? *Assessment & Evaluation in Higher Education*, *46*(4), 547-571. https://doi.org/10.1080/02602938.2020.1795617

Kibble, J. D. (2017). Best practices in summative assessment. *Advances in Physiology Education*, *41*(1), 110-119. https://doi.org/10.1152/advan.00116.2016

Kulick, G., & Wright, R. (2008). The impact of grading on the curve: A simulation analysis. *International Journal for the Scholarship of Teaching and Learning*, 2(2), n2. https://eric.ed.gov/?id=EJ1146678

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, 37(8), 919-931. https://doi.org/10.1080/02602938.2011.586991

Maloy, J. T. (1990). A useful grade-scaling equation. *Journal of Chemical Education*, *67*(5), 414. https://pubs.acs.org/doi/pdf/10.1021/ed067p414

Nelson, J.E., Varma-Nelson, P., & Kloempken, T.A. (1992). A graphic grade-scaling method. *Journal of Chemical Education*, *69*(6), 462. https://pubs.acs.org/doi/pdf/10.1021/ed069p462

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*(6), 778-803. https://doi.org/10.1002/tea.21026

Omar, N., Haris, S. S., Hassan, R., Arshad, H., Rahmat, M., Zainal, N. F. A., & Zulkifli, R. (2012). Automated analysis of exam questions according to Bloom's taxonomy. *Procedia-Social and Behavioral Sciences*, *59*, 297-303. https://doi.org/10.1016/j.sbspro.2012.09.278

Page, R. B., Espinosa, J., Mares, C. A., Del Pilar, J., & Shelton, G. R. (2018). The curvy road to student success in underserved populations. *Journal of College Science Teaching*, *47*(5), 6-7. https://eric.ed.gov/?id=EJ1178262

Rudolph, M. J., Daugherty, K. K., Ray, M. E., Shuford, V. P., Lebovitz, L., & DiVall, M. V. (2019). Best practices related to examination item construction and post-hoc review. *American Journal of Pharmaceutical Education*, *83*(7), 7204. https://doi.org/10.5688/ajpe7204

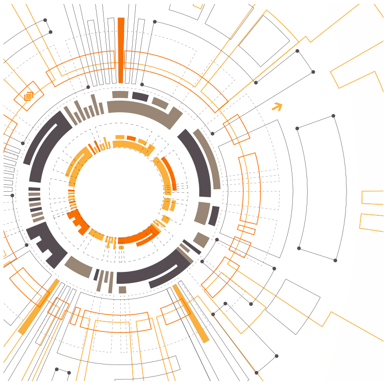
Stevens, S. P., Palocsay, S. W., & Novoa, L. J. (2022). Practical guidance for writing multiple-choice test questions in introductory analytics courses. *INFORMS Transactions on Education*, *24*(1), 51-69. https://doi.org/10.1287/ited.2022.0274

Tan Yuen Ling, L., Yuen, B., Loo, W. L., Prinsloo, C., & Gan, M. (2020). Students' conceptions of bell curve grading fairness in relation to goal orientation and motivation. *International Journal for the Scholarship of Teaching and Learning*, *14*(1), 7. https://digitalcommons.georgiasouthern.edu/ij-sotl/vol14/iss1/7/

Thelk, A. (2008). Detecting items that function differently for two-and four-year college students. *Research & Practice in Assessment*, *3*, 23-27. https://eric.ed.gov/?id=EJ1062685

Wellberg, S. (2023). Teacher-made tests: Why they matter and a framework for analysing mathematics exams. *Assessment in Education: Principles, Policy & Practice*, *30*(1), 53-75. https://doi.org/10.1080/0969594X.2023.2189565

Welch, A. C., Karpen, S. C., Cross, L. B., & LeBlanc, B. N. (2017). A multidisciplinary assessment of faculty accuracy and reliability with Bloom's Taxonomy. *Research & Practice in Assessment*, *12*, 96-105. https://eric.ed.gov/?id=EJ1168817



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

Stuart A. Miller, M.A.
Auburn University

Kendall M. McGoey, Ph.D.
Auburn University

Abstract

Priming incoming and second-year college students with questions about effort prior to completing low-stakes assessments has increased test-taking effort. We extended this research by randomly assigning college seniors to one of three priming conditions prior to completing low-stakes assessments: answering three questions about intended effort that infused positive self-identity, answering three questions about intended effort that incorporated the university's creed, or answering no priming questions. The self-identity questions resulted in higher self-reported effort than the control condition, the university creed questions resulted in higher testing time than the control condition, and neither priming condition increased test performance. However, Pell Grant eligibility moderated the priming effect on effort. Priming resulted in self-reported effort for Pell eligible students being the same or higher than noneligible students. Likewise, ethnicity moderated the priming effect on test scores. White students scored higher than underrepresented students in the control condition, but this difference disappeared with priming.

Increasing Expended Effort on Low-Stakes Accountability Tests via Priming: Effectiveness with Graduating University Students

Test scores reported for accountability mandates are often gathered in low-stakes contexts (Cole & Osterlind, 2008; Mathers et al., 2018; Roohr et al., 2016; Smith & Smith, 2004; Wise & DeMars, 2010). For example, in higher education, outcomes assessments typically yield group-level data (incoming student achievement versus senior-level student achievement; students who have versus have not experienced educational programming). Accreditors use this group-level data to evaluate institutional impact on student learning (Liu, 2017). Federal funding is contingent on accreditation (Council for Higher Education Accreditation, 2022). Yet, performance on these assessments may have no personal consequences for students completing them.

Student Effort on Low-Stakes Tests

The validity of the interpretation of outcomes assessment data collected in high- versus low-stakes contexts may not be equal because of the discrepancy in expended effort (Simzar et al., 2015; Sundre & Kitsantas, 2004; Sungur, 2007; Wise & Smith, 2016). Students in a high-stakes testing context (e.g., certification testing) have something personal to gain or lose as a result of their performance. Thus, students tend to put forth enough effort to demonstrate their ability. In contrast, students assessed in a low-stakes context with no personal consequence of test performance may not expend the effort necessary to display their ability. Given the difference in expended effort across high- and low-stakes assessment contexts, a potential simple solution is to make institutional accountability tests high-stakes

CORRESPONDENCE
Email
finneysj@jmu.edu

in nature. However, institutional accountability assessments are often not tied to a single course but rather to a set of courses or experiences. In turn, performance on these assessments cannot inform course grades. Moreover, assessments used for institutional accountability or improvement may not have the psychometric properties necessary for individual-level high-stakes decisions, such as graduation. Thus, many institutional accountability assessments must retain their low-stakes nature to students.

In turn, validity concerns associated with expended effort must be acknowledged when reporting and interpreting test scores gathered in low-stakes contexts (Wise, 2020). A meta-analysis found a positive relation between students' expended effort and test scores in both K-12 and higher education contexts (Silm et al., 2020). The average correlation between self-reported expended effort and test performance was r = 0.33 and the average correlation between response time (a behavioral measure of expended effort) and performance was r = 0.72, indicating these two measures are representing different forms of effort. Further, the educational level of the student was a significant moderator of the effect sizes, with the average relation between effort and test performance being stronger for university students than for K-12 students.

Students assessed in a low-stakes context with no personal consequence of test performance may not expend the effort necessary to display their ability.

Moreover, on average, effort is generally lower for students later in their educational experience. This difference in effort was found in K-12 (5th graders compared to 8th graders) and college contexts (incoming college students compared to second-year college students) when expended effort was measured using self-report measures and response time (Rios & Guo, 2020; Soland, 2018; Thelk et al., 2009). In turn, there is greater need to increase effort for more advanced students. If more advanced college students expend less effort than incoming students on accountability tests, value-added estimates will be biased downward (Finney et al., 2016).

Given concern about the accuracy of inferences from low-stakes assessments, the purpose of our study was to examine the effectiveness of a short priming intervention to increase test-taking effort from graduating university students. Our study built upon previous studies of this intervention by examining the effectiveness of the intervention with senior-level students and adapting the intervention to prime students' connection with their institution. Before detailing our results, we review studies of preemptive strategies to combat low test-taking effort.

Strategies to Increase Effort in Low-Stakes Testing Contexts

Four strategies have been examined to increase students' effort and improve test score accuracy (Wise & DeMars, 2005): (a) offering external incentives; (b) increasing test relevance; (c) modifying assessment design; and (d) promising feedback on performance. A meta-analysis evaluated the effectiveness of these strategies (Rios, 2021). Offering feedback (d = -0.01) or modifying the assessment design (e.g., remove mentally-taxing items; d = 0.13) had smaller impact on expended effort compared to offering external incentives (e.g., money; d = 0.36) and increasing the relevance of assessments for students (d = 0.21).

There are practical issues with providing external incentives and increasing test relevance. The financial burden of incentives may be prohibitive. Moreover, incentives for performance may be perceived as inappropriate by stakeholders (e.g., board of visitors, faculty; Wise & DeMars, 2005). With respect to increasing the test relevance (students' subjective value of assessments), modifying instructions to highlight the importance of results for institutional reputation and personal benefit increased students' effort (Hawthorne et al., 2015; Liu et al., 2015). However, modifying instructions may be prohibited, especially if messaging is not truthful (e.g., deceptively telling students that employers see test scores; Finney et al., 2018).

Because we could not justify the cost of external incentives or increase test relevance, we turned to a newly examined strategy to increase effort: the question-behavior effect (QBE). Asking people questions about their behavior toward a target action (e.g., volunteering) increases their likelihood of performing the behavior (Levav & Fitzsimons, 2006; Wilding et al., 2016). The QBE has been supported for behaviors such as increasing voting, helping, and exercising (Spangenberg et al., 2016; Wood et al., 2016).

In the initial study examining the QBE in a low-stakes assessment context (Finney & McFadden, 2023), incoming first-year college students completing low-stakes institutional accountability tests were randomly assigned to one of three question conditions: no questions (No Question Condition); answering five intended effort questions (e.g., "I will engage in good effort throughout the test"; Intended Effort Condition); and answering five intended effort questions with reference to a positive self-identity (e.g., "As a conscientious test-taker, I will engage in good effort throughout the test"; Self-Identity Condition). Students then completed two unproctored cognitive assessments (information literacy test and oral communication test). Finney and McFadden (2023) hypothesized they would find higher effort for both question conditions than the No Question Condition and that the Self-Identity Condition would elicit higher effort than the Intended Effort Condition. Moreover, based on previous research (Rios et al., 2014; Soland, 2018), they hypothesized that approximately 15% of incoming college students in the No Question Condition would be flagged to be filtered from the dataset due to low effort and a lower percentage would be filtered in both question conditions. Filtering non-effortful responses (i.e., motivation filtering) is a practical approach to produce more valid interpretations of assessment data. That is, low effort results in construct-irrelevant variance in test scores; scores do not only reflect ability but also motivation. Construct-irrelevant variance can be addressed by motivation-filtering. Two criteria have been used to filter scores provided by students who expend low effort: 1) self-identifying being unmotivated via self-report measures (Rios et al., 2014; Swerdzewski et al., 2011); and 2) rapidly responding to items as measured by response time (Wise & Kong, 2005). Finney and McFadden (2023) used both methods.

As expected, students primed with either set of questions exhibited higher self-reported effort, lower proportions of rapid responding to items, and a lower percentage of data filtered from the dataset due to low effort. When conducting motivation filtering based on self-reported effort, 15.3% of students in the No Question Condition were filtered from the dataset due to low effort. The percentage filtered was reduced to 11.8% and 10.3% for students in the Intended Effort and Self-Identity Conditions, respectively. These findings were replicated using response time, where 15.6% of incoming students in the No Question Condition were filtered due to low response time, but only 11.9% were removed in the question conditions. Because including a positive self-identity increased the QBE in the context of voting behavior (Bryan et al., 2011), Finney and McFadden (2023) hypothesized that the Self-Identity Condition (e.g., "conscientious test-taker") would result in greater expended effort in a testing context. Contrary to their hypothesis, there was no significant difference in effort across the two question conditions.

To extend the study of the QBE in a testing context, Finney et al. (2024) examined the QBE with second-year college students and explored if gender moderated the effect. These more advanced college students were randomly assigned to the same three conditions, then two cognitive assessments were administered, and self-reported effort and response time were collected. There was no effect of QBE condition for students identifying as male for either measure of effort. Recall, when examining incoming college students, Finney and McFadden (2023) found no difference in self-reported effort or response-time effort when questions included or excluded positive self-identity. Both question conditions prompted more effort than the No Question Condition for incoming students. Likewise, for second-year female students, Finney et al. (2024) found no difference in self-reported effort when questions included or excluded positive self-identity and both question conditions resulted in more effort than the No Question Condition. However, response times for female students were significantly higher in the Self-Identity Condition compared to the No Question Condition, with no difference in response time between the No Question Condition and the Intended Effort Condition. The moderating effect of gender and inclusion of positive self-identity wording needs further study.

Because we could not justify the cost of external incentives or increase test relevance, we turned to a newly examined strategy to increase effort: the question-behavior effect (QBE).

In a third study of the QBE intervention with incoming college students, McFadden and Finney (2025) examined whether administering a second "dose" of questions could combat the decrease in examinee effort later in a testing session. In the previous studies (Finney & McFadden, 2023; Finney et al., 2024), five priming questions were administered at the start of the testing session but never again once testing was underway. McFadden and Finney (2025) randomly assigned incoming students to one of three question conditions prior to completing two low-stakes tests: answering three questions about intended effort directly before the first test in a session; answering three questions about intended effort directly before each test in a session; and answering no priming questions. Administering a second dose of questions directly before the second test in a session significantly increased response-time effort and self-reported effort for the more difficult test. Moreover, the effects were found when reducing the priming questions from five to three questions. McFadden and Finney (2025) did not examine the utility of questions with self-identity wording or possible differential effects across gender.

In a fourth study of the QBE with first-year college students, Finney and Pastor (2025) examined if initially non-compliant students (those who completed testing after the testing deadline) would respond to the priming intervention to the same extent as compliant students (those who tested on time). Moreover, they examined if the priming effect would be similar when reducing the questions from three to one question. They randomly assigned first-year students to one of five priming conditions prior to completing a low-stakes test: answering one or three questions about intended effort, answering one or three questions about intended effort that infused positive self-identity, or answering no priming questions. Priming conditions were crossed by testing compliance status (students who tested on time versus late). Compliance status did not moderate the priming effect for self-reported effort; no questions resulted in significantly and practically lower self-reported effort than both three-question conditions. Compliance status moderated the priming effect for response time effort, with the three self-identity questions being effective for both compliant and non-compliant students. Thus, they demonstrated that priming with three questions is a quick and effective strategy to increase test-taking effort for first-year students, including those not initially compliant with testing requests.

Purpose of the Current Study

The current study extends the previous study of the QBE intervention in six important ways, aligned with the following six research questions.

1) Does the QBE emerge for effort and test scores with graduating senior students?

The QBE has been examined for first-year college students (Finney & McFadden, 2023; Finney & Pastor, 2025; McFadden & Finney, 2025) and second-year college students (Finney et al., 2024). We examined if the QBE intervention would be effective with university students who were completing assessments two to four months prior to college graduation. The graduating seniors were not exposed to the QBE intervention in prior low-stakes testing contexts. Given the limited study of the differential effectiveness of test-taking effort interventions across academic year of the student (O'Neil et al., 1995), we explored (rather than hypothesized) the effectiveness of the QBE intervention for these senior college students. It was unclear 1) if the intervention would have similar effectiveness as was found for incoming college students, 2) if the effectiveness would be differential across gender as was found with second-year students, or 3) if the effectiveness would be nil for students at this late point in college. Moreover, we further examined the utility of three rather than five priming questions.

2) Do differences in the self-identities primed in the questions impact the size of the QBE?

Similar to Finney and McFadden (2023), we framed priming questions in terms of general positive self-identities (e.g., "motivated student") and examined if this wording resulted in higher effort than no questions. However, we also created what may be more relevant self-identity priming questions. This study is the first to examine self-identity prompts that reflected a university's creed, which may be more pertinent to college students. Specifically, we explored whether there was a difference in effort if the self-identity primed in questions was generally positive (e.g., "conscientious test-taker) or specific to the positive characteristics of students noted in the university creed (e.g., "as someone who believes in hard work").

Administering a second dose of questions directly before the second test in a session significantly increased response-time effort and self-reported effort for the more difficult test.

3) Is the QBE moderated by student characteristics?

Finney et al. (2024) found the QBE for students identifying as female, but not male. Differential QBE associated with gender identity needs further examination. Moreover, other student characteristics (ethnicity, transfer status, first-generation status, Pell Grant eligibility) may moderate the QBE. Thus, we explored differential effects for different student sub-populations.

4) Could an increase in student motivation through priming decrease the amount of data removed, thus providing a cost benefit to an institution having to outsource its testing needs?

If priming students increases effort, a lesser amount of invalid data due to disengagement would need to be removed from the dataset. Given the commercial outcome measures administered at the university (i.e., at cost per student), any data removed due to low effort is wasteful. Priming students to give good effort may result in a cost benefit by increasing the amount of useable data.

5) Are results consistent when effort is operationalized using time and self-reported effort?

Finney and McFadden (2023) found the QBE to be similar when effort was operationalized using response-time effort and self-reported effort. Response-time effort was computed using the number of items for which the student did not rapidly respond. Finney et al. (2024) used the total time spent completing a test. They found different QBE results when operationalizing effort as total test time versus self-reported effort. For total test time from female students, the Self-Identity Condition resulted in more effort than the No Question Condition, whereas for self-reported effort, both the Self-Identity Condition and the Intended Effort Condition resulted in more effort than the No Question Condition. We employed total test time to operationalize effort, in addition to self-reported effort. Given that time-based measures of effort and self-reported effort have a low correlation and different nomological nets (Akhtar & Firdiyanti, 2023), we were prepared that the QBE may emerge for one measure of effort but not the other. We examined both effort measures since assessment practitioners may only be able to gather one measure.

6) Are test scores improved due to priming effort?

If priming enhances effort on the test, then it could result in increased test performance. Previous studies of the QBE (e.g., Finney & McFadden, 2023; Finney & Pastor, 2025; McFadden & Finney, 2025) examined its effects when gathering outcomes assessment data at pre-test or baseline. These incoming students had not received instruction in the test domain. Thus, even though effort was increased, there was no expectation that test scores would be increased. Increased effort would not translate to increased test scores if students do not have ability in the domain (Rios, 2021). However, for the current study of senior students, it was of great interest to evaluate if performance levels were impacted by priming and if this impact on test scores was differential across student characteristics (e.g., Pell Grant eligibility).

Priming questions that reflected a university's creed may be more pertinent to college students.

Methods

Procedures

In the current study, graduating seniors from a large R1 public southern university were enrolled in a zero-credit graduation course and asked to complete an assessment that would evaluate an institutional-level learning outcome focused on students' intercultural competence. To evaluate this learning outcome, the University used the HEIghten Intercultural Competency and Diversity (ICD) assessment, created by ETS© and administered by Territorium.

All graduating students enroll in a common graduation course where they must complete four assignments to receive their diploma after degrees are confirmed. Once the course is live for student participation, a hold is placed on the student's account until all four items are complete. The hold only impacts diploma delivery and does not impact a student's ability to graduate from the University. The hold does not impact financial aid disbursement, nor does it impact a student's ability to add or drop courses.

At the beginning of each semester, each enrolled student in the graduation course receives an e-mail informing them of said enrollment, the hold that has been applied to their student account, and the steps to take to have the hold removed by the end of the semester. All students are encouraged by the notification e-mail to navigate to the Canvas LMS platform where the graduation course assignments are located and to complete each assignment within the allotted time frame. The first assignment a student must complete within the course is a one-hour assessment, which is a part of the University's efforts to continuously improve its general education curriculum. Each year the University measures student learning associated with one of nine learning outcomes that are directly aligned to the core curriculum.

Although all students complete the assessment for graduation purposes, they had the choice of releasing their scores for the research study. That is, after the assessment was complete, students were provided a research statement with information about the study and IRB approval of the study. Once students confirmed that they read the research statement, they were then provided with the opportunity to opt in or out of the study.

The graduation course launched on January 11, 2023, and students were asked to complete this one-hour assessment by March 16, 2023. Beyond directions on how to access the assessment, students were only encouraged to "give their best effort" on the assessment.

Participants

A total of 3,457 graduating seniors were enrolled in the Spring 2023 graduation course. Of those enrolled, 3,413 were marked as either completing the ICD assessment or completing a waiver for a prior semester test completion. For the purposes of this study, 3,311 students completed the ICD assessment. Detailed instructions within the Canvas course directed students to this assessment based on the last digit of a student's ID number. Students were then randomly assigned to one of the three priming conditions: 1) Control Condition with no priming questions before completing the assessment; 2) Positive Self-Identity Condition with priming questions created by Finney and McFadden (2023) answered before completing the assessment; and 3) University Creed Condition with priming questions infused with the university's creed language answered before the assessment (see Appendix for priming questions).

Of importance, although 3,311 students completed the assessment, only students who gave consent to use their data for this study were analyzed. Therefore, after removing students who (1) did not give consent for the study, (2) had duplicate or missing data, or (3) took over the allotted one-hour time provided via the Territorium site, the final sample size was N = 2,204 (Control Condition = 656; Positive Self-Identity Condition = 852; University Creed Condition = 696). This sample of 2,204 students was used to evaluate the impact of priming on self-reported effort, response time on the test, and test performance.

The 2,204 students who completed the ICD assessment and provided consent represented the demographics of the University. The majority of students self-identified as female (54.6%) and White (85.9%). Other ethnicities included Black or African American (2.2%), American Indian or Alaska Native (0.2%), Asian (2.2%), Hispanics of any race (3.3%), Native Hawaiian or Other Pacific Islander (0.04%), Two or More Races (2.2%), Nonresident Alien (3.8%), and Race and Ethnicity Unknown (0.1%). The sample was predominately non-transfer students (84.4%), non-first-generation/continuing students (88.0%), and non-Pell eligible students (88.9%). Importantly, because student characteristic information was gathered through the University's system and not reported at the time of the assessment, not all 2,204 students had data for every demographic characteristic. Thus, our analyses that evaluate the following moderators are based on different sample sizes: gender (n = 2,135), ethnicity (n = 2,052), transfer status (n = 2,074), first-generation status (n = 2,004), and Pell eligibility (n = 2,135).

All graduating students enroll in a common graduation course where they must complete four assignments to receive their diploma after degrees are confirmed.

Measures

Analyze & Act Test Performance

The HEIghten Intercultural Competency and Diversity (ICD) assessment is a 74-item measure. The ICD assessment produces two scale scores: Analyze & Act scores, which range from 150 to 180 based on 40 cognitive (right or wrong) items and Approach scores, which range from 90 to 150 based on 34 noncognitive (Likert-type) items. The Approach scores reflect students' view of themselves, which is dispositional in nature and not used as a measure of learning at the university. For the Analyze & Act dimension, there are three proficiency levels: "Developing" (scores between 150 to 157), "Proficient" (scores between 158 to 174), and "Advanced" (scores between 175 to 180). Additionally, six Analyze & Act subscale scores can be computed: Self-Awareness, Cultural Knowledge Application, Suspending Judgment/Perspective Taking, Social Monitoring, Emotion Regulation, and Behavior Regulation. Subscale scores range from 1 to 10, with higher scores representing higher ability.

We only employed the total Analyze & Act score for our study. Specifically, students' average Analyze & Act performance serves as one of the three measures (in addition to self-report effort and response time) used to assess the impact of the QBE. This performance score was calculated and provided by Territorium.

Students with scores ≤ 15 on the Effort subscale were deemed unmotivated and identified as responses to be filtered in the dataset.

Self-Reported Effort

At the end of the ICD assessment, the Student Opinion Survey (SOS; Pastor et al., 2023; Thelk et al., 2009), a 10-item measure consisting of two subscales (Effort and Importance) was completed. The Effort subscale measures test-taker's reported effort put forth on a test ("I engaged in good effort throughout this test"), whereas the Importance subscale measures the degree to which students perceived the test to be important (e.g., "Doing well on this test is important to me"). Students responded to the SOS items using a 5-point Likert-type scale (1= Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly Agree).

We assessed the impact of the three priming conditions on the Effort subscale scores themselves and when the effort scores were used for motivation filtering. Based on Swerdzewski et al.'s (2011) suggestion, students with scores ≤ 15 on the Effort subscale were deemed unmotivated and identified as responses to be filtered in the dataset.

Response Time

Territorium provided students with one hour to complete the ICD assessment. Each students' time spent on the assessment was converted to minutes for the purpose of evaluating the effectiveness of the priming questions.

Results

Prior to conducting analyses, the three dependent variables (self-reported effort, response time, and test performance) were assessed for normality. All three variables were approximately normally distributed with skewness and kurtosis less than |2|: effort displayed -0.99 skew and 0.90 kurtosis; response time displayed -0.82 skew and 0.99 kurtosis; and test performance displayed -1.33 skew and 1.64 kurtosis.

Effect of Priming Condition across the Three Outcomes

To examine the effect of priming condition on the three outcomes of interest (self-reported effort, response time, and test performance), we conducted separate one-way ANOVAs for each outcome. Levene's test of equality of variance across groups was not significant for self-reported effort (p =.61), response time (p =.22) or test performance (p = .53), indicating the assumption of homogeneity of variance was not violated. As expected, there were significant effects of priming on self-reported effort [F (2, 2201) = 5.0, p = .007)] and response time [F (2, 2201) = 3.6, p = .027)]. For self-reported effort, the Positive Self-identity priming questions resulted in higher effort than the No Question condition (d = 0.16). For response time, the University Creed priming questions resulted in students spending more time on the test than the No Question conditions (d = 0.14). However, the effect on test performance was not significant [F (2, 2201) = 2.5, p = .085). Table 1 provides descriptive statistics and effect size information.

Group Differences by QBE Condition and Student Characteristics

To be diligent, we conducted several factorial ANOVAs to evaluate if priming condition was moderated by any student characteristic for each outcome of interest. In other words, we assessed the significance of the interaction between several student characteristics (gender, ethnicity, transfer status, first-generation status, and Pell eligibility) and the priming condition with respect to self-reported effort, response time, and performance.

First, gender, transfer status, and first-generation status did not interact with priming condition for any of the three outcome variables. Thus, general statements about the impact of priming on these outcomes can be made across these student groups.

Table 1
Descriptive statistics across priming conditions for the three outcomes of interest

		Self-reported Effort				Response Time			Test Performance	
Condition	*N*	Mean	SD	*d*	% to be filtered	Mean	SD	*d*	Mean	SD
No Questions	656	4.30_a	.62	-	5.34%	38.35_a	7.44	-	172.24_a	6.22
Positive Self-Identity	852	4.40_b	.60	.16	3.87%	38.68_{ab}	7.83	.04	171.95_a	6.24
University Creed	696	4.35_{ab}	.58	.08	3.30%	39.45_b	8.05	.14	171.49_a	6.56

Note. Effort scores can range from 1 to 5, with higher scores reflecting higher levels of self-reported effort. Response time was reported in minutes. Within a column, means with different subscripts are statistically significantly different. *d* = Cohen's *d* effect size when comparing the No Questions condition to each question condition. Self-reported effort "% to be filtered" is the percentage of students in that condition whose scores were flagged for removal due to low self-reported effort (at or below a summed score of 15 across five effort items as suggested by Swerdewski et al. (2011)). Mean test performance in each condition is based on all students without filtering low-effort students.

However, when examining self-reported effort, Pell eligibility interacted with priming condition. Prior to estimating and interpreting this effect, the assumption of homogeneity of variance was assessed and supported (Levene's test, p = .20). The interaction indicated that the question-behavior effect was stronger for students who were Pell eligible [interaction effect: F (2, 2129) = 5.90, p = .003]. When we did not prime students, Pell eligible students put forth significantly less effort (4.15) than those who were not Pell eligible (4.33). However, when we primed students with University Creed questions, the Pell eligible students put forth an equal (not significantly different) amount of effort (4.38) as students who were not Pell eligible (4.35). Notably, when we primed students with positive self-identity questions, the Pell eligible students put forth significantly more effort (4.54) than students who were not Pell eligible (4.39). As shown in Table 2, it appears that we have the opportunity to enhance expended effort from Pell eligible students via priming questions.

When we primed students with positive self-identity questions, the Pell eligible students put forth significantly more effort (4.54) than students who were not Pell eligible (4.39).

Table 2
Interaction of Pell Grant eligibility and priming condition on self-reported effort

	Pell Eligible				Not Pell Eligible			
Condition	*N*	Mean	SD	*d*	*N*	Mean	SD	*d*
No Questions	70	4.15_a	.65	-	566	4.33_a	.65	-
Positive Self-Identity	106	4.54_b	.50	.69	713	4.39_a	.60	.10
University Creed	60	4.38_b	.66	.35	620	4.35_a	.58	.03

Note. Effort scores can range from 1 to 5, with higher scores reflecting higher self-reported effort. Within columns, means with different subscripts are statistically significantly different. *d* = Cohen's *d* effect size when comparing the No Questions condition to each question condition.

Further, ethnicity moderated the effect of priming on test performance. Prior to estimating and interpreting this effect, the assumption of homogeneity of variance was assessed and supported (Levene's test, p = .08). The interaction effect [F (2, 2046) = 3.23, p = .040)] uncovered a positive effect for underrepresented students. Specifically, in the No Question condition (typical testing condition), underrepresented students scored significantly lower (170.49) than White students (172.91). However, when primed with either set of questions, there was no difference in test scores across the two student groups (see Table 3). It appears that priming students can

Table 3

Interaction of ethnicity and priming condition on test performance

	White				Underrepresented				
Condition	*N*	Mean	SD	d_W	*N*	Mean	SD	d_U	d_E
No Questions	551	172.91	5.74	-	65	170.49	7.10	-	.41
Positive Self-Identity	701	172.46	5.74	.08	92	172.33	5.41	.29	.02
University Creed	583	172.05	6.20	.14	60	172.05	5.10	.25	.00

Note. d_E = Cohen's *d* effect size comparing performance for the two ethnic groups within each priming condition (i.e., White students are .41 SDs higher on test performance than underrepresented students in the No Questions condition, which is statistically significant, but the two ethnic groups are not significantly different on test performance in the two priming conditions, *d* = .02 and .00). *dw* = Cohen's *d* effect size comparing the No Questions condition to each question condition for White students. *du* = Cohen's *d* effect size comparing the No Questions condition to each question condition for underrepresented students.

increase test scores for underrepresented students, erasing performance differences across the underrepresented and majority ethnic groups.

Each test costs the university $8 per student. When filtering based on self-reported effort, we found the following number of students would have been removed from the three groups for low effort. For the Control Group, 35 students out of 656 would have been removed, which fortunately is a low amount of filtering. The goal with priming is to reduce this number even though it is already low (which again is a positive outcome for this testing program). For the Positive Self-Identity Group, 33 out of 852 students would have been removed. For the University Creed Group, 23 out of 696 students would have been removed. To put this into a cost perspective, the Control Group removed $280 worth of test scores, the Positive Self-Identity Group removed $264 worth of test scores, and the University Creed removed $184 worth of test scores. Based on the low removal rates, we determined that for the purposes of our study, any cost benefits based around priming for low effort would only be warranted for an institution or office that had financial hardships or had much more disengagement on assessments. For institutions or offices that do not have financial constraints or have limited disengagement (as we did), priming for low effort is likely not going to be effective to impact costs.

Discussion

Priming students with questions about their intended effort prior to completing low-stakes assessments has been found to significantly increase self-reported effort and response time (e.g., Finney & McFadden, 2023). Although this strategy has been effective for students early in their college career, it had not been examined with graduating university students nor had possible moderating student characteristics been adequately examined (e.g., Finney et al., 2024). Thus, we randomly assigned college seniors to one of three priming question conditions prior to completing a low-stakes assessment for institutional accountability: answering three questions about intended effort that infused positive self-identity, answering three questions about intended effort that incorporated the University's Creed, or answering no priming questions (control).

It appears that priming students can increase test scores for underrepresented students, erasing performance differences across the underrepresented and majority ethnic groups.

When simply examining the impact of the priming conditions, one would infer that priming seniors with self-identity questions resulted in higher self-reported effort than the control condition, priming with the University creed resulted in higher time on the test than the control condition, and neither priming condition increased test performance over no priming. No student characteristics moderated the priming effect on response time. Moreover, we found that gender, transfer status, and first-generation status did not moderate the priming effect on the three outcomes. However, Pell Grant eligibility did moderate the impact of priming on self-reported effort with priming increasing effort from eligible students who were lower in effort than non-eligible students in the control condition. In short, priming resulted in self-reported effort for eligible students being equal or higher than non-eligible students.

Likewise, the effect of priming on test scores was moderated by ethnicity. White students scored higher on the test than underrepresented students in the non-priming control condition (typical testing condition). However, this difference in test performance disappeared with priming. That is, underrepresented students had test scores not significantly different than White students if they were primed with either set of questions, which was an unexpected but positive outcome of priming. In turn, without priming, when disaggregating data and reporting test performance differences by student group, inferences would have been made about White students performing better than underrepresented students. Likely, numerous discussions would have occurred postulating why underrepresented students performed 0.41 SDs lower than White students (e.g., opportunity to learn). However, priming with either question type would lead to very different conclusions about student performance differences (or lack thereof). Clearly, further study of the effect of priming on test performance is needed.

Recall, when examining effective motivation interventions, a meta-analysis (Rios, 2021) found the largest effect on effort ($d = 0.36$) involved paying for performance (may not be acceptable or financially feasible at an institution) and the next largest effect ($d = 0.21$) involved changing the relevance of the test (easier than incentives, but still difficult in many contexts and requires more testing time than priming). When collapsing across student characteristics, the priming intervention had smaller effect sizes ($d = 0.16$ for self-identity primes on self-reported effort and $d = 0.14$ for creed primes on response time) than the more costly interventions. However, the priming effect in this testing context was larger than prior QBE studies with prosocial behaviors, which is encouraging (Wilding et al., 2016: blood donation Hedges' $g = 0.06$, voting Hedges' $g = 0.06$). Importantly, when examining the priming effect by Pell eligibility status and ethnicity, the priming effects are much larger for some student populations. Specifically, for Pell eligible students, self-reported effort for self-identity primes ($d = 0.69$) and creed primes ($d = 0.35$) was over a third of a standard deviation higher than no primes. These effect sizes are similar to or exceed the effects associated with providing incentives and changing the test relevance, interventions that cost more money and time. Likewise, for underrepresented students, self-identity primes ($d = 0.29$) and creed primes ($d = 0.25$) resulted in test performance being at or over a quarter standard deviation higher than no primes and these effect sizes align with or exceed the effects of increasing test relevance ($d = 0.27$) and external incentives ($d = 0.21$) on test performance (Rios, 2021).

Priming students with questions about their intended effort prior to completing low-stakes assessments has been found to significantly increase self-reported effort and response time.

One may ask if these effect sizes ($d = 0.69, 0.35, 0.29, 0.25$) are non-negligible in an absolute sense; they are. Recently proposed benchmarks for effect sizes from causal studies of education interventions on student achievement in Pre-K to 12 are the following: less than 0.05 is "small", 0.05 to less than 0.20 is "medium", and 0.20 or greater is "large" (Kraft, 2020). These benchmarks were based on 1,942 effect sizes from 747 randomized control trials evaluating education interventions with standardized test outcomes. Although our study does not focus on Pre-K to 12 populations or educational interventions (e.g., interventions to increase reading or math), it does evaluate a motivational intervention using standardized test outcomes with a population of students. Moreover, echoing methodologists who focus on effect sizes (e.g., Kraft, 2020), cost matters when evaluating effect sizes for policy decisions. Effect sizes should be considered relative to the costs of implementing the intervention, strategy, or program. Administrators may want to observe a 0.50 effect size for important outcomes to justify implementing an expensive program. In our context of priming students, we would argue that smaller effects ($d = 0.15$) support implementing this low-cost, quick intervention, as it is a desired step toward more trustworthy data and in turn more valid interpretations and decisions. That is, any increase (big or small) in students' effort to show their true ability is desired to make valid inferences from test scores and in turn institutional decisions about curriculum and programming. Gathering test data that does not represent the construct of interest due to student disengagement can result in decisions that are wrong, inefficient, and that cause harm. Fortunately, priming not only resulted in non-negligible effect sizes, but it is also at no or low cost to implement.

Implications for Assessment Practitioners

Effect sizes should be considered relative to the costs of implementing the intervention.

Students approach testing with different perceptions and prior testing experiences, which impact their engagement and in turn their test scores and inferences from the scores. We need to try to understand how these personal characteristics interact with testing conditions, and how testing conditions can be altered to produce more valid inferences from test scores. Some of these alterations may be quite minor, such as priming students to expend the necessary effort to show their ability. We are not suggesting that a priming intervention (or any motivation intervention) can address systemic inequities in educational measurement (see instead Forzani et al., 2024; Randall et al., 2022, 2024; Russell, 2023; Sireci, 2020). Rather, if priming reduces or eliminates group differences in effort and performance (as found for Pell Grant eligibility and ethnicity), priming may offer more valid inferences about differences in test engagement or performance across groups. We believe this is particularly important given prior research showing incorrect inferences from test scores directly impact college students. Randall et al. (2024, p. 2) provide a powerful example regarding college-level placement testing: "Often based on placement test scores, Black (66%) and Hispanic (53%) students are placed in remedial courses (see Gilman, 2019; Nastal, 2019; Ngo & Melguizo, 2020) more frequently than White (36%) students (Chen, 2016). While some may argue that students need additional support for math and literacy, it has been shown that too often placement tests misplace students (Scott-Clayton et al., 2014). For too many students, misplacement becomes an academic death sentence (Klausman & Lynch, 2022)." Assessments used for student placement (e.g., math, world language) or for remediation programming (e.g., early alert assessments) may be perceived as low stakes to students, but decisions based on them can result in stigmatization. If priming students to give effort on low-stakes assessments results in more accurate test scores and thus more appropriate decisions for even a handful of students, we believe it is well-worth the few minutes to prime them. With respect to the institution, if priming results in more accurate interpretations about the ability levels of various student groups, then institutions may avoid unnecessary programming that (mis)targets specific student populations and the deficit narratives that often accompany group differences in performance.

Although not necessarily a purpose for the current study, priming questions were reduced from five in previous studies to only three. It was promising that priming for good effort was still effective with a reduction in questions, even with this graduating student sample. Thus, we recommend assessment practitioners use one to two minutes to ask students to answer three priming questions prior to engaging in low-stakes assessments. Priming can be done solely at the beginning of the testing session; however, priming the students before each test has been shown to be promising (McFadden & Finney, 2025).

We recognize that our method for priming students (whether primes are administered via computer prior to the test or via a test cover sheet if tests are administered paper/pencil) is a simple and quick strategy, yet unknown barriers could exist that prevent a desired outcome. Should institutions encounter any logistical barriers to priming, those barriers should be shared out to the assessment community. In our opinion, those couple of minutes used for priming are worth increasing the quality of inferences from the test scores reported for accountability and improvement purposes (i.e., low cost but high benefit). More specifically, if institutions are gathering measures of motivation (self-reported, response time) and motivation is quite high, then allocating a couple minutes for this intervention may not be worthwhile. However, if motivation is low or variable across students (as is common for low-stakes tests), priming to increase effort may have great benefit with very limited "cost".

Moreover, if institutions are not collecting motivation data, but instead assuming motivation is high, we warn against this dangerous assumption. If institutions are not collecting motivation data because it takes time to gather self-reported effort or a timing mechanism is not available, we strongly suggest these institutions err on the side of caution, assume that at least some students are disengaged, and employ a strategy to increase engagement during low-stakes testing. Here we offer a very quick and cheap strategy to enhance students' test-taking effort.

Given that time-based measures of effort and self-report effort have a low correlation and different nomological nets (Akhtar & Firdiyanti, 2023), we were not surprised by the different priming effect across these two operationalizations of effort. While we feel comfortable recommending the general strategy of priming to increase effort on low-stakes institutional accountability tests, we feel it premature to advise on self-identity versus creed questions given their differential effects on the two effort measures. If pressed, we would suggest the following. Based on this single study, either priming condition will positively impact test scores for underrepresented senior students, which is encouraging but needs to be replicated. If an institution operationalizes effort via self-report measures, we recommend using the self-identity priming questions as they had the largest effect. If an institution operationalizes effort via response time, we recommend the creed-infused questions. Again, these suggestions are tentative and future studies are needed to examine the stability of these effects, in addition to other considerations discussed below.

Limitations of the Current Study and Call for Future Research

Given we found the effect of priming was moderated by certain student characteristics, we hope future studies can better capture complete information on student characteristics. Notably, we recognize the extreme crudeness of our classification of White and underrepresented students. Our sample was representative of the University's demographics. Thus, we were limited in the comparison that could be made. Future research on the QBE in accountability testing contexts should be conducted with more diverse student populations. We perceive this study as the beginning of this line of research. Moreover, future studies should examine the generalizability of the priming effect across different types of institutions (very large, very small, community colleges).

Additionally, previous studies have explored the self-identity questions with incoming first-year students (Finney & McFadden, 2023; Finney & Pastor, 2025). Given our results comparing the self-identity questions and creed questions with graduating seniors, we encourage studies that explore if first-year students are impacted by creed-infused questions. It may be that a university's creed resonates more or less with incoming students depending on the university's culture (Miller & Finney, 2024). More generally, the creed-infused questions need more examination given this study was the first. Creed phrasing may be effective at institutions where the creed is ever-present and important to students but may have no impact at institutions where the creed is not highlighted or not perceived as important by students. Thus, we encourage future studies that evaluate creed phrasing to always include a control group (no priming), but also non-creed-infused priming questions. These types of designs will allow for a better understanding of what type of phrasing in the primes is most impactful and for who. Regarding prime phrasing more generally, we strongly encourage others to evaluate different phrasing of the primes, beyond self-identity or creed-infused wording. We believe that suggestions from current students may be particularly helpful to design the most impactful primes. These are research questions worth pursuing.

We also call for longitudinal studies of the priming effect. At some institutions, students complete the low-stakes assessments at multiple time points so value-added or growth estimates can be computed and reported to accreditors. Recall, when more advanced college students expend less effort on tests than incoming students, value-added estimates are biased downward (Finney et al., 2016). Future studies should evaluate if value-added estimates are less attenuated for students who were primed at both time points compared to students who were not primed.

We encourage future studies that evaluate creed phrasing to always include a control group (no priming), but also non-creed-infused priming questions.

Finally, future studies could examine if utilizing a simple "yes" or "no" response option rather than a Likert response scale impacts the QBE. Basic research in the domain of the QBE suggests "yes" or "no" responses may be effective (e.g., Spangenberg et al., 2016). If asking a few "yes" or "no" questions prior to low-stakes institutional testing produces significantly and practically higher effort and reduces differences in test performance across student subpopulations, this strategy would be an attractive option to increase the validity of inferences from assessment scores.

References

Akhtar, H. & Firdiyanti, R. (2023). Test-taking motivation and performance: Do self-report and time-based measures of effort reflect the same aspects of test-taking motivation? *Learning and Individual Differences*, *106*, 102323. https://doi.org/10.1016/j.lindif.2023.102323

Bryan, C. J., Walton, G. M., Rogers, T. & Dweck, C.S. (2011). Motivating voter turnout by invoking the self. *PNAS*, *108*(31), 12653-12656. http://doi.org/10.1073/pnas.1103343108

Chen, X. (2016). *Remedial coursetaking at U.S. public 2- and 4-year institutions: Scope, experiences, and outcomes* (NCES 2016-405). U.S. Department of Education. National Center for Education Statistics. https://nces.ed.gov/pubs2016/2016405.pdf

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, *57*(2), 119–130. https://doi.org/10.1353/jge.0.0018

Council for Higher Education Accreditation (2022, March 3). *Accreditation & recognition*. https://www.chea.org/about-accreditation

Finney, S. J. & McFadden, M. E. (2023). Examining the question-behavior effect in low-stakes testing contexts: A cheap strategy to increase examinee effort. *Educational Assessment*, *28*(4), 211-228. https://doi.org/10.1080/10627197.2023.2222588

Finney, S. J., Myers, A. J., & Mathers, C. E. (2018). Test instructions do not moderate the indirect effect of perceived test importance on test performance in low-stakes testing contexts. *International Journal of Testing*, *18*(4), 297–322. https://doi.org/10.1080/15305058.2017.1396466

Finney, S. J., & Pastor, D. A. (2025). Priming non-compliant students to expend test-taking effort: How many primes are needed? *Journal of Experimental Education*. Advance online publication. https://doi.org/10.1080/00220973.2025.2459392

Finney, S. J., Schaefer, K. E., & McFadden, M. E. (2024). Priming examinees to give good effort: Differential utility across gender identity. *Journal of Experimental Education*. Advance online publication. https://doi.org/10.1080/00220973.2024.2310678

Finney, S. J., Sundre, D. L., Swain, M. S., & Williams, L. M. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, *21*(1), 60–87. https://doi.org/10.1080/10627197.2015.1127753

Forzani, E., Corrigan, J., Slomp, D., & Randall, J. (2024) Prioritizing equitable social outcomes with and for diverse readers: A conceptual framework for the development and use of justice-based reading assessment. *Educational Psychologist*, *59*(4), 291-314. https://doi.org/10.1080/00461520.2024.2418400

Gilman, H. (2019). Are we whom we claim to be? A case study of language policy in community college writing placement practices. *Journal of Writing Assessment*, *12*(1). https://escholarship.org/uc/item/3kh925t2

Hawthorne, K. A., Bol, L., Pribesh, S., & Suh, Y. (2015). Effects of motivational prompts on motivation, effort, and performance on a low-stakes standardized test. *Research & Practice in Assessment*, *10*, 30–39. https://www.rpajournal.com/dev/wp-content/uploads/2015/06/A3.pdf

Klausman, J., & Lynch, S. (2022). From ACCUPLACER to informed self-placement at Whatcom Community College: Equitable placement as evolving practice. In J. Nastal, M. Poe, & C. Toth (Eds.), *Writing placement in two-year colleges: The pursuit of equity in postsecondary education* (pp. 59–83). The WAC Clearinghouse, University Press of Colorado. DOI: 10.37514/PRA-B.2022.1565.2.02

Kraft, M. A. (2020). Interpreting effect sizes for education interventions. *Educational Researcher*, *49*(4), 241 – 253. https://doi.org/10.3102/0013189X20912798

Levav, J., & Fitzsimons, G. J. (2006). When questions change behavior: The role of ease of representation. *Psychological Science*, *17*(3), 207-213. https://pubmed.ncbi.nlm.nih.gov/16507060/

Liu, O. L. (2017). Ten years after the spellings commission: From accountability to internal improvement. *Educational Measurement: Issues and Practice*, *36*(2), 34–41. https://doi.org/10.1111/emip.12139

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, *20*(2), 79–94. https://doi.org/10.1080/10627197.2015.1028618

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

McFadden, M. E., & Finney, S. J. (2025). Investigating the impact of multiple priming questions on examinee effort during low-stakes testing. *International Journal of Testing*, *25*(1), 109–133. https://doi.org/10.1080/15305058.2024.2414425

Miller, S.A. & Finney, S.J. (2024). Enhancing student effort for improved institutional accountability data: The impact of motivation priming interventions. *Assessment Update*, *36*(6), 8-9. https://doi.org/10.1002/au.30421

Nastal, J. (2019). Beyond tradition: Writing placement, fairness, and success at a two-year college. *Journal of Writing Assessment*, *12*(1). http://journalofwritingassessment.org/article.php?article=136

Ngo, F., & Melguizo, T. (2020). The equity cost of inter-sector math misalignment: Racial and gender disparities in community college Student outcomes. *The Journal of Higher Education*, *92*(3), 410–434. https://doi.org/10.1080/00221546.2020.1811570

O'Neil, H., Sugrue, B., & Baker, E. L. (1995) Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, *3*(2), 135-157. https://doi.org/10.1207/s15326977ea03022

Pastor, D. A., Patterson, C. R., & Finney, S. J. (2023). Development and internal validity of the Student Opinion Scale: A measure of test-taking motivation. *Journal of Psychoeducational Assessment*, *41*(2), 209-225. https://doi.org/10.1177/07342829221140957

Randall, J., Poe, M., Oliveri, M. E., & Slomp, D. (2024) Justice-oriented, antiracist validation: Continuing to disrupt White supremacy in assessment practices. *Educational Assessment*, *29*(1), 1-20. https://doi.org/10.1080/10627197.2023.2285047

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022) Disrupting White supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, *27*(2), 170-178. https://doi.org/10.1080/10627197.2022.2042682

Russell, M. (2023) Shifting educational measurement from an agent of systemic racism to an anti-racist endeavor. *Applied Measurement in Education*, *36*(3), 216-241. https://doi.org/10.1080/08957347.2023.2217555

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*(2), 85-106. https://doi.org/10.1080/08957347.2021.1890741

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, *2014*(161), 69–82. https://doi.org/10.1002/ir.20068

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential non effortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, *33*(4), 263–279. https://doi.org/10.1080/08957347.2020.1789141

Roohr, K. C., Liu, H., & Liu, O. L. (2016). Investigating student learning gains in college: A longitudinal study. Studies in *Higher Education*, *42*(12), 2284–2300. https://doi.org/10.1080/03075079.2016.1143925

Scott-Clayton, J., Crosta, P. M., & Belfield, C. R. (2014). Improving the targeting of treatment: Evidence from college remediation. *Educational Evaluation and Policy Analysis*, *36*(3), 371–393. https://doi.org/10.3102/0162373713517935

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-testing effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*, 100335. https://doi.org/10.1016/j.edurev.2020.100335.

Simzar, R. M., Martinez, M., Rutherford, T., Domina, T., & Conley, A. M. (2015). Raising the stakes: How students' motivation for mathematics associates with high-and low-stakes test achievement. *Learning and Individual Differences*, *39*, 49–63. https://doi.org/10.1016/j.lindif.2015.03.002

Sireci, S. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice*, *39*(3), 100–105. https://doi.org/10.1111/emip.12377

Smith, L. F., & Smith, J. K. (2004). The influence of test consequences on national examinations. *North American Journal of Psychology*, *6*(1), 13–25.

Soland, J. (2018). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Spangenberg, E. R., Kareklas, I., Devezer, B., & Sprott, D. E. (2016). A meta-analytic synthesis of the question-behavior effect. *Journal of Consumer Psychology*, *26*(3), 441-458. http://doi.org/10.1016/j.jcps.2015.12.004

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

Sungur, S. (2007). Contribution of motivational beliefs and metacognition to students' performance under consequential and nonconsequential test conditions. *Educational Research and Evaluation*, *13*, 127–142. https://doi.org/10.1080/13803610701234898

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Thelk, A., Sundre, D. L., Horst, J. S., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale (SOS) to make valid inferences about student performance. *Journal of General Education*, *58*, 131–151. https://doi.org/10.1353/jge.0.0047

Wilding, S., Conner, M., Sandberg, T., Prestwich, A., Lawton, R., Wood, C., Miles, E., Godin, G., & Sheeran, P. (2016). The question-behaviour effect: A theoretical and methodological review and meta-analysis. *European Review of Social Psychology*, *27*(1), 196-230. http://doi.org/10.1080/10463283.2016.1245940

Wise, S. L. (2020). Six insights regarding test-taking disengagement. *Educational Research and Evaluation*, *26*(5-6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

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

Wise, S. L., & DeMars, C. E. (2010). Examinee non-effort and the validity of program assessment results. *Educational Assessment*, 15(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S.L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Smith, L. F. (2016). The validity of assessment when students don't give good effort. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204–220). New York, NY: Routledge.

Wood, C., Conner, M. T., Miles, E., Sandberg, T., Taylor, N. J., Godin, G., & Sheeran, P. (2016). The impact of asking intention or self-prediction questions on subsequent behavior: A meta-analysis. *Personality and Social Psychology Review*, *20*, 245-268. https://doi.org/10.1177/1088868315592334

Appendix

Priming Questions with Positive Self-Identity

Please think about the test you are about to complete. Mark the answer that best represents how you feel about each of the statements below.

A =Strongly Disagree B =Disagree C =Neither Agree nor Disagree D =Agree E =Strongly Agree

1. As a conscientious test-taker, I will engage in good effort throughout the test.

2. I, a motivated student, will give my best effort on this test.

3. As a hardworking student, I will persist to completion of the test.

Priming Questions with Creed-Infused Self-Identity

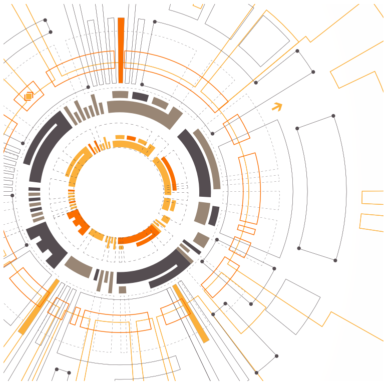
Please think about the test you are about to complete. Mark the answer that best represents how you feel about each of the statements below.

A =Strongly Disagree B =Disagree C =Neither Agree nor Disagree D =Agree E =Strongly Agree

1. As a someone who believes in education, which gives me knowledge to work wisely, I will engage in good effort throughout the test.

2. As someone who believes that this is a practical world and that I can count only on what I earn, I will give my best effort on this test.

3. As someone who believes in hard work, I will persist to the completion of the test.



Abstract

This study leverages data from direct assessments of learning (AoL) to build a dynamic model of student performance in competency exams related to computer technology. The analysis reveals three key predictors that strongly influence student success: performance on a practice exam, whether or not a student engaged in practice testing beforehand, and prior completion of an introductory course in computer applications. These findings offer valuable insights for enhancing technology competencies in assessment contexts. The authors suggest that future research should explore a more efficient alternative—offering brief, targeted instruction paired with a practice exam—rather than the traditional requirement of a full three-hour course. This approach could streamline the learning process while maintaining or improving performance outcomes.

AUTHORS
Melodie Philhours, Ed.D.
Arkansas State University

Kelly E. Fish, Ph.D.
Arkansas State University

Predictors of Student Technology Competencies in Assurance of Learning Assessment

CORRESPONDENCE
Email
kthompd@arizona.edu

Today's students are often categorized as digital natives, having grown up in a world dominated by smartphones, social media, and instant access to information (Davis, 2024). This familiarity with modern digital tools might suggest a seamless transition to using business technologies such as Microsoft Office. However, this assumption overlooks a significant gap: while these students excel in managing social media platforms, their experience with Microsoft Office—an essential suite used by 85-90% of employers in the private and public sectors—may be limited. This gap arises from the extensive use of Google Workspace in K-12 education, where over 50 million students globally have been trained primarily on Google's tools (Craig, 2024). As a result, students entering business schools often lack proficiency in Microsoft Office, which is crucial for their future careers. Addressing this discrepancy is vital for business schools to ensure that their graduates are well-prepared for the technological demands of their professional environments. This study explores how one business school has approached this issue, examining its assurance of learning (AoL) strategies for business technology skills and identifying the factors that influence the success of these initiatives. By aligning with the guidelines set forth by the Association to Advance Collegiate Schools of Business International (AACSB), the study aims to provide insights into how business schools can better prepare their graduates for the technological demands of their future careers.

In the AACSB-accredited business school used as the focus of this study, all Bachelor of Science (BS) degrees are guided by six student learning outcomes. These outcomes address communication skills, ethics, business knowledge, critical thinking and technology. Students are required to use technology appropriately to communicate, calculate, and present concepts and data. This study describes our assessment of this learning outcome and the conclusions we have drawn from this comprehensive data.

Literature Review

The Association to Advance Collegiate Schools of Business International (AACSB) accreditation standards require "…well-documented assurance of learning (AoL) processes that include direct and indirect measures for ensuring the quality of all degree programs that are deemed in scope for accreditation purposes." (AACSB 2020, page 44).

AoL is crucial for AACSB accreditation as it ensures that business programs meet quality standards and educational objectives (Nachouki 2017, Murray et al, 2008, Abraham 2013). The AoL system fostered by AACSB plays a critical role in validating the quality of business courses. A detailed analysis of this AoL system revealed its systemic, efficient approach to student learning, providing a double-loop process for course management. This system is based on the institutional mission, learning objectives, and strategies established by course stakeholders (Moraes et al, 2018). AoL is a globally accepted standard among colleges and universities for three reasons: 1) it influences curriculum development, 2) it signals to external stakeholders that a program is meeting and monitoring its educational objectives 3) it uses integrated processes for continuous improvement of student learning outcomes (Lawson et al., 2015).

Historically, most AACSB accredited institutions employ either direct and/or indirect AoL measures (Pringle & Michel, 2007). Direct AoL measures require students to represent, produce or demonstrate their learning. Examples include: standardized testing, portfolios, capstone semester projects, graded case studies, oral or written exams. Indirect measures use information about students' perceptions about their learning experiences and attitudes towards the learning process. Examples include: student focus groups, alumni surveys, exit interviews, and evaluation of retention rates (Minton & Lenz, 2019). Previously optional, 2020 AACSB criteria require indirect measurements. This change is largely due to the fact that schools which report their curriculum changes are often driven by indirect measures reflecting input from industry stakeholders (Fagnot 2023).

Students entering business schools often lack proficiency in Microsoft Office, which is crucial for their future careers.

Since much learning is delivered online the efficacy of digitally mediated learning must be assessed and measured. In a post-pandemic world where many business schools are expanding their reach via virtual delivery, new methodologies for AoL are now being developed (Fagnot 2023). Today as digital technologies are advancing, hybrid teaching frameworks where flexibility is highly valued by students are likely to be a growing area of research interest involving assessment methodologies (Tham et al, 2023). Several studies have compared face-to-face learning to virtual learning as well as blended learning in order to try to determine which results in the highest learning outcome, generates the most satisfied students or has the highest course completion rate. However, these studies often show that learning is swayed by more than teaching format alone. Among the manifold factors a few tend to have more effect, these include educator presence in virtual settings, interactions among students, teachers and content, and planned connections between online and offline activities (Nortvig, Petersen and Balle, 2018).

AoL relating to information systems (IS) is now of greater importance in a post-pandemic world and technological competency will likely become a learning goal of most business programs. Assessing student IS learning will become a task for many business programs (Zhou et al, 2022). Since AACSB accreditation standards mandate the AoL outcomes align with institutional missions and improve program quality, IS programs should assure that feedback from the AoL process along with input from internal and external sources are used to make curriculum adjustments (Emdad 2009). In this regard, Ducrot et al, (2008) proffer a structured framework for learning outcomes: the Learning Outcomes Management System (LOMS). It is used to integrate learning outcomes into both IS program-wide and IS course-specific curriculum development. The framework allows for ongoing monitoring and revision of the IS curriculum based on student performance and feedback.

In addition to developing a predictive model of student performance using results from a computer skills exam, this research analyzes rich data on student behaviors (e.g. taking an optional practice exam) and background (completion of an intro IT course, GPA, etc.). Through stepwise regression and ANOVA, it identified three dominant predictors of technology competency success – performance on a practice test, completion of an introductory computer applications course, and prior practice exam participation, By quantifying which factors most strongly influence learning outcomes, this research fills a gap in AoL literature regarding what drives student achievement in tech competencies. This study offers a data-driven model that isolates key contributors to learning and deepens our understanding of how specific inputs affect AoL results. This approach of analytically mining AoL data for predictors is a framework that other programs can adopt to diagnose and enhance learning in their own AoL processes.

We propose an alternative instructional approach in response to our findings: brief, targeted training paired with a practice exam as a potentially more efficient alternative to a full three-credit course. This idea challenges the traditional curriculum design by asking if a shorter intervention could achieve similar competency levels, a question that had not been explored in earlier AoL research on skill development.

The Study

The data for this study are comprised of the results from an AoL effort at a midsouth university in the United States. The AoL effort was a direct measure for a college of business competency in technology to communicate, calculate, and present concepts and data. Students were required to take an online, hands-on exam over the Microsoft Office environment involving database development (Access), spreadsheet operations (Excel), presentation software (PowerPoint) and word processing (Word). The college deployed the Cengage online environment for the assessment exam and optional practice exam. The remaining data were developed from student records and were scrubbed of any identifying information.

The final data file contains observations on whether or not each student had taken the optional practice exam (*PractExam*, categorical) and if so, the percentage score achieved (*PE%*, ratio). Also, the file contains information on whether or not the student has taken an introductory three-hour microcomputer course (*CIS101*, categorical) focusing on Microsoft Office (CIS101 is a disguised name for the course), the student's overall GPA (*OverallGPA*, ratio), students' GPA of any transferred hours (*TransferGPA*, ratio), overall college hours earned (*OverallHours*, ratio), hours transferred in (*HoursTrans*, ratio), year classification (senior, junior, etc., Class, categorical) and whether or not the student has a major in the College of Business (*COB*, categorical). Finally, the file contains each student's score on the assessment exam which is the dependent variable for the model. The cleaned data file contains results for 178 students.

We submitted the data to SPSS stepwise regression and the results are shown as *Model A* in Table 1. All nine variables enter the model that is statistically significant (< .001), with an R-square of .36. Three variables in the model are shown to be significant, *PE%* (< .001), *CIS101* (.005) and, *TransferGPA* (.019). No other variables come close to .05 significance with the exception of *HoursTrans* (.097). In examining the standardized coefficients, we see that *PE%* is largest (.510) followed by *CIS101* (.226) and finally *TransferGPA* (.177).

This approach of analytically mining AoL data for predictors is a framework that other programs can adopt to diagnose and enhance learning in their own AoL processes.

It is no surprise that a student's result on their practice exam is a good predictor of their results on the actual assessment exam. *PE%* bears this out as it is the most important variable in the model given its statistical significance and size of standardized coefficient. It is also a ratio variable and not in violation of the model assumption of normally distributed predictor variables. The ratio nature of *PE%* makes it richer in information than the categorical variable *PractExam* which is simply denoting whether or not a student took the exam. However, since all students do not have a score on the practice exam (*PE%*) and we are interested in the effect that taking the practice exam has on student performance, we explore *PractExam* more thoroughly.

We remove *PE%* and refit the model, the results are shown as *Model B* in Table 1. The new model is statistically significant (< .001) but the R-square drops to .25. The remaining eight variables enter the model with three of them being statistically significant, *PractExam* (<.001), *CIS101* (.004) and, *TransferGPA* (.050). *OverallGPA* comes close to the 90% confidence level but does not quite make it (.103). When we examine standardized coefficients, we see that these

Table 1
Stepwise Multiple Regression Models

	Model A		*Model B*		*Model C*	
Model	Standardized Coefficients	Significance	Standardized Coefficients	Significance	Standardized Coefficients	Significance
Constant		.001		.001		.008.
PE%	.510	< .001	N/A	N/A	.512	< .001
PractExam	-.094	.321	.288	< .001	-.097	.308
CIS101	.226	.005	.253	.004	.231	.004
OverallGPA	.110	.147	.133	.103	.155	.029
TransferGPA	.177	.019	.158	.050	N/A	N/A
OverallHours	.053	.641	.076	.530	.547	.585
HoursTrans	-.157	.097	-.113	.266	-.177	.075
Class	-.067	.521	-.096	.395	-063	.548
COB	.042	.542	.084	.255	.038	.580
Transfer	N/A	N/A	N/A	N/A	.173	.024

variables have similar patterns as the full nine variable model, *PractExam* has the largest standardized coefficient (.288) followed by *CIS101* (.253) and TransferGPA (.158). It is interesting that categorical variable *PractExam* is now arguably the most important variable in *Model B*, yet it was not even significant in *Model A* (.321).

CIS101 is a consistently significant predictor in both *Model A* and *B* and it merits further study for its effect on assessment outcomes. The last independent variable that is consistently significant is *TransferGPA*. We remove students that have not transferred hours from other institutions and refit the model (*N*=145, Sig < .001, R-Square = .36) interestingly *TransferGPA* is no longer a significant predictor (.724). Since non-takers are entered as zeros in the data file we feel that *TransferGPA's* previous importance may be due to the regression process focusing on the zeros for students that did not transfer hours. We develop a new variable *Transfer*, a categorical predictor denoting whether or not a student has transferred hours from another institution. We remove *TransferGPA* and add *Transfer* then refit the model with all students; it is significant (<.001) and has an R-Squared of .36, it is shown as *Model C* in Table 1. The new variable is significant (.024) with a standardized coefficient of .173. *OverallGPA* in *Model C* is also now significant (.029) with a standardized coefficient of .155.

Student performance on the assessment exam appears to be most influenced by the practice exam, the microcomputer course, whether or not the student transferred hours and potentially, a student's GPA.

Student performance on the assessment exam appears to be most influenced by the practice exam, the microcomputer course, whether or not the student transferred hours and potentially, a student's GPA. Since not every student in the sample took the practice exam (92 out of 178 did not take the practice exam) there is a lot of missing student data in *PE%*, however *PractExam* has data on every student and we choose to focus on it along with *CIS101* and *Transfer* for further analysis.

We run a Three-way ANOVA to check for interactions among the three variables in regard to their effect on assessment exam performance; the test is significant (.038) and results are shown in Table 2. In looking at the main effects we see that *Transfer* is not significant (.248) while *PractExam* (.003) and *CIS101* (<.001) are statistically significant. In trying to interpret the results of our Three-Way ANOVA, we choose to isolate any interaction between *PractExam* and *CIS101* across *Transfer* and run a Two-way ANOVA; our results are shown in Table 3. The new interaction between *PractExam* and *CIS101* appears to be dependent on a certain level of *Transfer*. There is no interaction when *Transfer* is equal to 0 (.279) yet when *Transfer* is at 1, there is a significant interaction (.041). We remove *Transfer* completely and run a Two-Way ANOVA with *PractExam* and *CIS101* and it shows no interaction (.295). Since *Transfer* is not significant as a main effect variable and it appears to be the culprit of the Three-Way interaction, we will focus our remaining analysis on *PractExam* and *CIS101*.

The new interaction between PractExam and CIS101 appears to be dependent on a certain level of Transfer.

In order to understand how *PractExam* and *CIS101* might affect students of different scholastic achievement, we divide the students into groups based on whether or not their *OverallGPA* score is above or below 3.0. We then examine assessment exam scores based on whether or not these students took the practice exam (*PractExam* = 1) or did not take the practice exam (*PractExam* = 0). We display the results in Table 4; *Score* is the average assessment score for that group and *OverallGPA* is the mean for that variable in the group. We run separate

Table 2
Stepwise Multiple Regression Models

Source	***F - Statistic***	***Significance***
Corrected Model	7.717	<.001
Intercept	972.158	<.001
PractExam	9.168	.003
Transfer	1.343	.248
CIS101	22.133	<.001
PractExam Transfer*	.001	.974
*PractExam*CIS101*	.016	.900
*Transfer*CIS101*	.582	.447
PractExam Transfer*CIS101*	4.384	.038

Table 3
Three-Way ANOVA

Source	***F - Statistic***	***Significance***
Corrected Model	3.316	.034
Intercept	223.199	<.001
PractExam	2.222	.147
CIS101	7.403	.011
*PractExam*CIS101*	1.220	.279
Corrected Model	14.290	<.001
Intercept	1635.393	<.001
PractExam	14.648	<.001
CIS101	24.298	<.001
*PractExam*CIS101*	6.060	.041

four group ANOVAs and find a significant difference among the groups on *Score* (<.001) and *OverallGPA* (<.001). Post hoc *Score* Tukey tests show significant differences between Group 1 and Group 2 (.008), Group 1 and Group 3 (.024) along with Group 1 and Group 4 (<.001). There were no other statistically significant differences in *Score* between the other groups. Post hoc Tukey tests on *OverallGPA* show significant differences between all groups, all at the <.001 level; with the exceptions of Group 1 and Group 2 (.201) as well as Group 3 and Group 4 (.966).

We also examine the effect of *CIS101* and display results in a similar fashion in Table 5. We again divide the students into groups based on whether or not their OverallGPA score is above or below 3.0. We then examine assessment exam scores based on whether or not a student took CIS101, then *CIS101* is coded as 1, otherwise 0. We run separate four group ANOVAs for *Score* and *OverallGPA* and both are significant at the <.001 level. Post hoc testing on *Score* reveals significant differences between Group 1 and Group 2 (.002) and Group 1 and Group 4 (<.001); no other significant differences between the groups are found. Tukey testing on *OverallGPA* shows significant differences between all groups except Group 3 and Group 4 (.540). Group 1 and Group 2 are found to be different at .032, with all other Group combinations different at the <.001 level.

Discussion of Results

To better understand what drives student performance on our technology assessment exam we built three different multiple regression models and focused on the explanatory aspects of these models rather than their predictive capabilities. The best predictor of a student's

Table 4
Analysis of PractExam

Group 1		**Group 2**		**Group 3**		**Group 4**	
Above 3.0		**Above 3.0**		**Below 3.0**		**Below 3.0**	
***PractExam* = 1**		***PractExam* = 0**		***PractExam* = 1**		***PractExam* = 0**	
Score	***OverallGPA***	***Score***	***OverallGPA***	***Score***	***OverallGPA***	***Score***	***OverallGPA***
72.07	3.54	62.13	3.44	58.53	2.58	51.53	2.54
***N*=57**		*N*=62		*N*=29		*N*=30	

Table 5
Analysis of CIS101

Group 1		**Group 2**		**Group 3**		**Group 4**	
Above 3.0		**Above 3.0**		**Below 3.0**		**Below 3.0**	
***CIS101* = 1**		***CIS101* = 0**		***CIS101* = 1**		***CIS101* = 0**	
Score	***OverallGPA***	***Score***	***OverallGPA***	***Score***	***OverallGPA***	***Score***	***OverallGPA***
71.51	3.54	58.04	3.40	61.67	2.63	52.24	2.53
***N*=69**		*N*=50		*N*=18		*N*=41	

score on the assessment exam was their score on the practice exam (*PE%*). However, in *Model B* we dropped that variable (*PE%*) since we did not have information on all students and found that a categorical variable involving the practice exam, *PractExam*, became very important in explaining student performance. In *Model C* we used a categorical variable, *Transfer*, in the place of *TransferGPA*, and it performed well and contained information on all students in the study. A third categorical variable *CIS101* always proved to be an important predictor of student scores across all models. By focusing our further study on the three best performing categorical variables, we could put our students in various groups to study the effects of the variables on student scores. Our Three-Way ANOVA testing and subsequent examination of main effects and two-way effects resulted in us removing *Transfer* due to its interaction with *PractExam* and *CIS101*.

Whether or not students took the practice exam has some interesting effects on their assessment scores. We see that students in Group 1, who have high GPAs and take the practice exam, clearly outperform all other groups of students. Their average score of 72.07 is higher than any other group's score and those differences are statistically significant. This result is not particularly surprising. However, when we examine Group 2 and Group 3, we see that students in Group 3, with significantly different and lower average GPAs (2.58) compared to Group 2 (3.44) score almost as well on the assessment exam to the point of no significant difference in the scores. It is also interesting to observe that although the difference in average GPAs for Groups 1 and 2 was not shown to be statistically significant, their respective scores on the assessment exam were. The results involving *PractExam* give evidence that the mere taking of the practice exam, a relatively inexpensive investment of time for a student could yield marked improvement in their assessment scoring. While other groups display some interesting trends, potential differences in performance do not have statistical significance, likely due to the smaller *N*s of the students below 3.0.

Whether or not a student took the introductory microcomputer course, *CIS101*, proves interesting in the study as well. High scholastic achieving students with above a 3.0 GPA that take CIS101 appear to outperform all other groups, although the difference with Group 3 is not statistically significant, potentially due to Group 3's small *N*. Similar to the results of *PractExam*, when we compare Groups 2 and 3, we have significantly different GPAs of 3.40 and 2.63, respectively, yet there is no significant difference in scores. We can conclude that lower scholastically achieving students that take *CIS101* can perform as well as higher achieving students that do not take the course.

Conclusion

This study involves assurance of learning data regarding a business school's student competencies in technology. The results provide evidence that interventions may substantially affect final results and assist student technology achievement.

In today's digitally charged environment business faculty may wonder if a three-hour course in something as fundamental as Microsoft Office is still appropriate for student learning. Given the understandable "googlification" of K-12 education due to cost, (Craig 2024) and the gap this creates to the ubiquitous use of Microsoft Office tools in business, clearly one of the key strategies of the business school is to bridge this gap and ready graduates for professional careers. These results show that indeed, the requirement of a three-hour course in basic Microsoft Office skills is needed. Lower scholastically achieving students do benefit from taking such a course; our group of students with a 2.63 GPA that took the course performed at the same level as the group of students with a 3.40 GPA that did not take the course. Also, even higher scholastically achieving students will likely have better technology competencies from taking such a course; our group with a 3.40 GPA that did not take the course scored well below the group with a 3.54 GPA that did take the course.

Lower scholastically achieving students that take CIS101 can perform as well as higher achieving students that do not take the course.

The economics of merely taking a practice exam cannot go unnoticed. High achieving students that took the practice assessment exam outperformed all others which is not unexpected. However, lower scholastically achieving students that took the practice exam (Group 3) performed at the same level of other groups that did not take the exam, including Group 2 with a 3.44 GPA. However, statistically there is no difference between Group 3 and Group 4 on assessment scores so it remains to be seen how much the exam helps.

An important component of an assurance of learning or assessment process is "closing the loop" to improve student learning and that this improvement be measurable in assessment data. The results of this study offer some very low effort and potentially high impact learning interventions. Given that we found higher technology competency in students who had completed CS101, we noted that 54 percent of the students had not taken CS101. This was surprising as CS101 is a freshman level required core course in the business school and this technology skill assessment was conducted in a junior level core business course. Investigation revealed that CS101 was offered infrequently in our online program and that students were progressing in the curriculum without it. Quickly and simply corrected, more online sections of CS101 were added and advisors notified to stress prerequisites and appropriate progression through the curriculum.

Other educational settings can benefit from these results as well. In STEM fields educators could apply the business AoL findings by incorporating optional practice tests to bolster student competencies. Research in engineering education shows that structured practice exam programs can improve student outcomes. For example, an engineering school that held guided practice exam sessions before each test saw participants achieve higher exam scores and reported more positive learning experiences than non-participants (Shew et al, 2019). These practice sessions function as brief, focused interventions – analogous to the optional practice exam in this study.

The mere taking of the practice exam, a relatively inexpensive investment of time for a student could yield marked improvement in their assessment scoring.

Many institutions are now integrating IS into humanities, having history and literature students learn text analysis software, GIS mapping, or media production tools as part of their coursework (e.g., University of North Carolina's Digital Humanities program). Direct AoL measures might include digital projects or portfolios, which require demonstrating competency with specific IS tools. To prepare students, brief targeted training sessions can be offered, much like this study's idea of a short module in lieu of a full course.

Rapid advances in AI mean that students across disciplines need to develop proficiencies that didn't even exist in curricula a few years ago. The AoL approach from the study of practice exams and targeted skill training can be directly applied to teaching these emerging technologies. For example, an instructor might create a brief tutorial on how to use a given AI platform, followed by a practice exercise where students must accomplish a task with it (such as training a simple machine learning model or prompting an AI to generate a desired output). This practice task serves as a formative assessment, revealing which students have grasped the tool and which may need more help. A subsequent performance-based AoL assessment of having students actually use the AI tool to solve a problem provides direct evidence of competency.

Despite the significant findings in this study, limitations exist that future assessments and research will address. While the overall sample size is robust (n=178), increasing the number of assessments could result in larger sub-groups within the data. Furthermore, the assessment was implemented in a required core business course that does not specifically address technology within the curriculum. It is possible that changing the assessment venue will change student engagement and performance. Additionally, data on course modality, student program, and student demographics may provide insights, e.g. online/face-to-face course/program, traditional/non-traditional students. This assessment was implemented in multiple sections of one course within one university. Collaboration with another institution could provide greater generalizability.

In future assessments of our technology outcome, we will configure the exam to require the practice exam as these results support this as a tool to improve assessment scores. While the goal of learning intervention is to obviously improve learning, perhaps the practice exam provides practice in these concepts and does indeed improve learning as well as assessment scores. Further research includes repeating this assessment to understand the effect of offering CS101 more frequently and accessibly, advisor effectiveness, and slight modification of assessment methodology to require the practice exam.

The AoL approach from the study of practice exams and targeted skill training can be directly applied to teaching these emerging technologies.

As a model of assessment and learning, this study provides a description and a baseline for evaluating both the methodology of assessment implementation and student learning. The aim of this study is to describe our process of measuring student learning in an effective and efficient methodology consistent with assessment best practices to improve such methods as needed, make data-driven changes to both curriculum and support services and to complete the cycle with longitudinal data. Through this cycle of continuous improvement, the aim of future research and assessment is to understand the long term impact of curricular changes on student learning.

References

AACSB (2020). *Guiding Principles and Standards for Business Accreditation* (updated 2-28-25) https://www.aacsb edu/-/media/documents/accreditation/_2020-aacsb-business-accreditation-standards-_final--july-1-2024.pdf?rev=e40931bf2adc4e37a3074c0e88453e5c&hash=D6C8A21B021E62F9E088471EDFE3539D

Abraham, S. (2013). Ten year assessment of learning outcomes of a computer information systems (CIS) program. *Information Systems Education Journal*, *11*, 50-58. https://isedj.org/2013-11/N6/ISEDJv11n6p50.html

Craig, R. (2024). College students are victims of the "googlification" of the classroom, *Forbes*, https://www.forbes.com/sites/ryancraig/2024/05/10/college-students-are-victims-of-the-googlification-of-the-classroom/

Davis, T. (2024). Helping students frame their futures in the AI age. *AACSB*, https://www.aacsb.edu/insights/articles/2024/05/helping-students-frame-their-futures-in-the-ai-age

Ducrot, J., Miller, S., & Goodman, P. (2008). Learning outcomes for a business information systems undergraduate program. *Communications for the Association for Information Systems*, *23*(6) https://doi.org/10.17705/1cais.02306.

Emdad, A. (2009). Integrating learning outcomes assessment in information systems curriculum revisions. *Review of Business Information Systems*,*13*(3), 43-48. https://doi.org/10.19030/RBIS.V13I3.4322.

Fagnot, I. (2023), Assurance of learning (AoL) and AACSB's 2020 business accreditation standards: a conversation with marine Condette, *Organization Management Journal*, *20*(2), 6-62.

Lawson, R., Taylor, T., French, E., Fallshaw, E., Hall, C., Kinash, S. and Summers, J. (2015), Hunting and gathering: new imperatives in mapping and collecting student learning data to assure quality outcomes, *Higher Education Research and Development*, *34*(3), 581-595.

Minton, C., & Lenz, A. (2019). Selecting assessment measures. In *Practical Approaches to Applied Research and Program Evaluation for Helping Professionals* (1st ed.) Routledge. https://doi.org/10.4324/9781315108933-15.

Moraes, M., Kalnin, G., & Karsten, M. (2018). An analysis of the assurance of learning system promoted by the AACSB american accreditation agency for business and management courses. Biotechnologie, Agronomie, *Société et Environnement*, *15*(1), 68-80. https://doi.org/10.4013/BASE.2018.151.06

Murray, M., Pérez, J., & Guimarães, M. (2008). A model for using a capstone experience as one method of assessment of an information systems degree program. *Journal of Information Systems Education*, *19*(2), 197-208 https://digitalcommons.kennesaw.edu/cgi/viewcontent.cgi?article=2379&context=facpubs

Nachouki, M. (2017). Assessing and evaluating learning outcomes of the information systems program. *4*(4) ,524. https://doi.org/10.22158/WJER.V4N4P524.

Nortvig, A.M., Petersen, A.K. and Balle, S.H. (2018) A literature review of the factors influencing e-learning and blended learning in relation to learning outcome, student satisfaction and engagement. *The Electronic Journal of e-Learning*, *16*(1), 46-55.https://eric.ed.gov/?id=EJ1175336

Pringle, C., & Michel, M. (2007). Assessment practices in AACSB-accredited business schools. *Journal of Education for Business*, *82*(4), 202-211. https://doi.org/10.3200/JOEB.82.4.202-211

Shew, D. P., & Maletsky, L. P., & Clark, G., & McVey, M. (2019, June), Practice Exam Program Impact on Student Academic Performance and Student Retention Paper presented at 2019 ASEE Annual Conference & Exposition , Tampa, Florida. 10.18260/1-2--33182

Tham, A., de Villiers Scheepers, M., Grace, A. and Ashton, A.S. (2023), Assurance of learning in business education – what exactly are we assuring, and whose business should it be?, *Quality Assurance in Education*, *31*(4), 616-636. https://doi.org/10.1108/QAE-03-2023-0051

Zhou, D., Morgan, D., Dwivedi, R., & Bai, S. (2022). A case of teaching and assessing an introduction to information technology course. *International Journal of Curriculum Development and Learning Measurement*, *3*(1), 16.