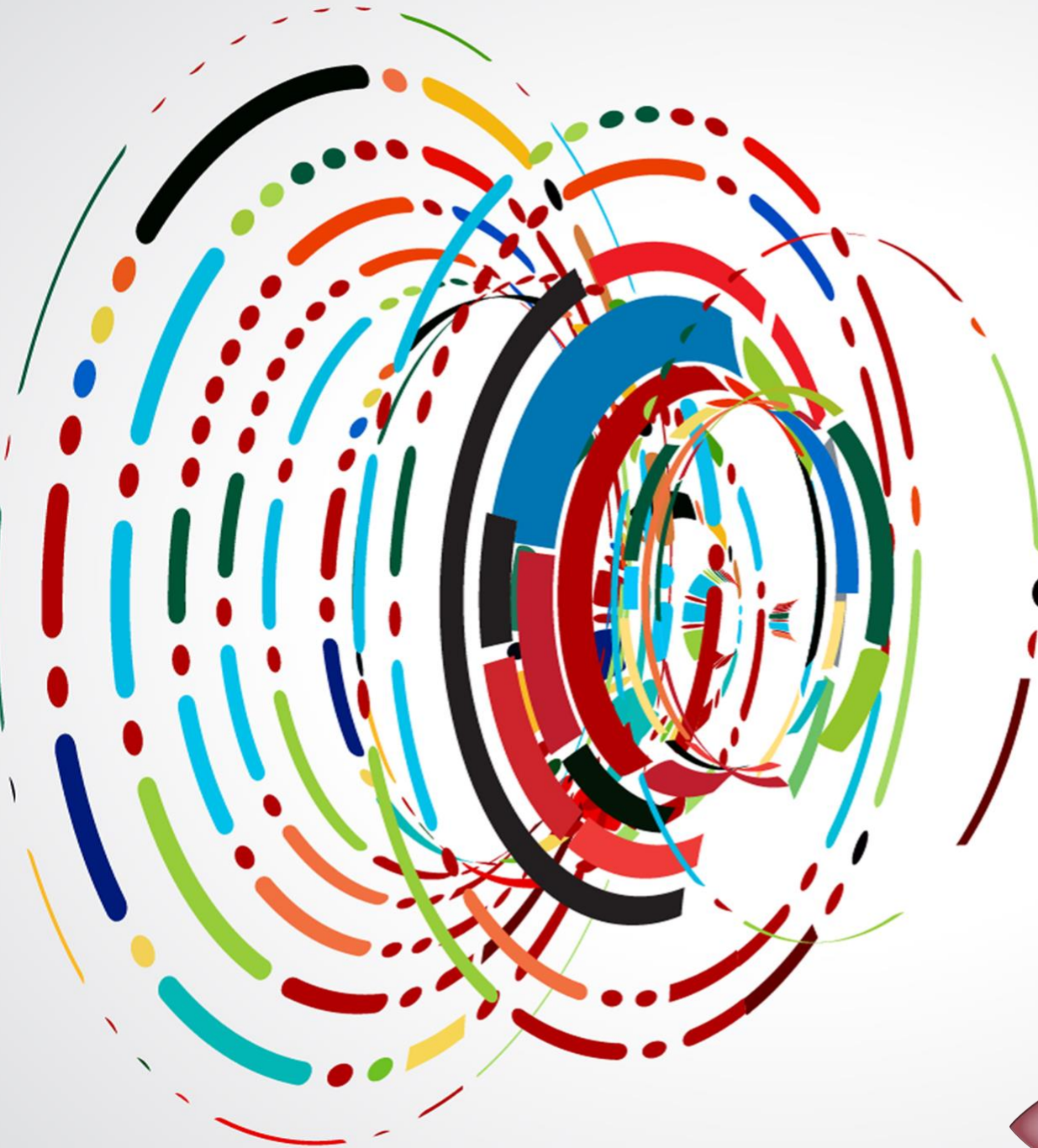


Research & Practice in Assessment

Volume Twenty-One | Issue 1 | rpajournal.com | ISSN #2161-4120



A Publication of the Virginia Assessment Group

Journal Overview

Research & Practice in Assessment (RPA) is a peer-reviewed journal dedicated to advancing scholarly and practical work in higher education assessment. Our mission is to foster dialogue between researchers and practitioners by publishing work that strengthens assessment practice, supports evidence-based decision-making, and contributes to the improvement of student learning. *RPA* originated within the Virginia Assessment Group and continues to serve the national and international assessment community.

Key Facts about RPA

- Peer-reviewed, double-blind review process
- Approximately 40% acceptance rate
- Published on a continuous cycle
- Issues are compiled periodically for archival purposes
- Indexed by EBSCO, ERIC, Gale, ProQuest, and Google Scholar
- Listed in Cabell's Directory

Call for Papers

RPA welcomes manuscripts on higher education assessment, including work grounded in assessment measurement, assessment policy, foundations, or applied practice. Submissions are accepted year-round and must follow *RPA* Submission Guidelines. All manuscripts are submitted through our online system: www.rpajournal.com/authors

Types of Submissions

RPA publishes a range of scholarly and practice-oriented work relevant to higher education assessment. We welcome:

- Empirical research studies on student learning, program assessment, and institutional effectiveness
- Practice reports that document applied assessment projects, innovations, and case studies
- Conceptual or theoretical analyses that explore foundations, models, or frameworks of assessment
- Policy-focused manuscripts examining assessment policy, governance, or accreditation
- Integrative reviews or literature syntheses related to assessment research and practice

History

Research & Practice in Assessment was founded in 2006 by the Virginia Assessment Group (VAG). The journal grew out of the organization's earlier newsletter, expanding to include peer-reviewed scholarly work that bridges assessment research and practice. Since its founding, *RPA* has expanded its editorial board, refined its review processes, and developed an international readership. The journal continues to be published by VAG. A full history and timeline for the journal can be found at www.rpajournal.com.

Join us at the **2026 VA Collaborative on Institutional Effectiveness**

Hosted by the Virginia Assessment Group in partnership with VRAS and VAIR
Learn more at www.virginiaassessment.org





Table of Contents

4

From the Editor

Nicholas A. Curtis

5

The Impact of Time of Day on Engagement and Performance for a University Low-Stakes Assessment

Jonathan M. Henriques, Brian C. Leventhal, Jack Gilmore, Yu Bao

19

The Communication and Classroom Climate Connection: Student Perceptions Based on Race, Gender, and Classification

Karen Santos Rogers, Jennie M. Carr, Gibbs Kanyongo

38

Evaluating Solutions for Improving Program Assessment

Nathan C. Anderson, Alaric Williams, Daniel Ringrose

64

Evaluating Accreditation: A Case Study of Student Views on the Value of Optional Program Accreditation

Erica Eckert

84

Creating and Implementing a Data Ecosystem for Assessing Sexual Violence Campus Climate

Sarah Socorro Hurtado, Caitlyn Potter Glaser

100

Implementation Fidelity in Higher Education: A Scoping Review

Sarah Kirkland, Elizabeth Hinton, Carley Dear, Xiaomei Song, Anthony Arciero

115

Measuring Meaning-Making Among College Students: The Living a Life of Meaning and Purpose—C Portfolio

Larry H. Ludlow, Theresa A. O'Keefe, Lauren Warner, Henry I. Braun, Christina Matz

Editorial Staff

Editor-in-Chief

Nicholas A. Curtis

University of Wisconsin-Madison

Senior Associate Editor

Robin D. Anderson

James Madison University

Associate Editor

John Moore

National Board of Medical Examiners

Associate Editor

Sarah Gordon

University of Central Oklahoma

Associate Editor

David Allen

Texas Christian University

Associate Editor

Laura Lambert

James Madison University

Copyeditor

Aleister Abercrombie
Valencia College

Designer

Melissa Curtis

"The important thing is not to stop questioning. Curiosity has its own reason for existing."

— *Albert Einstein*

Assessment exists to support better decisions, but better decisions depend on better evidence. As assessment has matured, our field has increasingly recognized that collecting data is only the beginning. We must also consider whether evidence is trustworthy, whether it is interpreted appropriately, and whether it ultimately leads to meaningful action. The articles in this issue of *Research & Practice in Assessment* each contribute to that ongoing conversation, examining different aspects of how evidence is generated, understood, and applied across higher education.

Henriques and colleagues begin by examining an often-overlooked question: does when students complete a low-stakes assessment matter? Their findings suggest that time of day influences both engagement and performance, reminding us that contextual factors deserve careful consideration when interpreting assessment results. Rogers, Carr, and Kanyongo then explore how race, gender, and classification shape students' perceptions of classroom communication and climate, illustrating the importance of understanding learning environments alongside learning outcomes.

Questions of institutional decision-making continue with Anderson, Williams, and Ringrose, who evaluate approaches for improving program assessment and offer practical guidance for strengthening assessment systems. Eckert examines student perspectives on optional program accreditation, providing valuable insight into how students perceive external quality assurance efforts and the value institutions place upon them.

At the institutional level, Hurtado and Glaser present the development of a data ecosystem supporting assessment of sexual violence campus climate. Their work demonstrates how thoughtful infrastructure can enable more coordinated, actionable, and sustainable institutional assessment. Kirkland and colleagues broaden the discussion through a scoping review of implementation fidelity, reminding us that evaluating outcomes without considering whether initiatives were implemented as intended provides only part of the story. Finally, Ludlow and colleagues explore the assessment of meaning-making through the Living a Life of Meaning and Purpose—C Portfolio, extending assessment beyond traditional measures of knowledge and skill toward constructs that are increasingly central to student development.

Taken together, these articles reinforce an enduring lesson for our profession: evidence is valuable only to the extent that we understand its context, its limitations, and its implications. Whether examining student engagement, classroom climate, institutional systems, accreditation, implementation, or personal development, each contribution encourages us to ask better questions before drawing conclusions. That commitment to thoughtful inquiry remains one of assessment's greatest strengths.

I hope these articles stimulate new ideas, productive conversations, and practical improvements within your own institutions. As always, thank you to our authors, reviewers, associate editors, and editorial staff for their continued commitment to advancing assessment scholarship and practice.



Warmly,

Nicholas Curtis

Nicholas Curtis

Editor-in-Chief

Research & Practice in Assessment

The Impact of Time of Day on Engagement and Performance for a University Low-Stakes Assessment



Authors:

Jonathan M. Henriques, M.A.
James Madison University

Jack Gilmore, M.A.
James Madison University

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

Yu Bao, Ph.D.
James Madison University

ABSTRACT

Student engagement on low-stakes assessments raises concerns about the validity of conclusions derived from those tests. With the growing prevalence of online administration at universities, students often have the flexibility to choose their testing time. This study explores how the self-selected testing time of day influences university students' test-taking engagement and performance in low-stakes assessments. A total of 1,397 university students were categorized into four time-of-day categories: morning, afternoon, evening, and nighttime. The number of rapid guesses was compared to assess engagement. Performance scores were also analyzed for comparison. The findings revealed that students who chose to test in the morning and afternoon exhibited greater engagement and better performance than those who tested in the evening and nighttime. These results underscore the importance of considering testing time when interpreting assessment outcomes, with implications for the scheduling of online assessments to enhance student engagement and performance.

Correspondence E-mail: Henri2jm@jmu.edu

Keywords: low-stakes assessment, test-taking engagement, test-taking effort, response time effort, rapid guessing

When interpreting assessment scores, it is critical to consider whether examinees appropriately engaged throughout the test administration (American Educational Research Association et al., 2014). Assuming high levels of examinee engagement is typically justified during high-stakes tests, where examinees receive benefits or face consequences based on their performance. However, for assessments with low stakes where performance results in no personal benefit or consequence, low test-taking engagement becomes a significant concern. When an examinee does not fully engage during a test to assure their highest possible score, the validity of our test score interpretations is threatened (Wise & DeMars, 2005). For example, higher education assessments conducted to evaluate program-level or institutional outcomes are commonly low-stakes. Therefore, student motivation to engage in the test tends to be low, consequently reducing their test-taking engagement and performance. In fact, disengaged students perform an average of .59 standard deviations lower than engaged students (Wise & DeMars, 2005).

Understanding the factors that drive an examinee's test-taking engagement is essential for identifying strategies to enhance it. One influential framework for explaining the variability in test-taking engagement is Expectancy-Value Theory (EVT). EVT posits that an examinee's expectations for success and their perceived task value significantly influence their effort (Eccles et al., 1983; Eccles & Wigfield, 2002). Expectations for success represent an individual's self-driven beliefs about likely task performance, while perceived task value reflects the importance they assign to succeeding or failing in a task (Eccles & Wigfield, 2002).

An individual's perceived task value is shaped by four key components: attainment value, intrinsic value, utility value, and cost. Attainment value pertains to the personal significance assigned to the task, intrinsic value relates to the enjoyment derived from task performance, utility value considers how well the task aligns with individual goals, and cost encompasses the negative aspects (such as time and effort) associated with task engagement (Eccles & Wigfield, 2002). When the costs of test-taking outweigh the expected benefits, disengagement often ensues, a pattern especially evident in low-stakes testing contexts, where the lack of meaningful consequences leads to consistently lower engagement compared to high-stakes assessments. Even when examinees derive little enjoyment from high-stakes tests, the potential outcomes often incentivize effort, aligning with EVT (Wise & DeMars, 2005). To explore this dynamic, Wise and DeMars (2005), relied on self-report data to assess examinees' perceived importance and effort, offering insight into the motivational process, but raising questions about the adequacy of self-report as a sole indicator of engagement.

While self-report data can provide valuable insights into overall effort indirectly from the test-takers, it assumes that examinees respond truthfully, and it only captures effort cross-sectionally at the end of the test (Wise & Kong, 2005). Examinees might also provide inaccurate information due to social desirability bias, confusion between poor performance and low effort, or low engagement with the self-report measure itself. Additionally, variation in effort during an exam (Pastor et al., 2019)—such as differential effort across different sections—may not be fully captured by self-report measures administered only at the end of the assessment.

To evaluate effort throughout a test, researchers have often relied on item-level behavioral indicators of disengagement. While effort refers to the mental resources an examinee expends toward success, engagement represents the behaviors that indicate such investment (Gorgun & Bulut, 2023). Two common behavioral indicators are solution behavior and rapid guessing behavior. Examinees exhibit solution behavior when they intentionally attempt to determine the correct answer (Schnipke, 1995; Wise & DeMars, 2006), while rapid guessing behavior occurs

when examinees answer an item more quickly than it would reasonably take to read, process, and conclude an answer (Wise & Kong, 2005). During high-stakes assessment, rapid guessing often arises when examinees run out of time and need to complete the test quickly (Wise, 2020). In contrast, during low-stakes assessments rapid guessing tends to appear more sporadically throughout the test and signals disengaged test-taking (Wise & Kong, 2005).

Determining Rapid Guessing Behavior

Rapid guesses can be identified by comparing the response time with a predetermined time threshold for each item. Responses faster than this threshold are labeled as rapid guesses. Several methods exist to establish rapid guess thresholds. The simplest approach involves selecting a fixed threshold time and applying it uniformly across all items (Wise et al., 2010). While easy to implement, this method overlooks variations among items as some may naturally require more or less time for examinees exhibiting solution behavior to answer. Alternatively, a unique threshold for each item can be determined through a visual inspection of the distribution of item response times for the specific item (Schnipke, 1995; Wise & DeMars 2006). If the distribution is bimodal, the threshold is set to the first local minimum. Unfortunately, item response times do not always exhibit a bimodal distribution. Another method, referred to as cumulative proportion correct thresholds, integrates both response time and response accuracy to determine the threshold, which is the maximum response time where the proportion of examinees answering the item in that duration or faster scored at chance level (Guo et al., 2016). However, this method is often plagued by complexity. A more practical approach involves setting an item-level Normative Threshold (NT), such as 10% or 20% of the mean response time of all examinees (Wise & Ma, 2012). These diverse methods allow for the identification of rapid guessing behavior on each item. By analyzing an examinee's behavior across items, we can then determine whether they are disengaged or not.

To comprehensively assess test-taking engagement, Wise and Kong (2005) proposed a metric called Response Time Effort (RTE), computed as the proportion of items that a test-taker did not rapidly guess. Examinees with an RTE below 0.90 are commonly flagged as disengaged (Swerdzewski et al., 2011; Wise & Kong, 2005). In a meta-analysis, Silm et al. (2020) found a strong correlation (0.72) between RTE and performance, surpassing the correlation (0.33) between self-reported effort and performance. This suggests that RTE may be a more reliable estimate of engagement than self-report.

Test-Taking Engagement Filtering

Indicators of disengagement can be used to adjust aggregate scores and enhance the validity of low-stakes assessment results through the application of filtering techniques either at the examinee or response level (Alahmadi & DeMars, 2024). Examinee-level filtering involves listwise deletion of response sets from disengaged examinees, such as those with an RTE below 0.90. In contrast, response-level filtering employs a pairwise deletion method, removing only rapid guesses and treating them as missing responses without excluding entire response sets. Consider an examinee who rapidly guesses on 11% of the test items (RTE of 0.89). Examinee-level filtering would exclude all their responses, while response-level filtering would only remove the 11% of responses that were rapid guesses. Although response-level filtering has its advantages, Alahmadi and DeMars (2024) argue that examinee filtering is easier to explain to stakeholders and tends to be more politically acceptable and interpretable.

Predictors of Test-Taking Disengagement

Test-taking engagement may be influenced by the following: personal characteristics, item characteristics, and test-taking characteristics. Personal characteristics relate to the individual taking the test. Notably, gender significantly impacts engagement, with more males likely to rapid guess than females (Rios & Soland, 2022). Additionally, language background (Kroehne et al., 2020) and socioeconomic status (Rios & Soland, 2022) also play roles in test-taking engagement. Item characteristics refer to specific attributes of test items, including item length (Wise, 2006), item position (Rios & Soland, 2022; Wise, 2006), and response type, such as multiple choice or constructed response (DeMars, 2000; Michaelides & Ivanova, 2022). Lastly, examinees' test-taking engagement varies under different test-taking conditions or situational factors. For example, rapid guessing tends to be higher on paper-based tests compared to computer-based tests (Chua & Don, 2013; Kroehne et al., 2020), but the absence of a proctor for computer-based tests can also relate to higher instances of rapid guessing (Kroehne et al., 2020; Wise et al., 2019). Recently, researchers have investigated how the time of day at which a student takes a test influences test-taking engagement (Wise et al., 2010; Wise et al., 2024).

Time of Day

In today's educational landscape, many low-stakes tests are conducted remotely, allowing students greater flexibility in choosing the time of day (TOD) to complete their assessments. This shift has prompted practitioners to consider whether TOD influences test-taking engagement. To date, we identified two studies that investigated the relationship between TOD and test-taking engagement: Wise et al. (2010) and Wise et al. (2024).

In 2010, Wise et al. examined the reading and mathematics sections of a large-scale computer adaptive test administered to students in grades 3 through 9. Students were assigned test start times between 7:00 a.m. and 2:00 p.m. by their local educators, with most students tested two to three times throughout the year. Wise et al. (2010) found no significant relationship between rapid guessing behavior and time of year or day of the week. However, they observed that rapid guessing behavior was significantly more prevalent later in the testing day. The percentage of examinees flagged for disengaged test-taking consistently increased across grade levels and tests. For 9th graders taking the math test, 1.7% were flagged at 7:00 a.m., while 4.6% were flagged at 2:00 p.m. Similarly, for 9th graders taking the reading test, 4.1% were flagged at 7:00 a.m., while 11.6% were flagged at 2:00 p.m. These findings suggest that TOD may predict test-taking engagement. Notably, performance scores on both math and reading tests decreased throughout the day, likely due in part to reduced engagement. In 2024, Wise et al. conducted a similar study using the same large-scale tests for students in grades 2 through 8. Consistent with the original study, they found a significant relationship between TOD and rapid guessing, with rapid guessing being more prevalent later in the day for both the math and reading tests.

TOD Effects in Higher Education Assessment

The impact of TOD on rapid guessing behavior has primarily been studied in K-12 testing contexts. Although no direct research has investigated this relationship in higher education settings, studies on cognitive capabilities and performance can provide insights. Notably, college students exhibit better processing speed (Allen et al., 2008), executive control performance (Allen et al., 2008), and semantic processing (Smith, 1987) in the afternoon. Academic performance also varies throughout the day, with some research suggesting improved performance later in the day and inferior performance in morning classes, with lowest performance in the earliest classes (e.g.,

8:00 a.m. start time; Dills & Hernandez-Julian, 2008). However, conflicting findings make it challenging to pinpoint the optimal TOD for performance (Kaur et al., 2021; Onyper et al., 2012).

The shift toward technology-based assessment in higher education has introduced greater flexibility in testing outside the classroom, further distinguishing low-stakes assessments in higher education from those in K-12 settings. Unlike K-12 education, where standardized school hours apply to all students, college students follow individualized daily schedules, making it difficult for institutions to establish fixed assessment times. As a result, it is now common for college students to self-select their testing times and complete assessments remotely (e.g., Alahmadi & DeMars, 2022), whereas K-12 students are typically assigned testing times and complete assessments in school settings (e.g., Wise et al., 2024). While this added flexibility benefits college students, it may come at the cost of testing during suboptimal times for engagement and performance. For example, although many college students have more free time in the evening, they may experience reduced mental stamina after attending classes or working throughout the day. Choosing to test in the evening may be more convenient but could lead to decreased performance compared to morning or afternoon testing. Additionally, the standardized schedules in K-12 education create a highly structured daily routine, with each class or event following the previous one. In contrast, college students often have more dispersed schedules, which may contribute to different patterns of engagement decline as the day progresses.

The Current Study

In the current study, we investigate the relationship between TOD and both engagement and performance among test-takers during a low-stakes higher education assessment. Specifically, we address two primary research questions:

1. To what extent does time of day relate to test-taking engagement?
Given that the limited literature on TOD and motivation has focused on K-12, we refrain from specifying a hypothesis for potential differences in engagement across different times of day.
2. To what extent does time of day relate to performance?
Research on the relationship between time of day and performance has yielded conflicting results, especially concerning university students' performance in the morning as compared to the afternoon. Although few studies directly compare performance across morning, afternoon, evening, and nighttime, it is likely that fatigue and reduced mental stamina later in the day impacts performance. Therefore, we hypothesize that student performance will vary between those who test in the morning or afternoon versus those who test in the evening or nighttime.

Method

Participants

We analyzed first-year students from a large mid-Atlantic university who participated in a remote computer-based ethical reasoning assessment (ERA). All first-year students (~4,900) attending the university were mandated to complete a series of assessments prior to arriving at the institution. The battery of assessments was randomly assigned to each student. After

employing our inclusion criteria¹, we retained a sample of 1,397 students who completed the ERA. While we did not collect demographic information for our specific sample, institutional data for the undergraduate population at the university indicated 57% of students identified as female and 43% identified as male. Additionally, 74.5% of students identified as White, 7.9% as Hispanic/Latino, 5.1% as Asian, 4.8% as identifying with two or more ethnicities, 4.5% as African American, and 3.2% as another ethnicity.

Achievement Test

The computer-based assessment used in this study consisted of 50 multiple choice items that measured students' understanding of eight key components of ethical reasoning: fairness, outcomes, responsibilities, character, liberty, empathy, authority, and rights (Sanchez et al., 2017). Each item presented an ethical issue or dilemma, and students were required to select the component that best aligned with the prompt. All items had eight options in the same order, corresponding to each of the ethical reasoning components. Only one of the eight options was correct for each item, and items were scored as either correct or incorrect. Students received the ERA as their first assessment in the battery, with no completion time restrictions. Importantly, the ERA was low stakes, as students were explicitly informed that their individual performance did not impact their grade point average (GPA) or transcript.

Rapid Guesses

Item response times were recorded for each student and calculated as the time between a student starting an item and moving on to the following item. Initially, response time thresholds were determined individually by each of the four researchers analyzing the response time distribution, the cumulative proportion correct, and the mean NT10, NT15, and NT20 for each individual item. The estimated interrater reliability between researchers' separate thresholds was high (coefficient alpha = .945). The researchers' collective set of thresholds was used to determine the final rapid guess threshold for each item.

Time of Day

Originally, students were given a 12-day window to complete their assessment, but this was later extended to approximately 1 month. Students had the flexibility to complete the ERA at any TOD during the testing period. Each student's local TOD was recorded through the online testing software (Qualtrics) at the start of the ERA. Subsequently, students were categorized into four general periods: morning (8:00 a.m. - 11:59 a.m.), afternoon (12:00 p.m. - 4:59 p.m.), evening (5:00 p.m. - 9:59 p.m.), and nighttime (10:00 p.m. - 7:59 a.m.). These TOD periods were established based on previous psychological TOD research classifications (e.g., Smith, 1987; Allen et al., 2008) and the structure of a typical workday. Although nighttime had the longest interval, we anticipated that fewer students would be testing during those hours.

Data Analyses

To answer Research Question 1, we conducted a Kruskal-Wallis H test to investigate whether the number of rapid guesses is significantly different across TOD. The Kruskal-Wallis H test was chosen because rapid guessing distributions tend to be positively skewed; many students

¹ We only considered students who consented to research and completed the ethical reasoning assessment in full (without skipping questions) in a typical testing environment.

typically do not engage in rapid guessing at all (Wise et al., 2024). To further investigate Research Question 1 at the test level, we performed a chi-square test of independence to explore whether there was a relationship between TOD and the proportion of students filtered using the RTE cutoff .90.

To address Research Question 2, we conducted a one-way ANOVA to compare performance scores across different times of day. We specifically tested the hypothesis that students who began their test in the morning or afternoon would score higher than those who started in the evening or nighttime. Additionally, to examine the impact of TOD on performance after accounting for motivation, we conducted the same analyses after filtering students at an RTE cutoff of 0.90.

Results

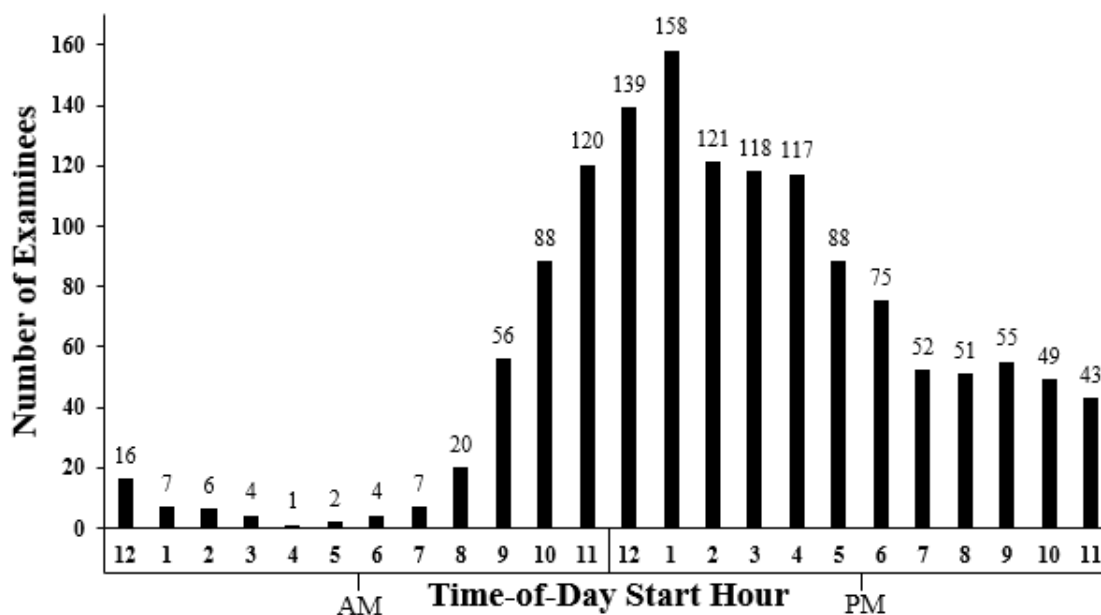
Time of Day of Self-Selected Testing

Students were most likely to self-select to test in the afternoon, with approximately 47% of students starting the ERA between 12:00 p.m. and 4:59 p.m. (see Figure 1). The proportions of students who opted for morning and evening testing were similar, accounting for 20% and 23% of students, respectively. The nighttime group had fewer participants, representing only 10% of students, with 6.5% starting between 10:00 p.m. and 11:59 p.m.

Engagement and Time of Day

Students who began the ERA in the morning or afternoon had one fewer rapid guess, on average, than students who began in the evening or nighttime (see Table 1). Although this finding may have important practical implications, a Kruskal-Wallis H test revealed no statistically significant differences in the number of rapid guesses across the TOD periods ($\chi^2(3) = 2.00, p = .571, \eta H^2 = 0$).

Figure 1. Frequencies of student time-of-day start hour on the ERA



Note: The start times for students were recorded in local time zones.

Time-of-Day Effects on Filtering Students

We examined potential differences in student filtering across different times of day using an RTE of 0.90. Students who began the ERA in the evening and nighttime exhibited a higher prevalence of disengagement, resulting in more frequent filtering compared to those who started in the morning and afternoon (refer to Table 2). Approximately 14% of students who tested in the evening and nighttime were filtered out, while 11.6% of students who tested in the morning and afternoon were filtered. A chi-square test of independence conducted between TOD and students filtered at the RTE 0.90 level revealed no significant relationship, $\chi^2(3) = 1.80$, $p = .61$, Cramér's $V = .036$. Although no statistically significant findings emerged, there appears to be a practical association between motivation and TOD, with students exhibiting greater disengagement during the evening and nighttime.

Table 1. Descriptive statistics for rapid guesses, performance, and filtered performance by time of day.

Time of Day	N	# of Rapid Guesses		Performance Before Filtering (Out of 50)		Performance After Filtering (Out of 50)	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Morning	284	3.00	9.25	30.05	9.73	32.29	7.37
Afternoon	653	3.22	9.33	30.28	9.55	32.53	7.16
Evening	321	3.75	9.95	29.02	9.33	31.52	6.90
Nighttime	139	4.03	10.91	28.65	9.84	30.83	7.59

Table 2. Percentage test events classified as disengaged for different filtering criteria across day.

Filtering criteria based on number of rapid guesses	RTE filter threshold	Morning Percent Filtered	Afternoon Percent Filtered	Evening Percent Filtered	Nighttime Percent Filtered
5 or more	.90	11.6	11.6	14.3	13.7

Performance and Time of Day

Before filtering, students who began the ERA in the morning or afternoon, on average, answered one more question correctly than those who started in the evening or nighttime (see Table 1). The mean score of students in the morning and afternoon was significantly higher than the mean score of students in the evening and nighttime, $t(1393) = 2.242$, $p = .025$, [95% CI (0.33, 4.98)], $d = .28$. This finding suggests that student performance is higher during workday hours. After filtering using an RTE cutoff of 0.90, students who took the ERA in the morning or afternoon still scored over one point higher on average than those who started in the evening or nighttime. The difference was significant for both the omnibus one-way ANOVA, $F(3, 1219) = 2.617$, $p = .0496$, $\eta^2 = .006$ and a priori test of the average of morning and afternoon compared to average of evening and nighttime $t(1219) = 2.593$, $p = .010$, [95% CI (0.60, 4.36)], $d = .34$.

Discussion

In this study, we investigated the relationship between test-taking engagement and the TOD when university students chose to complete a low-stakes assessment. Although we did not propose specific hypotheses regarding engagement due to limited prior research, we hypothesized that students testing in the morning and afternoon would perform better than those testing in the

evening or nighttime. Our findings revealed a subtle relationship between TOD and engagement: students exhibited greater engagement during morning and afternoon hours compared to evening and nighttime. Additionally, significant performance differences aligned with our hypothesis, with higher scores observed in the morning and afternoon, both before and after adjusting for engagement.

Relationship between Time of Day and Disengagement

The relationship between TOD and student disengagement appears to differ between higher education and K-12 contexts. Among college students, our findings indicated that test-taking engagement remained consistent across the morning and afternoon, whereas previous research with K-12 populations has shown engagement tends to decline in the afternoon (Wise et al., 2010; Wise et al., 2024). This discrepancy may be attributed to the greater scheduling flexibility in college, where students' classes are more spread out, potentially preserving cognitive capacity into the afternoon. Our results also revealed increased disengagement during evening and nighttime hours. This decline may be attributable to fatigue and listlessness becoming more pronounced later in the day, possibly due to the cumulative depletion of energy and cognitive resources.

These findings reflect common student experiences: initiating a test in the evening or nighttime often coincides with reduced energy and a stronger desire to complete tasks quickly, especially as students approach their habitual sleep or social time. In such cases, students may prioritize task completion over sustained, effortful engagement. This behavior aligns with EVT, which suggests that motivation to engage in the test is influenced by perceived value and cost. For low-stakes assessments, the cost of engagement, such as sacrificing sleep or delayed social activities, may outweigh the perceived value, leading to disengagement. Conversely, morning or afternoon testing offers several advantages: students have been awake for fewer hours, retain more cognitive capacity, and face fewer direct costs such as lost sleep or missed social time. These conditions may foster higher levels of engagement, suggesting that scheduling assessments during these periods could support improved student performance and effort.

Relationship between Time of Day and Performance

The lack of performance differences between students who tested in the morning and afternoon aligns with prior research showing mixed findings across these periods (Dills & Hernandez-Julian, 2008; Kaur et al., 2021; Onyper et al., 2012). However, our study revealed a clear decline in performance during the evening and nighttime compared to the morning and afternoon hours. This decline can be partially attributed to increased disengagement, as students were more likely to rapidly guess and answer fewer questions correctly during later testing periods. Notably, even after adjusting for test-taking effort, students who tested in the morning and afternoon consistently outperformed those who tested in the evening and nighttime. Although the mean number of correct responses increased slightly across all four time periods after filtering for effort, the approximate one-question gap between the morning/afternoon and evening/nighttime persisted. This suggests that the self-selected TOD may serve as a proxy for other underlying factors that influence performance beyond what is captured by our engagement measures alone.

Implications, Limitations, and Future Directions

In this study, we aimed to better understand the impact of self-selected TOD on test-taking engagement and performance among college students. While previous research has explored these effects in K-12 settings, differences in administration structure and scheduling flexibility between K-12 and higher education warranted further investigation. As higher education increasingly embraces remote testing, the flexibility it offers has become especially important for non-traditional students, such as adult learners, those returning after gap years, and part-time students, who often face scheduling constraints due to employment or family responsibilities. For these students, the ability to self-select testing times is highly desirable.

However, our findings indicate that testing in the evening is associated with lower engagement and reduced performance. While flexibility remains essential, alternative approaches that preserve choice while optimizing outcomes merits further investigation. For example, universities might consider implementing testing windows that mimic typical workday hours (8 a.m. to 5 p.m.). Yet, restricting testing to these hours could disadvantage students with daytime obligations (e.g., job or family responsibilities). A more balanced solution may involve slightly extending testing hours (e.g., 7 a.m. to 7 p.m.), which could maintain flexibility while mitigating the engagement and performance declines observed during late evening hours. Multi-day or weekend testing options could also support students with weekday conflicts. Although we did not investigate the impact of the day of the week on performance and engagement, prior research suggests this factor is a weak predictor in K-12 populations (Wise et al., 2010). Given the differences in college student demographics and routines, further research is needed before drawing practical conclusions about weekday effects in higher education.

A more practical and student-centered approach may be for universities to recommend, but not require, testing during optimal windows (e.g., 8 a.m. to 5 p.m.). Framing this guidance as evidence-based could encourage students to self-select times that support better performance and engagement. Communicating the rationale behind these recommendations may also enhance perceived value of the assessment, signaling institutional care for student success and outcomes.

These results have important implications for longitudinal studies of curricular effectiveness. If students choose different testing times during pre- and post-assessments, observed performance changes may reflect TOD effects rather than learning gains. For example, students who completed the ERA in our study will be retested later in their academic careers. It will be valuable to examine whether their self-selected testing times remain consistent or vary as such differences could influence the validity of longitudinal score interpretations.

Despite observing practical differences in performance across different TODs, these disparities persisted even after accounting for engagement. The self-selection of testing times limits experimental control and introduces potential confounding variables. Beyond engagement, factors such as fatigue and listlessness likely contribute to performance declines, as previously demonstrated (Weirich et al., 2017). These effects may be moderated by individual characteristics, including gender, employment status, family responsibilities, and socioeconomic status. For example, a student working full-time from 9:00 a.m. to 5:00 p.m. may only be able to test in the evening, potentially leading to lower engagement and performance compared to a peer with more flexible availability. Future research should explore these individual differences to better understand the mechanisms behind TOD-related performance declines and inform equitable testing practices.

Overall, our findings highlight the complex relationship between TOD and assessment outcomes, suggesting that factors such as fatigue, daily schedules, and personal responsibilities

significantly influence engagement and performance. Recognizing and accommodating these influences is essential for optimizing testing conditions and improving the validity of academic assessments.

Conclusion

Our study demonstrates that test-taking engagement varies across TOD, and that TOD can significantly influence performance. Students who choose to take tests in the evening or nighttime tend to disengage and achieve lower scores. These findings suggest that institutions should consider limiting evening and nighttime testing to enhance student performance and engagement. However, because this study did not employ experimental controls, TOD may act as a proxy for other underlying characteristics that influence outcomes.

Our research contributes valuable insights into how TOD relates to effort and performance among university students—a population that has been underexplored in this context. By highlighting performance variations based on TOD, our study lays the foundation for future investigations into how flexible testing schedules may affect academic outcomes assessment results. These findings also prompt important considerations for educational policy, particularly regarding the scheduling of low-stakes assessments and implications of allowing students to self-select their testing times.

Overall, this study advances our understanding of TOD's relationship with test-taking engagement and performance in higher education. It also paves the way for further research into the individual and contextual factors that shape these outcomes, with potential implications for assessment design, scheduling practices, and the interpretation of scores in flexible testing environments.

References

- Alahmadi, S., & DeMars, C. E. (2022). Large-scale assessment during a pandemic: Results from James Madison University's remote Assessment Day. *Research & Practice in Assessment*, 17(1), 5–15.
- Alahmadi, S., & DeMars, C. E. (2024). Comparing examinee-based and response-based motivation filtering methods in remote low-stakes testing. *Applied Measurement in Education*, 37(1), 43–56. <https://doi.org/10.1080/08957347.2024.2311927>
- Allen, P. A., Grabbe, J., McCarthy, A., Bush, H., & Wallace, B. (2008). The early bird does not get the worm: Time-of-day effects on college students' basic cognitive processing. *The American Journal of Psychology*, 121(4), 551–564. <https://doi.org/10.2307/20445486>
- 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.
- Chua, Y. P., & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior* 29(5), 1889–1895. <https://doi.org/10.1016/j.chb.2013.03.008>
- DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, 13(1), 55–77. https://doi.org/10.1207/s15324818amei301_3
- Dills, A. K., & Hernandez-Julian, R. (2008). Course scheduling and academic performance. *Economics of Education Review* 27(6), 646–654. <https://doi.org/10.1016/j.econedurev.2007.08.001>
- Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 74–146). San Francisco, CA: W. H. Freeman.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53(1), 109–132. <https://doi.org/10.1146/annurev.psych.53.100901.135153>
- 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(27). <https://doi.org/10.1186/s40536-023-00177-5>
- Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, 29(3), 173–183. <https://doi.org/10.1080/08957347.2016.1171766>
- Kaur, P., Kumar, H., & Kaushal, S. (2021). Affective state and learning environment based analysis of students' performance in online assessment. *International Journal of Cognitive Computing in Engineering*, 2, 12–20. doi:10.1016/j.ijcce.2020.12.003
- Kroehne, U., Deribo, T., & Goldhammer, F. (2020) Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling* 62(2), 147–177. <https://doi.org/10.25656/01:23630>
- Michaelides, M. & Ivanova, M. (2022). Response time as an indicator of test-taking effort in PISA: country and item-type differences. *Psychological Test and Assessment Modeling*, 64(3), 304–338.
- Onyper, S. V., Thacher, P. V., Gilbert, J. W., & Gradess, S. G. (2012). Class start times, sleep, and academic performance in college: A path analysis. *Chronobiology International*, 29(3), 318–335. <https://doi.org/10.3109/07420528.2012.655868>

- Pastor, D. A., Ong, T. Q., & Strickman, S. N. (2019). Patterns of solution behavior across items in low-stakes assessments. *Educational Assessment*, 24(3), 189–212.
<https://doi.org/10.1080/10627197.2019.1615373>
- 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. & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis. *Large-Scale Assessments in Education*, 9(1), 18.
<https://doi.org/10.1186/s40536-021-00110-8>
- Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2016). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing* 17(1), 74–104. <https://doi.org/10.1080/15305058.2016.1231193>
- Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*, 22(2), 154–184.
<https://doi.org/10.1080/15305058.2022.2036161>
- Sanchez, E. R. H., Fulcher, K. H., Smith, K. L., Ames, A., & Hawk, W. J. (2017). Defining, teaching, and assessing ethical reasoning in action. *Change: The Magazine of Higher Learning*, 49(2), 30–36. <https://doi.org/10.1080/00091383.2017.1286215>
- Schnipke, D. L. (1995). *Assessing speededness in computer-based tests using item response times* [Unpublished doctoral dissertation]. Johns Hopkins University.
- 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, 100335. <https://doi.org/10.1016/j.edurev.2020.100335>
- Smith, A. P. (1987). Activation states and semantic processing: A comparison of the effects of noise and time of day. *Acta Psychologica*, 64(3), 271–288. [https://doi.org/10.1016/0001-6918\(87\)90012-6](https://doi.org/10.1016/0001-6918(87)90012-6)
- 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>
- Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, 41(2), 115–129. <https://doi.org/10.1177/0146621616676791>
- 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. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, 43(1), 19–38.
<https://doi.org/10.1111/j.1745-3984.2006.00002.x>

- 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. R., & Lindner, M. A. (2024). Don't test after lunch: The relationship between disengagement and the time of day that low-stakes testing occurs. *Applied Measurement in Education*, 37(1), 14–28. <https://doi.org/10.1080/08957347.2024.2311925>
- Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education*, 32(2), 183–192. <https://doi.org/10.1080/08957347.2019.1577248>
- 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 2012 Annual Meeting, Vancouver, BC, Canada.
- Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010, May). An investigation of the relationship between time of testing and test-taking effort [Paper presentation]. National Council on Measurement in Education 2010 Annual Meeting, Denver, CO, United States.

The Communication and Classroom Climate Connection: Student Perceptions Based on Race, Gender, and Classification



Authors:

Karen Santos Rogers, Ph.D.
Coppin State University

Jennie M. Carr, Ph.D.
Bridgewater College

Gibbs Kanyongo, Ph.D.
Duquesne University

ABSTRACT

This quantitative study examined students' perceptions of factors considered important in creating a positive classroom climate to add to the literature focused on classroom climate determinants. It sought to answer the following research question: Do students of different race, gender, or classification (freshman, sophomore, junior, senior) perceive classroom dynamics uniformly? The Education School Climate Survey (EDSCLS) was streamlined and administered to 198 students. Data were analyzed to measure student perceptions of classroom climate. Both student gender and classification failed to account for differences in perception of classroom climate; however, statistically significant differences were found in six instances when comparing student responses based on race. In these instances, White students ranked the indicated classroom environment factors more positively than students of other races. Four out of six of these factors relate to faculty-student communication and interactions.

Correspondence E-mail: jcarr@bridgewater.edu

Keywords: Communication, Classroom climate, Race, Gender, Classification

Funding: This work was supported by a \$1200.00 Institutional Faculty Research Grant.

As colleges and universities nationwide fight to increase enrollment and grapple with retention issues, focusing on classroom climate is more important than ever. Classroom climate is a broad construct made up of students' perceptions and feelings about their instructor, peers, and environment. It is affected by classroom structure, power dynamics, pedagogical styles, curriculum, and faculty-student and student-student relationships. Creating a positive classroom climate is essential to the success of students in higher education. Research indicates a positive classroom climate is associated with student achievement and motivation (Brand et al., 2008; Ellis, 2004; Church et al., 2001; Fraser & Fisher, 1982) and is the best predictor of students' overall satisfaction with their college experience (Graham & Gisi, 2000).

Faculty strive to create a classroom environment that is conducive to learning, including creating a positive climate; however, the higher education classroom is multi-dimensional, and the determinants of classroom climate and its effect in terms of student satisfaction are often difficult to identify. Adding to the difficulty of creating a positive classroom climate may be the fact that students' experiences can lead to differing perceptions of the same classroom environment. According to Steele (1997), different individuals can—and do—experience the same school in dramatically different ways on the basis of race. In addition, earlier research has suggested women, older students (commonly defined as 24+), and minority students often encounter a classroom climate that is “chillier” than other demographic groups, and faculty are frequently unaware they systematically treat students differently based on gender (Wilkie, 2000; Sadker, 2000; Sadker et al., 1991; Trujillo, 1986; Hall & Sandler, 1982). For example, women may be less likely to be called on, more frequently interrupted, or given less praise. Faculty, even those concerned about discrimination, may inadvertently communicate limiting preconceptions about behavior, abilities, career directions, and personal goals based on gender rather than student interest or ability (Sadker, 2000; Handler & Sandler, 1982). A student's tenure at the university and previous interactions with the faculty may also play a key role in their perception of the classroom climate. Recent research by Carr et al., (2021) indicates the longer or more times a student interacts with a faculty member, regardless of the circumstances, the more likely they are to say they felt the faculty cared for them.

Therefore, it stands to reason that each of these variables, race, gender, and classification (freshman, sophomore, junior, senior), will likely affect students' perceptions of their environment. The subtle (or overt) ways in which students feel they are treated differently can create a learning environment that hinders academic achievement. Although faculty are responsible for designing and promoting supportive classroom climates, they are rarely given the tools or training to do this effectively (Fertig, 2012). If faculty can gain a better understanding of how students perceive their classes, that information can be used to improve classroom environment and foster a climate that ensures equity for all students.

In stating the need for ongoing investigation of classroom environment preferences, Byrne et al., (1986) said, “A central question is whether there is one ideal environment that is perceived uniformly by all students, or whether the ideal depends on moderators such as student age, gender, and actual environment” (p. 10). He asserts the measurement of student perceptions is a superior approach to observational techniques in assessing classroom environment because in many cases students' perceptions are more important determinants of their behavior than the competing objective reality. The

current study examined students' perceptions of 19 factors considered important in creating a positive classroom climate. It sought to answer the following research question: Do students of different race, gender, or classification (freshman, sophomore, junior, senior) perceive classroom dynamics uniformly?

Theoretical Perspectives

Students perceive deep and meaningful learning occurs at the intersection of the teaching, cognitive, and social presence (Huang et al., 2018). The term "classroom climate" can be used to describe the environment that encompasses these three factors. Therefore, the current research study was influenced by several theories of learning. Social-cognitive learning theory provides an understanding of the social and mental processes in achieving learning outcomes by weaving emotional capacities with cognitive capabilities in the social context (Bandura, 2005). A key focus of the current study is on faculty-student relationships; therefore, social-cognitive theory helps provide an explanation for the climate-achievement link as it relates to faculty and student interactions (Bandura, 1993, 1997). Additionally, this study is also informed by the affective-cognitive teaching and learning theory. The cognitive domain is concerned with thinking and reasoning. Affective learning is focused on how students feel while they are learning, as well as with how learning experiences are internalized, so they can guide a student's attitudes, opinions, and behavior in the future (Miller, 2005). More specifically, the affective domain highlights the outlook and emotional state of students toward other learners and/or faculty. This study also drew on the tenets of the learning theories of Vygotsky (1962) and Piaget (1965). These theories assert the primary purpose of language is communication and learning results from social interaction.

Underlying theoretical influences applicable to classroom climate can also relate to the shared mental image (isomorphism) students have of the classroom construct (Kozlowski & Klein, 2000; Lüdtke et al., 2006; Morin et al., 2014), and student success may be heavily influenced by the degree to which they feel accepted and a sense of belonging. Social Identity Theory posits that a group (e.g., class), system, or organization (e.g., school) influences individual behavior when an individual feels psychologically part of that group, system, or organization (Tajfel & Turner, 1979). Membership is defined by a feeling of psychological inclusion, identification, and connectedness (Maxwell et al., 2017). The more a student feels accepted (and acceptable), the more they will be able to express themselves, act authentically, and be fully present (Osterman, 2000). This is closely connected to Identity Theory, whereby one's self-concept can be defined as "the totality of a complex, organised, and dynamic system of learned beliefs, attitudes and opinions that each person holds to be true about his or her personal existence" (Purkey, 1988). Terry et al., (1999) identify a close link between self-identity (or self-concept) and behaviors that are demonstrated in the wider social context. Academic self-concept has been identified as influencing a number of educational outcomes, including academic achievement (Whannell & Whannell, 2015; Choi, 2005; Marsh & Yeung, 1997; Reynolds, 1988).

Literature Review

What is Climate?

Research in educational settings is guided by the assumption that the environment in which students are immersed plays an influential role in shaping their cognition, behavior, motivation, and emotion (Lüdtke et al., 2009). The term “climate” has long been used to describe how students perceive their environment (Lewin et al., 1939). More specifically, climate is a reflection of students’ perception of their academic experience (Reid & Radhakrishnan, 2003), including a variety of factors, such as faculty-student interactions and relationships, student-student interactions and relationships, an instructor’s teaching presence (such as pedagogical style, tone, humor, rapport, instances of stereotyping or tokenism), and perceptions related to the curriculum and course content (Ambrose et. al, 2010; Barr, 2016; Frisby & Martin, 2010; Meyers, 2009; Sidelinger & Booth-Butterfield, 2010). Classroom climate is fundamentally interpersonal in nature (as opposed to factors out of faculty control, such as equipment), which is why it is generally expressed through perceptions of students and demonstrated to have such a profound impact on student outcomes (Barr, 2016).

Although students develop their own sense of the classroom environment, there can also be a community sense of the environment shared by all those in the class (Fraser & Treagust, 1986). Students’ perceptions often define the classroom climate because their exposure to multiple learning environments allows for many opportunities to form impressions and gives them a credible vantage point from which to make judgements (Fraser & Treagust, 1986). According to Reid and Radhakrishnan (2003), the concept of climate explains how environment variables (i.e., observations) can affect psychological ones (i.e., perceptions).

Classroom Climate Research

One of the earliest researchers of classroom climate, Rudolph Moos (1979), asserted that “the social-ecological setting in which students function can affect their attitudes and moods, their behavior and performance, their self-concept and general sense of well-being” (p. 3). Moos’ (1979) theory forms the basis of some of the most commonly used classroom climate instruments in higher education, including Fraser and Treagust’s (1986) College and University Classroom Environment Inventory (CUCEI) and the College Classroom Environment Scale (CCES) developed by Winston et al. (1988). Research based on these inventories primarily focuses on differences between faculty and student perceptions (Fraser & Treagust, 1986; Letizia, 1989; Ghanatabadi, 1991).

Communication scholars have contributed significantly to the study of climate, focusing on both faculty-student interactions and student-student interactions. Research on faculty-student interactions shows that instructors who engage in rapport building actions help create a positive classroom climate (Barr, 2016; Catt et al., 2007; Schrodtt et al., 2006). Students report rapport is an essential characteristic of an effective instructor (Catt et al., 2007, Faranda & Clarke, 2004). Using the Professor-Student Rapport Scale (Wilson et al., 2010), instructor rapport was found to account for 54% of the variance in student ratings of instruction on evaluations completed at the end of the semester (Richmond et al., 2015).

Research focused on student-student interactions indicates students benefit from a connected classroom in a variety of ways, including motivation and increased

preparedness and participation (Bouville, 2010; Sidelinger & Booth-Butterfield, 2010; Frisby & Martin, 2010; Johnson, 2009; Osterman, 2000). Additionally, a connected classroom with positive student-student interactions leads to an increased feeling of belonging for students. Myers (1995) stated that for students "... climate may be affected by the need to establish and defend personal worth and social stability in the eyes of both teachers and peers" (p. 193). Lee and Robbins (1995) conceptualized belongingness as a construct of self in relation to the larger [classroom] community. Freeman et al., (2007) emphasized students develop a sense of belonging when they perceive they are accepted, respected and valued. Bosma and Kunnen (2001) describe the process of [academic] identity development as "an iterative process of person-context transactions" (p. 39) that requires individual openness to change and appropriate environmental support. Further research shows positive classroom peer relations, including supportive and cooperative student-student communication, affect a students' perception of climate and results in greater engagement with and enjoyment of the classroom experience (Johnson, 2009; Dwyer et al., 2004; Anderson & Carta-Falsa, 2002).

Climate research at both the school and class level has established a strong correlation between student satisfaction and achievement. Positive outcomes indicating classroom environment have a major influence on students' interest, satisfaction, and achievement (Ellis, 2004; Church et al., 2001; Brand et al., 2008; Fraser, 2000; Fraser, 1993; Fraser, 1994; Fraser & Treagrust, 1986; Haertel et al., 1981, Walberg, 1969) have been found in the US, Canada, Australia, and India (Haertel et al., 1981). Maxwell et al., (2017) assert school climate is a leading factor in explaining student learning and achievement. One of the main reported causes of college dropout is a low level of students' sense of school belonging (Gizir, 2019). Students who perceive themselves to have a history of poor academic outcomes may have difficulty conceiving of themselves as a successful university student and may be more susceptible to attrition due to negative comparisons with other students, an inability to perform the more challenging and independent role associated with being a university student, and the nature of the institution they are attending (Whannell & Whannell, 2015). Additionally, Graham & Gisi (2000) found the instructional or academic climate was the best predictor of students' overall satisfaction with their institution.

Although some studies have sought to compare student perceptions of climate based on gender (Wilkie, 2000; Sadker et al., 1991) and classification (Wilkie, 2000; Hadley & Graham, 1987) limited research has been conducted that disaggregates data to determine if students perceive classroom climate uniformly based on varied individual characteristics. The purpose of this study is to add to the body of research on classroom climate by analyzing and comparing perceptions of students in the same course(s) based on race, gender, and classification.

Method

Participants

Convenience sampling was used to identify 8 faculty members at a private, liberal arts college in the Mid-Atlantic region of the United States. Faculty were invited to attend a technology-focused professional development training and participate in the study. Participating faculty agreed to disseminate a voluntary perceptions-based survey at the end of the semester. Faculty participants were from varied disciplines, including

Mathematics (1), Science (1), Communication (2), Literature (1), and World Languages (1). Several faculty taught multiple courses; therefore, 15 courses were included in the study. The 198 undergraduate students who completed the perceptions-based survey (out of 307 total) enrolled in these classes are identified as the participants. The study population is primarily White (70.7%) and female (68.7%) from a traditional college student age (18-24) range (99%). For the purposes of this study, gender was operationalized as a self-identified variable, allowing participants to select the gender identity that best represents them. The survey provided the following response options: male, female, or prefer not to answer. Due to the small sample size, all non-White students were combined into a single group. The demographic characteristics of student participants can be found in Table 1 (Demographic characteristics of student participants).

Table 1. *Demographic characteristics of student participants.*

Demographic Categories	n=198
Race	
African American	31
Caucasian	140
Hispanic/Latino	12
Asian/Pacific Islander	6
Other Race	9
Gender	
Male	61
Female	136
Prefer not to answer	1
Classification	
Freshman	147
Sophomore	18
Junior	12
Senior	18
Prefer not to answer	3

Design and Data Collection

The United States Department of Education and National Center for Educational Statistics has prioritized research on school climate to develop policies on effective school practices. The Education School Climate Survey (EDSCLS) series is a reliable, nationally validated public survey tool available for use by any school seeking to utilize feedback from its stakeholders (NCES, 2015; NCES, 2016). The EDSCLS is a 73-item survey using a Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). For the purposes of the study, only the 19 questions related to classroom climate designated for students were included and analyzed. As defined by Ambrose, Bridges, DiPietro & Lovett, (2010), classroom climate refers to the intellectual, social, emotional and physical environment in which students learn. The survey was distributed to participants via email.

Data Analysis

This study sought to determine if students of different race, gender, or classification perceive classroom dynamics uniformly to add to the literature focused on

identifying determinants of classroom climate. Participating students completed a modified version of the EDSCLS. The variables used in this study were proxies derived from the items in the NCES Climate Survey, and all items were measured on a 1 (strongly disagree) to 5 (strongly agree) point scale. Because climate is closely related to other concepts that are often used to describe the character of a learning environment (e.g., culture and context), and also because classroom climate is based on perceptions of classroom experiences, it is difficult to measure directly, therefore the need to use proxies.

Researchers calculated the internal consistency of the instrument using data from the current study (Cronbach's $\alpha = .92$). Each of the variables are described below, including items from the survey that constituted it. Even though the data were measured on a Likert scale, we treated the data as containing parametric properties. This rationale was based on the work by Lubke & Muthen (2004) who found that it is possible to find true parameter values with Likert item data.

Data for students were analyzed through the independent samples t test using the statistical software SPSS, version 27.0. Since the sample sizes of the racial groups of African American ($n = 31$), Latino ($n = 12$), Asians/Pacific Islander ($n = 6$), and Other ($n = 9$) were very small relative to the Caucasian group ($n = 140$), these groups were not compared separately; instead, they were combined into one group, and this group was then compared with the Caucasian group using the independent samples t test. Each demographic variable (race, gender, classification) originally contained more than two categories and was dichotomized for analysis. One-way Anova was deemed inappropriate to analyze these data since the minimum recommended sample size per group for one-way ANOVA is 15-20 participants (Stevens, 2002). Stevens further recommends that the ratio of the sample sizes of the smallest to the largest group in ANOVA should not be too extreme.

Before conducting the test, the following statistical assumptions of the independent t test were checked: normality assumption and the homogeneity of variance assumption. The normality assumption was checked using Q-Q plots, and based on the Q-Q plot, the normality assumption was violated; however, according to the central limit theorem, the sampling distribution of the mean is always normally distributed, as long as sample size is large enough ($n > 30$), and in this study the sample size was $n = 198$. The homogeneity assumption was conducted using Levene's test ($p < .05$), indicating that this assumption was not met, which led us to interpret the 'Equal variances not assumed' set of results for the independent t test.

In conducting the independent samples t test, a comparison was made of each of the 19 items on the classroom climate survey. To control for the inflation of a Type I error, the Bonferroni adjustment was performed by dividing the alpha level of 0.05 by the number of tests to yield an altered alpha of ($\alpha_{\text{altered}} = .05/19 = .003$ (rounded)). However, to account for the multiple tests which were conducted, we elected to use a more conservative alpha of .001. The results of the analysis are presented in the tables below organized according to the comparisons based on race, gender, and classification.

Results

The results of the analysis based on race are presented in Table 2 (Differences in classroom climate based on race). Based on the results in the table, statistically significant racial differences in classroom climate were observed at alpha level of .001 on the

following items: The instructor of the course cares about me: Caucasian ($M = 3.79$; $SD = .424$), Other Races ($M = 3.48$; $SD = .682$), $t(196) = -3.875$, $p < .001$, Cohen's $d = .61$; I have many opportunities to be a part of class discussions and activities: Caucasian ($M = 3.81$; $SD = .478$), Other Races ($M = 3.53$; $SD = .655$), $t(196) = -3.26$, $p < .001$, Cohen's $d = .51$. My instructor gives me individual attention when I need it: Caucasian ($M = 3.81$; $SD = .390$), Other Races ($M = 3.53$; $SD = .706$), $t(196) = -3.26$, $p < .001$, Cohen's $d = .56$; My instructor expects me to do my best all times: Caucasian ($M = 3.90$; $SD = .301$), Other Races ($M = 3.71$; $SD = .496$), $t(196) = -3.36$, $p < .001$, Cohen's $d = .52$; I can communicate with my instructor about problems or questions I have in class: Caucasian ($M = 3.87$; $SD = .357$), Other Races ($M = 3.60$; $SD = .647$), $t(196) = -3.26$, $p < .001$, Cohen's $d = .58$; I feel like I belong at this institution: Caucasian ($M = 3.54$; $SD = .735$), Other Races ($M = 3.12$; $SD = .993$), $t(196) = -3.26$, $p < .001$, Cohen's $d = .51$. A comparison of the whole construct of classroom climate was also conducted between the two racial groups of 'Other Races' and 'Caucasian', the results of which are presented in the final row in Table 2. The results show that there is no statistically significant difference in the whole construct of classroom climate, at the alpha level of .001, $t(196) = -2.32$, $p < .011$, Cohen's $d = .43$, with Caucasian group scoring higher ($M = 3.71$) than the 'Other' group ($M = 3.54$).

Comparison was also made to determine differences on the overall construct of classroom climate based on gender and based on students' classification (presented in the last row of each table). Results show that there were no statistically significant differences at the alpha level of .001 in classroom climate based on either gender or classification. The results of the comparison based on gender are presented in Table A3: Differences in classroom climate based on gender, while the results of the comparison based on classification are presented in Table A4: Differences in classroom climate based on classification in Appendix A. The results show that the two gender groups did not differ on the overall construct of classroom climate, $t(195) = -2.67$, $p = .009$, Cohen's $d = .41$. Additionally, the results show that there was no statistically significant difference between students' classification on the overall construct of classroom climate, $t(193) = -1.675$, $p = .095$, Cohen's $d = .28$.

Discussion

In this study, we examined students' perceptions of 19 factors considered important in creating a positive classroom climate. Two general conclusions emerged from the analysis. First, the results of this study showed neither gender nor classification had an impact on student perceptions of classroom climate in any statistically significant way. This was surprising given that earlier research has suggested women often feel a classroom's climate is less hospitable than some other demographic groups, and faculty are frequently unaware they systematically treat students differently based on gender, particularly in some disciplines (Wilkie, 2000; Sadker et al., 1991; Trujillo, 1986; Hall & Sandler, 1982). Additionally, it was hypothesized undergraduates of a higher classification (sophomore, junior, senior) would indicate they feel more positive (comfortable) communicating with their professor and participating in class than freshmen. Freshmen are more apt to be intimidated when navigating a college classroom for the first time, since they are not likely to know the professor or most, if any, of the other students. However, data indicate classification had no bearing on the classroom climate factors targeted in this study.

Second, our findings indicate there are several statistically significant differences when comparing White students with students of other races. This supports Steele's (1997) assertion that different individuals can—and do—experience the same school in dramatically different ways on the basis of race and aligns with prior research that suggests racial minority students are more likely to report a negative academic climate than White students (Reid & Radhakrishnan, 2003; Hurtado & Carter, 1997; Hurtado, 1994; Allen & Haniff, 1991). In all six instances, White students ranked the indicated classroom environment factors more positively than students of other races. Most of these (4/6) relate to faculty-student communication and interactions. The following factors were ranked lower (less positively) by students of other races than by White students:

- The instructor of this course cares about me.
- I have many opportunities to be a part of class discussions and activities.
- My instructor gives me individual attention when I need it.
- My instructor expects me to do my best at all times.
- I can communicate with my instructor about problems or questions I have in class.
- I feel like I belong at this institution.

The current research shows that when compared to White students, students of other races do not believe their instructors care about them to the same extent, feel they have equal opportunities to participate, or feel as comfortable communicating with their professors about problems related to their classes. The disparity between these groups suggests faculty-student communication and interactions need to be addressed in a direct and purposeful way. These outcomes almost certainly have implications for students' success as numerous studies demonstrate a strong correlation between academic achievement and both positive faculty-student communication and students feeling cared for by their professors (Carr et al., 2021; Beattie & Thiele, 2016; Cox et al., 2010; Kim & Lundberg 2016; Kuh 2010; Laird & Cruce, 2009; Meyers, 2009). Therefore, recognizing a divide in perceptions can help faculty make adjustments to foster a positive classroom climate that ensures equity for all students.

Instructional Implications

The current research implies classroom climate could be improved if faculty were more explicit in building rapport, demonstrating care for students and their success, and being more proactive regarding faculty-student communication and interactions. Many institutions have begun to recognize the power of prioritizing relationships and building soft skills on college campuses as key (Schmier, 1997; Supiano, 2018). Couros (2018) claims, "Relationships in every position in education are non-negotiable. Build them, and you can do anything. Don't build them, and you will struggle to do everything" (para. 1).

Both faculty and students affect the characteristics of the classroom environment, which influences student learning (Hirshy & Wilson, 2002). Each enters the classroom with relational goals that must be met for a positive outcome (Frymier, 2007), and faculty play a crucial role in shaping the interactions in the classroom (Fassinger, 2000). Supportive faculty-student relationships can encourage students to be more invested in their learning (Meyers, 2009).

Table 2. Differences in classroom climate based on race

Racial differences	Race	N	M	SD	t	p-value	Cohen's d
In this course, males and females are treated equally.	Other Races	58	3.86	.398	-.639	.523	.10
	Caucasian	140	3.90	.403			
People of different cultural backgrounds, races, or ethnicities get along well in this course.	Other Races	58	3.83	.464	-.674	.501	.10
	Caucasian	140	3.87	.395			
The instructor understands my academic concerns.	Other Races	58	3.53	.731	-1.585	.115	.25
	Caucasian	140	3.69	.599			
The instructor is available when I need to talk with them.	Other Races	58	3.68	.506	-1.659	.099	.26
	Caucasian	140	3.81	.448			
It is easy to communicate with the instructor of this course.	Other Races	58	3.69	.568	-2.512	.013	.39
	Caucasian	140	3.86	.383			
The instructor of this course cares about me.	Other Races	58	3.48	.682	-3.875	<.001	.61
	Caucasian	140	3.79	.424			
The instructor makes me feel good about myself.	Other Races	58	3.47	.777	-1.676	.095	.26
	Caucasian	140	3.64	.590			
The students in this course respect each other.	Other Races	58	3.66	.608	-2.428	.016	.38
	Caucasian	140	3.83	.378			
Students in this course like one another.	Other Races	58	3.40	.771	-2.930	.004	.46
	Caucasian	140	3.66	.489			
If I am absent, the instructor notices my absence.	Other Races	58	3.55	.680	-1.147	.253	.18
	Caucasian	140	3.66	.547			
I have many opportunities to be a part of class discussions and activities.	Other Races	58	3.53	.655	-3.260	.001	.51
	Caucasian	140	3.81	.478			

Table 2. Continued from Previous Page

I feel like I belong in this course.	Other Races	58	3.34	.849	-2.093	.038	.33
	Caucasian	140	3.59	.719			
I feel socially accepted in this course.	Other Races	58	3.57	.752	-1.274	.204	.20
	Caucasian	140	3.69	.561			
My instructor praises me when I work hard in this class.	Other Races	58	3.21	.874	-2.985	.003	.47
	Caucasian	140	3.55	.671			
My instructor gives me individual attention when I need it.	Other Races	58	3.53	.706	-3.562	<.001	.56
	Caucasian	140	3.81	.390			
The things I am learning in this course are important to me.	Other Races	58	3.18	.909	-1.360	.175	.21
	Caucasian	140	3.35	.777			
My instructor expects me to do my best at all times.	Other Races	58	3.71	.496	-3.356	<.001	.52
	Caucasian	140	3.90	.301			
I can communicate with my instructor about problems or questions I have in class.	Other Races	58	3.60	.647	-3.727	<.001	.58
	Caucasian	140	3.87	.357			
I feel like I belong at this institution.	Other Races	58	3.12	.993	-3.273	.001	.51
	Caucasian	140	3.54	.735			
Classroom Climate	Other Races	58	3.54	.504	-2.32	.011	.43
	Caucasian	140	3.71	.328			

When faculty do not care about building relationships and setting the climate for students, the result is students who do not care about faculty or the course (Hawk & Lyons, 2008). Students can become disinterested in a subject area due to the professor's lack of attention to classroom climate. Trust, respect, and communication develop with relationships, and relationships are essential to the collegiate teaching and learning experience (Supiano, 2018). It is not surprising students prefer courses with a supportive and positive classroom climate (Rosenfeld, 1983).

Research indicates building rapport with students is a fundamental characteristic of a successful instructor (Catt et al., 2007; Faranda & Clarke, 2004; Perkins et al., 1995), and engaging in rapport-building behaviors has been shown to positively influence students' opinions towards their instructor (Schrodt et al., 2006). To help build rapport and earn student trust, faculty must demonstrate care for students and their success. Students report faculty convey caring by expressing concern for how well they learn, creating an atmosphere that encourages expectations, and stimulating their interest in and enthusiasm for the subject (Hoyt & Eun-Joo, 2002). When faculty engage in interpersonal interactions that demonstrate students are valued and cared for, it has a profound impact on student outcomes (Barr, 2016; Noddings, 2005). Classroom interaction dynamics are central to instructional communication (Johnson, 2009). Recent research also shows the more often a student interacts with a faculty member, the more they feel the faculty member cares for them (Carr et al., 2021), and students believe they learn best when opportunities are provided for personal contact with the instructor (Wilkie, 2000). Therefore, faculty should consider scheduling time to conference with students, rather than waiting passively for students to reach out or come to office hours.

Additionally, rapport is built and a positive classroom climate is developed when faculty and students co-construct a learning environment that encourages active student participation (Sidelinger & Booth-Butterfield, 2010). Therefore, class norms and expectations should be discussed and agreed upon from the beginning. Further, faculty must implement practices that ensure equitable opportunities for engagement from all students. To promote interaction and participation, students should be encouraged to engage in one-on-one conversations, move to small group interactions, and eventually to whole group involvement (Sidelinger & Booth-Butterfield, 2010).

Student-student rapport may also play a critical role in classroom climate (Frisby & Martin, 2010). Evidence exists students desire positive classroom peer relationships and student-student communication influences students (Johnson, 2009). Dwyer et al. (2004) assert a connected classroom is one in which students perceive a supportive and cooperative communication environment between students allowing them to express themselves freely. Nadler & Nadler (1990) found in a supportive communication climate, "students felt more comfortable participating in class, disagreeing with instructors, and meeting with faculty outside of class" (p. 61). Although faculty cannot control student behavior, one way they can facilitate positive student-student interactions is by modeling supportive communication. Waldeck et al., (2001) encourage faculty to intentionally focus on how college students communicate and interact with peers and faculty inside and outside the classroom as a way to identify their needs and tendencies. Additionally, creating ample occasions for student-student interactions may help students feel more comfortable interacting with their peers and provide additional opportunities for collaboration and participation.

As faculty pursue ways to balance day-to-day responsibilities and sort through innovative ways to communicate and interact with students, Schmier (1997) reminds us to persist despite frustrations and setbacks, to tolerate feelings of vulnerability that sometimes occur when emotion is evident or addressed, and to focus more on students than on subject matter at times. This may require educators to deliver an appropriately differentiated program that considers individual differences and the challenges that some students may face (Whannell & Whannell, 2015). Ultimately, a student's positive experience and academic self-identity are more important than any content objective.

Limitations

This study has several limitations. Data were collected at a small, liberal arts college in the mid-Atlantic where faculty-to-student ratio is generally small. The data may be swayed because faculty voluntarily participated in professional development focused on more progressive and innovative teaching methods. A majority of students (70%) and professors (83%) in the study self-identify as White. Additionally, students completed the survey anonymously; therefore, this study was limited by its inability to further examine the reasons why students rated classroom climate factors as they did. Also, due to a small number of non-White students, all other races were combined into one simplistic racial category. Examining perceptions in differing arenas may produce different results. Finally, based on the nesting effect of our data, the intraclass correlation coefficient is likely to be high, which came as a result of grouping multiple sections together and likely led to high intraclass correlation.

Conclusion and Future Research

This research demonstrates the disparity of numerous classroom climate factors when comparing students by race. The differences reported are important in informing classroom practice because they may indicate a source of tension between individuals and within the classroom environment based on race previously unexamined. Addressing differences in perception could lead to better instruction and greater student satisfaction related to classroom interactions and environmental preferences. Sensitivity and responsiveness to students' preferences may lead to improved faculty-student communication, opportunities for more equitable engagement, and ultimately, student achievement.

Disparity in the perceptions between Whites and students of other races suggests students have differing experiences based on race. This could be based on subtle (or overt) faculty practices related to language, communication style, faculty-student interactions, accessibility, expectations, or assumptions. When students feel they are treated differently, the result can be a learning environment that hinders academic achievement. Faculty cannot assume they have created a positive classroom climate or that all students perceive the environment uniformly. They must examine their practices and solicit explicit feedback to determine students' experiences in their classes in order to improve students' satisfaction, ensure equitable opportunities for engagement, and improve their classroom climate for all students.

The findings from this research raise several questions that could be addressed in future research. One of the major issues to address is why students of other races perceive a classroom climate differently from White students. Further research that includes

follow-up qualitative interviews to determine why students rated classroom climate factors as they did and how race may affect their perceptions is warranted. Additionally, future studies could examine populations made up of different race ratios to determine if outcomes are the same. Considering one of the factors non-White students rated lower than White students was “I feel like I belong at this institution,” a study examining the link between classroom climate and overall college climate could be valuable in determining how the culture of the college and factors beyond the classroom affect perceptions within the classroom.

Overall, results of the current study suggest students of different races do not perceive the classroom environment uniformly. Exposing and discussing differences in perceptions is likely to increase positive academic outcomes for students. Faculty must work to determine student perceptions of their classroom environment to improve the climate for all students. These results support the contention by West (1994) and Reid and Radhakrishnan (2003) that race continues to matter on college campuses.

Appendices

[Appendix A: Differences in classroom climate based on gender](#)

[Appendix B: Differences in classroom climate based on classification](#)

Acknowledgments

We would like to thank each student and faculty member who participated in this study.

References

- Allen, W. R., & Haniff, N. Z. (1991). Race, gender, and academic performance in U.S. higher education. In W. R. Allen & E. G. Epps (Eds.), *College in Black and White: African American students in predominantly White and in historically Black public universities* (pp. 95–109). State University of New York Press.
- Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How learning works*. John Wiley & Sons.
- Anderson, L. E., & Carta-Falsa, J. (2002). Factors that make faculty and student relationships effective. *College Teaching*, 50(4), 134–138.
<https://doi.org/10.1080/87567550209595894>
- Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist*, 28, 117–148.
https://doi.org/10.1207/s15326985ep2802_3
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. Macmillan.
- Bandura, A. (2005). The evolution of social cognitive theory. In K. G. Smith & M. A. Hill (Eds.), *Great minds in management* (pp. 9–35). Oxford University Press.
- Barr, J. J. (2016, October 1). *Developing a positive classroom climate* (IDEA Research Report No. 61). https://www.ideaedu.org/idea_papers/developing-a-positive-classroom-climate/
- Beattie, I. R., & Thiele, M. (2016). Connecting in class? College class size and inequality in academic social capital. *The Journal of Higher Education*, 87(3), 332–362.
<https://doi.org/10.1353/jhe.2016.0017>
- Bosma, H., & Kunnen, W. (2001). Determinants and mechanisms in ego identity development. *Developmental Review*, 21(1), 39–66.
<https://doi.org/10.1006/drev.2000.0514>
- Bouville, M. (2010). Why is cheating wrong? *Studies in Philosophy and Education*, 29(1), 67–76. <https://doi.org/10.1007/s11217-009-9148-0>
- Brand, S., Felner, R. D., Seitsinger, A., Burns, A., & Bolton, N. (2008). A large-scale study of the assessment of the social environment of middle and secondary schools. *Journal of School Psychology*, 46, 507–535.
<https://doi.org/10.1016/j.jsp.2007.12.001>
- Byrne, D. B., Hattie, J. A., & Fraser, B. J. (1986). Student perceptions of preferred classroom learning environment. *The Journal of Educational Research*, 80(1), 10–18. <https://doi.org/10.1080/00220671.1986.10885714>
- Carr, J. M., Rogers, K. S., & Kanyongo, G. (2021). Improving student and faculty communication. *Research in Learning Technology*, 29.
<https://doi.org/10.25304/rlt.v29.2463>
- Catt, S., Miller, D., & Schallenkamp, K. (2007). You are the key: Communicate for learning effectiveness. *Education*, 127, 369–377.
- Choi, N. (2005). Self-efficacy and self-concept as predictors of college students' academic performance. *Psychology in Schools*, 42(2), 197–205.
<https://doi.org/10.1002/pits.20048>
- Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perceptions of classroom environment. *Journal of Educational Psychology*, 93, 43–54. <https://doi.org/10.1037/0022-0663.93.1.43>

- Couros, G. (2018, May 10). Don't add. Make better. *Medium*.
<https://gcouros.medium.com/dont-add-make-better-4bdeda74feb8>
- Cox, B. E., McIntosh, K. L., Terenzini, P. T., Reason, R. D., & Lutovsky Quayle, B. R. (2010). Pedagogical signals of faculty approachability. *Research in Higher Education*, 51(8), 767–788. <https://doi.org/10.1007/s11162-010-9178-z>
- Dwyer, K. K., Bingham, S. G., Carlson, R. E., Prisbell, M., Cruz, A. M., & Fus, D. A. (2004). Communication and connectedness in the classroom. *Communication Research Reports*, 21, 264–272. <https://doi.org/10.1080/08824090409359988>
- Ellis, K. (2004). The impact of perceived teacher confirmation. *Communication Education*, 53(1), 1–20. <https://doi.org/10.1080/0363452032000135742>
- Faranda, W. T., & Clarke, I. (2004). Student observation of outstanding teaching. *Journal of Marketing Education*, 26, 271–281.
<https://doi.org/10.1177/0273475304268782>
- Fassinger, P. A. (2000). How classes influence student participation. *Journal of Classroom Interaction*, 35, 38–47.
- Fertig, J. (2012, March 12). How do professors learn to teach (or do they)?
<https://www.jamesgmartin.center/2012/03/how-do-professors-learn-to-teach-or-do-they/>
- Fraser, B. J. (1993). Classroom environments in post-compulsory education. *Youth Studies in Australia*, 12(4), 41–46.
- Fraser, B. J. (1994). Research on classroom and school climate. In D. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 493–541). Macmillan.
- Fraser, B. J. (2000). Instructional environment. In A. E. Kazdin (Ed.), *Encyclopedia of psychology* (Vol. 4, pp. 308–311). American Psychological Association.
- Fraser, B. J., & Fisher, D. L. (1982). Predicting students' outcomes. *American Educational Research Journal*, 19(4), 498–518.
<https://doi.org/10.2307/1162539>
- Fraser, B. J., Treagust, D. F., & Dennis, N. C. (1986). Development of an instrument for assessing classroom psychosocial environment. *Studies in Higher Education*, 11, 43–54. <https://doi.org/10.1080/03075078612331378451>
- Freeman, T. M., Anderman, L. H., & Jensen, J. J. (2007). Sense of belonging in college freshmen. *Journal of Experimental Education*, 75(3), 203–220.
<https://doi.org/10.3200/jexe.75.3.203-220>
- Frisby, B. N., & Martin, M. M. (2010). Instructor-student rapport. *Communication Education*, 59(2), 146–164. <https://doi.org/10.1080/03634520903564362>
- Frymier, A. B. (2007). Teachers' and students' goals. Paper presented at NCA, Chicago, IL.
- Ghanatabadi, J. L. (1991). *Relationship of perceived classroom social climate and course performance* (Doctoral dissertation). *Dissertation Abstracts International*, 52, 3145A.
- Gizir, S. (2019). The sense of classroom belonging. *European Journal of Educational Research*, 8(1), 87–97. <https://doi.org/10.12973/eu-jer.8.1.87>
- Graham, W. W., & Gisi, S. L. (2000). Effects of instructional climate. *Journal of College Student Development*, 41, 279–291.

- Hadley, T., & Graham, J. W. (1987). Cognitive development and environmental press. *Journal of College Student Personnel*, 28, 388–394.
- Haertel, G. D., Walberg, H. J., & Haertel, E. H. (1981). Socio-psychological environments. *British Educational Research Journal*, 7(1), 27–36.
<https://doi.org/10.1080/0141192810070103>
- Hall, R. M., & Sandler, B. R. (1982). *The classroom climate: A chilly one for women?* Association of American Colleges.
- Hawk, T. F., & Lyons, P. R. (2008). Please don't give up on me. *Journal of Management Education*, 32(3), 316–338. <https://doi.org/10.1177/1052562908314194>
- Hirshy, A. E., & Wilson, M. E. (2002). Sociology of the classroom. *Peabody Journal of Education*, 77(3), 85–100. https://doi.org/10.1207/s15327930pje7703_5
- Hoyt, D. P., & Eun-Joo, L. (2002). *Teaching "styles" and learning outcomes*. IDEA Research Report 4.
- Huang, W., Hurt, A., Richardson, J. C., Swan, K., & Caskurlu, S. (2018). *Community of inquiry framework*. Purdue Repository.
- Hurtado, S. (1994). Graduate school racial climates. *American Journal of Education*, 102, 330–351. <https://doi.org/10.1086/444072>
- Hurtado, S., & Carter, D. F. (1997). Latino students' sense of belonging. *Sociology of Education*, 70, 324–345. <https://doi.org/10.2307/2673270>
- Johnson, D. I. (2009). Connected classroom climate. *Communication Research Reports*, 26, 146–157. <https://doi.org/10.1080/08824090902861622>
- Kozlowski, S. W. J., & Klein, K. J. (2000). Multilevel approaches. In *Multilevel theory, research, and methods in organizations* (pp. 3–90). Jossey-Bass.
- Kuh, G. D. (2010). *Student success in college*. Jossey-Bass.
- Laird, T. F. N., & Cruce, T. M. (2009). Part-time enrollment effects. *Journal of Higher Education*, 80(3), 290–314. <https://doi.org/10.1080/00221546.2009.11779014>
- Lee, R. M., & Robbins, S. B. (1995). Measuring belongingness. *Journal of Counseling Psychology*, 42, 232–241. <https://doi.org/10.1037/0022-0167.42.2.232>
- Letizia, D. M. (1989). *Effects of age-integration on the social climate of nursing school classes*. Dissertation Abstracts International, 50, 3825A.
- Lewin, K., Lippett, R., & White, R. K. (1939). Patterns of aggressive behavior. *Journal of Social Psychology*, 10, 271–299.
<https://doi.org/10.1080/00224545.1939.9713366>
- Lubke, G. H., & Muthén, B. O. (2004). Multigroup CFA on Likert data. *Structural Equation Modeling*, 11, 514–534. https://doi.org/10.1207/s15328007sem1104_2
- Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Learning environments. *Contemporary Educational Psychology*, 34(2), 120–131.
<https://doi.org/10.1016/j.cedpsych.2008.12.001>
- Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Student ratings reliability. *Learning Environments Research*, 9(3), 215–230.
<https://doi.org/10.1007/s10984-006-9014-8>
- Marsh, H., & Yeung, A. (1997). Academic self-concept. *Journal of Educational Psychology*, 89(1), 41–54. <https://doi.org/10.1037/0022-0663.89.1.41>
- Maxwell, S., Reynolds, K. J., Lee, E., Subasic, E., & Bromhead, D. (2017). School climate and achievement. *Frontiers in Psychology*, 8, 2069.
<https://doi.org/10.3389/fpsyg.2017.02069>

- Meyers, S. A. (2009). Do students care whether you care? *College Teaching*, 57(4), 205–210. <https://doi.org/10.3200/ctch.57.4.205-210>
- Miller, M. (2005). Affective domain. In M. Orey (Ed.), *Emerging perspectives on learning, teaching, and technology*.
- Moos, R. H. (1979). *Evaluating educational environments*. Jossey-Bass.
- Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, F. L. (2014). Doubly latent multilevel analyses. *Journal of Experimental Education*, 82(2), 143–167. <https://doi.org/10.1080/00220973.2013.769412>
- Myers, S. A. (1995). Affinity-seeking and classroom climate. *Communication Research Reports*, 12, 192–199. <https://doi.org/10.1080/08824099509362056>
- Nadler, L. B., & Nadler, M. K. (1990). Sex differences in classroom communication. *Women's Studies in Communication*, 13, 46–65. <https://doi.org/10.1080/07491409.1990.11089740>
- National Center for Education Statistics. (2015). *EDSCLS: National benchmark study 2016*.
- National Center for Education Statistics. (2016). *Education Department School Climate Survey (EDSCLS)*.
- Noddings, N. (2005). *The challenge to care in schools* (2nd ed.). Teachers College Press.
- Osterman, K. F. (2000). Students' need for belonging. *Review of Educational Research*, 70(3), 323–367. <https://doi.org/10.2307/1170786>
- Perkins, D., Schenk, T. A., Stephan, L., Vrungos, S., & Wynants, S. (1995). Rapport and learning. *Psychological Reports*, 76, 627–635. <https://doi.org/10.2466/pro.1995.76.2.627>
- Piaget, J. (1965). *The moral judgment of the child*. Free Press.
- Purkey, W. (1988). *An overview of self-concept theory for counselors*. ERIC.
- Reid, L. D., & Radhakrishnan, P. (2003). Race and campus climate. *Cultural Diversity and Ethnic Minority Psychology*, 9(3), 263–275. <https://doi.org/10.1037/1099-9809.9.3.263>
- Reynolds, W. (1988). Academic self-concept. *Journal of Personality Assessment*, 52(2), 223–240. https://doi.org/10.1207/s15327752jpa5202_4
- Richmond, A. S., Berglund, M. B., Epelbaum, V. B., & Kelin, E. M. (2015). Rapport and student engagement. *Teaching of Psychology*, 42(2), 119–125.
- Rosenfeld, L. B. (1983). Communication climate and coping. *Communication Education*, 32, 167–177. <https://doi.org/10.1080/03634528309378526>
- Sadker, D. (2000). Gender equity. *Equity & Excellence in Education*, 33(1), 80–83.
- Sadker, M., Sadker, D., & Klein, S. (1991). Gender in education. In G. Grant (Ed.), *Review of research in education* (pp. 269–334). AERA.
- Schmier, L. (1997). Teaching is love. <http://www.halcyon.com/arborhts/rt/>
- Schrodt, P., Turman, P., & Soliz, J. (2006). Perceived understanding. *Communication Education*, 55, 370–388.
- Sidelinger, R. J., & Booth-Butterfield, M. (2010). Student involvement. *Communication Education*, 59(2), 165–184.
- Steele, C. M. (1997). A threat in the air. *American Psychologist*, 52, 613–629.
- Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Lawrence Erlbaum.

- Supiano, B. (2018, January 19). Relationships and student experience. *Chronicle of Higher Education*, 1.
- Tajfel, H., & Turner, J. (1979). Intergroup conflict theory. In *The social psychology of intergroup relations* (pp. 33–47). Brooks & Cole.
- Terry, D., Hogg, M., & White, K. (1999). Planned behavior and group norms. *British Journal of Social Psychology*, 38(3), 225–244.
- Trujillo, C. M. (1986). Classroom interactions. *American Educational Research Journal*, 23(4), 629–642.
- Vygotsky, L. S. (1962). *Thought and language*. MIT Press. (Original work published 1934)
- Walberg, H. J. (1969). Social environment and learning. *Journal of Educational Psychology*, 60(6), 443–448.
- Waldeck, J. H., Kearney, P., & Plax, T. G. (2001). Instructional communication theory. In *Communication yearbook 24* (pp. 207–230). Sage.
- West, C. (1994). *Race matters*. Vintage.
- Whannell, R., & Whannell, P. (2015). Identity theory and attrition. *Student Success*, 6(2), 43–52.
- Wilkie, C. J. (2000). Preferred college classroom environment scale. *Journal of the First-Year Experience*, 12(2), 7–32.
- Wilson, J. H., Ryan, R. G., & Pugh, J. L. (2010). Professor-student rapport scale. *Teaching of Psychology*, 37, 246–251.
- Winston, R. B., Jr., Vahala, M. E., Nichols, E. C., Gillis, M. E., Wintrow, M., & Rome, K. D. (1988). College classroom environment scales. *Journal of College Student Development*, 35, 12–18.

Evaluating Solutions for Improving Program Assessment



Authors:

Nathan C. Anderson, Ph.D.
Minot State University

Alaric Williams, Ed.D.
Minot State University

Daniel Ringrose, Ph.D.
Minot State University

ABSTRACT

This paper demonstrates how a small, public, master's level university in the midwestern United States identified and responded to areas for improvement related to their implementation of program assessment. An overview of a program evaluation framework is provided based on five categories of assessment that the university uses as a frame of reference to guide its program development and assessment work. The university's methods are described for assessing needs to identify areas for program assessment improvement, assessing the theory of potential solutions to ensure that they would be appropriate to meet demonstrated needs, and assessing the efficiency of suitable solutions to ensure that they could be implemented within the parameters of the institution's available resources. Results are discussed through the lens of a newly proposed model, ALTMAP, which synthesizes established, practical, and transferable frameworks and tools to collectively support a comprehensive approach to program planning and evaluation.

Correspondence E-mail: nathan.c.anderson@minotstateu.edu

Keywords: program evaluation, program assessment, needs assessment, theory assessment, efficiency assessment

In a society that often encourages the use of quality evidence to guide important decisions (e.g., Stufflebeam & Coryn, 2014; Rossi et al., 2004; Yarbrough et al., 2011), higher education institutions generate evidence through ongoing program assessment to inform decisions that promote student development (Allen, 2003; Banta & Palomba, 2015; Miller & Leskes, 2005; Tucker et al., 2020). Effective assessment can help a program understand whether it is being implemented as intended, its intended outcomes are being achieved, and its resources are utilized responsibly (Banta & Palomba, 2015). Implementing an assessment model can be challenging, however, as several tasks are required to ensure it is feasible and produces meaningful results (Banta, 2002; Hutchings, 2019). Additional barriers may result from inconsistent interpretations of assessment purposes, definitions, and methods (Ewell, 2002), including pessimistic perspectives that potentially impede the effectiveness of its implementation (Greene, 2023). Evaluating program assessment practices and systems at an institution can help identify and alleviate these and other types of barriers to ensure assessment provides evidence for logical decisions that support high-quality, relevant learning opportunities for students.

Purpose

This study demonstrates how a small, public, master's level university in the Midwest operationalized three categories of assessment within a comprehensive program evaluation framework to generate and use multiple forms of evidence to improve their program assessment system. It provides an overview of a comprehensive evaluation framework comprising five assessment categories related to program needs, theory, processes, outcomes, and efficiency. Then, it explains how the university applied needs, theory, and efficiency assessment methods to generate evidence related to areas for improvement to their program assessment system, potential solutions that would be appropriate to meet their needs, and available resources that could be utilized to implement the solutions. The goal was to develop a common model that would be appropriate to the needs of diverse programs and to identify available resources that could be utilized to implement possible programmatic changes. The study is guided by three research questions:

- RQ1: What are the university's needs related to program assessment?
- RQ2: What solutions could be appropriate responses to the needs?
- RQ3: Which of the appropriate solutions would be the most feasible to implement within the parameters of the university's existing resources?

RQ1 addresses the needs assessment category of assessment. RQ2 addresses the theory assessment category. RQ3 addresses the efficiency assessment category.

Significance

Effective program assessment is essential for universities to continually provide relevant, high quality academic programs to their students. However, assessment can sometimes seem burdensome because it requires substantial allocation of time, attention, energy, and resources toward the generation and utilization of useful evidence related to programmatic needs, design, implementation, outcomes, and resources. Assessment

systems can help alleviate some of the burdens by leveraging technology to improve the efficiency of processes related to planning and reporting.

Several assessment system options are available for universities to purchase or build on their own. Many factors are important to consider when identifying which system will be the best fit for the institution. Considering potential assessment systems through a comprehensive evaluative lens can help a university choose a relevant system that addresses their specific needs and is feasible to implement within the scope of its available resources.

Theoretical Framework

Evaluation may be implemented as a systematic means of generating evidence to guide improvements to something and to inform a judgment on the merit or worth of something (Fitzpatrick et al., 2011; Stufflebeam & Coryn, 2014; Yarbrough et al., 2011). This study is guided by Anderson's (2022) *Program Evaluation Framework*, which describes where program development, implementation, assessment, and reflection phases may occur within a program's life cycle. The framework subsumes five categories of assessment: needs, theory, process, outcome, and efficiency (Figure 1). The assessment categories are directly informed by Rossi et al.'s (2004) five types of assessment questions. The categories also align with common elements of logic models (e.g., Kekahio et al., 2014; W.K. Kellogg Foundation, 2004; Wyatt Knowlton & Phillips, 2013) and Banta and Palomba's (2015) definition of assessment: "the process of providing credible evidence of resources, implementation actions, and outcomes undertaken for the purpose of improving the effectiveness of instruction, programs and services in higher education" (p. 2). Table 1 depicts alignment between the five assessment categories in Anderson's (2022) *Program Evaluation Framework*, five types of assessment questions (Rossi et al., 2004), common elements of logic models, and Banta and Palomba's (2015) definition of assessment.

Figure 1. *Five Categories of Assessment for Comprehensive Program Evaluation*



Table 1. Assessment Category Alignment

Anderson's (2022) Program Evaluation Framework	Rossi et al.'s (2004) five types of assessment questions	Common elements of logic models	Banta and Palomba's (2015) definition of assessment
Needs assessment	Needs assessment	Needs	for the purpose of improving the effectiveness of instruction, programs, and services
Theory assessment	Assessment of program theory	Activities	
Process assessment	Assessment of program process	Outputs	credible evidence of implementation actions
Outcome assessment	Impact assessment	Outcomes/ Impact	credible evidence of outcomes
Efficiency assessment	Efficiency assessment	Resources	credible evidence of resources

A needs assessment can help reveal gaps to provide rationale for implementing a program (Fitzpatrick et al., 2011; McCawley, 2009; Witkin & Altshuld, 1995). It reflects “a series of procedures for identifying and describing both present and desired states in a specific context” (Witkin & Altshuld, 1995, p. 4). Needs assessments focus on “(a) establishing whether a problem or need exists and describing that problem, and (b) making recommendations for ways to reduce the problem” (Fitzpatrick et al., 2011, p. 26). They are “often conducted for a specific group, organization, or business in order to improve effectiveness or productivity of the group related to its mission” (McCawley, 2009, p. 4). A couple of questions that may be considered when assessing needs include “What are the needs of the population?” and “What are the nature and magnitude of the problem to be addressed” (Rossi et al., 2004, p. 77).

A theory assessment can help reveal which potential solutions may be appropriate responses to the identified needs (Calley, 2011; Halpern, 1998; Main, 2011; Rossi et al., 2004). “The theory of a program is the set of imputed links between program operations and program results” (Halpern, 1998, p. 6). Assessing theory is “necessary to establish an empirical basis for the program” (Calley, 2011, p. 13). “When a program design is rooted in a strong empirically-supported theory, it increases the probability that the desired outcomes will occur for clients and that the program will be the most obvious explanation for the achievement of those outcomes” (Main, 2011, p. 13). A few questions that may be considered when assessing program theory include, “What services should be provided? What are the best delivery options for the services?” and “How should the program be organized?” (Rossi et al., 2004, p. 77).

A process assessment can help reveal whether a selected solution is being implemented as intended (Berkel et al., 2011; Durlak & DuPre, 2008; Saunders et al., 2005). “Different dimensions of implementation influence program outcomes” (Berkel et al., 2011, p. 29), and accurately interpreting a program’s outcomes requires knowing which of its aspects “were delivered and how well they were conducted” (Durlak & DuPre, 2008,

p. 328). Therefore, process assessment “is used to monitor and document program implementation and can aid in understanding the relationship between specific program elements and program outcomes” (Saunders et al., 2005, p. 134). A few questions that may be considered when assessing program processes include “Are administrative and service objectives being met? Are the intended services being delivered to the intended persons?” and “Are there needy but underserved persons the program is not reaching” (Rossi et al., 2004, p. 77).

An outcome assessment can help reveal whether an implemented solution is having its desired effects (Kekahio et al., 2014, p.2; Main, 2011; Rossi et al., 2004; U.S. Government Accountability Office, 2011). Outcome assessments focus on “the extent to which a program achieves its outcome-oriented objectives” (U.S. Government Accountability Office, 2011, p. 2). Outcomes may include “changes in program participants’ knowledge, beliefs, and behavior,” “adjustments in organizational practices or system design,” or “a program’s lasting influences” (Kekahio et al., 2014). If a program is not designed with outcomes in mind, it becomes difficult for a program to prove its effectiveness” (Main, 2011, p. 13). A few questions that may be considered when assessing program outcomes include “Do the services have beneficial effects on the recipients? Do the services have adverse side effects on the recipients?” and “Is the problem or situation the services are intended to address made better?” (Rossi et al., 2004, p. 77).

An efficiency assessment can help reveal whether resources for implementing the solution are being utilized responsibly (Levin & Belfield, 2015; Rossi et al., 2004; U.S. Government Accountability Office, 2011; Yates, 2009). Efficiency assessments can help “decision makers select the most effective interventions they can afford for a given budget, or the least costly interventions that result in a desired outcome” (Levin & Belfield, 2015, p. 401). An efficiency assessment may occur in the form of cost-benefit and cost-effectiveness analyses, which “compare a program’s outputs or outcomes with the costs (resources expended) to produce them” (U.S. Government Accountability Office, 2011, p. 2). When conducting an efficiency assessment, “the value of resources used to implement a program may be compared to the value or amount of outcomes for that program” (Yates, 2009, p. 54). A few questions that may be considered when assessing program efficiency include “Are resources being used efficiently? Is the cost reasonable in relation to the magnitude of the benefits?” and “Would alternative approaches yield equivalent benefits at less cost” (Rossi et al., 2004, p. 77).

The five categories of assessment can support evaluation of a program, project, or other evaluand by collectively informing its improvements or judgments to its merit or worth. For example, if there is a need for a program (needs assessment), the program is theoretically appropriate to meet the need (theory assessment), the program is being implemented as intended (process assessment), the program is achieving the desired results (outcome assessment), and the program is operating responsibly within the parameters of its available resources (efficiency assessment), then the program could likely be deemed to have adequate merit or worth to warrant its implementation. If any of the inverses are true (i.e., there is not a need for a program, the program is not theoretically appropriate to meet a need, the program is not implemented as intended, the program is not producing the desired outcomes, or the program is not operating efficiently), then its merit or worth is likely to be weakened. This study specifically addresses the needs, theory,

and efficiency categories of assessment to help inform a judgment on the merit or worth of a new assessment system.

Method

This section describes methods related to the three research questions and their corresponding assessment categories. Methods related to the first research question reflect needs assessment processes. Methods related to the second research question reflect theory assessment processes. Methods related to the third research question reflect efficiency assessment processes.

Needs Assessment Methods

The university's Director of Institutional Assessment (DIA) facilitated a mixed methods needs assessment to gain a better understanding of strengths and weaknesses related to its annual program assessment procedures as well as recommendations for improvements. Methods included interviewing department chairs, engaging with assessment leaders on campus, analyzing internal and external content related to assessment, and surveying faculty.

The needs assessment was initially conducted for the university's internal purposes to support its continuous improvement efforts. After the needs assessment methods were implemented and synthesized with other program evaluation methods (theory assessment and efficiency assessment), the authors realized that adapted versions of the methods may be useful in other settings and, therefore, would be worthwhile to disseminate via scholarly channels. The university's Institutional Review Board reviewed the needs assessment study and approved it with exempt status (protocol number 2463).

Department Chair Interviews

The DIA conducted one-on-one interviews with fifteen Department Chairs representing all but one department at the university over a two-week period during September 2021. The purpose of the interviews was to gain a better understanding of their perceptions related to strengths and weaknesses of program assessment as well as recommendations for improving program assessment. Each interview lasted around 30-60 minutes. The DIA recruited participants for the interviews by sending individual emails to each chair. The interviews were guided by a semi-structured script that allowed for relaxed, conversational dialogue with probing questions as needed (Adams, 2015). The semi-structured interview script included the following questions:

- What is your approach to program assessment?
- Which attributes of the assessment process seem to be working well?
- Is there anything about the assessment process that is unclear?
- Is there anything about the assessment process that frustrates you?
- Do you have any recommendations for changes to the assessment process?
- How can I support the assessment process for your program?

The DIA collected qualitative data by collecting field notes (Phillippi & Lauderdale, 2017) during the interviews and then compiling and analyzing more detailed notes in a Microsoft Word document after the interviews concluded.

Assessment Leadership Observations

The DIA observed assessment leaders at the university through formal and informal interactions between September 2021 and January 2022. Individuals were not formally recruited for participation, as opportunities for observation naturally presented themselves through everyday interactions. Assessment leaders observed included current and former members of the Academic Assessment and General Education Committees, the former Director of Academic Assessment (DAA), and the Vice President of Academic Affairs (VPAA), among others. Formal ways of engaging with assessment leaders included participation in Academic Assessment and General Education Committee meetings. Informal ways of engagement included email, phone, virtual, and in-person correspondence. The DIA collected qualitative data by collecting field notes (Phillippi & Lauderdale, 2017) related to program assessment strengths, weaknesses, and improvement strategies as appropriate during the interactions and later reflected on the notes in greater detail.

Content Analysis

The DIA implemented content analysis techniques between September 2021 and January 2022 by compiling, reading, and interpreting several documents to draw inferences in relation to the research questions (Krippendorff, 2004). He utilized a two-pronged approach to analyzing content: the first prong focused on reviewing internal documentation relevant to academic assessment, while the second prong focused on reviewing external documentation relevant to academic assessment. The purpose of analyzing internal and external resources was to gain a better understanding of whether assessment guidance strategies implemented or recommended by other professionals or organizations may be appropriate to consider at the institution where he worked to improve the ways in which academic program assessment guidance is organized and disseminated internally. Internal content analysis included reviews of several resources, such as institutional webpages, informational documents, assessment reports, and other resources relevant to assessment (Table 2).

External content analysis included reviews of program assessment manuals from higher education institutions, assessment guides prepared by professional organizations, journal articles, reports, videos, podcasts, and other sources to gain a better understanding of guidance that is available in the field of higher education program assessment as well as how guidance may be organized.

Survey

The university's Academic Assessment Committee recruited participants using a convenience sampling strategy (Cohen et al., 2017) by inviting faculty to complete a survey at the institution's annual spring assessment day in February 2022 to gain a better understanding of the extent to which faculty perceive potential areas for improvement to academic program assessment to be priorities. The potential areas for improvement were based on an initial list of areas revealed through the department chair interviews, assessment leadership engagement, and internal and external content analysis. The survey asked: "In terms of academic yearly program assessment at the university, mark the extent to which you consider making improvements to each of the following six areas to be a

Table 2. *Examples of Resources for Internal Content Analysis*

Type of Source	Name of Source
Institutional Webpages	Academic Assessment Academic Assessment Liaisons Assessment of General Education
Informational Documents	Academic Assessment Policy and Procedure Assessment Terminology Academic Assessment Calendar Assessment 101 Basics Assessment Planning Diagnostic Arranging Assessment of Student Learning Outcomes University Student Learning Goals
Assessment Reports	Yearly Program Assessment plans and reports Ad Hoc Student Learning Assessment Committee report General Education Assessment report
Other Resources	Assessment planning and reporting templates Minutes of assessment-related committees

priority: template consistency, efficiency, relevance, comprehensiveness, institutional memory, assessment guidance.” Participants rated each priority area on a seven-point scale ranging from “Not a Priority” to “Essential Priority.” An online version of the survey was created in Microsoft Forms. Faculty were recruited to participate in the survey in person toward the conclusion of an academic assessment update presentation delivered by the DIA to faculty on the university’s spring assessment day. The survey remained open through March 2022. Twenty-five of the institution’s 150 (17%) full-time faculty members representing 14 of its 16 academic departments participated in the survey. It should be noted that the low response rate and convenience sampling technique illuminate non-response and self-selection biases that weaken the generalizability of the survey results (Groves et al., 2009).

Theory Assessment Methods

After the university established needs related to program assessment, the DIA—along with support from the institution’s Vice President for Academic Affairs and members of assessment-related committees—implemented theory assessment strategies to identify potential solutions that could help respond to the needs. The team identified a few vendor-based program assessment systems for consideration. Then, they explored the features and functionality of each system in relation to the identified areas for improvement and the university’s yearly program assessment (YPA) requirements, which reflect a program’s annual documentation of its assessment plan, results, and utilization of results. This exploration included meeting with vendors, viewing product demonstrations by vendors, meeting with end-users of systems, and viewing product training videos, manuals, and other related resources. As the team explored potential solutions, they identified attributes of the systems that could support the university’s YPA planning and reporting requirements and help respond to the areas for improvement identified through the needs assessment. For example, specific areas in the system were

identified where the following student learning and operational assessment details from their MS Word YPA template could be entered:

- Goals
- Outcomes
- Methods
- Target
- Gather, review, and discuss
- Recommendations and implementation

Efficiency Assessment Methods

After the university established a list of potential solutions that aligned with the institution's program assessment requirements and could help respond to identified needs, officials implemented efficiency assessment strategies to identify which solution would be the most feasible to implement within the parameters of their available resources. In their exploration of efficiency, they examined costs of product implementation, networks of users within the state's university system, existing contracts between vendors and the state's university system, and technical assistance available through the vendor and the state's university system.

Results

This section presents the results of the needs, theory, and efficiency assessments.

Needs Assessment Results

The needs assessment results revealed evidence of program assessment areas that are working well, areas that could be improved, and recommendations for strategies that could potentially support improvements.

Working Well

Results revealed that several improvements had been made to the university's culture of academic assessment in recent years. The institution had made substantial strides to advance support for program assessment and the relevance of assessment. Examples of support include common templates, guidance documents, and technical assistance for YPA requirements and processes. Establishing a common language and a YPA template that was the same for all academic programs had been considered helpful. Feedback from some Department Chairs indicated the template clearly outlined the requirements of details to be provided in the report and plan. They expressed gratitude for having a YPA document that allowed them to write the report and plan in a single file, the prompts for documenting their closure of the assessment loop by describing decisions that had been informed by assessment results, and efforts that had been made to simplify the YPA template. Previous Academic Assessment Directors worked extensively on cultivating a collective energy across the institution that tended to understand the value of assessment. In addition to providing templates, guidance documents, and technical assistance, they created sustainable opportunities for the broader campus community to inform the development and refinement of academic program assessment methods, which led to the establishment and implementation of the Academic Assessment Liaisons group and Student Learning Assessment Committee (SLAC). Interviews with department chairs

revealed that the YPA process seemed to be much more intentional than it was in the past. For accredited programs, the details provided on the YPA tended to align well with the assessment requirements of their accrediting bodies. Some feedback suggested that the annual program assessment process used to feel like checking a box, but it now seems to serve a relevant purpose.

Areas for Improvement

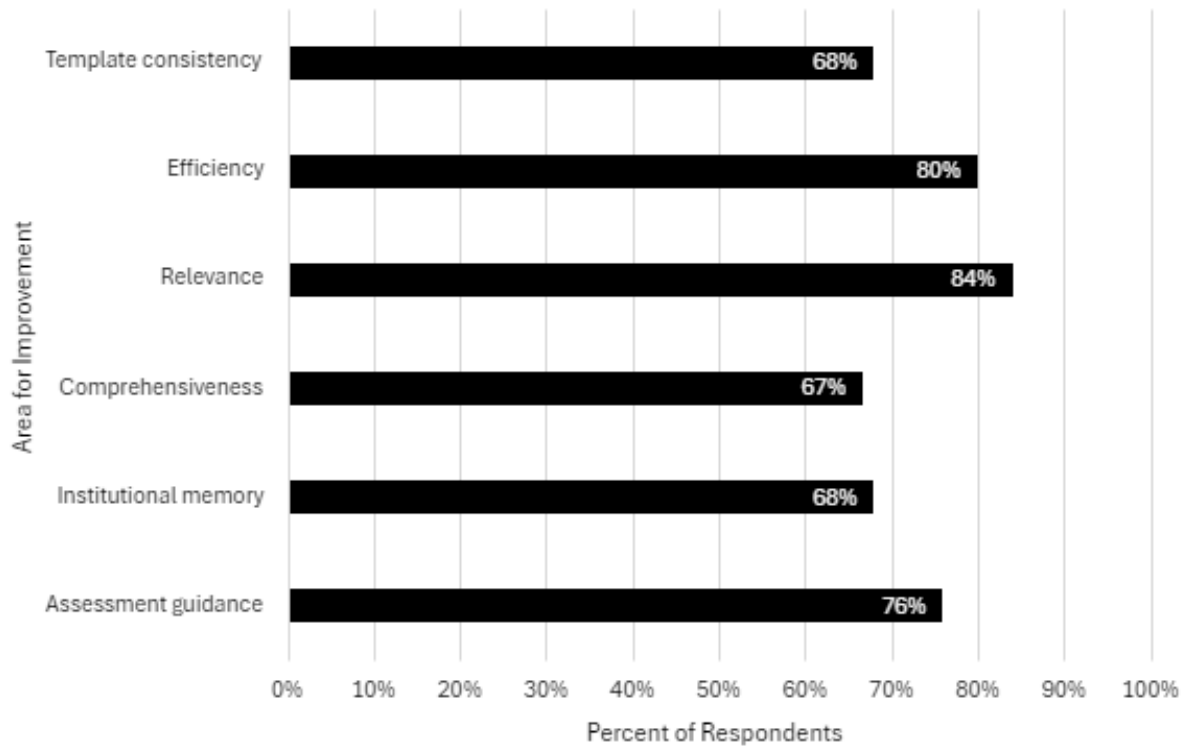
Results of the Department Chair interviews, assessment leadership observations, and internal and external content analysis revealed potential areas for improvement through themes of template consistency, efficiency, relevance, comprehensiveness, institutional memory, and assessment guidance. The survey findings ($n=25$) helped validate these results as each item was considered one of the three highest priority levels (moderate, high, or essential) on a seven-point priority-level scale by at least two-thirds of the respondents (Figure 2).

Template Consistency. Template consistency refers to the uniformity of the YPA template format requirements across time. Use of a common template for YPA reporting and planning has worked well for many programs; however, it has changed multiple times over the past few years. These changes were implemented as needed based on feedback from faculty, but frequent revisions have led to some frustration and confusion about which template is the correct version and whether reporting and planning requirements have changed from year to year.

Efficiency. Efficiency refers to the amount of time, energy, and resources required to prepare, submit, and organize YPAs. In terms of preparing YPAs, locating YPAs and supplemental documentation from prior years can be cumbersome, and copying and pasting details from the prior year to the current year is monotonous. In terms of submitting and organizing the YPAs, the current process requires authors to email the YPAs to their Chair. Then, the Chair reviews and emails the YPAs to the DIA to be uploaded to SharePoint folders that the DIA creates each year. The DIA replies to Chairs via email to confirm receipt of the files. If changes are made to any of the files, similar steps are taken to submit and organize the files. This is effectively a manual and rather technical process that relies on multiple people (e.g., Department Chairs, DIA) executing it correctly every year. While free, the cost lies in incomplete reports, lost files, use of email as a transfer mechanism, and reliance on many individuals to know the system and execute it correctly every year. Other institutions have implemented program assessment information systems to increase efficiency by streamlining the preparation, submission, and organization of annual reports and plans.

Relevance. Relevance refers to the relationship between YPAs and new knowledge or decisions that support continuous improvement of a program. The relevance of program assessment can be unclear at times. For example, faculty do not always understand how to analyze the data they collect, connections between assessment results and budgetary decisions sometimes seem vague, and some programs struggle with quantifying learning in meaningful ways (e.g., establishing performance benchmarks based on criteria for success related to what they intend for students to learn and then identifying how many students demonstrated the success criteria). Furthermore, although

Figure 2. Moderate, High, or Essential Priority by Improvement Area



a balanced program assessment strategy includes both quantitative and qualitative methods, some faculty seem to perceive there to be an overemphasis on quantitative measures that impedes their opportunity to address critical qualitative measures.

Comprehensiveness. Comprehensiveness refers to the thoroughness of the methods, targets, and results for each outcome, including the instrument title, type of measure (e.g., locally developed exam, standardized exam, behavioral intervention, focus group, survey, portfolio, etc.), indication of whether the measure is direct or indirect, courses associated with each measure, criteria for success on each measure, quantifiable target relevant to each criterion, actual results compared to targets, and other related details. The comprehensiveness of information provided in the reports and plans varies. Although the template includes sections for describing methods, targets, and results, some programs do not always sufficiently address all required areas.

Institutional Memory. Institutional memory refers to the accessibility of knowledge generated through YPAs and the transferability of knowledge to new employees. The university has been using SharePoint to compile and organize YPAs. Navigating the system to retrieve historic reports, plans, and other related information, as well as compare YPA information between years, is not always intuitive or feasible. In addition to presenting challenges related to the university's operational uses of the resources, inadequate institutional memory has a significant hidden cost when preparing materials for external accreditation review.

Assessment Guidance. Assessment guidance refers to technical assistance and instructional resources for writing mission, goal, and outcome statements; describing

methods, targets, and results; preparing and submitting YPAs; and implementing other related tasks. Program assessment guidance exists, but it is often siloed in separate files and, consequently, may be challenging to locate and review. Other institutions have developed manuals as central access points to improve end-user navigation of resources with definitions, templates, examples, and other assessment-related guidance and tools. Feedback from chairs and other faculty suggests that additional guidance may need to be developed on clarifying assessment language, condensing data into reportable metrics, establishing targets, writing operational goals, completing the YPA forms, and clarifying due dates.

Recommendations

Transitioning to an online assessment system for preparing and submitting YPAs could support improvements in consistency, efficiency, relevance, comprehensiveness, institutional memory, and assessment guidance. Recommendations stemming from the identified potential areas for improvement included: creation of a YPA dashboard as a single point of access to efficiently retrieve YPA plans and reports; development of an academic program assessment manual; and transition to a vendor-based online program assessment system for preparing and submitting YPA plans and reports. A YPA dashboard should improve institutional memory by facilitating efficient access to historical files. The program assessment manual will clarify program assessment guidelines and improve consistency.

Theory Assessment Results

As the university explored potential assessment systems, assessment leadership considered the congruency between the features of the systems and the university's YPA requirements as well as the areas for improvement identified in the needs assessment. The DIA conducted a detailed alignment between the university's YPA MS Word template and three assessment systems and shared the results with members of its assessment-related committees. All three systems included areas for entering information that would otherwise be written in the YPA MS Word template. A preliminary review of other systems suggested that they would also be likely to include areas for entering information required by the university's template.

All the assessment systems they reviewed seemed as if they could help alleviate, to at least some extent, the program assessment needs related to consistency, efficiency, comprehensiveness, relevance, institutional memory, and assessment guidance. The consistency of the reporting process would be likely to improve because the interface of an online assessment system would be less likely to change from year to year than the structure of a MS Word template. Efficiency would be likely to improve because information would be entered directly into the system and there would be automated workflows to facilitate submission, review, and approval processes. Comprehensiveness of information would be likely to improve because data entry fields in an established online system would help ensure that required details are provided for each outcome. Relevance would be likely to improve because there would be a common format for reporting targets and actual values that would facilitate answers to the following key questions that were not possible to efficiently answer at various levels within the institution using its MS Word template reporting solution:

- To what extent are targets being met?
- Which areas of learning appear to be the strongest and weakest?
- Which areas of learning appear to be showing positive, negative, or neutral trends across time?
- Which areas of learning appear to be meeting and not meeting targets?

Reporting target and actual values as percentages of students meeting the criteria of success in an online assessment system could facilitate efficient answers to these questions at course, outcome, program, department, and institution levels. Furthermore, relevance would be likely to improve in an online system that allows connections to be established between student learning, program planning, accreditation criteria, and budget requests. Institutional memory would be likely to improve by having a single system for accessing current and historic YPA plans and reports. Assessment guidance would likely improve because assessment systems tend to include training resources that align with best practices in the field. Discussions with a variety of assessment professionals at other institutions revealed that some institutions utilized homegrown assessment systems while others utilized vendor-based systems. Most seemed to be able to identify strengths and challenges associated with their current system. Some of the institutions experienced challenges comparable to those addressed in the Needs Assessment Results section in this paper. Existing end users and potential end users of the three vendor-based systems under consideration indicated that the respective systems were supporting or had potential to support effective implementation of their program assessment policies.

All three systems appeared to be able to help respond to the university's needs that emerged through utilization of its current MS Word template. Therefore, they were considered, in theory, to be more appropriate solutions than the MS Word template.

Efficiency Assessment Results

After the institution identified solutions that would be theoretically appropriate to meet their needs, they identified which of the solutions would be the most efficient to implement within the scope of its resources. Table 3 summarizes the efficiency assessment results in terms of product implementation costs, networks of users, existing contracts, and technical assistance. As a standalone piece of evidence, a system with a lower product implementation cost is considered more efficient than one with a higher cost because it requires fewer financial resources for the institution to implement the solution. A system that has a network of users within the state's university system is considered more efficient than a system that does not have a network of users because it provides more opportunities for the institution to learn best practices for implementation and other relevant lessons from institutions that operate within the scope of similar academic policies and procedures. A system that has an existing contract between its vendor and the state's university system that includes the institution is considered more efficient than one that does not have an existing contract because it reduces the volume of administrative resources required by the institution to procure and navigate the contract. A system with technical assistance provided by the state's university system is considered more efficient than a system without because it provides additional opportunities for training and support. A system with technical assistance provided by the system's vendor is considered

more efficient than a system without vendor-provided technical assistance because it would be likely to reduce the amount of time required for troubleshooting unforeseen challenges encountered by users.

Table 3. *Efficiency Assessment Evidence*

Evidence	System 1	System 2	System 3
Product implementation costs for the institution	\$0	Approximately \$30,000/year + onboarding	Approximately \$30,000/year + onboarding
Network of users within the state's university system	Yes	Yes	No
Existing contract with the state's university system that includes the institution	Yes	No	No
Technical assistance provided by the state's university system	Yes	No	No
Technical assistance provided by the system's vendor	Yes	Yes	Yes

The institution did not assign formal weights to each efficiency factor; however, they did consider certain factors to be more important than others. Technical assistance provided by the vendor was considered among the most essential factors to ensure that the system could be rolled out smoothly and that issues could be troubleshooted effectively. Given the context, implementation costs and the existing contract with the state's university system seemed to be the next most important factors. They felt like they were being better stewards of their financial resources by selecting a theoretically appropriate solution that costs \$0 per year through an existing contract compared to other theoretically appropriate solutions that would cost tens of thousands of dollars per year. Even if the difference in costs between systems would not have been so steep (i.e., \$0 to tens of thousands of dollars), the university would likely have still considered cost one of the most important factors. Technical assistance provided by the state's university system and an existing network of users in the state were added bonuses.

Taking all the presented efficiency assessment evidence into consideration, System 1 seems to be the most efficient, System 2 appears to be the second most efficient option, and System 3 seems to be least efficient of the three options. It should be noted that, although the amount of effort required to initially set up programs in the systems was not compared, all three systems would require extra time and energy to be spent building the programs. Continuing to use the university's MS Word template would not require extra time and energy for building programs in a new system; however, the one-time expenditure of effort to build the programs in any of the systems seemed to be worthwhile because they would be more efficient after the initial build and help alleviate needs that would likely persist through continued utilization of the MS Word template.

Discussion

Successful program assessment promotes the ongoing development and implementation of relevant, high-quality programs through intentional generation and utilization of multiple forms of evidence to support continuous improvement (Banta & Palomba, 2015). The effectiveness of program assessment processes may be impeded by various barriers, which could inhibit faculty buy-in for the practice (Ewell, 2002; Greene, 2023). Identifying and appropriately responding to the barriers could improve program assessment buy-in and effectiveness.

The institution in this study addressed three categories of assessment within a comprehensive program evaluation framework (Anderson, 2022; Rossi et al., 2004) to identify barriers to optimal program assessment implementation and a solution to help reduce the prevalence of the barriers. Its needs assessment revealed areas for improvement related to template consistency, efficiency, relevance, comprehensiveness, institutional memory, and assessment guidance. The needs assessment process also produced recommendations for improvement. Theory assessment revealed three vendor-based assessment systems (System 1, System 2, and System 3) that aligned with the institution's program assessment requirements and that could help meet its identified needs. Efficiency assessment revealed one assessment system (System 1) that seemed notably more efficient than the others (System 2 and System 3). Considering the results of the needs, theory, and efficiency assessments, the DIA—along with support from the institution's Vice President for Academic Affairs and members of assessment-related committees—decided to proceed with transitioning programs to System 1 on a voluntary basis. The institution intends to implement comprehensive planning strategies to help strengthen the likelihood of a successful transition to the new system (Herroelen, 2005; Meredith & Shafer, 2022).

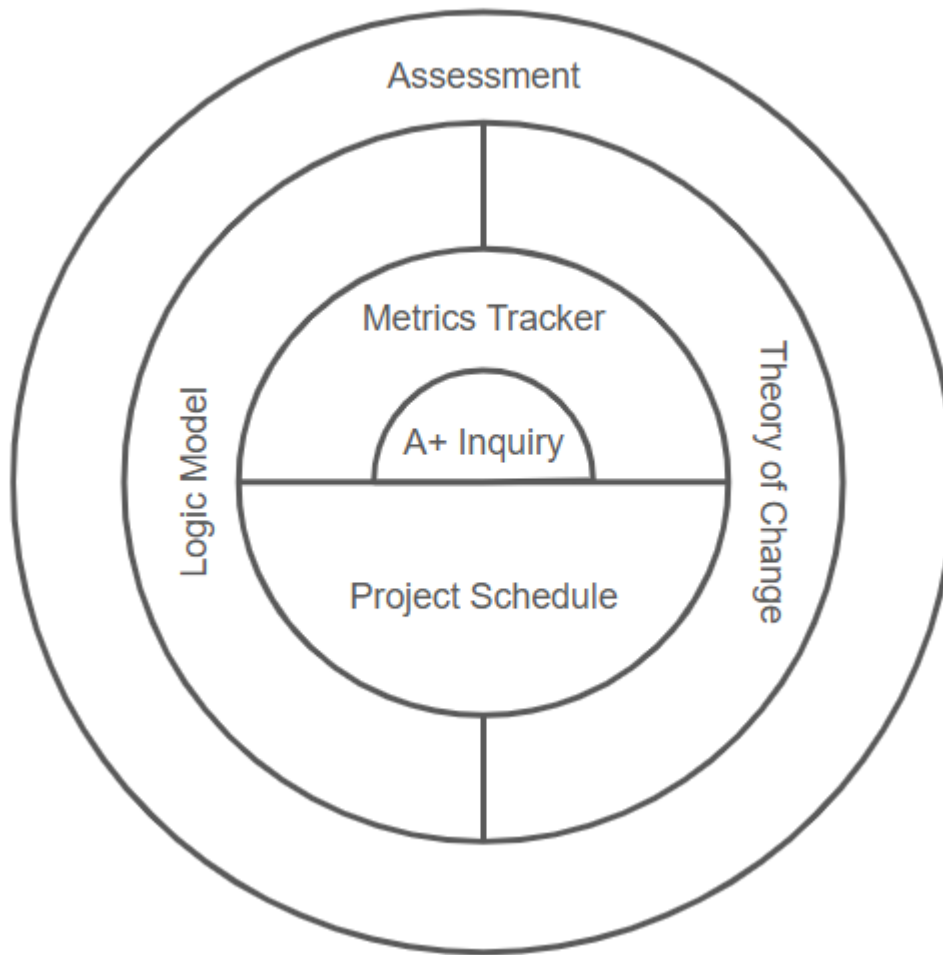
ALTMAP

The authors of this paper propose the ALTMAP model, based on Anderson's (2022) synthesis of established tools and frameworks for developing an evaluation plan, as a lens for comprehensively planning, implementing, and evaluating assessment systems or other purposeful initiatives. In the case of this paper, ALTMAP serves as a framework for supporting the planning, implementation, and evaluation of a relevant, efficient, and effective assessment system. Its components include (A)ssessment, (L)ogic model, (T)heory of change, (M)etrics tracker, (A)+ Inquiry, and (P)roject schedule (Figure 3). ALTMAP components relate to the five categories of assessment mentioned in this paper (i.e., needs, theory, process, outcome, efficiency).

Assessment

Assessment refers to the five categories of assessment (needs, theory, process, outcome, efficiency) previously introduced in this paper (Anderson, 2022; Rossi et al., 2004). This study addressed the needs, theory, and efficiency categories. The needs assessment relied on Department Chair interviews, assessment leadership observations, content analysis, and a survey to generate evidence of gaps in the institution's program assessment consistency, efficiency, relevance, comprehensiveness, institutional memory, and assessment guidance. The theory assessment relied on meetings with vendors, viewing product demonstrations by vendors, meeting with end-users of systems, and

Figure 3. *ALTMAP Diagram*



viewing product training videos, manuals, and other related resources to generate evidence of three systems being theoretically appropriate responses to the institution's assessment needs related to consistency, efficiency, relevance, comprehensiveness, institutional memory, and assessment guidance. The efficiency assessment relied on product implementation costs, a network of users within the state's university system, an existing contract with the state's university system, technical assistance provided by the state's university system, and technical assistance provided by the system's vendor. Other institutions may consider using similar or other types of sources to generate evidence for their needs, theory, and efficiency assessments.

This study did not focus on the process or outcome categories of assessment. However, moving forward, the institution intends to utilize a variety of data sources to generate evidence related to the assessment system being implemented as planned and producing the desired effects. Process assessment evidence of implementation may include the number of assessment systems selected for utilization, number of system guidance materials, number of training sessions conducted, the number and percentage of faculty trained on the system, the number and percentage of faculty with access to the system, the number and percentage of faculty signing in to the system, the number of

programs with assessment plans in the system, and the number of programs with assessment reports submitted in the system. Outcome assessment evidence addressing effects may include changes in knowledge of assessment, skill in assessment implementation, satisfaction with program assessment processes, time required to complete assessment plans and reports, timeliness of submissions, comprehensiveness of assessment plans and reports, institutional memory, utilization of data to improve student learning, and student achievement. Qualitative data related to implementation and effects may be collected and analyzed to help contextualize quantitative process and outcome assessment indicators. Other institutions may consider using similar or other types of evidence sources to support their process and outcome assessment efforts.

Logic Model

A logic model depicts an overview of a program or project, typically in a one-page snapshot. Logic model elements often include the title of a project or program, its context and related needs, its expected goal, the target population, resources, activities, outputs, and short-, mid-, and long-term outcomes (Kekahio et al., 2014; W.K. Kellogg Foundation, 2004; Wyatt Knowlton & Phillips, 2013). Some of the logic model elements are related to the five categories of assessment.

The context/need element, related to the needs category of assessment, helps establish a foundational rationale for a program or project. The resources element, related to the efficiency category of assessment, outlines resources required for its implementation. The activities element, related to the theory category of assessment, represents strategies intended to be implemented with the resources. The outputs element, related to the process category of assessment, represents verification of activity implementation. The short-, mid-, and long-term outcomes elements, related to the outcome category of assessment, represent the effects of implementing the project's activities. Figure 4 depicts a logic model related to the needs, theory, and efficiency assessment evidence in this study and the process and outcome assessment evidence that the institution could track moving forward. The figure is color-coded to demonstrate its alignment with the colors depicted in the five categories of assessment in Figure 1.

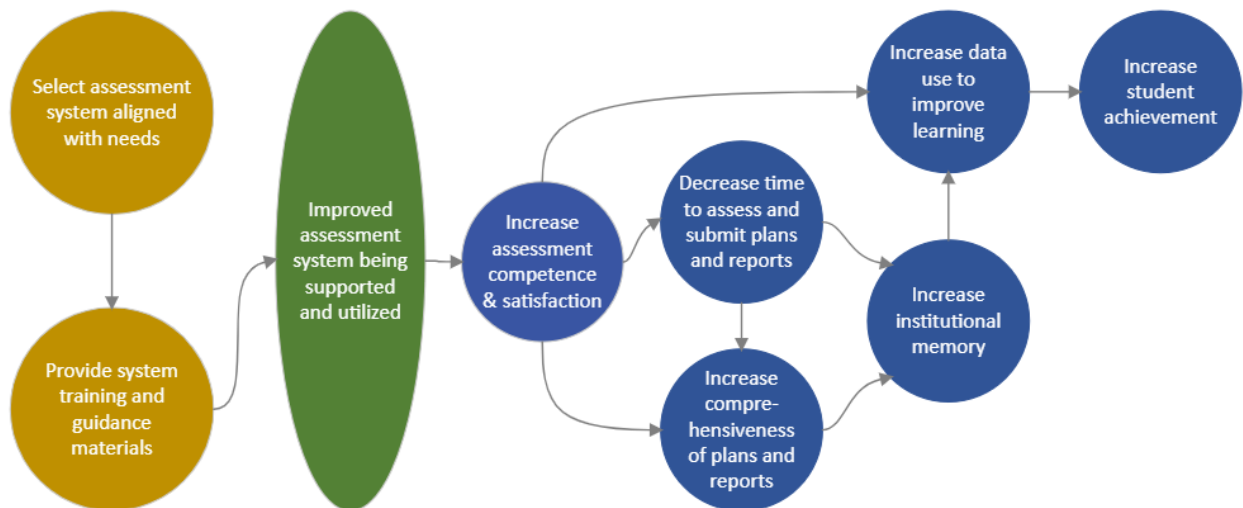
Theory of Change

A theory of change is a high-level depiction of how change is likely to occur through a project (Dhillon & Vaca, 2018; Wyatt Knowlton & Phillips, 2013). Theories of change may depict causal relationships that help further contextualize the information provided in logic models. The design of a theory of change is likely to be different for each project. Figure 5 illustrates an example theory of change related to the logic model in Figure 4. Its elements are color-coded to align with the theory, process, and outcome categories of assessment in Figure 1 as well as with the activities, outputs, and outcomes elements of the logic model in Figure 4. The theory of change suggests that selecting an assessment system aligned with needs and then providing training and guidance materials on the system will lead to an improved assessment system being supported and utilized, which will lead to increased assessment competence and satisfaction. Higher levels of competence and satisfaction are expected to decrease the time required to complete assessments and to submit plans and reports, increase comprehensiveness of reports, and increase data use to improve student learning. Increasing comprehensiveness of plans and

Figure 4. *Logic Model Example*

Project Title		Expected Goal	Target Population		
Program Assessment System Enhancement		Improve Program Assessment	Faculty		
Context/Need					
Academic programs are required to regularly implement assessment and to submit yearly program-level assessment plans and reports. Various barriers weaken the institution's assessment system and inhibit assessment effectiveness. Evidence suggests that the institution has room for improvement in the areas of template consistency, efficiency, relevance, comprehensiveness, institutional memory, and assessment guidance.					
Resources	Activities	Outputs	Outcomes		
			Short-Term	Mid-Term	Long-Term
Technology infrastructure	1 Select assessment system	1 # assessment systems	Increase faculty assessment knowledge	Decrease time to complete assessment	Increase institutional memory
Funding	2 Provide system guidance materials	2 # system guidance materials	Increase faculty assessment skill	Decrease time to submit plans and reports	Increase use of data to improve student learning
System user network	3 Provide system training	3 # training sessions # and % faculty trained	Increase faculty satisfaction with assessment	Increase comprehensiveness of plans and reports	Increase student achievement
Technical assistance by university system	4 Utilize system	4 # and % faculty with access # and % faculty signing in # plans entered in system # reports entered in system			

Figure 5. *Theory of Change Example*



reports and decreasing time required to assess and submit plans and reports is expected to increase institutional memory, which is expected to also increase data use to improve student learning. Increases in data use to improve student learning are expected to promote increases in student achievement.

Metrics Tracker

A metrics tracker is utilized to compile values reflecting measurable output and outcome indicators (Anderson, 2022). It may include fields to indicate whether a metric is an output or outcome, identify the metric's associated activity category, provide a description of the metric, and enter target and actual values for the metric. Figure 6 shows an example metrics tracker with a few indicators aligned with the outputs and outcomes in the logic model in Figure 4. It is color-coded to align with the process and outcome assessment categories in Figure 1, output and outcome logic model elements in Figure 4, and related theory of change elements in Figure 5. Sample target values that the institution could consider are provided. Actual values are to be determined (TBD).

Figure 6. *Metrics Tracker Example*

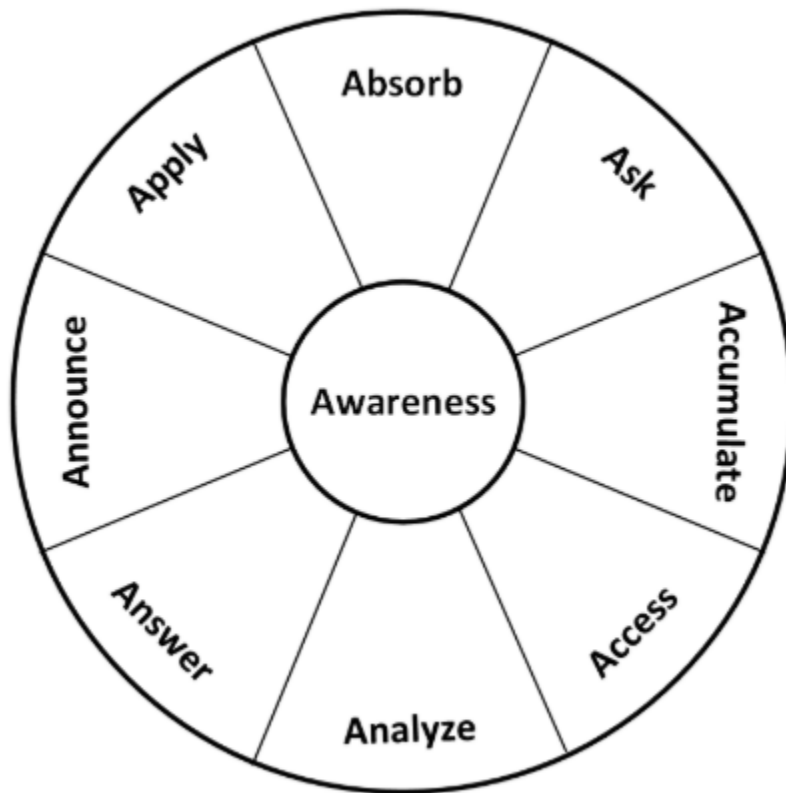
Logic model section	Activity category	Metric description	Target value	Actual value
Output	3 Provide system training	Number of training sessions	5	TBD
Output	3 Provide system training	Percentage of faculty program assessment report authors trained on system	100%	TBD
Output	4 Utilize system	Percentage of faculty program assessment report authors with system access	100%	TBD
Short-term outcome	N/A	Percentage of faculty self-reporting satisfaction with assessment processes	90%	TBD
Mid-term outcome	N/A	Percentage of program assessment plans and reports submitted on time	90%	TBD

A+ Inquiry

A+ Inquiry is a broadly applicable inquiry framework that may be utilized to operationalize output and outcome metrics as needed (Anderson, 2022; Anderson et al., 2014). A+ Inquiry uses visualization (Gilbert, 2008; Tufte, 1990; Ware, 2000) and alliteration (Bryant et al., 1990; Lea et al., 2008; Stoll, 1940) to promote understanding of common stages in various disciplined inquiry cycles (e.g., evaluation, assessment,

research) and how the stages support one another. It may be utilized as a conceptual lens for planning or implementing a study, establishing common language for discussing inquiry methods, identifying gaps in inquiry processes, telling an evidence-based story, or providing rationale for action. A few examples of its use include training teachers in methods of utilizing data (Anderson et al., 2018), facilitating discussions on data utilization in schools (Statewide Longitudinal Data Systems Grant Program, 2015), developing a thesis-alternative capstone course for M.Ed. students (Anderson et al., 2025), conducting a study to enhance the development of a holistic teacher evaluation tool (Tenam-Zemach et al., 2024), assessing program assessment workload needs of faculty (Pontenila et al., 2024), and developing outcome assessment plans (Anderson, 2022). The cyclical framework includes eight stages (Absorb, Ask, Accumulate, Access, Analyze, Answer, Announce, Apply) connected in the center by a hub of Awareness (Figure 7).

Figure 7. *A+ Inquiry Diagram*



In the Absorb stage, identify what is already known about a context and identify a need to know or understand more about something. In the Ask stage, formulate one or more questions that, if answered, could help close the knowledge gap established in the Absorb stage. In the Accumulate stage, collect new data or verify the collection of existing data that could be analyzed to help answer the question(s) posed in the Ask stage. In the Access stage, retrieve the collected data from where it is stored in preparation for analysis. In the Analyze stage, analyze the data using applicable quantitative and/or qualitative methods. In the Answer stage, respond to the questions posed in the Ask stage based on the analysis results. In the Announce stage, distribute the results to applicable

stakeholders. In the Apply stage, make decisions based on the results. The Awareness hub serves as a reminder to ensure that when someone is working on tasks in one stage of the process, they are aware of how it integrates with the other stages. Table 4 depicts an example of how A+ Inquiry could be utilized to operationalize the output metric in Figure 6 that focuses on the percentage of faculty program assessment report authors trained on the system.

Table 4. *A+ Inquiry Example*

A+ Inquiry Stage	Example
Absorb	<i>What is known?</i> The university is implementing a new assessment system. They set a target to train 100% of their faculty program assessment report authors on the system. <i>What is not known?</i> They do not know the extent to which their target is being achieved.
Ask	To what extent is the target being achieved for training faculty program assessment report authors?
Accumulate	Compile a list of program assessment authors by requesting program assessment report author names from department chairs. Add the names to a spreadsheet as they are submitted. Compile a list of faculty participating in the training by asking them to complete an online sign-in form. As participants submit the form, their names are automatically compiled in an online data source connected to the report.
Access	Retrieve the list of program assessment authors from the spreadsheet where they are compiled. Retrieve the names of training participants from the online data source where the names are automatically compiled upon submission of the form.
Analyze	Cross-reference the list of program assessment report authors with the list of training participants. Identify which authors participated in training. Count the total number of authors and the number of authors who completed training. Divide the number of authors who completed training by the total number of authors to calculate the percentage of authors who received training. Calculate the difference between the target percentage and the actual percentage of authors receiving training.
Answer	The percentage point difference between the target and the actual percentages illuminates the extent to which target is being achieved. Note the names of authors who participated and names of those who did not participate.
Announce	Share the results with the team responsible for implementing the new system. Also, share the results with the department chairs to inform them of the extent to which training has been completed by report authors for the programs in their department. Include names of the training participants and non-participants for reference.
Apply	Follow up with training participants to inquire about their needs for further support. Plan and implement a strategy for training report authors who have not yet been trained on the system.

Project Schedule

Project scheduling helps facilitate optimal implementation of a project's activities across time (Herroelen, 2005; Meredith & Shafer, 2022). A project schedule tool can be utilized to detail a sequential scope of specific activities required for operationalizing the project and to track progress toward completion of the activities. Project schedules may evolve over time as new information becomes available and tasks need to be added, revised, or removed. Components of a project schedule may include a code for each item in the schedule, descriptions of activities with associated tasks and subtasks, the implementation status, target start and end dates, responsible parties, and comments related to progress and challenges. Figure 8 depicts an example of a project schedule template related to the third activity category (3 provide system training) in the logic model in Figure 4. It is color-coded to demonstrate its alignment with the theory assessment category in Figure 1, the activities element of the logic model in Figure 4, and related theory of change elements in Figure 5.

Figure 8. *Project Schedule Example*

Code	Description	Status	Start Date	End Date	Responsible	Comments (progress, challenges, etc.)
3	Provide system training	Not started	mm/dd/yy	mm/dd/yy	Assessment Director	
3.1	Develop training	Not started	mm/dd/yy	mm/dd/yy	Assessment Director	
3.1.1	Develop draft content	Not started	mm/dd/yy	mm/dd/yy	Assessment Director	
3.1.2	Collect feedback on content	Not started	mm/dd/yy	mm/dd/yy	Assessment Director	
3.1.3	Analyze feedback	Not started	mm/dd/yy	mm/dd/yy	Assessment Director	
3.1.4	Make edits based on feedback	Not started	mm/dd/yy	mm/dd/yy	Assessment Director	
3.2	Recruit participants	Not started	mm/dd/yy	mm/dd/yy	Assessment Director	
3.3	Deliver training	Not started	mm/dd/yy	mm/dd/yy	Assessment Director	

The ALTMAP examples in this section are specific to the study presented in this paper; however, the tools and frameworks reflected in the ALTMAP model are highly transferrable and may support other institutions' adaptations of the methods in this study for promoting assessment quality within their unique contexts.

Conclusion

Consideration of various pieces of evidence through lenses of needs, theory, and efficiency helped the institution in this study determine which assessment system would be the most appropriate to implement. The next steps will be to formalize and implement a plan to begin transitioning programs to the new system on a voluntary basis, to provide general and targeted support for faculty using it, and to critically evaluate the system's value. A successful transition will depend on a variety of factors characterized in Venkatesh et al.'s (2003) *Unified Theory of Acceptance and Use of Technology* model, including performance expectancy, effort expectancy, attitude toward using technology,

social influence, facilitating conditions, self-efficacy, anxiety, and behavioral intention to use the system. As programs move to the new system, assessment leadership will need to be prepared for challenges that may accompany a transition, such as resisting change, learning terminology of a new system, and acquiring new skills for navigating the system.

The institution in this study should continue to periodically assess its needs related to program assessment, the theory of its assessment system, and the support required to ensure the system works efficiently. It will also be important for the DIA to implement process assessment to know whether the institution's assessment system is functioning as intended and outcome assessment methods to provide a better understanding of whether the system's implementation is helping meet its identified needs and supporting achievement of assessment policy requirements. If their assessment system effectively organizes assessment plans and results, faculty will have readily accessible and analyzable assessment information for decision-making related to providing high quality, relevant programming that promotes student learning.

In addition to the institution in this study continuing to utilize evidence generated through multiple categories of assessment to evaluate its program assessment system and supports, other institutions may consider adapting the methods and frameworks presented in this paper to support the evaluation of their program assessment tools and practices. The newly proposed ALTMAP model provides a lens for discussing and organizing essential components of comprehensive evaluations by synthesizing common evaluation-related frameworks and tools into a cohesive package. While institutions may find value in addressing each ALTMAP component in the planning, implementation, and evaluation of their own program assessment platforms, the ALTMAP model could help inform programming and judgments on the merit or worth of a wide array of initiatives beyond the scope of program assessment and the field of higher education.

References

- Adams, W. C. (2015). Conducting semi-structured interviews. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (4th ed., pp. 492-505). Jossey-Bass.
- Allen, M. J. (2003). *Assessing academic programs in higher education*. Wiley.
- Anderson, N. (2022). Synthesizing frameworks and tools to develop a plan for evaluating an online data utilization curriculum for teachers. *Evaluation and Program Planning*, 94. <https://doi.org/10.1016/j.evalprogplan.2022.102148>
- Anderson, N. C., Brockel, M. R., & Kana, T. E. (2014). Disciplined inquiry: Using the A+ Inquiry framework as a tool for eliminating data hoarding, mindless decision-making, and other barriers to effective ESA programming. *Perspectives: A Journal of Research and Opinion About Educational Service Agencies*, 20(3). https://www.researchgate.net/publication/377656559_Disciplined_Inquiry_Using_the_A_Inquiry_Framework_as_a_Tool_for_Eliminating_Data_Hoarding_Mindless_Decision-Making_and_Other_Barriers_to_Effective_ESA_Programming
- Anderson, N. C., Conn, D. R., & Weigel, K. D. (2025). Cultivating graduate-level assessment skills through an A+ Inquiry thesis alternative. *Assessment Update: Progress, Trends, and Practices in Higher Education*, 37(3), 1-2, 15-16. <https://doi.org/10.1002/au.30438>
- Anderson, N., Ova, A., Oliver, W., & Greer, D. (2018). *Develop your data mindset: Essentials of educational data use* [Online curriculum]. North Dakota Department of Public Instruction, Information Technology Department, Statewide Longitudinal Data System.
- Banta, T. W. (2002). *Building a scholarship of assessment*. Jossey-Bass.
- Banta, T. W., & Palomba, C. A. (2015). *Assessment essentials: Planning, implementing, and improving assessment in higher education* (2nd ed.). Jossey-Bass.
- Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the pieces together: An integrated model of program implementation. *Prevention Science*, 12(1), 23-33. <https://doi.org/10.1007/s11121-010-0186-1>
- Bryant, P. E., MacLean, M., Bradley, L. L., & Crossland, J. (1990). Rhyme and alliteration, phoneme detection, and learning to read. *Developmental Psychology*, 26(3), 429-438. <https://doi.org/10.1037/0012-1649.26.3.429>
- Calley, N. G. (2011). *Program development in the 21st century: An evidence-based approach to design, implementation, and evaluation*. Sage.
- Cohen, L., Manion, L., & Morrison, K. (2017). *Research methods in education* (8th ed.). Routledge.
- Dhillon, L., & Vaca, S. (2018). Refining theories of change. *Journal of MultiDisciplinary Evaluation*, 14(30), 64-87. <https://doi.org/10.56645/jmde.v14i30.496>
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41, 327-350. <https://doi.org/10.1007/s10464-008-9165-0>
- Ewell, P. T. (2002). An emerging scholarship: A brief history of assessment. In T. W. Banta (Ed.), *Building a scholarship of assessment* (pp. 3-25). Jossey-Bass.
- Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. L. (2011). *Program evaluation: Alternative approaches and practical guidelines* (4th ed.). Pearson.

- Gilbert, J. K. (2008). Visualization: An emergent field of practice and enquiry in science education. In J. Gilbert, M. Reiner, & M. Nakhleh (Eds.), *Visualization: Theory and practice in science education* (pp. 3-24). Springer.
- Greene, G. (2023, January 24). The terrible tedium of 'learning outcomes': Accreditors' box-checking and baroque language have taken over the university. *The Chronicle of Higher Education*. <https://www.chronicle.com/article/the-terrible-tedium-of-learning-outcomes>
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Wiley.
- Halpern, G. (1998). From hubris to reality: Evaluating innovative programs in public institutions. *The Innovation Journal: The Public Sector Innovation Journal*, 3(3), article 3. <https://innovation.cc/document/1998-03-3-3-from-hubris-to-reality-evaluating-innovative-programs-in-public-institutions/>
- Herroelen, W. (2005). Project scheduling—theory and practice. *Production and Operations Management*, 14(4), 413-432. <https://doi.org/10.1111/j.1937-5956.2005.tb00230.x>
- Hutchings, P. (2019, January). What faculty need to know about assessment (Assessment Brief). University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA). <https://www.learningoutcomesassessment.org/wp-content/uploads/2019/08/Assessment-Brief-Faculty.pdf>
- Kekahio, W., Cicchinelli, L., Lawton, B., & Brandon, P. R. (2014). *Logic models: A tool for effective program planning, collaboration, and monitoring* (REL 2014–025). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Pacific.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage.
- Lea, R. B., Rapp, D. N., Elfenbein, A., Mitchel, A. D., & Romine, R. S. (2008). Sweet silent thought: Alliteration and resonance in poetry comprehension. *Psychological Science*, 19(7), 709-716. <https://doi.org/10.1111/j.1467-9280.2008.02146.x>
- Levin, H. M., & Belfield, C. (2015). Guiding the development and use of cost-effectiveness analysis in education. *Journal of Research on Educational Effectiveness*, 8(3), 400-418. <https://doi.org/10.1080/19345747.2014.915604>
- Main, K. (2011). *Program design: A literature review of best practices*. United Way of Calgary.
- McCawley, P. F. (2009). *Methods for conducting an educational needs assessment: Guidelines for extension system professionals* (BUL 870). University of Idaho Extension.
- Meredith, J. R., & Shafer, S. M. (2022). *Project management: A strategic managerial approach* (11th ed.). Wiley.
- Miller, R., & Leskes, A. (2005). *Levels of assessment: From the student to the institution*. Association of American Colleges and Universities.
- Phillippi, J., & Lauderdale, J. (2017). A guide to field notes for qualitative research: Context and Conversation. *Qualitative Health Research*, 28(3), 381-388. <https://doi.org/10.1177/1049732317697102>
- Pontenila, S., Stephens, E., & Anderson, N. C. (2024). Using A+ Inquiry as a framework for exploring faculty needs related to program assessment workload. *Intersection: A Journal at the Intersection of Assessment and Learning*. <https://doi.org/10.61669/001c.126752>
- Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (2nd ed.). Sage.
- Saunders, R. P., Evans, M. H., & Joshi, P. (2005). Developing a process evaluation plan for assessing health promotion program implementation: A how-to guide. *Health Promotion*

- Practice*, 6(2), 134-147. <https://doi.org/10.1177/1524839904273387>
- Statewide Longitudinal Data Systems Grant Program. (2015, July). *SLDS Data Use Standards: Standards in Practice*. U.S. Department of Education. Washington, DC: National Center for Education Statistics.
- Stoll, E. (1940). Poetic alliteration. *Modern Language Notes*, 55(5), 388-390. <https://doi.org/10.2307/2910998>
- Stufflebeam, D. L., & Coryn, C. L. (2014). *Evaluation theory, models and applications* (2nd ed.). Jossey-Bass.
- Tenam-Zemach, M., Conn, D., Anderson, N. C., & Alstad, E. D. (2024). Utilizing an inquiry-agnostic framework to advance the development of a holistic teacher evaluation tool. *Holistic Education Review*, 4(2). <https://her.journals.publicknowledgeproject.org/index.php/her/article/view/3214>
- Tucker, C., Drummond, S., & Ostrogorsky, T. L. (2020). ASSESS-IT: The development story of an institutional rubric for evaluating programmatic assessment plans and reports. *Research & Practice in Assessment*, 15(2), 1-9. <https://files.eric.ed.gov/fulltext/EJ1293308.pdf>
- Tufte, E. R. (1990). *Envisioning information*. Graphics Press.
- U.S. Government Accountability Office. (2011, May). *Performance measurement and evaluation: Definitions and relationships* (Report No. GAO-11-646SP). Author.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425-478. <https://doi.org/10.2307/30036540>
- Ware, C. (2000). *Information visualization: Perception for design*. Morgan Kaufmann Publishers.
- W.K. Kellogg Foundation. (2004). *Logic model development guide*. Author.
- Witkin, B. R., & Altshuld, J. W. (1995). *Planning and conducting needs assessments: A practical guide*. Sage.
- Wyatt Knowlton, L., & Phillips, C. P. (2013). *The logic model guidebook* (2nd ed.). Sage.
- Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Sage.
- Yates, B. (2009). Cost-inclusive evaluation: A banquet of approaches for including costs, benefits, and cost-effectiveness and cost-benefit analyses in your next evaluation. *Evaluation and Program Planning*, 32(1), 52-54. <https://doi.org/10.1016/j.evalprogplan.2008.08.007>

Evaluating Accreditation: A Case Study of Student Views on the Value of Optional Program Accreditation



Authors:

Erica Eckert, Ph.D.
Kent State University

ABSTRACT

This concurrent triangulation mixed methods study explores undergraduate students' perspectives on programmatic accreditation, a rigorous process that demands significant resources from faculty and staff. Few studies have investigated students' perspectives on accreditation at institutions of higher education (IHEs) based in the United States. Most students were unaware their program was accredited and knew little about accreditation in general. However, when presented with simplified versions of the accreditation standards, students ranked most components as important and identified additional elements that evidence academic program quality. They indicated programmatic accreditation validated quality and demonstrated institutional and faculty commitment. This is one of few studies of students' perspectives on accreditation in the American context—programmatic or otherwise. This study serves as a model for faculty and administrators seeking to determine if the work for optional accreditation is “worth it” and to determine how well accreditation standards align with students' notions of program quality.

Correspondence E-mail: eeckert@kent.edu

Keywords: Programmatic accreditation, student perceptions, academic quality, mixed methods, higher education

Accreditation is a voluntary external review process that serves as a signal of accountability, quality assurance, and continuous improvement for colleges and universities (Alstete, 2007; Dodd, 2004; Eaton, 2009; Kinser, 2014; Wheelan, 2016). The standards and processes of United States-based accreditors are structured to appraise how well institutions are carrying out their mission; formal recognition by a U.S. Department of Education-recognized accreditor is a prerequisite for access to federal financial aid, among other benefits (Ewell et al., 2010; Flores, 2018; Wheelan, 2016). Accreditation can be pursued at the institution and individual program level (Alstete, 2007; Eaton, 2009; Gaston, 2014; Romanowski, 2022; Wheelan, 2016). Most universities pursue both institutional accreditation for the entire enterprise and programmatic accreditation for individual disciplines, although not all disciplines have a programmatic accreditor. Accreditation at both levels is an intense undertaking (Alstete, 2007; Bardo, 2009; Dassa et al., 2017; Dodd, 2004; Elderman, 2013; Romanowski, 2022; Wheelan & Elgart, 2015). While standards vary from accreditor to accreditor, they usually include mission, integrity (e.g., policies, conflicts of interest), resource sufficiency, curricula, student support, assessment of student learning, use of data for improvement, and planning for the future (Dodd, 2004). The accreditation process—typically described as the self-study document and site visit—requires substantial faculty and staff time and expertise along with a technical infrastructure to produce the required evidence (Hoare & Goad, 2022).

Despite the costs, accreditation is perceived positively by many within higher education (Elderman, 2013; Gaston, 2014; Kelly, 2023; Navarro, 2021; Oden, 2009; Schall, 2023; Wolff, 2014; Ulker & Bakioglu, 2019). The reasons are many, but center mostly around the planning, assessment, self-examination, and collaboration required by the process. Critics of accreditation claim accreditors have too much (or little) power, their level of scrutiny is too high (or low), their standards are too prescriptive (or vague), they rely on site visitors drawn from the institutions they accredit, and that accreditation drives up costs and encourages a compliance mindset (Alstete, 2007; Burke & Butler, 2012; DeMillo, 2013; Ewell, 2015; Gillen et al., 2010; Harvey, 2024; Romanowski, 2022; Romanowski & Karkouti, 2024; Urwitz, 2023). Speaking to decades of scholarship on accreditation, Harvey (2024) describes a “tension... between accountability and improvement” (p. 360) which may lead institutions to focus on compliance and meeting reporting requirements rather than transformation and innovation. In 2023, *The Chronicle of Higher Education* surveyed 415 administrators about their perspectives on accreditation in the United States. Eighty-four percent of respondents agreed accreditation did a “good job ensuring academic quality in higher education,” but only 41% felt that accreditation fostered academic innovation (Wheeler, 2023, p. 7).

According to the same *Chronicle of Higher Education* survey, 90% of academic administrators believe the American public does not understand what accreditors do and 55% believe that the accreditation process does not produce useful information for students and families for their college-going decisions (Wheeler, 2023). The narrative about the benefits, drawbacks, and impact of accreditation are largely set by academic leaders and policymakers, not students—ostensibly the beneficiary of academic quality. Andreani et al. (2020) notes students’ perspectives on quality assurance are understudied, as is the impact of quality assurance on higher education more generally. Inspired by these findings, this study uses a concurrent triangulation mixed methods design approach to explore students’ perspectives on accreditation at a large research institution

in the midwestern United States. Although the inquiry will be limited to a single programmatic accreditor, this study can serve as a blueprint for others who wish to ascertain the value students attach to accreditation more broadly.

Existing Literature: Programmatic Accreditation

Impact on Outcomes

Scholarship on the impact of programmatic accreditation spans enrollment through post-completion outcomes across many disciplines. Accreditation by the Association to Advance Collegiate Schools of Business (AACSB) has been linked to increased enrollment in business programs at historically black colleges and universities (HBCUs) in the United States (Williams, 2021) and improved teaching practices at a business school in Japan (Ito et al., 2024). Other scholars have examined the impact of accreditation on student achievement. Volkwein et al. (2007) identified significant improvements in engineering students' knowledge and skills after the implementation of new ABET standards at 40 institutions based in the United States. They avoided assigning all credit to the new standards, however, and surmised the change created a space for the innovation and collaboration necessary to improve a program.

Studies of medical education accreditation also tend to explore the link between student performance and accreditation. Researchers at King Saud University in Saudi Arabia found students' exam scores increased significantly after the introduction of accreditation (but pass rates did not). The authors suggested the improvements were achieved through engaging in the process, which compelled "coherence, administrative buy-in, and coordinated quality improvement actions" that arose through the accreditation process (Alenezi et al., 2023, p. 6). The results at King Saud University align with an earlier study that found Mexican and Philippine students who attended accredited medical schools were more likely to pass the United States Medical Licensing Examination (USMLE) than those who did not (van Zanten et al., 2012). Student performance on licensing examinations may also fluctuate in relation to an institution's accreditation cycle, as was the case in a study of 17 Canadian medical schools (Roy et al., 2020). Recognition by an accreditor may also be associated with post-graduation outcomes; a recent study identified a significant difference in employment rates and early career income between alumni of accredited graduate-level healthcare programs and those from programs without accreditor recognition (Fisher et al., 2025).

Faculty and Administrator Perspectives

Scholars have also conducted studies of faculty and administrators' perspectives on programmatic accreditation around the world. For example, Ulker and Bakioglu (2019) surveyed self-study coordinators and other administrators in several countries. The authors concluded that the accreditation process enhanced program quality, but the effect was more pronounced for initial accreditor recognition than continuing recognition. Staub (2019) interviewed faculty and administrators at a Turkish university to consider the relative value of two sequential foreign language programmatic accreditation pursuits. Participants in the study named several benefits that emanated from the process (the creation and maintenance of documents and policies, increased focus on teaching and learning, the perspective of outsiders) but warned that the benefits may be experienced unevenly across the organization and sustaining momentum is a challenge. A similarly-

oriented study of faculty and administrators at three of the four AACSB-accredited Canadian business schools suggested accreditation promoted organizational learning by prompting discussions about alignment to mission, increasing the school's focus on quality and research productivity, and by serving "as a catalyst for change" which led to meaningful improvement (Elliott and Goh, 2013, p. 737).

Although they may be hesitant to engage in accreditation, faculty members may benefit from the process, according to Garrison et al. (2013), who studied faculty that participated in an American Association of Family and Consumer Sciences programmatic accreditation process. The faculty in this study recognized the accreditation process as a validating learning experience that triggered small but useful changes to their program. Although myriad factors can influence how an organization experiences an accreditation process, coherence with an accreditors' aims, organizational buy-in, and effective, intentional collaboration can influence organizational quality (Desveaux et al., 2017). Summarizing scholarship from several authors attesting to the value of programmatic accreditation as a tool for continuous improvement, Dassa et al. (2017) noted, "the process is equal to the product" (p. 2).

Student Perspectives

A similar search of scholarship on student perspectives on American accreditation published since 1995 returned few empirical studies—programmatic or otherwise. A study of graduates of an accredited engineering program reported greater levels of what are commonly called high-impact practices: collaboration and active learning, interaction with instructors, instructor feedback, and intentional student engagement (Volkwein et al., 2007). Internationally, several researchers have examined the relationship between program accreditation and student satisfaction or perceptions of quality. Architecture students at the University of Lima associated the introduction of a new programmatic accreditor with program quality (Acevedo-De-los-Ríos & Rondinel-Oviedo, 2022). Pharmacy students in Qatar associated recognition by the Canadian Council for Accreditation of Pharmacy Programs (CCAPP) with an increase in educational quality and workload (Wilby et al., 2019). The self-reported satisfaction of undergraduate medical students at King Saud University increased significantly after it obtained initial recognition from the National Center for Academic Accreditation and Evaluation (NCAAA), a Saudi Arabian accreditor (Al-Eyadhy & Alenezi, 2021). Abou Hashish et al. (2025) studied nursing students' perceptions of accreditation through the NCAAA and found they considered it valuable even if they weren't all familiar with accreditation. The students in their study believed accreditation ensured their program met rigorous standards and provided verification that the program was both relevant and similar to other nursing programs. They also indicated a desire to be more involved and more informed about the process.

Not all studies have linked accreditation with positive attitudes or outcomes. Butt et al. (2024) attempted to connect international programmatic accreditation to undergraduate students' notions of quality and reputation; the results weren't significant. However, they did note nearly two-thirds of their respondents knew little about accreditation. Perhaps this is why studies of accreditation are uncommon—it's challenging to study a phenomenon about which people are unfamiliar. A broader search of literature on college students' perceptions of quality (rather than accreditation) returned several

results, mostly from authors outside the United States. Researchers based in the United Kingdom explored students' characterizations of quality in higher education and concluded faculty, curriculum, social and emotional support systems, and resources were essential (Hill et al., 2003). López (2005) identified similar variables in a study of Spanish students; their general satisfaction was of paramount importance, followed by teaching, physical facilities and resources, and a comprehensive and theory-to-practice approach that attends to both “professional and personal components” (p. 331). López also found students in different specialization areas had different ideas on what constituted quality. Informed by López's work, Heffner (2013) studied the relationship between 21 components of academic quality and based students' intentions to persist at American colleges and universities. Their analysis produced several models, one of which identified as academic engagement and academic expectations, opportunities to provide feedback, student satisfaction, and “life-long learning” as predictors of intention to persist (p. 96). In a study of a soon-to-be AACSB-accredited program, Clemons and Jance (2024) explored seven dimensions of service quality identified in a prior study at a regional university in the Midwestern United States. Their recommendations provided insights on what the institution could do to improve students' satisfaction with their educational experiences.

A comprehensive analysis of scholarship on university quality is beyond the scope of this review, but these studies suggest students' perspectives on institutional quality are of interest. However, most of these studies do not address accreditation. Colleges and universities in the United States are competing against one another in a competitive marketplace; many institutions are using accreditation to establish their legitimacy and to boost their reputation and value in the eyes of stakeholders (Burnett, 2022; Kinser, 2014; Romanowski, 2022; Staub, 2019). Accreditation shows students, potential students, employers, and other institutions that a college or university is of sufficient quality in a competitive market. However, this assertion is often made by citing the common wisdom of higher education leaders and media rather than through evidence collected empirically. Much of the value and purpose of accreditation is structured on a foundational belief that it considers and promotes quality, which ultimately serves to benefit students. To better understand what this means, students can and should be part of the conversation about accreditation at the programmatic level (Abou Hashish et al., 2025).

Method

The researcher used a concurrent triangulation mixed methods design to conduct this study. Case studies are bounded in a specific place, entity, or time (Merriam, 1998; Yin, 2009). Data collection for a case study is flexible; the instrumentation and procedures should be selected to best address the research questions (Merriam, 1998; Yin, 2009). A concurrent triangulation approach incorporates multiple complementary data collection approaches to create a broader understanding (Creswell, 2003). Quantitative and qualitative data are considered equally important and were considered together during analysis. The researcher collected qualitative data via focus groups and quantitative data via survey for this study of undergraduate students' perspectives on programmatic accreditation. Surveys allow researchers to collect data in a cost-effective, standard format that can be analyzed quantitatively (Creswell, 2003; Henning & Roberts, 2024) while focus groups allow researchers to collect qualitative data about participants' experiences

and feelings without making assumptions on which aspects are most salient. Focus groups also promote elaboration because participants can build upon each other's responses (Henning & Roberts, 2024; Krueger & Casey, 2009). The questions guiding this study were: (1) what are students' perspectives on a specific programmatic accreditation in relation to program selection and quality, (2) what value do students assign to a programmatic accreditor's standards, and (3) how would students feel if a program changed how it engaged with its accreditor?

Positionality

The researcher has worked extensively with programmatic accreditation and through that experience formed a deep understanding of the benefits and challenges of undertaking the accreditation process. As an instructor of graduate coursework about assessment and accreditation, she regularly discusses the purpose, role, and perception of accreditation presented by accreditors, colleges and universities, the media, and other organizations. Collectively, these experiences informed this study. While she strives to identify and minimize bias in her scholarship, she acknowledges that her experiences guide how her research questions are framed and may influence data interpretation.

Participants

The researcher collected data for this study during regularly scheduled class meetings at the invitation of the program coordinator (faculty member with administrative responsibilities) at a large public university in the Midwestern United States. This constituted a purposefully selected sample (Creswell, 2003) of undergraduate students enrolled in two different courses in a degree program with optional programmatic accreditation. One course enrolled mostly first- and second-year students who were either majoring in the degree program or interested in the field (hereinafter referred to as the introductory (intro) course). The other course served as a capstone experience for fourth-year students in the degree program. The program will not be named to maintain confidentiality. It is important to note that this program was not required by the state or another body to pursue accreditation. All participants were 18 years of age or older and most majored in the degree program.

Although the data were collected during class meetings, participation was not mandatory. Students were given the opportunity to opt out of the study during the consent process without penalty. One student declined to participate in the study; the remaining 33 signed the informed consent document—21 in the intro course, and 12 in the capstone course. Eleven of the intro course participants (52.3%) and all capstone participants indicated they were majoring in the program. Data from the institution's Institutional Research office indicates the program enrolled 24 freshmen, 22 sophomores, 30 juniors, and 32 seniors at the time of data collection. It is challenging to evaluate participant counts against the enrollment in the program; however, a student may have enrolled in the major in their first or second year but wasn't taking the intro course, or a student may have taken the intro course without declaring the major. With this caveat, it appears the capstone group ($n = 12$) represented approximately 37.5% of the seniors enrolled in the major. Eleven self-reported program majors participated from the introductory course, representing 23.9% of freshmen and sophomores enrolled in the program.

Instrumentation and Data Collection Procedures

The researcher conducted two 90-minute focus groups and administered a 45-item survey designed for this study at the midway point during each focus group session. Dividing a single session into three phases (focus group, survey, focus group) allowed the researcher to gauge students' baseline understanding accreditation before providing information about accreditation through the survey and documenting their reaction to that information. The accreditor maintains nine standards with multiple elements and sub-elements within each. The researcher developed 18 items to encompass the eight standards related to program administration, and 27 items to account for the student knowledge and skills standards. Based on their prior experience engaging students in discussion of accreditation standards, the researcher decided to reword the standards to make the language more accessible for undergraduate students while maintaining their original intent as closely as possible. For example, the original strategic planning standard included language about what the strategic plan must contain: mission, vision, goals, implementation details, and provisions for reporting. These were condensed into a single survey item, which read: *The program engages in strategic planning (setting objectives and measuring if they have been achieved).*

In several cases, multiple inter-related elements within a standard were condensed for brevity. For example, two elements about instructional support and resources were blended together into a single item. The inverse was also true. For example, within the student knowledge and skills standards, communication skills were listed together. The researcher chose to break these into multiple items on the survey so each piece (written, oral, interpersonal, digital) could be evaluated separately; a similar approach was taken for operations management knowledge. Rewording the standards into language undergraduate students could understand and review in a short period of time often meant sacrificing specificity, but the intent of the standards were retained as much as possible. It also means that the findings will describe the items as “components” of standards, because they may describe entire standards, single or multiple elements, or single or multiple sub-elements. In accordance with guidance from Gliner et al. (2017), the researcher asked a colleague to review the tool for content prior to deployment and incorporated all recommended adjustments. The survey relied upon a five-point scale (Not at all important (1) to Extremely important (5)). Questions about participants' identities were not included because the number of participants prevented any meaningful post-hoc analysis. Participants returned the surveys to the researcher upon completion and the focus group resumed.

The researcher conducted focus groups using guidance from Krueger and Casey (2009). The first four prompts focused on 1) why they selected the program, 2) what they knew about its accreditation, 3) how knowing a program was accredited would make them feel about it, and 4) how they personally determined program quality. As anticipated, most participants knew little about accreditation, so the researcher provided a general description of accreditation between the third and fourth prompts. The description was vague to avoid biasing participants' responses to the third and fourth prompts. The researcher then paused the focus group discussion to administer the paper survey designed to align with the program accreditor's standards. The survey was designed for two purposes—to collect participants' ratings of the importance of each standard and to show participants what programs are required to demonstrate to accreditors. The final

four focus group prompts focused on the content of the standards, ideas around indicators of quality, and the relative weight of accreditation in considering quality. The final prompt asked students how they would feel about the program if it voluntarily complied with accreditation standards (but did not complete the formal process) and gave up its accreditation so the faculty could devote more time to other endeavors. The focus group concluded with an opportunity to share final comments, new information, or conflicting opinions, but none were offered.

The number of focus groups was limited to the aforementioned class meetings; this was a practical decision made by the researcher in collaboration with the program faculty and deemed acceptable because the data were being collected primarily for internal purposes. Therefore, it was not possible to check for saturation. As Krueger and Casey (2009) note, it is not always feasible to collect data to the point of saturation and the researcher did not attempt to draw comparisons across the two focus groups. The focus group data were used to better understand and extend the data collected with the survey tool. The first focus group (capstone course) was audio-recorded, but the recording was unusable due to sound quality and the second group (intro course) was not recorded. The participants usually spoke at a pace the researcher could transcribe in real time; the data were a mix of direct quotes and paraphrased comments. Although 33 students elected to participate in the study, approximately half contributed verbally to the focus group discussion. Participants in the capstone course ($n = 11$) were far more engaged in the focus group than those in the intro course ($n = 18$). Participants who did not contribute to the discussion appeared to be following the conversation, but did not choose to share any comments even when prompted. Their body language and gestures suggested they didn't know enough to contribute or simply didn't have any feelings about the topic.

Data Analysis

The quantitative and qualitative data were analyzed concurrently. To mitigate bias related to her positionality and to explore emerging findings, the researcher composed memos at the conclusion of each focus group and during the theming process (Merriam, 1998). The two focus groups yielded a total of 70 statements across eight prompts. The focus group data were arranged in Microsoft Excel by prompt and coded by the researcher using a simultaneous eclectic approach. Saldaña (2021) describes eclectic coding as a flexible approach for exploratory analysis that allows the use of multiple code types—in this case concept, evaluation, and descriptive codes. A simultaneous approach allows a researcher to assign multiple codes to a datum (Saldaña, 2021). Initially, the researcher coded the focus group data according to how participants conceptualized accreditation (its purpose or meaning), the relative importance of accreditation, whether participants framed accreditation in reference to themselves, outsiders (the institution, employers), or both, and what participants considered to be valid indicators of quality (e.g., accreditation standards, other indicators). The last category was further defined in a second phase of analysis, described in the Findings section.

The survey responses were manually entered into Microsoft Excel; 29 of the 33 surveys were included in the analysis. The responses from four participants were discarded due to incongruous answers (all 5 or 1 responses) or substantial omissions. The survey data were highly reliable; both scales exceeded the typical 0.8 threshold for internal consistency (Gliner et al., 2017). The program administration scale (18 items)

demonstrated acceptable internal consistency ($\alpha = .86$) and the academic standard scale (27 items) demonstrated excellent internal consistency ($\alpha = .93$). Because this study was not intended to be generalized to other contexts, the analysis was descriptive in nature. The item-level median scores were evaluated visually to draw conclusions about the relative importance of each standard. Although the research questions did not call for analysis by group (intro and capstone), the researcher presented median-level data by group for transparency. Because this case study employed a concurrent triangulation mixed methods design, the researcher went back and forth between the focus group data and survey item results to make meaning of the data and to identify points of alignment, convergence, or correspondence.

Limitations and Delimitations

This study has three important limitations that must be considered. First, the number of participants ($N = 29$) is relatively small, and although many participants were present for the focus group, several did not actively engage in the process. However, the participants represented roughly 30% of students enrolled in the major. Members of the capstone group were more engaged than those in the intro group, which could have been a product of the number of students, the classroom culture, their sense of agency, or even the time of day. Second, the instrumentation itself may serve as a limitation—the researcher created the survey items by rewording accreditation standards to remove jargon and technical language. The researcher felt this was important to ensure students unfamiliar with accreditation could decipher the survey items, but it also meant students did not evaluate the accreditor's original language. Third, the duration of data collection could also be considered a limitation, as it was the only time available to conduct the study. However, the 90-minute period for each focus group was sufficient; no new points were raised by either group by the conclusion of each session.

The study also has three primary delimitations. First, this study was conducted at the invitation of a faculty member at a specific university and thus represents a single institution. Second, the participants consisted of introductory and capstone students, not students near the mid-point of the program and their perspectives are thus excluded. Third, the data collection was designed to obtain as much information as possible in a confined period of time. Although the researcher tried to minimize the consequences as much as possible, these tradeoffs likely led to the omission of elements of accreditation standards in order to promote participant engagement.

Results

Few participants reported knowing anything about programmatic accreditation prior to the focus group; many gestured nonverbally that they had nothing to add to the conversation. The sparse discussion, in a way, represented an important finding about students' lack of familiarity with programmatic accreditation. The survey deployed at the halfway point of the focus group provided participants with a summary of program accreditation standards, which informed the second half of the focus group discussion. The survey data indicate participants felt the standards addressed many important elements, but the focus group conversation suggests the standards omit a few elements they associate with program quality.

Perspectives on programmatic accreditation, program selection, and program quality

Participants ascribed low importance to accreditation in part because they felt other factors served as better indicators. They saw accreditation mostly as a means for external validation—a way to verify “the lessons we’re getting are valid... that [students] are benefiting from it.” Participants equated accreditation to proof of a program having “been tested” and used words like “official” and “honest” when describing what accreditation meant to them, which aligns in part with findings from Abou Hashish et al. (2025). Although they knew they could get a job in the industry with a degree from an unaccredited program, they seemed to believe that the program’s accredited status would help in their search for employment. One indicated the program’s accredited status meant “the program is more official [and] that we have outer connections that will take us further at graduation” and another felt it would result in “more job opportunities.” Although they did not draw a connection to income, these findings partially align with actual employment outcomes observed by Fisher et al. (2025).

When asked to consider what the status of “accredited” would convey about a program, participants indicated accreditation was synonymous with quality, reliability, trust, and reputation. They also felt that it showed the program was a priority. One participant indicated, accreditation shows a program is “kind of like a focus for the college, more than the other programs, that you’re getting a better education in that program.” Studies of faculty and staff who engaged in accreditation work by Elliott and Goh (2013), Garrison et al. (2013), Ito et al. (2024), and Volkwein et al. (2007) give credence to this idea. However, just because they viewed accreditation positively, that didn’t mean they considered it of primary importance. Participants repeatedly referenced other aspects of the program that they felt superseded accreditation—such as the breadth of the curriculum, their relationships with program faculty members, having a network of well-connected and dedicated alumni, and the quality of experiential education opportunities. Many of these ideas align with factors identified by other researchers who have studied students’ perspectives on quality (Heffner, 2013; Hill et al., 2003; López, 2005).

Only two of 27 participants (7.4%) indicated they knew the program was accredited before enrolling, and both were in the intro group. This aligns with a study that concluded that nearly two thirds of students knew little about accreditation (Butt et al., 2024). Considering program selection, a focus group participant explained, “I didn’t know that the program was accredited, it wasn’t something I knew mattered... that’s not the reason that I came here.” Programmatic accreditation simply didn’t factor in their decision-making. This finding partially aligns Williams’ (2021) study of accredited business schools, which found AACSB accreditation to be associated with increased HBCU enrollment, while ACBSP (Accreditation Council for Business Schools and Programs) accreditation was not. A few participants in the capstone group mentioned having been interviewed as part of a recent accreditation process, but didn’t feel that they knew what it really meant to be accredited or what it took to obtain accreditation. When asked directly how accreditation might influence their decision to enroll if they had it to do all over again, a participant indicated accreditation was a “plus” and could serve as a tipping factor—all else equal, accreditation might lead them to “pick the better school.” The researcher, attempting to engage a broader swath of the participants to join the discussion, asked them to raise their hands to indicate if they would choose an accredited program over a non-

accredited program, all other things being equal between them—most participants indicated that they would.

The value of programmatic accreditation standards in relation to professional preparation

Twenty-nine participants rated the importance of their accreditor's standards. The standards were obtained directly from the accreditor's website but were simplified and paraphrased on the survey to promote content validity with the target population. Because the data were ordinal level, they are presented in Table 1 as median scores by component (using keywords) to avoid identifying the accreditor. Nearly every median component ranking was between 4 (highly important) and 5 (extremely important), indicating that participants felt the accreditor was choosing to focus on important things. The table also provides the difference between the first and third quartile rankings (interquartile range) for each item by group as evidence of the small variability in the scores. Most items had an interquartile range of approximately 1. The median rankings for the knowledge and skills components were slightly higher than the program administration components, which is not surprising. Participants in the capstone group ranked many of the components at a higher level of importance than their counterparts in the intro group.

Although interpreting the data is fairly challenging because of the small variability between the groups and the item-level responses, it appears that the capstone group ranked the applied knowledge and skill components slightly higher than intro group; for example, critical thinking, problem solving, diversity, ethical leadership and culture, and all four components related to communication had median rankings of 5, while accounting, economics, human resources, marketing, and management theory were ranked as level 4. That said, not all theory-based components were ranked lower; financial management, leadership theory, and legal issues all had median ranks of 5. The intro group provided slightly lower median rankings to the knowledge and skill components, regardless of focus. Compared to the capstone group, the intro group assigned higher rankings to accounting and human resources. In short, everything was considered important.

Participants also ranked program administration-related components highly. The components with the lowest median rankings were *culminating experience* and *library materials and access*, but even these were ranked between “moderately” and “highly” important. Only one component, *website/promotional materials* was ranked higher by the intro group, which aligned with the conversation during the intro focus group about the importance of the website in the program selection process. Two components were assigned a median rank of 5 (extremely important) by both groups; *faculty sufficiency* and *use of learning assessment data*. Several components had the same median rankings across groups, but as with the knowledge and skill components, the capstone group tended to assign higher ranks. The capstone group assigned a median rank of 5 to *advising*, *curriculum review/improvement*, *experiential education*, *faculty control*, and *instructional support and resources* which aligned closely with the focus group conversation. The capstone group strongly felt the faculty in the program should be the ones making decisions about the program and that they should be well-supported in that process. Many of these factors align with the findings of other studies of students’

Table 1. Median Scores and Interquartile Range of Component Importance by Group

Knowledge and Skills Components	Group Median		Interquartile Range	
	Intro	Capstone	Intro	Capstone
Accounting	5	4	0.75	1.00
Analytical skills	4	5	1.00	1.50
Communication, Digital	4	5	0.75	1.00
Communication, Interpersonal	4	5	1.00	2.00
Communication, Oral	5	5	1.00	1.50
Communication, Written	4	5	0.00	1.00
Critical thinking	4.5	5	2.00	1.00
Diversity	4	5	1.00	1.50
Economics	4	4	1.00	1.50
Ethical conduct, politics	4	4	1.50	1.00
Ethical leadership and culture	4	5	1.00	1.50
Financial management	5	5	1.00	1.00
Human resources	4.5	4	1.75	0.50
Industry/profession	5	5	1.00	2.00
Integration, industry/profession	4	5	1.00	1.00
Leadership skills	4.5	5	1.00	1.00
Leadership theory	4	5	1.00	1.00
Legal Issues	4	5	1.00	1.50
Management theory	4	4	1.00	0.50
Marketing	4	4	1.00	1.50
Operations in domain A	4	4	1.00	2.00
Operations in domain B	4	5	2.00	1.00
Operations in domain C	4	5	1.00	0.00
Problem solving	5	5	1.75	1.00
Strategy	4	5	1.75	1.00
Subtypes of industry/profession	5	5	1.00	0.50
Technology	4	5	0.75	1.00

Note. Scale: Not at all important (1), Slightly important (2), Moderately important (3), Highly important (4), Extremely important (5)

Table 1. *Continued from previous page*

Program Administration Components	Group Median		Interquartile Range	
	Intro	Capstone	Intro	Capstone
Advising	4	5	1.75	1.00
Stakeholder data use	4	4	1.00	0.50
Culminating experience	3	4	1.75	1.00
Curriculum review and improvement	4	5	1.00	1.00
Experiential education	4.5	5	1.00	1.00
Faculty control	4	5	1.00	0.50
Faculty evaluation, development	4	4	1.00	1.00
Faculty sufficiency	5	5	1.00	0.00
Financial well-being	4	4	0.00	0.00
Instructional support	4	5	1.00	0.00
Learning outcomes	4	4	1.00	0.50
Library materials, access	3	4	1.00	0.50
Mission	4	4	1.75	1.00
Policies and procedures	4	4	0.00	1.00
Strategic planning	4	4.5	2.00	2.00
Use of assessment data	5	5	1.75	1.00
Validity, assessment tools	4	4	1.00	0.50
Website/PR materials	5	4	1.00	1.00

Note. Scale: Not at all important (1), Slightly important (2), Moderately important (3), Highly important (4), Extremely important (5)

perspectives on quality (e.g., Heffner, 2013; Hill et al., 2003; López, 2005), although faculty control of program and use of learning assessment data findings are somewhat novel. This is likely due in part to the survey being crafted around accreditation standards—but it is a worthy finding, nonetheless.

Participants were asked to suggest indicators of program quality that may or may not be reflected in the accreditation standards (as phrased in the survey). Focus group participants' reflections on program quality indicators were coded into four sub-categories: post-graduation outcomes, student/program support, curriculum, and evaluations. The post-graduation indicators included: alumni accomplishments, positions/roles, and network; connections upon graduation, and a variety of employment metrics (salary, percent employment in field, percent employed within a specific timeframe). The student/program support indicators included: guest and alumni speakers, industry connections, resources for students, policies and procedures for students (including advising), graduation rate, evidence the program is trying to provide a good experience, sense of community, institutional support for the program, and program faculty (quality, relationships with, feeling cared for). The curriculum indicators included: quality of the in-class experience, hands-on experience/experiential education, curriculum variety, and program content. The evaluation indicators included accreditation and reputation, which encompassed student and alumni feedback. Accreditation standards address many of these ideas, but not necessarily in the way they

were framed by the participants in this study. Many accreditors' standards make reference to having outcomes during and after the program, while students are interested in very specific outcomes (e.g., employment in field within 12 months of graduation). A few of these findings align with existing scholarship, but most extended beyond previous studies cited in this manuscript.

Perspectives on changing engagement or abandoning the program accreditation

The focus group concluded with prompts asking participants how they would feel about a program claiming alignment with accreditation standards without completing the external review process or abandoning program accreditation entirely. The conversation on this topic was brief and pragmatic—one participant asked, in essence, 'if you're bothering to abide by the standards, why wouldn't you just do the external review?' The researcher did not attempt to explain the amount of work required to generate a self-study for fear of showing bias or creating undue influence on the group. Another participant stated that abiding by the standards with or without formal recognition from the accreditor "makes the program look more honest" but this was the only positive comment. Another participant felt that accreditation was an important verification process, and if the program wasn't going through the external verification of accreditation, they could be "half-assing" it. The value in the external review process was the site visit team— "talking to students, making sure [things] are happening... is beneficial."

The final prompt about abandoning program accreditation generated more conversation. Participants' comments were a bit more divided but still reflected a pragmatic outlook. For example, one commented, "If nothing changed besides the accreditation... it wouldn't matter. It probably would change how many choose that program over the others. Personally, I wouldn't... it wouldn't matter to me... as long as the program stays the same." This idea that the program would need to maintain its quality or 'stay the same' was echoed by several students. Thinking about how the cessation of accreditation would be perceived by potential employers, one student said, "if they were to take away our accreditation it wouldn't hurt at all as long as the university continues to fund us and give... the program the same amount of attention." One participant was concerned that giving up accreditation could impact recruitment of new students if they were using accreditation as a factor in their enrollment decision-making.

Other participants were concerned about how ceasing efforts to maintain accreditation would be perceived by others. One suggested that "it gives the program a bad look... it makes the program look shady." Similarly, another felt "if universities aren't willing to go through the work [of accreditation], they obviously don't think the work is that important... they can find the money somewhere" and to not use the resources says something [bad] about the program. Another participant, speaking to a middle ground, suggested, "If they do that, they need to explain it... explain how they are reallocating the funds... the benefits of stopping doing it." Many participants around the room nodded in agreement to this comment.

Discussion

After learning of the program's accreditation during the focus groups, participants appeared to assign significant value to it. They felt accreditation recognition was an

important third-party verification of program quality that included many elements associated with academic quality. Further, accreditation showed the program had institutional support. Given that 93% of participants didn't know the program was accredited before enrolling, it is unlikely that accreditation status would influence new or continuing enrollment in the program. However, focus group participants expressed concern about abandoning the program's accreditation and worried its removal would have negative consequences.

The researcher presented the survey and focus group data to program faculty and their advisory board comprised of alumni and industry professionals. Most advisory board members agreed that accreditation served as a signal of quality and attested that programmatic accreditation did not influence employers' hiring decisions. This was a major concern expressed during the student focus group sessions. Advisory board members felt the benefits of programmatic accreditation did not outweigh their costs—but agreed that the rollout of a decision to disaffiliate would be critical. Further, participants indicated that they wanted assurance that the program was continuing to engage in careful self-examination and improvement and expected transparency about these processes. It was clear from the comments that participants relied on information presented on institutional websites—so presenting the data in these spaces would be prudent.

Just as accreditation occurs in cycles, programs holding optional programmatic accreditation should also periodically consider whether their accreditors' standards and processes align with the program's best interests. This form of inquiry can inspire productive conversations and provide students with a way to express concerns (and unknown bright spots) about a program. Those wishing to replicate this study may find it useful for decision-making with two suggestions. First, provide participants with copies of the survey for consultation during the latter half of the focus group to support recall during the conversation. Second, extend the length of the focus groups to allow for rapport-building with hesitant participants or lengthier conversations with engaged participants.

Conclusion

Although programmatic accreditation was not a factor in participants' enrollment decisions, they viewed it as a signal of academic quality and integrity that promoted a program's reputation. When asked to elaborate, participants explained that the validity of accreditation rested in its external nature—that unbiased experts deemed the curriculum, faculty, and support resources available to students sufficient. This suggests that a voluntary, internal review process may not convey the same validation as an external review process. Participants also suggested that undergoing the process demonstrates the program is working to maintain its good position and is a priority for the larger college or university in which it operates. Students in this study indicated the program accreditor's standards encompassed many important academic and program administration elements but identified several additional elements they considered essential to program quality. The results from this study suggest college faculty associated with programmatically accredited programs should educate students more intentionally about the merits of holding and maintaining recognition. It may increase students' pride in their program and reassure them they made a responsible choice. The results also suggest programmatic accreditors should consider including students in conversations about accreditation standards and practices. If accreditation is truly about ensuring programs provide

students with high-quality educational experiences, students should be part of the conversation about what quality means.

References

- Abou Hashish, E. A., Alnajjar, H., & Rawas, H. (2025). Voices on academic accreditation: Lived experiences of nurse educators, administrators, students, and alumni in nursing education. *BMC Medical Education*, 25(1), 64. <https://doi.org/10.1186/s12909-025-06657-2>
- Acevedo-De-los-Ríos, A., & Rondinel-Oviedo, D. R. (2022). Impact, added value and relevance of an accreditation process on quality assurance in architectural higher education. *Quality in Higher Education*, 28(2), 186-204. <https://doi.org/10.1080/13538322.2021.1977482>
- Al-Eyadhy, A., & Alenezi, S. (2021). The impact of external academic accreditation of undergraduate medical program on students' satisfaction. *BMC Medical Education*, 21(1), Article 1. <https://doi.org/10.1186/s12909-021-03003-0>
- Alenezi, S., Al-Eadhy, A., Barasain, R., AlWakeel, T. S., AlEidan, A., & Abohumid, H. N. (2023). Impact of external accreditation on students' performance: Insights from a full accreditation cycle. *Heliyon*, 9(5), e15815. <https://doi.org/10.1016/j.heliyon.2023.e15815>
- Alstete, J. W. (2007). College accreditation: Managing internal revitalization and public respect. Palgrave Macmillan.
- Andreani, M., Russo, D., Salini, S., & Turri, M. (2020). Shadows over accreditation in higher education: Some quantitative evidence. *Higher Education*, 79(4), 691-709. <https://doi.org/10.1007/s10734-019-00432-1>
- Bardo, J. W. (2009). The impact of the changing climate for accreditation on the individual college or university: Five trends and their implications. *New Directions for Higher Education*, 2009(145), 47-58. <https://doi.org/10.1002/he.334>
- Burke, L. M. & Butler, S. M. (2012). Accreditation: Removing the barrier to higher education reform. *The Heritage Foundation* (Report No. 2728). ERIC Number: ED535877 <https://eric.ed.gov/?id=ED535877>
- Burnett, C. A. (2022). Accreditation sanctions as quality signals in a competitive market for students. *Educational Policy*, 36(6), 1507-1528. <https://doi.org/10.1177/0895904820983034>
- Butt, M. M., Huisman, J., Hussain, D. H., Alam, M., & Amin, M. (2024). Antecedents and consequences of students' attitudes towards internationally accredited business schools: A signalling theory perspective. *Journal of Marketing for Higher Education*, 34(1), 1-17. <https://doi.org/10.1080/08841241.2021.1942388>
- Clemons, R., & Jance, M. (2024). Defining quality in higher education and identifying opportunities for improvement. *SAGE Open*, 14(3), 21582440241271155. <https://doi.org/10.1177/21582440241271155>
- Creswell, J. W. (2003). *Research design: Quantitative, qualitative, and mixed methods approaches* (2nd ed.). Sage.
- Dassa, L., Ridener, B., & Sampson, G. (2017). An overview of topics on accreditation: A beginning. *SAGE Open*, 7(3), 215824401773379. <https://doi.org/10.1177/2158244017733791>
- DeMillo, R. (2013, January 16). Accreditation—or real quality assurance? *The James G. Martin Center for Academic Renewal*. <https://jamesgmartin.center/2013/01/accreditation-or-real-quality-assurance/>
- Desveaux, L., Mitchell, J. I., Shaw, J., & Ivers, N. M. (2017). Understanding the impact of accreditation on quality in healthcare: A grounded theory approach. *International Journal for Quality in Health Care*, 29(7), 941-947. <https://doi.org/10.1093/intqhc/mzx136>

- Dodd, A. H. (2004). Accreditation as a catalyst for institutional effectiveness. *New Directions for Institutional Research*, 2004(123), 13-25. <https://doi.org/10.1002/ir.116>
- Eaton, J. S. (2009). Accreditation in the United States. *New Directions for Higher Education*, 2009(145), 79-86. <https://doi.org/10.1002/he.337>
- Elderman, E. (2013). Accreditors now find themselves under critical review. *Chronicle of Higher Education*, 60(14), A23-A25.
- Elliott, C. J., & Goh, S. C. (2013). Does accreditation promote organizational learning? A multiple case study of Canadian university business schools. *Journal of Management Development*, 32(7), 737-755. <https://doi.org/10.1108/JMD-03-2011-0028>
- Ewell, P., Boeke, M., & Zis, S. (2010). State uses of accreditation: Results of a fifty-state inventory. CHEA Occasional Paper. *Council for Higher Education Accreditation*. <https://eric.ed.gov/?id=ED544332>
- Ewell, P. (March, 2015). Transforming institutional accreditation in US higher education. National Center for Higher Education Management Systems [White Paper]. Author Copy.
- Fisher, D., Opoku-Agyeman, W., Stanowski, A., & Tuttle, W. (2025). Accredited vs non-accredited graduate healthcare administration programs: Are there differences in measurable outcomes. *The Journal of Health Administration Education*, 41(1), 15-30. <https://www.aupha.org/publications/journalofhealthadministrationeducation>
- Flores, A. (2018, November 27). Fact sheet: A legislative history of the federal government's reliance on accreditation. *Center for American Progress*. <https://www.americanprogress.org/article/fact-sheet-legislative-history-federal-governments-reliance-accreditation/>
- Garrison, S., Herring, A., & Hinton, W. J. (2013). Learning through accreditation: Faculty reflections on the experience of program evaluation. *Journal of Family & Consumer Sciences*, 105(2), 8-13. <https://doi.org/10.14307/JFCS105.2.4>
- Gaston, P. L. (2014). *Higher education accreditation: How it's changing, why it must*. Routledge.
- Gillen, A., Bennett, D. L., Vedder, R. (2010). The inmates running the asylum? An analysis of higher education accreditation. Center for College Affordability and Productivity [Policy Paper]. <https://files.eric.ed.gov/fulltext/ED536283.pdf>
- Gliner, J. A., Morgan, G. A., & Leech, N. L. (2017). *Research methods in applied settings: An integrated approach to design and analysis* (3rd ed.). Routledge.
- Harvey, L. (2024). What have we learned from 30 years of Quality in Higher Education: Academics' views of quality assurance. *Quality in Higher Education*, 30(3), 360-375. <https://doi.org/10.1080/13538322.2024.2385793>
- Heffner, C. L. (2013). *Does student perception of quality matter? A study of perceived academic quality and intention to persist* [Doctoral dissertation, Capella University]. <https://www.proquest.com/pqdtglobal/docview/1431912164/abstract/C82EC5D377DC4717PQ/1>
- Henning, G. W. & Roberts, D. M. (2024). *Student affairs assessment: Theory to practice* (2nd ed.). Routledge.
- Hill, Y., Lomas, L., & MacGregor, J. (2003). Students' perceptions of quality in higher education. *Quality Assurance in Education*, 11(1), 15-20. <https://doi.org/10.1108/09684880310462047>

- Hoare, A., & Goad, P. (2022). The quality continuum: Perceptions of institutional accreditation. *Quality Assurance in Education*, 30(1), 102-117. <https://doi.org/10.1108/QAE-08-2021-0135>
- Ito, H., Takeuchi, S., Yokoyama, K., Makita, Y., & Ischii, M. (2024). Impact of AACSB accreditation on education quality: Perceptions of faculty in an accredited school in Japan. *International Journal of Educational Management*, 38(3), 769-785. <https://www-emerald-com.proxy.library.kent.edu/insight/content/doi/10.1108/ijem-05-2023-0258/full/html>
- Kelly, J. (2023, March 31). The right-wing war on accreditation. *The Chronicle of Higher Education*. <https://www.chronicle.com/article/the-right-wing-war-on-accreditation>
- Kinser, K. (2014). Questioning Quality Assurance: Questioning Quality Assurance. *New Directions for Higher Education*, 2014(168), 55-67. <https://doi.org/10.1002/he.20113>
- Krueger, R. A. and Casey, M. A. (2009). *Focus groups: A practical guide for applied research* (4th ed.). Sage.
- López, I. G. (2005). Building universities of quality: An analysis of the views of university students concerning their academic training. *Higher Education in Europe*, 30(3/4), 321-334. <https://doi.org/10.1080/03797720600625275a>
- Merriam, S. B. (1998). *Qualitative research and case study applications in education* (2nd ed.). Jossey-Bass.
- Navarro, M. A. (2021, June 3). How a college accrediting agency failed to protect students from a decade of fraud. *Center for American Progress*. <https://www.americanprogress.org/article/college-accrediting-agency-failed-protect-students-decade-fraud/>
- Oden, R. A. (2009). A college president's defense of accreditation. *New Directions for Higher Education*, 2009(145), 37-45. <https://doi.org/10.1002/he.333>
- Romanowski, M. H. (2022). The idolatry of accreditation in higher education: Enhancing our understanding. *Quality in Higher Education*, 28(2), 153-167. <https://doi.org/10.1080/13538322.2021.1948460>
- Romanowski, M. H., & Karkouti, I. M. (2024). United States accreditation in higher education: Does it dilute academic freedom? *Quality in Higher Education*, 30(1), 97-111. <https://doi.org/10.1080/13538322.2022.2127165>
- Roy, M., Wood, T. J., Blouin, D., & Eva, K. W. (2020). The relationship between accreditation cycle and licensing examination scores: A national look. *Academic Medicine*, 95(11S) S103-S108.
- Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). Sage.
- Schall, L. (2023, February 26). What the accreditation naysayers don't understand. *Inside Higher Ed*. <https://www.insidehighered.com/views/2023/02/27/accreditation-misunderstood%E2%80%94and-essential-opinion>
- Staub, D. (2019). 'Another accreditation? What's the point?' effective planning and implementation for specialised accreditation. *Quality in Higher Education*, 25(2), 171-190. <https://doi.org/10.1080/13538322.2019.1634342>
- Ulker, N., & Bakioglu, A. (2019). An international research on the influence of accreditation on academic quality. *Studies in Higher Education*, 44(9), 1507-1518. <https://doi.org/10.1080/03075079.2018.1445986>

- Urwitz, J. (2023, January 24). Accreditors are sleeping on the job. *Inside Higher Ed*.
<https://www.insidehighered.com/views/2023/01/25/accreditors-must-do-job-congress-assigned-them-opinion>
- van Zanten, M., McKinley, D., Durante Montiel, I., & Pijano, C. V. (2012). Medical education accreditation in Mexico and the Philippines: Impact on student outcomes. *Medical Education*, 46(6), 586-592. <https://doi.org/10.1111/j.1365-2923.2011.04212.x>
- Volkwein, J., Lattuca, L., Harper, B., & Domingo, R. (2007). Measuring the impact of professional accreditation on student experiences and learning outcomes. *Research in Higher Education*, 48(2), 251-282. <https://doi.org/10.1007/s11162-006-9039-y>
- Wheelan, B. (2016). Accreditation: The elephant in the room. *Journal of Nonprofit Education and Leadership*, 6(4), Article 4. <https://doi.org/10.18666/JNEL-2016-V6-I4-7897>
- Wheelan, B. S. & Elgart, M. A. (2015, October 22). Accreditation's real cost (and value). *Inside Higher Ed*. <https://www.insidehighered.com/views/2015/10/22/real-costs-accreditation-and-processs-value-essay>
- Wheeler, D. (2023). The road ahead for accreditation. *The Chronicle of Higher Education* [Special report]. https://connect.chronicle.com/rs/931-EKA-218/images/Research%20Brief%20-%20The%20Road%20Ahead%20for%20Accreditation_Watermark.pdf
- Wilby, K. J., Alamri, M., & Monfared, S. (2019). Student and graduate perceptions regarding Canadian-based accreditation of a pharmacy program in Qatar. *American Journal of Pharmaceutical Education*, 83(6), 6805. <https://doi.org/10.5688/ajpe6805>
- Williams, Z. B. (2021). The impact of business program accreditation on ranking and enrollment for HBCU schools. *Research Journal of Business and Management*, 8(4), 233-242. <https://doi.org/10.17261/Pressacademia.2021.1465>
- Wolff, R. A. (2014, January 14). Exploding myths: What's right with regional accreditation. *ACE: Higher Education Today*. <https://www.higheredtoday.org/2014/01/14/exploding-myths-whats-right-with-regional-accreditation/>
- Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Sage.

Creating and Implementing a Data Ecosystem for Assessing Sexual Violence Campus Climate



Authors:

Sarah Socorro Hurtado, Ph.D.
University of Denver

Caitlyn Potter Glaser, M.A.
University of Denver

ABSTRACT

In this case study we examined how an institution engages in the development and implementation of a data ecosystem for assessing campus climate related to sexual violence. We found that while the institution was active in collecting different types of data, these data were limited to certain dimensions of campus climate (psychological and behavioral), which offer a limited understanding of campus climate. Further, this institution faced several challenges in engaging in data synthesis due to tensions related to trust, transparency/confidentiality, and perception/reality. These tensions created further challenges in collecting and using the data they were collecting, which in turn exacerbated the tensions. We conclude the piece by offering recommendations for more intentionally engaging in data synthesis that will benefit other institutions in similar situations.

Correspondence E-mail: sarah.hurtado@du.edu

Keywords: sexual violence, campus climate, data ecosystem, data synthesis, institutional trust

As part of their responsibilities under Title IX, institutions of higher education (IHE) are required to assess their campus climates related to the issue of sexual violence (McMahon et al., 2022). The rationale behind this mandate is that, by understanding climate, institutions are better positioned to prevent and respond to incidents of sexual violence because they have a better understanding of why the issue persists within their institution (McMahon et al., 2022). Most institutions administer large-scale, quantitative campus climate surveys to fulfill this requirement (McMahon et al., 2022). However, these typically only focus on perpetration rates, which offer a limited understanding of campus climate (McMahon et al., 2022). For this reason, Driver-Linn & Svensen (2017) recommend the implementation of a data ecosystem to understand more robustly the nature of an institution's climate for sexual violence. Additionally, most institutions rely on a limited understanding of climate focusing solely on behaviors and perceptions of the issue of sexual violence. For this reason, Moylan et al. (2021) proposed a broader conceptualization of campus climate informed by Hurtado et al.'s (2012) Diverse Learning Environments (DLE) model.

Understanding the climate of an institution, including the elements of the climate that contribute to the perpetration of sexual violence, is extremely important. This understanding builds an accurate sense of perpetration, affects programming, and assists program evaluation (Moylan et al., 2018). Yet, whether and how institutions assess a full picture of the campus climate as it relates to sexual violence remains unclear. Therefore, the purpose of this case study was to explore how institutional actors engage with data concerning campus sexual violence to implement a data ecosystem as informed by a robust climate framework. We sought to answer the following research questions:

1. What are the institutional factors that shape whether and how an institution engages in a data ecosystem related to sexual violence campus climate?
2. How do faculty, staff, and administrators understand a data ecosystem related to sexual violence campus climate?
3. How can institutions implement a data ecosystem to make context-specific, data-driven, system-level changes to address the issue of sexual violence?

A note on language: there are many terms to describe the type of violence described by this climate and conceptualized in this ecosystem. We will use the term “sexual violence” to refer to instances of assault, sexual harassment, gender-based violence, relationship violence, stalking, sexual misconduct, and sexual contact or attempted contact that is not consensual.

Conceptual Frameworks

In this study, we relied on two conceptual frameworks. First, Driver-Linn and Svensen's (2017) data ecosystem framework offers a comprehensive plan for collecting various types of campus climate data. Within their data ecosystem they include counts of behaviors, sampling of attitudes and experiences, comprehensive surveys of prevalence and climate, question-based data collection, and program evaluation (Driver-Linn & Svensen, 2017). They present both categories of data collection and suggest timelines of assessments; for example, they recommend staggering large-scale comprehensive climate

surveys every four years, with annual surveys to a narrower population, ongoing program evaluations with data-informed changes implemented and reviewed, staggered mixed-methods studies running parallel, and ongoing collection of prevalence data to be analyzed every term.

Following data collection, they articulate the need for data synthesis to provide a fuller picture of the campus climate. Unfortunately, this type of data ecosystem has not been documented as implemented by institutions (McMahon et al., 2022). One of the limitations of the data ecosystem is it focuses on collecting data from students and omits major campus populations that both play a significant role in shaping the climate and experience the climate themselves: employees including faculty, staff, and administrators (Moylan & Javorka, 2020).

Although Driver-Linn & Svensen (2017) push institutional actors to think beyond simply administering campus climate surveys, their recommendations are not informed by a campus climate framework. Therefore, we combined their concept of data ecosystem with our second framework proposed by Moylan et al. (2021) and informed by Hurtado et al.'s (2012) Diverse Learning Environments (DLE) model. In this expanded conceptualization of campus climate there are five dimensions: the historical, or context and legacy of the institution's past; the organizational, or structural components of institutional operation; the compositional, or the demographic makeup of student and employee populations; the behavioral, which encompasses observable interactions between individuals and groups; and the psychological, which includes feelings and perceptions of the climate.

While Hurtado et al. (2012) developed this model for fostering a climate of diversity, equity, and inclusion on campuses, Moylan et al. (2021) have found its consideration of holistic climate and reduction of harm useful for conceptualizing campus climates relating to sexual violence. We are using these dimensions as a model for campus climate for sexual violence and applying them as a lens for examining the synthesis component of Driver-Linn and Svensen's (2017) data ecosystem, where institutional actors collaborate to understand the implications of their collected data. Table 1 shows how these two concepts integrate into one framework.

Method

We conducted a case study using Yin's (2014) definition of case study as "empirical inquiry that investigates a contemporary phenomenon in-depth and within its real-world context" (p. 16) at one institution to understand how such a data ecosystem could be created, the importance and challenges of including faculty, staff, and administrators in an ecosystem, and to understand institutional factors that affect whether and how the institution engages in a data ecosystem related to campus climate. We chose a medium-sized private institution as the case study site because of its recent history regarding addressing concerns about the campus climate related to sexual violence. In early 2020, public activism called attention to the prevalence of sexual violence at the institution. These activists demanded the institution take active steps to address their concerns. Following this, the institution was part of the NASPA Culture of Respect cohort, which led to the administration of a campus climate survey. This demonstrated the institution was in the process of assessing their campus climate, which made them an appropriate site for

Table 1. *A Data Ecosystem Model within Diverse Learning Environments*

Ecosystem Data Type(s)	DLE Dimension(s)	Examples/Instruments
Counts of behaviors	Behavioral	Campus police reports, Student conduct reports, Title IX reports, Clery data collection
Localized sampling of attitudes and experiences	Psychological	Surveys of first-year students
Comprehensive, comparative surveys of prevalence and climate	Compositional	Campus-Wide Climate Surveys (e.g., HEDS, ACHA-NCHA)
Question-based data collection	Psychological	Focus groups for LGBTQIA+ students
Program evaluation	Organizational	Bystander prevention training evaluation, campus wide evaluations (e.g. CORE)
Synthesis	Historical Organizational	Campus-wide study of all data collected, Add individual institution's context (e.g. past social media activism); presentations to stakeholders

the study. Additionally, the institution experienced significant staff turnover in key positions related to sexual violence prevention and education over the previous few years.

Data Collection

We used two methods of data collection: artifact review and interviews. Participants also offered other artifacts during interviews, and after receiving access we reviewed these artifacts to further contextualize how the institution was thinking about campus climate data.

Artifact Review

We reviewed the following artifacts: 2022 HEDS Sexual Assault Campus Climate Survey ("HEDS Survey"), also called the Campus Perceptions of Unwanted Sexual Experiences, and the institution's report of the findings, 2023 American College Health Association National College Health Assessment survey ("ACHA NCHA"), 2021 NASPA Culture of Respect Evaluation ("CORE"), 2020 student activism social media posts, and a student-produced research report.

We reviewed the HEDS Survey by mapping each question to its corresponding climate dimensions as outlined in our framework. The institution's website did not explain which stakeholders were using the data outside of the Title IX Office. We found nearly all of the 63 undergraduate survey items related to the compositional, behavioral, and psychological dimensions. A summary of this question mapping is in Table 2.

Table 2. *Mapping the Undergraduate HEDS Survey Instrument to the DLE dimensions*

DLE Dimension	Number of questions	Example
Compositional Dimension	19 questions	“Since starting at [institution] have you been sexually assaulted [on or near campus]?”
Behavioral Dimension	32 questions	“Did this incident of sexual assault involve your drinking alcohol?”
Psychological Dimension	12 questions	“What stopped you from telling anyone about the sexual assault?”
Organizational Dimension	3 questions	“Have you received information or education from [institution] about who on campus is required to report instances of sexual violence to campus authorities?”
Historical Dimension	0 questions	No related questions

Note. Three survey questions addressed multiple dimensions, so the table reflects 66 questions.

Similarly, a study participant recommended we review the ACHA NCHA. This survey assesses health and wellness behaviors to inform health-related programming and services (American College Health Association, 2025). A summary of this question mapping is in Table 3.

Table 3. *Mapping the ACHA NCHA Survey Instrument to the DLE dimensions*

DLE Dimension	Number of questions	example
Compositional Dimension	12 questions	“Did you experience the following in an intimate relationship: Were a victim of stalking?”
Behavioral Dimension	6 questions	“Have you experienced the following while drinking alcohol: Had sex with someone without their consent?”
Psychological Dimension	9 questions	“Select your level agreement with the following: We are a university where we look out for each other.”
Organizational Dimension	0 questions	No related questions
Historical Dimension	0 questions	No related questions

The CORE included a thorough assessment of survivor supports and services, staffing and employee training protocol, training for students and student employees, collection and publication of prevalence data, engagement of faculty, staff, and students school-wide, and evaluation including campus climate surveys and other assessments. The institution’s CORE report contained recommendations to expand on data collection and assessment efforts. Additionally, the CORE recommended that the institution report to

the students and employees updates and other communication related to data collection efforts and use.

We reviewed the January 2020 student activism social media posts. The posted demands for the institution included increased diversity and intersectionality among campus security personnel, more safety resources, trauma-informed training for campus employees, increased support resources for survivors, more survivor-focused policies and procedures, increased prevention programming, and better involvement in Greek Life regarding preventing and responding to violence. Institutional leadership responded to the demands in two memos praising the student activists and outlining actions to be taken in response to the demands. These documents provided insight into the institution's historical and organizational elements of campus climate.

Lastly, a study participant shared the report from a qualitative student research project completed in a recent undergraduate social science course. This study was conducted to examine feelings and perceptions of students, and whether the institution is making progress in preventing and addressing sexual violence on campus. Their findings relate to the psychological dimension of climate, as they remark on perceptions of a lack of support for students who have experienced sexual violence.

Interviews

To understand the nature of the data ecosystem on campus, we spoke to university employees to learn how they collect, access, share, and use data regarding the campus climate for sexual violence. Our inclusion criteria for participation was twofold: 1) participants were currently employed at the institution and 2) based on title and scope of work would likely have insight into sexual violence data collection and use. Using purposeful sampling we identified several employees who met the inclusion criteria and were also likely to be primary intended users of the climate data. These employees represented different units across campus, including the Title IX Office, Clery, Campus Security, Survivor Advocacy, Health Education, Student Conduct, and Institutional Research. We then used snowball sampling by asking our participants to recommend others who could speak to a data ecosystem. These recommendations resulted in interviews with employees from the Office of Diversity and Inclusion, Student Health, and a second employee from Health Education, Student Involvement, and Diversity initiatives within an academic unit. We also invited other employees from Student Involvement, the Dean of Student Office, Assessment, and recommended faculty who all declined to participate. We maintain participant confidentiality in our findings by using pseudonyms and not including identifiable or attributable details related to their personal identities and roles on campus.

In total, we interviewed 12 employees via Zoom to provide convenience for the participants, and for the ease of the recording and transcription of each interview. The interviews were semi-structured to be a conversation between researchers and participants. Interview questions focused on the type of data each employee collects and uses related to sexual violence, how they engage with others related to these data, what challenges they have with collecting/using/sharing data, what data they wish they had to inform their jobs, and how historical records of their roles/data use are maintained.

Data Analysis

After conducting the artifact review, we identified all the types of data being collected by the institution as informed by our study participants and identified the types of data participants expressed wanting but not having. We used these lists to map each data type to the corresponding component of the data ecosystem framework as recommended by Driver-Linn and Svensen (2017) and campus climate framework (Moylan et al., 2021).

Immediately following each interview, we shared our reactions together and conducted some pre-analysis and memoing. Once all interviews were completed and transcribed, we conducted a largely inductive analysis. To ensure intercoder reliability, we started by independently coding the same interview transcript. We then came together to discuss similarities and differences in our approaches. We discussed the language and use of different codes to ensure consistency moving forward. We then separately coded different interviews. We found emerging codes related to ways participants gathered their own climate data, but also how they used existing data (e.g. from surveys). We also discovered codes relating to the relationships among employees and departments on campus, as well as tensions between transparency and confidentiality, challenges trusting data, processes, or leadership, and challenges associated with staff turnover or lack of funding for departments or projects. These codes align with themes from the artifact review. We also used deductive coding for codes associated with the DLE dimensions of climate (historical, organizational, compositional, behavioral, and psychological). This coding helped us conceptualize the data ecosystem, how the ecosystem is supported at the institution, where challenges are embedded in the ecosystem, and how employees think about the data ecosystem.

Criteria for Goodness & Delimitations

We intentionally chose a single-site case study, because we believe the institutional context is the heart of our understanding their approaches to implementing a data ecosystem. While we cannot claim generalizability of this study, we took several steps to ensure trustworthiness and transferability. The study was approved by the Institutional Review Board, and we received informed consent from participants. As noted in our analysis section, we took steps to ensure intercoder reliability. Additionally, protecting participant confidentiality by using pseudonyms and masking demographic information and departmental affiliation helped ensure trustworthiness of their responses. Lastly, we offered thick description of the institutional context, our methodological choices, and the finding. Our hope is that others will be able to apply what we learned from this study to their own institutional context.

Results

Overall, we found the institution in our study collects many of the types of data outlined in Driver-Linn and Svensen's (2017) model. However, we found the biggest challenge with the ecosystem was a lack of institution-wide synthesis of collected data. Therefore, our central finding is that a data ecosystem is more than the collection, storage, and dissemination of data. An effective ecosystem connects divisions across campus and is represented in not only the collection efforts but also the relationships across the entire institution, especially concerning synthesis. These ecosystems require institution-wide

synthesis, trust, and continuity. They balance tensions created between transparency and confidentiality, and perception and reality.

A Data Ecosystem is as much about relationships as it is about data, and staffing challenges impede the sharing and synthesis of campus climate data

Driver-Linn and Svensen's (2017) ecosystem model includes synthesis, or "look[ing] at the whole picture of data collected and present to stakeholders and community...to build robust, interconnected data practices that allow for systematic and relatively automated examination of trends around important metrics" (Driver-Linn & Svensen, 2017, p. 2–3). We discovered that synthesizing and sharing data between departments in this case is not institutionalized and instead relies on relationships between individuals. When those relationships suffer, or when someone leaves their position, the sharing of data suffers. We found synthesis is challenging because of issues related to personnel, including employee turnover and tensions between balancing transparency and confidentiality.

Staff turnover plays a big role in the breakdown or lack of relationships and sharing between departments. In some instances, the employee someone had a relationship with left the institution and they need to start over building a relationship with a new person. As Dana explained:

They have new staff, so I think we have to redevelop that [relationship]. But I had been really close with the previous director...and talked a lot and shared a lot of information with that staff, and so now we're going to rebuild those relationships.

In other instances, a department is now understaffed and cannot continue to collaborate regularly due to capacity. Taylor explained the impact of this turnover:

Staff keep transitioning pretty quickly in and out of departments. I think they just don't feel that they have (and I believe they don't have) the capacity to focus on something else that's not their main job objectives.

Sam reflected on how their ability to work was interrupted due to a turnover of nearly their entire department:

Our team lost a lot of institutional knowledge...we have obviously some transition documents that exist that do allow us to get some information from those people...but in terms of an explicit 'Here's the information you need to do your job the best that you can, and here are the places for you to go in order to access more information to be able to better inform your work,' I don't think we had that.

Part of this team's turnover meant that new members did not have access to the data from the most recent climate survey. When asked if they had seen or used data from the climate survey, Robin answered, "That is not data that has come across my email, to my knowledge." They explained that the data they were using to inform their work were

what had been shared directly by others engaged in similar work. This means that Robin's data access is limited by their individual relationships with others, and that valuable climate survey data is inaccessible.

Some employees try to mitigate the disruption of turnover internally. Dana explained how their department maintains an ongoing shared document to onboard new employees faster and to ensure work can continue despite changes in personnel. They referred to this practice as their "standard operating procedures" and explained how they use this method to share information among the team and with future new members. This is an answer to Sam's concern of not having enough guidance to continue the work of a department after turnover, yet this procedure is voluntary and department specific. However, these procedures do not include specific recommendations related to ensuring continued sharing and use of important campus climate data.

While staffing challenges impede the relationships on campus that fuel the data ecosystem, staffing also disrupts the work of preventing violence. Morgan reflected:

We need to get [our efforts] way upstream rather than reacting [after violence occurs]. The challenge is we just don't have capacity to be upstream...But in a perfect world we would have all of the necessary resources, however we want to define that, to be upstream all the time.

Since employees in understaffed departments often undertake the work of two positions, they are unable to dedicate time to prevention efforts, including using resources like climate data or conducting their own smaller-scale assessments as recommended by Driver-Linn and Svensen (2017).

A major asset to the collection, synthesis, and sharing of climate data, and therefore an important component of the data ecosystem, is a robust employee base with time to build systematic relationships, the means to preserve institutional knowledge, and the staffing to ensure necessary work can be accomplished. Without this operationalized network, access and sharing of data relies on arbitrary relationships among staff, and efforts for smaller-scale data collection and assessment are limited.

Relationships are challenged by tensions with trust, transparency/confidentiality, and perception/reality

For the interdepartmental relationships that do exist, we found several tensions in trying to balance efforts of sharing data, protecting the privacy of both students and survey responders, and making progress on efforts despite staffing challenges. These tensions represent different themes: One is the feelings of trust around data collection and use, one is the tension from balancing transparency and the need for confidentiality, and one is the tension between perceived work and processes and the reality of the work being conducted on the campus.

Tensions with Trust

A common sentiment from participants was a lack of trust with either the data collected, the processes and effort to address sexual violence, or both. Participants explained that faculty and students do not trust results of climate surveys, do not trust the

process of collection and analysis, do not trust the institution is working to prevent and respond to violence, and do not trust that preventing and responding to violence is an institutional priority.

First, participants discussed a lack of trust in the data being collected. A few admitted they were reluctant to put much faith in the data from the previous campus climate survey due to the low response rate, with Alex explaining:

We didn't want to rely too much on that data because we knew that the response rate was fairly low...We didn't want to put as much emphasis on the reliability of the data and generalizing to the whole campus community, because we knew it wasn't particularly good response rate. It wasn't particularly representative, either.

Others were explicitly cautioned from putting too much stock in the results. Some participants believed they were encouraged to dismiss the survey results due to unsatisfactory results. Taylor explained how the data interpretation could seem arbitrary:

People like the data when it shows them what they what they like to see. And then they're very happy to share it when it does. And they're less happy to share when the data is obviously a little more negative.

Taylor, Morgan, and Rory each expressed distrust in the institution, and a feeling that preventing and responding to sexual violence on campus is not currently a priority for the administration. To them, this was evidenced by a lack of funding for initiatives, understaffed offices, and a general lack of attention paid to the issue. Rory elaborated further:

At the leadership level it's not a priority. So...and I say that across the board, from a sexual violence perspective, gender identity violence, to, you know, racial violence, to you know all sorts of different types of harm that we see across campus, it's not a priority for our leadership right now.

This lack of trust in the institutional commitment affects the ability to collect data in the future, as students and employees are reluctant to participate because they do not believe the data will be used to advance meaningful work.

Some participants observed a hesitancy by the institution to prioritize data collection and response to sexual violence. Taylor explained, "I think people are afraid of backlash or additional questions, or you know, 'What are you doing about this [violence]?'". Multiple participants implied the institution is not always transparent about its sexual violence data and efforts because the institution is afraid of how it will be perceived. It is almost like the institution does not trust its students or employees will believe it is working to prevent and respond to sexual violence.

Trust with the data ecosystem is important because if students and employees do not trust the institution with their data, or do not trust the institution to act on the data collected, then they are less likely to participate in data collection efforts. This continues the cycle of low response rates and therefore less trust in future data.

Tensions with transparency versus confidentiality

Participants pointed to a tension between transparency and confidentiality with data collection, utilization, and investigation efforts. For example, many reported not understanding why students and employees cannot learn the outcome of some Title IX investigations, or about sexual violence prevalence trends. Some departments mentioned sharing de-identified information with other departments, but others limited access to even de-identified data due to its confidential nature. Pat acknowledged that some information should be kept confidential to protect individuals' privacy, especially with the campus climate survey, but that the institution needed to be transparent about information needing this protection: "I know some of the data you're going to collect can't be [shared], so I think that's the other [important] thing is full transparency about what you can and cannot share from the get-go." Clarification on what information can and cannot be shared would be helpful for them but is not available.

Participants understood that confidentiality and privacy are necessary due to the sensitive nature of this data, but believed some information should be available. Alex said, "Even knowing confidentiality is one component of things...it'd be really helpful to see themes and patterns and general de-identified case studies, and communicate [all] together as a team." There is a strong desire to understand the climate better using data, but those data are inaccessible due to privacy concerns.

While many participants spoke on the trust in data availability and use, others considered trust in procedures. Pat framed their thinking by explaining the idea of procedural trust among employees:

Faculty and staff want procedural trust, procedural justice. They want to know what is happening [within Title IX processes], how it's going to happen, where's the transparency, where is the outcome, who's accountable, what are the accountability structures? Those are the data points they actually want to see. It may not be a specific incident but it's the accumulation of incidents. So, '[Title IX Office] got this many complaints: Where did they all go? How were they handled? Who made the decisions? Why was this one decided this way, and this one decided a different way?' They want to see the scoping, right? They want to see the ecosystem. They want to see the levers behind to be able to trust that the university is also caring for them, not just looking out for its bottom line.

Other participants reflected on their understanding of student perspectives regarding transparency, and student desire to understand more about processes and outcomes. Taylor stated:

Students want to know that they're being heard ... [that] what they say isn't going into some nebulous cloud that they'll never see again, and they don't know how it impacts current programming or anything that they're being asked about on campus.

Several participants asked for more transparency with how the institution uses data to inform initiatives, and for more sharing about changes that were made in response

to data. This sharing extends to general outcomes of investigations and reports made by students. If students and faculty are aware that violence occurs, but are not made aware of the resolution, then the perception remains that harm continues with no action by the institution. Pat summed up the necessity of this sharing of prevalence and action data:

A student [believes], 'I want to know that what happened to me will change something,' and maybe the student gets that. But what if that [resolution] never loops back to [a] faculty member? So those faculty members, all they know is 20 years of harm. And so that further potentially further replicates this notion of distrust...Faculty are like, 'Yeah, I've heard these stories over and over again; [the institution has] done nothing'.

This is the historical dimension of climate, as offered by the DLE. Pat went on to explain institutional memory:

If you're going to change something you have to change [it for] the faculty and staff, not the students, actually. The students come and go...an institution only has a memory of four years. So, if you can navigate the waters of a really crappy incident within four years, it's gone for students... But what helped me in my career is like, "Oh, if I want to change things I have to work with faculty and staff. They're here longer. They have the memory."

To create and maintain a functional data ecosystem for climate, institutions must find ways to safely share specific data when it can, share other de-identified information, and set clear expectations with rationale with the campus community for what cannot be shared.

Tensions with perception versus reality

We discovered tensions between the perceptions and the reality of campus climate, processes, and work being done. One prominent example is how the Title IX Office is perceived. Several participants commented on how students or employees feel when they make reports to the Title IX Office. Pat observed the office "feels like...a black hole to a lot of people. Like, things just get sucked into it and they never show back up, including, people feel, their souls." The perception is Title IX issues are not resolved adequately, leading to anxiety.

There is also a perception that collected data are not being used to improve the campus climate. According to Alex:

It's great that we're telling the deans what this data is, but how are we actually changing our programs and adapting to student needs if we're only getting bogged down in the presentations to the leadership that aren't actually implementing the programs?

Their implication is that data collection and sharing here is a performative act and is not actually informing decision-making.

Finally, there is a perception that campus members do not collaborate across divisions or departments in their work despite examples to the contrary. Robin spoke to this:

Something that I've heard a couple of times now, in some of my campus partner meetings, is reflection on how sometimes how siloed [the institution] can be in its work. And in this position...it's requiring me to not be siloed. I can't do this position in a vacuum. I haven't experienced the siloing as much at [the institution] quite yet, but it's reflection that I've got from a couple different folks of people feeling kind of as though they're on their own island.

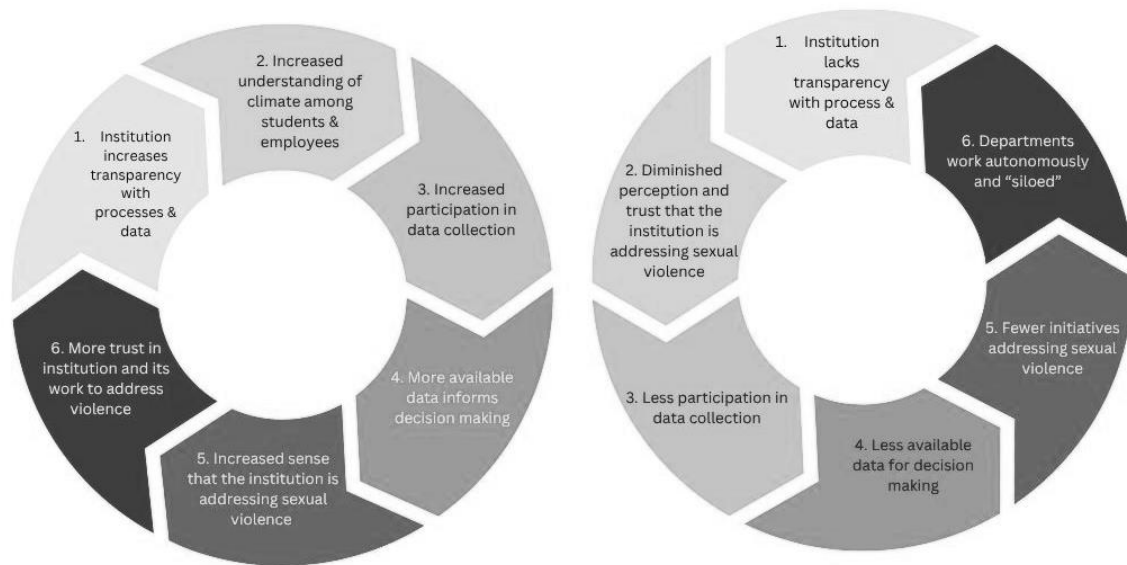
Whether employees' perceptions of the campus climate align with reality is not as important as understanding how these perceptions inform behavior. When students or employees perceive a lack of institutional commitment, they are less likely to report incidents. If they perceive the institution does not act upon its collected data, they are less likely to participate in data collection. This compromises the institution's assessment efforts across all sectors, not simply as it relates to campus climate for sexual violence.

Figure 1 illustrates these tensions, and how we found they contribute to a cycle of added or diminished trust, transparency, and perceptions on campus.

Discussion

Beyond merely collecting campus climate data, institutional actors need to attend to the data synthesis challenges and tensions we named above. Collecting data is not enough to position institutional actors to address the issue of campus sexual violence. And, in using a robust understanding of campus climate (Moylan et al., 2021), institutional actors might recognize that these challenges in relationships, trust, transparency/confidentiality, and perception/reality, are part of the climate as well. Staff turnover, loss of institutional memory, and inability to implement change using previously collected data is part of the historical dimension of climate. The lack of institutionalized processes in facilitating data synthesis is part of the organizational dimension. The perceptions of our participants that their institution does not prioritize sexual violence prevention (organizational dimension), and that data collection is performative (psychological dimension) then influences their ability to engage in synthesis of data in their work (behavior dimension). These experiences from the participants represent an important understanding of the campus climate, which ultimately shapes the experiences and perceptions of students. With this in mind, we make the following recommendations for the case institution that can serve as an example for others looking to improve their assessment of campus climate for sexual violence. The appendix offers a guide for applying these themes to different contexts.

Figure 1. *Cycles of Trust and Transparency*



Note. The left cycle reflects increased transparency increasing trust and perceptions of institutional action. The right cycle reflects decreased transparency diminishing trust and creating a perception of inaction.

1. Rebuild trust by increasing transparency with planning the next climate survey

In preparing the next plan for surveying campus climate, this institution should take active steps to increase its transparency. It should create a plan that outlines both a clear timeline (including data analysis and reporting) and for the use of collected data. The institution can also regain trust by articulating how data from the previous survey informed decision-making and policy changes. Several participants shared how they used data from the previous climate survey in programming and education, so those efforts should be communicated broadly.

2. Be transparent about how data will be shared and what will be kept confidential

Several participants expressed a lack of clarity on privacy procedures and a suspicion that patterns and trends from the survey could have been easily de-identified and shared but simply were not. This frustrated and confused them, and diminished their trust. Some employees believe the institution purposefully withheld useful data about the climate. Therefore this institution must be clear when it collects data on how privacy will be ensured, how data will be shared, and whether there is any data that cannot be shared (e.g. student demographic data) in the event such data compromises student privacy.

3. Share and use collected data

Once survey data are compiled, this institution must follow through to both share results and then use the data (regardless of response rate) to inform decision-making. Several participants wished for the institution to act using data, and if action had been taken to communicate which actions were informed by survey data. This communication, explanation, and sharing will help to build trust in the survey, in the administration, and

in the institutional culture, because there will be more transparency in what the data say and how they are used.

While the synthesis step of the data ecosystem is happening on a small scale on this campus, this synthesis must occur on a institutional scale. Furthermore, data must be easily available, and employees must have capacity to use data and collaborate with others in their work.

Conclusion

As noted, this institution is engaging in a data ecosystem by collecting a variety of data types, over time, from different campus constituents. Multiple offices are involved in this ecosystem, which establishes significant potential for thorough synthesis. However, the institution has an opportunity to engage more robustly in data synthesis by attending to the issues of trust and transparency related to sexual violence campus climate data.

If an institution is not synthesizing its collected data regarding sexual violence, then it is not truly understanding its campus climate. Since the goal of the campus climate survey and other data collection efforts is to inform efforts to prevent and address sexual violence, an institution not meaningfully understanding its climate is not equipped to do this work. Institutions have an obligation to prevent harm to their students and employees, and by not meaningfully engaging in these assessments then an institution wastes time and resources on data that do not protect the community. Institutions must act to create more meaningful data ecosystems to conduct this critical work.

Appendices

[Appendix A: Ecosystem Development Guide for Practitioners](#)

References

- American College Health Association. (2023). *NCHA III Codebook*. https://www.acha.org/wp-content/uploads/ACHANCHA_III_Spring_2023_Codebook_2_8_2023.pdf
- American College Health Association. (2025). *National college health assessment*. <https://www.acha.org/ncha/>
- Driver-Linn, E., & Svensen, L. (2017). Moving toward a “data ecosystem” to assess campus responses to sexual assault and misconduct. A resource for college and university decision-makers. *Association of American Universities*, (1-9). <https://www.aau.edu/sites/default/files/AAU-Files/Key-Issues/Campus-Safety/Data-Ecosystem-Approach-Resource-Guide>.
- Hurtado, S., Alvarez, C. L., Guillermo-Wann, C., Cuellar, M., & Arellano, L. (2012). A model for diverse learning environments. In J. C. Smart & M. B. Paulsen (Eds.) *Higher Education: Handbook of Theory and Research*, 27, Springer.
- McMahon, S., Cusano, J., Buttner, C., Snyder, S., Ast, R. S., & Camerer, K. (2022). Evaluating efforts to address campus sexual violence: Developing a data ecosystem. *Journal of Interpersonal Violence*, 37(23–24), NP23563–NP23586. <https://doi.org/10.1177/08862605221078817>
- Moylan, C. A., Hatfield, C., & Randall, J. (2018). Campus sexual assault climate surveys: A brief exploration of publicly available reports. *Journal of American College Health*, 66(6), 445–449. <https://doi.org/10.1080/07448481.2018.1431914>
- Moylan, C. A., Javorka, M., Maas, M. K., Meier, E., & McCauley, H. L. (2021). Campus sexual assault climate: Toward an expanded definition and improved assessment. *Psychology of Violence*, 11(3), 296. <https://doi.org/10.1037/vio0000382>
- Moylan, C. A., & Javorka, M. (2020). Widening the lens: An ecological review of campus sexual assault. *Trauma, Violence, & Abuse*, 21(1), 179–192, <https://doi.org/10.1177/1524838018756121>
- Yin, R. K. (2014). *Case study research: Design and methods*. SAGE.

Implementation Fidelity in Higher Education: A Scoping Review



Authors:

Sarah Kirkland, EdD
Columbia Southern University

Elizabeth Hinton, EdD, MSIS, AHIP
University of Mississippi Medical Center

Carley Dear, EdD
University of Mississippi Medical Center

Xiaomei Song, PhD
Case Western Reserve University

Anthony Arciero, PhD
Teaching Happiness, Inc.

ABSTRACT

Assessment in higher education seeks to improve program outcomes through rigorous evaluation. Implementation fidelity measures the degree to which educational interventions are carried out as planned, which is critical for ensuring program effectiveness. This scoping review investigates implementation fidelity in higher education assessment, focusing on definitions, frameworks, measurement approaches, and their integration into assessment practices. A comprehensive search strategy yielded 55 relevant articles from various databases and gray literature sources. The inclusion criteria included studies that investigated implementation fidelity in postsecondary education in the United States, using both qualitative and quantitative methodologies. The review identifies several definitions of implementation fidelity, focusing on factors such as adherence, exposure, delivery quality, participant responsiveness, and program differentiation. Several frameworks for assessing implementation fidelity were identified, including component-based and structured models, which helped to understand intervention execution. Measurement methods ranged from traditional observations to novel approaches, including self-reports and multi-step iterative processes. Implementation fidelity shows promise in improving the quality and reliability of educational assessments in higher education. However, challenges persist, including the need for standardized definitions and validated measurement instruments. Future research should focus on refining implementation fidelity frameworks, validating measurement tools across various contexts, and integrating implementation fidelity seamlessly into assessment practices to promote continuous improvement in educational interventions.

Correspondence E-mail: Sarah.Kirkland@columbiasouthern.edu

Keywords: assessment process, implementation fidelity, measuring results

Within an assessment framework, analyzing program outcomes leads to continual quality improvement. Traditional assessment models in higher education identify outcomes and establish measurement criteria and metrics for determining success, analyzing results, and developing strategies to improve ongoing performance (Fisher et al., 2014). However, while outcomes data indicate whether a program "worked," they do not explain why it did or did not succeed. Implementation fidelity (IF) examines the extent to which an assessment program or educational initiative is carried out as designed, ensuring that evaluation findings accurately reflect program performance (Carroll et al., 2007; Finney & Smith, 2016; Fisher et al., 2014; Gerstner & Finney, 2013; Mihalic, 2004; O'Donnell, 2008). Simply put, IF connects program design to program outcomes, providing insight into whether disappointing outcomes are due to flaws in the intervention or its implementation (Carroll et al., 2007; O'Donnell, 2008).

For example, a retention strategy might have little effect on student persistence. Without IF data, one could conclude that the project failed. However, fidelity data may indicate that important advising sessions were delivered inconsistently, suggesting that the intervention was not fully implemented. Conversely, if an intervention produces good outcomes while maintaining high fidelity, assessors can be more certain that the results were driven by program design rather than chance or other influences.

Calls to include IF in higher education assessments are not a new phenomenon. In 2013, Gerstner and Finney recommended including fidelity checks into assessment processes, claiming that outcome data should not be evaluated in isolation. Despite this call, IF remains underutilized in higher education ten years later. While some academics have begun to identify frameworks and instruments, there is currently no agreement on definitions, measuring methods, or practice standards. This gap makes it difficult for institutions to compare findings from different studies or adapt lessons learned in one environment to another.

Professional standards have also acknowledged the value of IF. The current CAS Standards (Council for the Advancement of Standards in Higher Education, 2019) expressly require assessing both learning and program effectiveness using evidence of IF. In other words, institutions are now expected not only to measure outcomes but also to assess whether initiatives were implemented as intended. Aligning higher education assessment with these aspirations necessitates a more thorough synthesis of existing literature.

This scoping study meets that requirement by outlining the landscape of IF research in higher education—what has been done, what is still inconsistent, and where the field needs to go next. By doing so, it addresses both previous scholarly calls and professional standards that require greater integrity in assessment. The study examines 55 papers, focusing on definitions, frameworks, and measurement methodologies, and suggests a more evidence-based and actionable assessment culture.

Aims of This Review

The purpose of the IF scoping review was to:

1. Identify and compare the various definitions of IF used in higher education evaluation, noting points of convergence and divergence.

2. Examine existing frameworks that incorporate IF into higher education assessment and evaluate how they have been implemented.
3. Examine and analyze techniques of measuring IF, including their strengths, limitations, and compliance with professional standards (e.g., CAS).

Method

This scoping review followed established methodological guidance for conducting scoping reviews. We adhered to the framework proposed by Arksey and O'Malley (2005), which outlines five stages: identifying the research question, identifying relevant studies, selecting studies, charting the data, and collating, summarizing, and reporting the results. This framework was further refined following recommendations from Levac et al. (2010), emphasizing iterative team-based screening, clear inclusion criteria, and the use of data charting tools to enhance rigor and transparency. To ensure comprehensive and reproducible reporting, we also aligned the review process with the PRISMA extension for scoping reviews (Tricco et al., 2018). The PRISMA-ScR checklist guided the reporting of our search strategy, study selection, and data charting procedures, including the presentation of the PRISMA flow diagram (Appendix B).

Search Strategy

The search strategy aimed to find both published and unpublished studies. A three-step strategy was used and included electronic and manual searches of reference lists, as well as journal and professional association websites for gray literature. The initial step involved a limited search of PubMed and ERIC, followed by an analysis of the text words in the title and abstract, as well as the index terms used to describe the articles. This review informed the development of the search strategy tailored for each information source. The second step was a search using all keywords and index terms identified in the electronic databases listed below. Finally, the reference lists of the included studies were reviewed for relevant citations. The search was limited to studies published in English, and each database was searched from inception. The databases searched included:

- PubMed (U.S. National Library of Medicine)
- Web of Science Core Collection (Web of Science/Clarivate Analytics)
- Teacher Reference Center (EBSCO)
- ProQuest Dissertations & Theses Global: The Sciences & Engineering Collection
- Educational Administration Abstracts (EBSCO)
- ERIC (EBSCO)
- Academic Search Premier (EBSCO)
- APA PsycARTICLES (EBSCO)
- APA PsycINFO (EBSCO)
- ScienceDirect (Elsevier)
- JSTOR
- OECD iLibrary (<https://www.oecd-ilibrary.org/>)
- MedEdPORTAL (<https://www.mededportal.org/>)
- Wiley Online Library (<https://onlinelibrary.wiley.com/>)

and the websites for:

- About Campus
- Intersection
- Research and Practice in Assessment
- Journal of Educational Research
- American Evaluation Association
- AALHE
- AERA
- National Conference on Student Assessment
- Assessment in Higher Education Network.

All searches were performed in July and August 2022. The complete search strategies for all databases are presented in Appendix A.

Study Selection

All identified citations were imported into EndNote (Version 20), and duplicates were removed. A total of 2,343 citations were identified through searching, and 1,923 unique citations remained after duplicates were removed. The remaining citations were imported into Rayyan (Ouzzani et al., 2016) for screening. Two independent reviewers conducted a screening of the titles and abstracts of citations to ensure alignment with the inclusion criteria for this scoping review. The 128 potentially relevant studies were retrieved in full text, and their citation details were imported into a data extraction spreadsheet in Google Sheets. Two independent reviewers assessed the full text of the selected citations in detail against the inclusion criteria. Full-text studies that did not meet the inclusion criteria were excluded. Disagreements between the reviewers at each stage of the study selection process were resolved through consensus or by a third reviewer. A total of 55 articles were included in the study. The search results were reported using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram (Appendix B).

Inclusion Criteria

Participants

Participants included students, faculty, or staff in higher education institutions and programs.

Concept

This review included scholarly articles on interventions measuring IF, essays, dissertations, commentary, and other works specifically addressing the inclusion of IF in assessment within the higher education context.

Context

The context of this review focused on postsecondary-level institutions of higher education within the United States.

Types of Sources

A broad range of literature was included in this scoping review to capture the definition, framing, and measurement of IF in higher education assessment. Both conceptual and empirical works were taken into account. Multiple types of studies were gathered: quantitative studies (single-group, experimental, and quasi-experimental designs that are common in higher education practice such as course-based interventions, retention initiatives, and pre/posttest program evaluations), qualitative studies (action research, feminist research, phenomenology, grounded theory, ethnography, and qualitative description), and mixed-methods studies that incorporate both qualitative and quantitative techniques. Additionally, non-empirical sources, including dissertations, technical reports, conceptual papers, literature reviews, and commentaries, were also used.

This range was important because much of the higher education literature on IF consists of single-group evaluations, conceptual analyses, or descriptive reports rather than controlled experiments. Excluding such sources would have left out a significant percentage of the field's present practice and emerging debates.

Data Extraction and Analysis

Data were extracted independently using Qualtrics, with questions resolved by a second reviewer. Extracted data included source type, context/program in focus, assessment methodology, data collection method, data analysis approach, units of analysis (e.g., program-level, course-level, student-level), findings, and key insights. The full data extraction survey instrument is included in Appendix C.

For data analysis, we examined the articles from three distinct perspectives. We analyzed how IF is defined within assessment contexts. We captured the various frameworks used to incorporate fidelity into program evaluation. Finally, we examined the methods for measuring IF in assessment practice. We conducted a thematic analysis for each of these areas to identify trends, gaps, and innovative approaches. The findings offer guidance for higher education practitioners on integrating IF into their assessment practices.

Results

This section presents the findings of the scoping review. To ensure clarity, we first define key terms related to IF and distinguish between programs and interventions, levels of analysis, and IF descriptors. These definitions provide a foundation for understanding subsequent analyses and interpretations.

Definitions

IF investigates the extent to which an intervention is carried out as intended (Carroll et al., 2007; Dhillon et al., 2015; O'Donnell, 2008). Dhillon et al. (2015) define IF as "the degree to which a program model [educational intervention] is instituted as intended" (p. 9). Other terminology for IF includes implemented curriculum, program integrity, treatment integrity, and clinical effectiveness (Dhillon et al., 2015; Mellard, 2010).

In this review, programs are defined as structured curricular or co-curricular initiatives designed to achieve educational outcomes, often spanning multiple courses or

activities (e.g., first-year experience programs, student affairs initiatives) (Banta & Palomba, 2015). Interventions, by contrast, are narrower instructional or support strategies implemented within a course or specific learning environment (e.g., supplemental workshops, redesigned assignments, or flipped-classroom approaches) (O'Donnell, 2008).

IF can be assessed at different levels of analysis. Course-level fidelity refers to the consistency with which an instructor or facilitator delivers a particular intervention within a course setting (Stains et al., 2017). Program-level fidelity refers to whether multiple components of a broader initiative are implemented as designed across courses, departments, or student populations (Banta & Palomba, 2015).

Several key factors, known as IF descriptors, are frequently used to define fidelity. Program differentiation refers to the distinctiveness of an intervention compared to other programs or conditions. Exposure refers to the frequency and duration with which participants receive the intervention. Adherence refers to how closely implementers adhere to the intervention's design. Delivery quality is the ability or proficiency with which implementers carry out the intervention. Participant responsiveness is the extent to which participants interact with or respond to the intervention (Carroll et al., 2007).

Smith et al. (2017) categorize these dimensions into five components for evaluation. First, evaluators must identify the core components of the intervention. Next, they determine the extent to which each component was implemented. Evaluation also considers the quality of implementation, followed by an assessment of student response and engagement. Finally, evaluators examine the duration of implementation to capture whether the intervention was sustained as intended.

Table 1. *Definitions of Implementation Fidelity Across Sources*

Source	Definition of IF	Key Components Highlighted
Dhillon et al., 2015	“Degree to which a program model is instituted as intended”	Adherence, exposure, delivery quality, participant responsiveness
Dane & Schneider, 1998	Fidelity encompasses program differentiation, exposure, adherence, delivery quality, participant responsiveness	Program differentiation, exposure, adherence, delivery quality, participant responsiveness
Smith et al., 2017	Focus on operationalizing fidelity for evaluation	Intervention components, implementation, quality, engagement, duration

Applications and Frameworks of Implementation Fidelity

IF is a useful tool for understanding and enhancing educational interventions, as well as a definitional notion. Collecting IF data guarantees that evaluators do not base their judgment of a program's performance exclusively on outcomes. For example, an intervention may appear ineffective if key components are inconsistently delivered, whereas high fidelity combined with positive outcomes provides confidence that the program design, rather than external factors, drove the results (Swain et al., 2013; Dumas et al., 2001; Gerstner & Finney, 2013).

Integration with Outcomes Assessment

Despite its importance, IF is sometimes overlooked in outcome assessment. Few researchers use fidelity data to interpret outcomes (Berman & McLaughlin, 1976; Dhillon et al., 2015). When applied, IF data enable researchers and practitioners to tailor interventions to the context and better understand why a program succeeds or fails (Smith et al., 2017).

Applications in Higher Education and Beyond

Programs in higher education, particularly student affairs initiatives, should be deliberately conceived and faithfully implemented to achieve intended results (Banta & Palomba, 2015; Durlak & DuPre, 2008; Finney & Smith, 2016). Beyond schooling, IF has been used in health interventions (Dusenbury et al., 2003; Readdean, 2022) and psychology (Cook & Shadish, 1986; Dane & Schneider, 1998; Reinke et al., 2012). A common thread throughout these domains is that assessing fidelity aids in determining which program components are beneficial, which require adjustment, and whether outcome data are reliable (Cook & Shadish, 1986; Gerstner & Finney, 2013).

Methodological Considerations

Methodological issues persist in fidelity assessment, including inconsistency and lack of precision between studies (Dane & Schneider, 1998). Data collection tactics include observations, fidelity checklists, and numerous data sources to ensure that program effectiveness interpretations are based on actual implementation rather than preconceptions.

Frameworks and Execution

An IF framework can help identify the alignment between planned and implemented educational programs. In higher education, programs are structured sets of learning experiences designed to achieve specific goals, such as degree programs, credential tracks, or student affairs initiatives. Interventions are more targeted tactics or activities used within a program to achieve specific goals, such as workshops, tutoring sessions, mentoring, or new instructional approaches. IF frameworks help higher education professionals discover potential differences in process factors, which are conditions or activities that influence delivery (e.g., facilitator expertise, participant engagement, or resource availability) and contribute to desired outcomes.

This scoping review found that the IF literature adopted various frameworks. One frequently mentioned framework is the component-based approach, which focuses on measuring structural and process aspects of implementation (Offerdahl et al., 2018; Stains & Vickrey, 2017). Structural components encompass the expected organizational features of the interventions, whereas process components pertain to how the intervention is expected to be implemented, including the presenter's and participants' anticipated behaviors (Century et al., 2010). For example, Chia (2012) examined how two performance tasks designed for American students were adapted for use in the languages and cultures of five participating countries. Informed by this framework, Chia (2012) examined several factors related to the IF of translation and adaptation procedures, including the materials provided, review stages included, the theoretical mindset of country team members, available expertise, and external pressures within the country.

Another frequently referenced IF framework was developed at James Madison University (Finney et al., 2021; Fisher et al., 2014; Smith & Finney, 2020; Smith et al., 2017), and outlines five components to make data-informed decisions about a program's performance:

1. Adherence: Did implementers do what was expected?
2. Exposure: Did participants receive as much as expected?
3. Quality of Delivery: Did implementers perform activities in the manner expected?
4. Participant Responsiveness: Did participants follow through as expected?
5. Program Differentiation: Did the treatment condition differ from the control condition as expected?

For instance, Gerstner and Finney (2013) used this framework to measure IF in a student affairs program. They placed IF on the assessment planning chart after defining objectives and designing instruments to measure effectiveness, and before reviewing outcome data.

Nevertheless, this scoping review acknowledges that the program may not be implemented exactly as intended in every context. Factors such as participant expertise, available resources, and support systems can significantly influence the quality of implementation. Viennet and Pont (2017) elaborated on the multidirectional process that entails continuous interaction between policymakers, the public, and implementers at different levels. In their work, Rhodes et al. (2021) summarized four primary steps to creating a fidelity model: (1) articulating the change and logic model; (2) identifying critical components; (3) selecting fidelity variables and definitions of measures; and (4) applying fidelity scores.

Nelson et al. (2010) introduced a five-step model for implementation fidelity assessment: (1) specify the intervention model; (2) identify appropriate fidelity indices; (3) determine index reliability and validity; (4) combine indices where appropriate; and (5) link fidelity to outcomes where possible. The researchers acknowledged that the intervention-as-implemented often differs from the intervention-as-designed, underscoring the importance of researchers' efforts to adapt to changes and enhance fidelity. Armed with these insights, well-informed decisions can be made to improve and optimize the delivery of educational interventions.

A well-accepted framework of IF describes the components that have been considered essential for the effective implementation of the intervention or program within the specific context. It identifies the critical elements of the intervention that must be enacted for the intervention to succeed and measures whether they were enacted as part of the intervention, along with the quality of their delivery. The length of time spent on each critical element, including resources and support systems, participants' beliefs and values, and their engagement during the intervention, is also crucial (Smith et al., 2017; Hamre et al., 2011). If the IF results do not reflect the program outcomes, there are at least two issues for researchers to address: (1) Were the correct items included in the measure? (2) Can the intervention results be trusted? Table 2 outlines the major IF frameworks found in the review.

Table 2. *Comparison of Major Implementation Fidelity Frameworks*

Framework	Core Components	Measurement Focus	Example Use
Offerdahl et al., 2018	Structure, Process	Observations, participant behaviors	Cross-cultural adaptation studies
James Madison University	Adherence, Exposure, Delivery Quality, Participant Responsiveness, Program Differentiation	Alignment with planned intervention	Student affairs program evaluation
Rhodes et al., 2021	Logic model articulation, component identification, fidelity variables, scoring	Multi-step iterative analysis	Policy and program implementation fidelity
Nelson et al., 2010	Specify model, identify indices, determine reliability/validity, link to outcomes	Quantitative indices, linkage to outcomes	Intervention outcome studies

Measurement

The primary reason for measuring IF is to facilitate the interpretation of study results. High fidelity increases confidence that outcomes reflect the intervention itself. For example, if a program successfully achieves its outcomes and is implemented with high fidelity, the results suggest that the intervention worked as intended. Conversely, if outcomes are positive but fidelity is low, other causal factors may explain the results (Mokher et al., 2016), which must be further explored. This example highlights the importance of accurately measuring IF in interpreting study results.

A key concept in IF measurement is understanding how each intervention element was enacted (Mokher et al., 2016; Readdean, 2022). If some elements were poorly implemented while others were delivered effectively, those poorly implemented elements are candidates for improvement. Measuring IF can therefore support continuous refinement of both interventions and assessment practice.

Most Common Approaches to Measuring Implementation Fidelity

Most rigorous IF measurement efforts begin with a basic framework (Smith et al., 2017). Many of these frameworks originated in K–12 education, clinical psychology, or public health, but have been adapted for higher education research to account for differences in context, participants, and program structure (Smith et al., 2017; Mokher et al., 2016). When using these frameworks in higher education, researchers typically modify intervention components, measurement tools, and analytic strategies to reflect institutional settings, adult learners, and disciplinary differences.

Common methods include observations (in-person or video), self-reports by facilitators and participants, interviews, focus groups, logs, document review, and sometimes covert observations. Resource availability and study complexity influence which methods are feasible. IF research generally asks, “*Was the intervention carried out as intended?*” Using multiple methods provides triangulation, balancing the strengths and weaknesses of each method, and increasing measurement validity (Mokher et al., 2016; Reinke et al., 2012; Smith et al., 2017; Zoblotsky & Gallagher, 2020). When using

observational or rating instruments, establishing inter-rater reliability ensures consistency across raters.

Innovative Approaches to Measuring Implementation Fidelity

Innovative approaches often use multi-step or iterative processes. For instance, researchers may develop a measure based on a framework, collect initial IF data, and analyze correlations with program outcomes. Measures may then be refined to improve accuracy and relevance in subsequent interventions (Offerdahl et al., 2018; Rhodes et al., 2021).

Similarly, when dividing intervention elements into “structural” and “process” components (Farr et al., 2009; Hill & Erickson, 2019; Stains & Vickrey, 2017), it is important to note that these categories were originally developed for broader educational interventions but are frequently adapted in higher education studies to reflect course design, faculty facilitation, and student engagement.

Linking Implementation Fidelity Components to Outcomes

IF data are paired with outcomes data to interpret program effectiveness (Mokher et al., 2016; Smith et al., 2017; Reinke et al., 2012). For example, a faculty professional development program in a higher-education setting may achieve high adherence and delivery quality but moderate participant responsiveness. Outcome data may indicate modest gains in faculty adoption of active learning strategies. The interpretation logic suggests that while the program was mainly delivered as intended, lower participant engagement may have limited its overall impact. Similarly, in an academic advising initiative aimed at improving student retention, fidelity data may indicate that advisors complete nearly all required check-ins with students (high adherence), but many sessions are rushed and lack individualized goal setting for the participating students (low delivery quality). When paired with outcomes data showing only modest improvements in retention, the fidelity findings clarify that although the program is implemented as planned, the limited depth of advising interactions likely diminishes its impact. In short, the implementation fidelity data help to understand why the outcomes occurred as they did (low quality of delivery), and guide both program refinement and research conclusions (Mokher et al., 2016; Smith et al., 2017). Building on these examples, Table 3 links the primary components of IF with common measurement approaches and key citations used in higher education and program evaluation research.

Discussion

Integrating IF into higher education assessment practices is a crucial yet underutilized strategy. This scoping review demonstrates that, while definitions and frameworks exist, their practical application across institutions varies. IF provides critical context for interpreting program outcomes, allowing assessment professionals to determine whether observed results reflect the intervention itself or variations in its implementation.

The literature shows that, while the terminology varies, definitions of IF consistently emphasize the alignment of planned and implemented interventions. Frameworks promote systematic assessment, but their integration with outcome

Table 3. Linking Implementation Fidelity Components to Measurement Approaches

IF Component	Measurement Method	Key Citations
Adherence	Observations, self-reports	Smith et al., 2017; Mokher et al., 2016
Exposure	Logs, attendance records, session tracking	Smith et al., 2017; Mokher et al., 2016
Delivery Quality	Observations, rating scales	Offerdahl et al., 2018; Rhodes et al., 2020
Participant Responsiveness	Surveys, engagement metrics, interviews	Smith et al., 2017; Reinke et al., 2012
Program Differentiation	Document review, control group comparisons	Finney et al., 2021; Gerstner & Finney, 2013

measurement is frequently incomplete. Measurement methods are evolving, moving away from traditional observations and self-reports and toward multi-step, iterative approaches that improve reliability and enable continuous refinement of both interventions and assessment tools.

Despite these advancements, challenges remain. The literature reviewed here is current only up to 2022. Although two independent reviewers conducted the screening of titles and abstracts to align with the inclusion criteria, the possibility of bias in interpretation during data extraction remains. To mitigate this risk, reviewers' disagreements were resolved through consensus and discussion, with a third reviewer available to make decisions as needed. Furthermore, while definitions, frameworks, and measurement approaches were systematically extracted, their interpretation is left to the reviewer's discretion. These limitations underscore the need for ongoing research to enhance IF frameworks, develop validated and reliable measurement instruments, and evaluate fidelity across various institutional and programmatic contexts.

In practice, incorporating IF into assessment processes can help improve evidence-based decision-making. Educators and program evaluators are encouraged to use clear definitions, systematically measure all core components of IF (such as adherence, exposure, delivery quality, participant responsiveness, and program differentiation), and employ multiple methods to capture fidelity accurately. Piloting and refining instruments prior to large-scale implementation ensures alignment with program objectives and maximizes utility. Ultimately, incorporating IF into routine assessments fosters continuous improvement, enhancing the interpretability and credibility of program evaluation results.

Conclusion

Implementation fidelity is critical for improving the effectiveness of educational interventions in higher education. By systematically defining and measuring implementation fidelity, practitioners and researchers can ensure that programs are delivered as intended and outcomes are correctly interpreted. Integrating implementation fidelity into assessment processes enhances evidence-based decision-making and provides actionable insights for program improvement. Future efforts should focus on the creation and validation of measurement tools, the improvement of assessment frameworks, and the routine incorporation of implementation fidelity data into evaluation

practices. When used consistently, implementation fidelity has the potential to enhance the quality, reliability, and impact of educational interventions, while also fostering a culture of continuous improvement across higher education programs.

Appendices

[Appendix A: Complete Search Strategies](#)

[Appendix B: PRISMA Flow Diagram](#)

[Appendix C: Data Extraction Survey Instrument](#)

References

- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19–32.
<https://doi.org/10.1080/1364557032000119616>
- Banta, T. W., & Palomba, C. A. (2015). *Assessment essentials: Planning, implementing, and improving assessment in higher education* (2nd ed.). Jossey-Bass/Wiley.
- Berman, P., & McLaughlin, M. W. (1976). Implementation of educational innovation. *The Educational Forum*, 40(3), 345–370. <https://doi.org/10.1080/00131727609336469>
- Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, 2, 40.
<https://doi.org/10.1186/1748-5908-2-40>
- Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, 31(2), 199–218.
<https://doi.org/10.1177/1098214010366173>
- Chia, M. Y. (2012). *Fidelity of implementing an assessment translation and adaptation framework in a study of an emerging international assessment* (Publication No. ED551885) [Doctoral dissertation, University of Colorado at Boulder]. ProQuest Dissertations Publishing.
- Cook, T. D., & Shadish, W. R. (1986). Program evaluation: The worldly science. *Annual Review of Psychology*, 37(1), 193–232. <https://doi.org/10.1146/annurev.ps.37.020186.001205>
- Council for the Advancement of Standards in Higher Education. (2019). *CAS releases 10th edition of professional standards book* [Press release]. <https://www.cas.edu/news--notes/cas-releases-10th-edition-of-professional-standards-book>
- Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, 18(1), 23–45. [https://doi.org/10.1016/S0272-7358\(97\)00043-3](https://doi.org/10.1016/S0272-7358(97)00043-3)
- Dhillon, S., Darrow, C., & Meyers, C. V. (2015). Introduction to implementation fidelity. In C. V. Meyers & W. C. Brandt (Eds.), *Implementation fidelity in education research* (pp. 8–22). Routledge.
- Dumas, J., Lynch, A., Laughlin, J., Smith, E., & Prinz, R. (2001). Promoting intervention fidelity: Conceptual issues, methods, and preliminary results from the early alliance prevention trial. *American Journal of Preventive Medicine*, 20(1 Suppl), 38–47.
[https://doi.org/10.1016/S0749-3797\(00\)00272-4](https://doi.org/10.1016/S0749-3797(00)00272-4)
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41(3–4), 327–350.
<https://doi.org/10.1007/s10464-008-9165-0>
- Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, 18(2), 237–256. <https://doi.org/10.1093/her/18.2.237>
- Farr, B., Bradby, D., Hartry, A., Sipes, L., Hall, L., & Tasoff, S. (2009). *Evaluation of the demonstration sites in the ConnectEd Network*. MPR Associates.
- Finney, S. J., & Smith, K. L. (2016, January). Ignorance is not bliss: Implementation fidelity and learning improvement. *Viewpoint - National Institute for Learning Outcomes Assessment (NILOA)*.

- Finney, S., Wells, J., & Henning, G. (2021). *The need for program theory and implementation fidelity in assessment practice and standards*. Occasional Paper No. 51. National Institute for Learning Outcomes Assessment. <https://eric.ed.gov/?id=ED612091>
- Fisher, R., Smith, K., Finney, S. J., & Pinder, K. (2014). The importance of implementation fidelity data for evaluating program effectiveness. *About Campus*, 19(5), 28–32. <https://doi.org/10.1002/abc.21171>
- Gerstner, J. J., & Finney, S. J. (2013). Measuring the implementation fidelity of student affairs programs: A critical component of the outcomes assessment cycle. *Research and Practice in Assessment*, 8, 15–28.
- Hamre, B., Henry, A., Locasale-Crouch, J., Downer, J., Pianta, R., Burchinal, P., Howes, C., LaParo, K., & Scott-Little, C. (2011). Implementation fidelity and teachers' engagement in a course on effective teacher-child interactions: Effects on teacher beliefs, knowledge, and observed practice. *Society for Research on Educational Effectiveness*. <https://eric.ed.gov/?id=ED518177>
- Hill, H. C., & Erickson, A. (2019). Using implementation fidelity to aid in interpreting program impacts: A brief review. *Educational Researcher*, 48(9), 590–598. <https://doi.org/10.3102/0013189X19891436>
- Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, 5(69). <https://doi.org/10.1186/1748-5908-5-69>
- Mellard, D. (2010). *Fidelity of implementation within a response to intervention framework* [White paper]. New Mexico Public Education Department. <https://webnew.ped.state.nm.us/wp-content/uploads/2018/03/Fidelity-of-Implementation-guidev5-1.pdf>
- Mihalic, S. (2004). The importance of implementation fidelity. *Emotional and Behavioral Disorders in Youth*, 4(4), 83–105.
- Mokher, C., Lee, S., & Sun, C. (2016). *Final findings from impact and implementation analyses of the Northeast Tennessee College and Career Ready Consortium* [White paper]. CNA Corporation.
- Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2010). *A procedure for assessing fidelity of implementation in experiments testing educational interventions* [White paper]. Society for Research on Educational Effectiveness.
- O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, 78(1), 33–84. <https://doi.org/10.3102/0034654307313793>
- Offerdahl, E. G., McConnell, M., & Boyer, J. (2018). Can I have your recipe? Using a fidelity of implementation (FOI) framework to identify the key ingredients of formative assessment for learning. *CBE—Life Sciences Education*, 17(4), 1–9. <https://doi.org/10.1187/cbe.18-02-0029>
- Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1), 210. <https://doi.org/10.1186/s13643-016-0384-4>
- Readdean, K. C. (2022). *Evaluating the implementation of primary care behavioral health systems in college health: Development of a theory-based understanding of the organization of integrated care* [Doctoral dissertation, Rutgers University]. ProQuest Dissertations and Theses Global. <https://rucore.libraries.rutgers.edu/rutgers-lib/67525/>
- Reinke, W. M., Stormont, M., Webster-Stratton, C., Newcomer, L. L., & Herman, K. C. (2012). The Incredible Years teacher classroom management program: Using coaching to support

- generalization to real-world classroom settings. *Psychology in the Schools*, 49(5), 416–428. <https://doi.org/10.1002/pits.21608>
- Rhodes, L. A., Marciniak, M. W., McLaughlin, J., & Pinelli, N. R. (2021). Leveraging an implementation fidelity framework to assess quality in experiential education. *American Journal of Pharmaceutical Education*, 85(4), 297–303. <https://doi.org/10.5688/ajpe8302>
- Smith, K. L., & Finney, S. J. (2020). Elevating program theory and implementation fidelity in higher education: Modeling the process via an ethical reasoning curriculum. *Research & Practice in Assessment*, 15(2), Article 2. <https://www.rpajournal.com/dev/wp-content/uploads/2020/09/Elevating-Program-Theory-and-Implementation-Fidelity-in-Higher-Education.pdf>
- Smith, K. L., Finney, S. J., & Fulcher, K. H. (2017). Actionable steps for engaging assessment practitioners and faculty in implementation fidelity research. *Research and Practice in Assessment*, 12, 71–86.
- Stains, M., & Vickrey, T. (2017). Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices. *CBE—Life Sciences Education*, 16(1), 1–11. <https://doi.org/10.1187/cbe.16-03-0113>
- Swain, M. S., Finney, S. J., & Gerstner, J. J. (2013). A practical approach to assessing implementation fidelity. *Assessment Update*, 25(1), 5–7, 13. <https://doi.org/10.1002/au.20028>
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473. <https://doi.org/10.7326/M18-0850>
- Viennet, R., & Pont, B. (2017). *Education policy implementation: A literature review and proposed framework* [White paper]. Organisation for Economic Co-operation and Development. <https://doi.org/10.1787/fc467a64-en>
- Zoblotsky, T., & Gallagher, B. (Principal Investigators). (2020). *Get the picture?! Final evaluation report* (U411C140061) [Grant]. University of Memphis, Center for Research in Educational Policy (CREP). <http://ies.ed.gov/ncee/wwc/Study/90063>

Measuring Meaning-Making Among College Students: The Living a Life of Meaning and Purpose—C Portfolio



Authors:

Larry H. Ludlow, Ph.D.
Boston College

Henry I. Braun, Ph.D.
Boston College

Theresa A. O'Keefe, Ph.D.
Boston College

Christina Matz, M.S.W., Ph.D.
Boston College

Lauren Warner, Ph.D. Candidate
Boston College

ABSTRACT

This article describes the development and initial validation of a scenario-based instrument to assess meaning-making in college students, grounded in constructive developmental theory (O'Keefe et al., 2025). Using Rasch/Guttman Scenario (RGS) scale methodology (Ludlow et al., 2020a), we constructed two parallel instruments employing school and family contexts. Each instrument comprised an ordered sequence of scenarios designed to increase in complexity across four facets of the construct: cognitive ability, relational awareness, conflict resolution, and sense of responsibility. Results from a diverse undergraduate student sample supported the hypothesized scenario hierarchy and proper functioning of the response categories. The instruments also provide substantive interpretations of respondent score locations that can inform longitudinal tracking of meaning-making growth in emerging adults. These findings offer preliminary support for using the RGS framework to model meaning-making as a structured, developmental construct, offering a theory-aligned alternative to qualitative interviews.

Correspondence E-mail: ludlow@bc.edu

Keywords: meaning-making, constructive developmental theory, Rasch/Guttman scenario scales

The college years represent a critical period of developmental transitions as students move from adolescence into early adulthood. While postsecondary education is often evaluated in relation to intellectual growth and career readiness, there is a growing recognition that institutions of higher learning also bear a responsibility to support the formation of students as whole persons capable of leading lives of meaning and purpose (Dalton & Crosby, 2010; Glanzer et al., 2017). For educators and administrators who consider formation a worthwhile educational outcome, the need for psychometrically sound tools to measure such development becomes both practical and urgent.

Yet the measurement of formative outcomes remains inherently challenging. Constructs like meaning-making are conceptually rich, developmentally dynamic, and not easily captured by traditional survey instruments. In response to this measurement problem, we developed the Living a Life of Meaning and Purpose–C (LAMP-C), a self-report instrument designed to assess meaning-making capacity among traditional-age college students. LAMP-C comprises two context-specific instruments—school and family—each composed of developmentally ordered scenarios.

The design of LAMP-C is grounded in two complementary frameworks. The first is constructive developmental theory (CDT), which provides the conceptual foundation for the meaning-making construct. The second is the Rasch/Guttman Scenario (RGS) scale methodology, which guides scenario development, response scaling, and item calibration (Ludlow et al., 2014, 2020b).

This paper outlines our theoretical and methodological rationale for the development of LAMP-C. We begin by summarizing the relevant scholarship on meaning-making and offer a reframing of CDT for the purpose of construct operationalization. We then describe the application of RGS methodology to instrument construction and conclude with a discussion of the psychometric characteristics of the school and family scales, next steps, and implications for future assessment research.

Constructive development theory (CDT)

The literature on meaning-making, cognitive development, self-authorship, and leadership within educational contexts, particularly among college students and emerging adults, presents a rich tapestry of interconnected ideas that deepen our understanding of individual development. Building upon the cognitive development work of Jean Piaget (1952) theorists such as Kegan (1982, 1994), Parks (2000), Baxter Magolda (2001, 2009), King (2009), McAdams (2013) and Helsing & Howell (2014) have explored how individuals construct and evolve their understanding of themselves and their place in the world. In the aggregate, this work is called constructive developmental theory (CDT).

Briefly, CDT posits that human development is a process of progressively more complex meaning-making with individuals potentially moving from childhood to mature adulthood toward increasingly complex capacities for understanding themselves and the world. Enhancement of meaning-making impacts and is impacted by cognitive capacity, identity, and relationships. Longitudinal research that follows traditional-age college students into mature adulthood has been crucial for understanding what contributes to young adults achieving self-authorship, a key developmental milestone. “Self-authorship” is understood as the capacity to internalize and ‘own’ one’s values, beliefs, and identities, while acknowledging connections with and responsibility to others (Kegan, 1994; Baxter Magolda, 2001).

Studies of cognitive and moral development within higher education settings suggest that the cognitive and moral reasoning capacities of students evolve in response to both educational and social experiences. In particular, narrative identity formation illustrates how individuals continually revise their life stories as they encounter new experiences and challenges, and highlights how critical incidents and supportive relationships can facilitate or hinder the journey toward self-authorship (King, 2009; McAdams, 2013; Pizzolato, 2003). The extension of CDT into the realm of leadership illustrates how the principles of CDT can inform leadership training and assessment (Helsing & Howell, 2014). Finally, aligning educational experiences with the developmental capacities of emerging adult learners exemplifies how theoretical insights can be translated into practical strategies for enhancing a broad range of educational outcomes (Steward & Wolodko, 2016). In sum, the literature suggests that meaning-making, self-authorship, and cognitive development are deeply interconnected processes that are central to personal and educational growth; CDT provides a robust framework for understanding these processes.

In nearly each case, the research cited above involved some form of qualitative interviews, such as the Subject-Object Interview developed by Kegan and colleagues (Leahy et al., 2011). Qualitative interviews, while potentially highly informative, are problematic for large-scale administration because they are time consuming, both in administration and the interpretation of data. Accordingly, they are usually employed in small samples. A few self-report questionnaires and scales have been created to assess self-authorship as it pertains to career readiness (Creamer et al., 2010; Fallar et al., 2019). However, these measures have limitations, particularly in their ability to capture the dynamic and developmental nature of meaning-making across contexts. Our goal was to create a practical, quantitative assessment of meaning-making, drawing on CDT to inform instrument development.

Facets of CDT

As explained in greater detail below, the Rasch/Guttman Scenario psychometric approach operationalizes complex constructs, like meaning-making, by breaking the construct down into its constituent characteristics, or facets. These facets are then combined in a specific manner to construct an ordered sequence of scenarios that describe responses to situations occurring in common contexts in which college students must exercise meaning-making. Within a context, the scenarios are incrementally modified to reflect gradations of the facets, representing different responses to the situations that reflect different capacities of meaning-making. We offer warrants for both the facets and their gradations by drawing from our analysis of the theories of Kegan, Parks, and Baxter Magolda. Below, each facet is described and associated with at least one theorist's framework.

With regard to facet gradations, all of the theorists cited above identified developmental steps in which individuals exhibit qualitatively different capacities for interpreting and responding to their worlds. Kegan (1994) identifies five possible "orders of consciousness" across the lifespan (p. 314-315). Parks (2000) identifies four "forms" spanning from adolescence to mature adulthood (p. 90). Baxter Magolda (2001) names four "phases," starting in college and extending into mature adulthood (p. 40). With each shift of order, form or phase, a person does not forget or reject what was previously known,

but subsumes it and organizes it in a broader framework of meaning. In our focus on traditional-age college students, we propose five gradations, which we call “positions” (O’Keefe et al., 2025).

Across the developmental span between late adolescence and early adulthood, traditional college students are actively shifting how they make sense of their lives and relationships. The theories of Kegan, Parks, and Baxter Magolda each describe this transition in slightly different terms, but the general pattern is consistent: students move from externally guided, socially defined ways of knowing—what Kegan calls the Third Order and Baxter Magolda describes as Following External Formulas—toward internally defined, self-authored frameworks such as Kegan’s Fourth Order or Baxter Magolda’s Self-Authorship.

The LAMP-C model uses five positions to capture this developmental progression. Positions 2 and 3 represent earlier, socially driven forms of meaning-making, while Positions 4 and 5 capture the emergence and consolidation of self-authorship. Position 1 anchors the continuum, reflecting more concrete, categorical ways of thinking often observed at entry to college. This structure allows us to see not just where students begin and where they might ultimately arrive, but also the shifts that occur in between.

There are two reasons why this level of detail matters. Theoretically, five positions reflect the nuanced changes described by each framework, aligning with the complexity of meaning-making growth during the college years. Practically, the five-position structure enables greater measurement precision. Our preliminary Rasch calibrations demonstrated that five distinct positions allow clearer discrimination among student responses and support the construction of vertically ordered scenarios that increase in developmental complexity. Table 1 shows how these five positions correspond to the developmental stages identified by Kegan, Parks, and Baxter Magolda.

We elaborate on the five positions by identifying four facets, or characteristics, that reflect the cognitive, interpersonal, and intrapersonal characteristics of meaning-making capacity (O’Keefe et al., 2025). They are: cognitive ability, relational awareness, conflict resolution, and sense of responsibility. The first facet of CDT is cognitive ability. Appendix A presents the range in cognitive ability across the five positions.

Table 1. Correspondence Between Five LAMP Positions and Three Constructive Developmental Frameworks

LAMP Position	Kegan (1994)	Parks (2000)	Baxter Magolda (2001)
1	2nd Order	—	—
2	Emerging 3rd Order	Adolescent/Conventional	Following external formulas
3	3rd Order	Adolescent/Conventional	Following external formulas
4	Emerging 4th Order	Emerging adult	Crossroads
5	4th Order	Tested adult	Self-authorship

Note. The five LAMP meaning-making positions are aligned with comparable stages from three developmental theories: Robert Kegan’s orders of consciousness, Sharon Daloz Parks’ model of adult meaning-making, and Marcia Baxter Magolda’s theory of epistemological development. Dashes (—) indicate no directly corresponding classification in the respective framework.

With development comes the ability to better recognize and make sense of the world as students move from concrete-operational, to early-, then later-formal-operational cognitive capacity. For example, whereas those in position 1 are not guided by values as such, those in positions 2 and 3 are able to recognize and talk about the values they hold and try to live by, but may be challenged to prioritize among competing values. Finally, those in positions 4 and 5 are not only comfortable in discussing values, but also are able to prioritize among competing values. Similarly, increased cognitive complexity is also reflected in their ability to recognize the relationship between cause and effect, and among behaviors, ideas and worldviews.

The second facet is relational awareness (Appendix B). With appropriate challenges and support, it becomes increasingly apparent to the developing person that they have been and are in relationships with others. It is not just that they are more able to engage in relationships with mutuality but that they are beginning to recognize the relationships of which they are already a part. The person in Position 1 is unselfconsciously self-interested. They may recognize that others have a point of view, just as they do, but they do not recognize they have relationships, with implicit expectations, with others. The move towards Position 2 brings a capacity to recognize relationships, and to begin to take others into consideration. In time and with encouragement (Position 3), tacit belonging is replaced by conscious alignment, whereby one takes on and even can become a proponent for the values and expectations of particular relationships. Position 4 marks a shift in one's sense of dependence on, and obligation to, one's relationships, with the awakening recognition that one's perspective is distinct from that of important others, and one may risk sharing that perspective. By Position 5 one is more confident in claiming their own voice and, in general, is not distressed by diverse perspectives within their relationships.

Conflictual situations are spaces in which intrapersonal meaning-making is made apparent, for how one chooses to resolve conflict reflects a sense of self in relation to others. In conflict resolution (facet 3, Appendix C), the person in Position 1 is able to think logically about their perceptions, but unable to simultaneously consider the perceptions of others; they tend to focus on their own interests and preferences. Therefore, in a conflict, they usually resolve in favor of their own needs and desires, but not for any principled reason. Those in Position 2 have become tacitly aware of the relationships and expectations around them, and so tend to conform to what "everyone" expects. In time, it may develop into a greater sense of identity, grounded in particular relationships and social alignments (Position 3), and so begin to resolve conflicts based on explicit values or relational expectations. As one develops the capacity to distinguish between oneself and one's relationships (Position 4), one becomes more likely to risk resolving a conflict in a manner different from what "everyone expects," even while mindful of those expectations. Finally, the person in Position 5 is more integrated, bravely resolving conflicts in a manner that reflects their principles and values, even as the relationship or context varies.

Facet 4, sense of responsibility, reflects both the interpersonal and the intrapersonal dimension where a person's sense of responsibility can be a strong indicator of meaning-making capacity (Appendix D). In Position 1, since relationships and obligations to others are not yet acknowledged, there is no felt sense of responsibility to others. On the other hand, the person in Position 2 has become tacitly aware of their relationships and social connections, resulting in a move to feel responsible to those relationships. As one moves to Position 3, there is a stronger alignment with particular

relationships, and the values they represent, thus enabling an ability to be more mindful of, and responsible to, those relationships and values. As the person comes to see themselves as having relationships, but not being defined by them, they can begin (Position 4) to determine how they will be in them; that is, they increasingly determine their own rules of engagement for their relationships. As they become more consistently able to name who they are in their relationships (Position 5), they can better see how their attitudes and actions shape the relationships for themselves and others. As such, they become more responsible for those relationships; they recognize they are part of determining the rules of engagement for everyone.

Across these four facets (cognitive ability, relational awareness, conflict resolution, and sense of responsibility), students move from concrete, self-focused interpretations of experience toward increasingly complex capacities to integrate diverse perspectives and act with principled autonomy. Cognitive ability reflects how individuals understand ideas and systems, progressing from concrete, then categorical thinking to the ability to manage competing priorities. Relational awareness captures a shift from unexamined self-interest to mutual, reciprocal relationships. Conflict resolution traces how individuals engage differences, moving from avoidance or compliance to actively balancing self-expression and relational integrity. Finally, sense of responsibility highlights how individuals come to first recognize then potentially shape the “rules of engagement” in their communities. These gradations informed scenario writing by providing a developmental blueprint: each scenario systematically combines facet levels to portray a single, coherent position along the meaning-making continuum.

Scenario construction

For each context, the development of the ordered scenario sequence representing the five levels of meaning-making capacity adhered to the Rasch/Guttman Scenario (RGS) scale construction framework as detailed in Ludlow et al. (2014, 2020a, 2021). This framework consists of the following iterative and theory-driven stages: (a) specifying the target construct; (b) identifying relevant facets and drafting initial facet-level descriptions; (c) articulating developmental gradations within each facet; (d) determining the structural template for scenario construction; (e) formulating and integrating mapping sentences to build theory-anchored scenarios; (f) selecting response formats and drafting respondent instructions; and (g) empirically evaluating the degree of alignment between theoretical expectations and model-based estimates. In the present application, meaning-making capacity serves as the target construct, operationalized through four integrated facets: cognitive complexity, relational awareness, conflict resolution, and sense of responsibility. Each scenario within a given context reflects a fixed position on the developmental continuum beginning with Position 1 and ascending through Position 5. The resulting scenario sets are thus designed to function as a vertically ordered scale with increasing complexity across the meaning-making continuum, as defined by constructive developmental theory and later supported through Rasch model calibration.

The scenarios were deliberately constructed to be brief yet semantically rich portrayals of context-specific situations that were intended to be both ecologically valid and experientially recognizable to college students. Each scenario is aligned with one of five hypothesized meaning-making positions and is written in language representative of that developmental level. This design feature serves a dual function: first, to anchor the

narrative in familiar domains of student experience; and second, to differentiate the scenarios linguistically and conceptually across levels of developmental complexity. At the lower end of the scale (Position 1), the character's language is rooted in concrete, immediate preferences and personal advantage. By contrast, by Position 5 the language reflects the capacity to articulate abstract principles, consider competing values, and make intentional commitments. This progression in linguistic and conceptual sophistication across the scenario sequence is essential to supporting the scale's developmental structure and interpretive utility.

The current implementation includes two parallel Rasch/Guttman Scenario (RGS) instruments—one for the school context and one for the family context—each designed to assess meaning-making capacity within a contextually grounded, potentially conflictual situation. These two contexts were purposefully selected due to their salience in the developmental lives of college students.

The instructions were the same for both instruments.

Below you will find a description of two contexts (school and family) and a situation within each of those contexts. Following the contextual description, you will find 5 scenarios in which a character is making sense of the situation in that context. Within each context, a single character is named for all scenarios. Obviously, one character cannot see a situation from all perspectives simultaneously. The use of the same character is to emphasize that all elements of the context and situation remain the same for all scenarios; what changes is how the character makes sense of the situation. Your task is to read each scenario and determine how similarly or differently from the character you would approach the situation if you were in it.

For both instruments we utilized a “comparative response option” approach (Ludlow et al., 2014). Here the respondent is asked to compare themselves to the character in each scenario. More specifically, the instructions direct the respondent “to imagine yourself in each situation and consider the degree of similarity between your approach to that of the character's approach.” They are then offered five progressively higher-level meaning-making capacity options:

1. In the past I might have approached this kind of situation similarly to X, but I don't now.
2. Sometimes I approach this kind of situation similarly to X, but not regularly anymore.
3. Now, I regularly approach this kind of situation similarly to X.
4. Sometimes I approach this kind of situation similarly to X, but not as regularly as I would like to.
5. Someday I might approach this kind of situation similarly to X, but I don't now.

These response options reflect plausible changes over time, and require the respondent to situate themselves vis-a-vis each scenario in a manner that reflects their current meaning-making capacity. Tables 2 and 3 present the situation and meaning-making scenarios for the school and family contexts, respectively.

Table 2. *Context and Scenarios for School*

Scenario	Description
Context	John is a student in a college and a program that he likes. He is enrolled in a course required for his program of study and he gets on fine with the professor. There has never been any indication from this professor that there are consequences for holding diverse perspectives from the professor. In a question on an exam, the professor instructs students to offer their perspective on an issue, applying the course material to support their answer. Each scenario below reflects a different way John might approach the situation.
John1	“I don’t like questions like this. It feels like a trap. Do they want my opinion or the course material? Which is it? If they know what answer they want, they should just tell us what it is and not try to mess with us.”
John2	“I really don’t see how my opinion matters on this. There’s got to be a right way to answer this question, whether it’s naming the right steps or the right theorists. My job is to figure out what the right answer is.”
John3	“There are multiple ways to answer this question, and I really think differently about this than the professor does. But there’s no way in the world I’m going to offer anything different from what I think the professor wants.”
John4	“There are multiple ways to answer this question, and some are better than others. On this question, I really think differently than the professor does. It feels a little risky, but I think if I make a strong, well-supported argument using the course material, I should be alright.”
John5	“There are multiple ways to answer this question, and some are better than others. If I have the space and time, I think I will lay out a few options and argue the merits and limits of each. But in the end, it will be really important that I make a clear argument for which one I think is most effective. I like questions like this; they offer a chance to really engage with the material.”

Note to the reader: This table presents a developmental sequence of five possible responses by John, illustrating increasingly complex ways of engaging with the exam question and course material. Scenario 1 reflects the least complex meaning-making perspective, whereas Scenario 5 reflects the most complex meaning-making perspective.

Table 3. *Context and Scenarios for Family*

Scenario	Description
Context	Maria comes from an intact family made up of immediate and extended members who have occasion to interact with some regularity and are considered a relatively close family. By and large, everyone gets on well. There is no history of abuse or trauma. When it comes to questions of conflict or differences of opinion among family members or between Maria and family members, these are possible ways Maria might make sense of the situation.
Maria1	“When it comes to disagreements, it’s just a matter of what you like and don’t like. If I am doing something that my family disagrees with, I just try not to get caught. But sometimes when I get caught, they say I ‘should have known better’. I don’t get it. What should I have known?”
Maria2	“In my family, we really know what family stands for, which is important to me. For the most part, I’m happy to contribute to making things go smoothly. But on the rare cases we disagree, I just keep that to myself. I don’t want to make a scene or get anyone upset. Voicing a different opinion from my family is not that important to me.”
Maria3	“Sometimes I just want to shout from the rooftop that I don’t agree with what they are all thinking, but I never do, because it would be too upsetting. Other times I try and it’s a major fail; it comes out all wrong. I think I’m right and don’t understand why they don’t agree with me. They are my family; shouldn’t they get me? Most of the time, I just keep things to myself.”
Maria4	“It may seem easier to keep my mouth shut, but sometimes the issues are too important to let go. As risky as it feels, I’m getting better at expressing myself to my family on things that are important to me. It’s important to me that we get along as a family, which means I have to contribute to how we listen and speak to each other.”
Maria5	“Even though we’re family and we love one another, we have very different perspectives and personalities. Talking to each other about difficult issues takes real care. Not only do I have to recognize their point of view, I have to express mine in a way they can understand.”

Note. This table presents a developmental sequence of five possible responses by Maria, illustrating increasingly complex ways of navigating conflict and differences of opinion within her family. Scenario 1 reflects the least complex meaning-making perspective, whereas Scenario 5 reflects the most complex meaning-making perspective.

Preliminary field trials of the initial LAMP-C instrument yielded encouraging evidence of construct alignment (O’Keefe et al, 2022); however, the hypothesized “ladder-like” hierarchy of scenario difficulties did not materialize. In addition, we observed a few student responses that departed from model expectations—most often on the lowest complexity scenarios. Some students at higher locations on the scale responded in ways that indicated disagreement or disengagement with scenarios that had been written to be low-complexity. In one case, a student provided a pattern of responses that represented severe model misfit and raised questions about either respondent interpretation or scenario design. Rather than simply dismiss these as outliers, we treated them as important indicators of how respondents were making sense of the task. These response patterns invited closer scrutiny of scenario clarity, response instructions, and response options.

We collaborated with a *Student Voices Advisory Group*—a demographically and experientially diverse group of undergraduate students—who reviewed not just the scenarios themselves, but also the scoring prompts and overall framing of the instrument. This collaborative effort led to several revisions intended to enhance clarity and developmental coherence: (a) the CDT scenario framework was recast from seven to five positions to more clearly reflect psychometrically distinguishable stages; (b) the conflictual situation within each context was standardized rather than varied; (c) a single recurring character (John or Maria) was introduced to anchor each scenario set; and (d) the response options were restructured to align more closely with a developmental self-assessment format. These changes were not trivial; they represented an attempt to make the task more ecologically valid and more interpretable for students, while preserving the structure needed for Rasch calibration. Following additional cognitive interviews with SVAC members, the revised instrument was administered in full-scale form.

Sample

Data were collected through an anonymous online survey, administered via Qualtrics, and disseminated to students at a private university in the northeastern United States. They were recruited through multiple channels, including email invitations sent via university listservs, direct outreach from faculty and staff, and physical flyers distributed in high-traffic student areas such as dining halls, residence halls, and the athletic center. The research team collaborated with university departments and student organizations to maximize outreach. Recruitment materials emphasized the voluntary nature of participation and highlighted the opportunity to enter a raffle for a \$50 Amazon gift card, providing a 1 in 10 chance of winning. Upon completion, they were directed to a separate webpage where they could enter the raffle without their contact information being linked to their responses, ensuring anonymity.

The study was approved by the Boston College Institutional Review Board (Protocol 23.074.01e) (Ludlow, 2023). Inclusion criteria required participants to be at least 18 years old and currently enrolled. There were no exclusions based on gender, race, or ethnicity. Recruitment efforts targeted a diverse student population to ensure broad representation and to assess potential demographic differences in responses. These efforts yielded 114 undergraduate and graduate students—Appendix A presents the demographic breakdown.

While these procedures yielded a somewhat diverse group of respondents, the reliance on a single-institution sample represents an important limitation. This private university has distinctive characteristics in terms of institutional culture, demographics, and developmental supports, which may influence meaning-making processes. Consequently, our findings should be interpreted with caution and not assumed to generalize to students in other institutional settings, such as public universities, community colleges, or institutions with different missions or student populations.

Analysis

The Rasch rating scale model (Andrich, 1978; Wright & Masters, 1982) was used to analyze the extent to which data conformed to model-based expectations; that is, if the empirical results provide construct validity evidence congruent with the theorized scale structures. For a particular context, the model may be expressed as:

$$\pi_{nix} = \frac{e^{\sum_{j=0}^{x_{ni}} [\beta_n - (\delta_i + \tau_j)]}}{\sum_{k=0}^m e^{\sum_{j=0}^k [\beta_n - (\delta_i + \tau_j)]}}$$

For the present analysis, the key parameters of the model are: π_{nix} denotes the probability that person n , when presented with scenario i , responds in category x ; x_{ni} is the observed score of person n on item i ; m is the maximum possible rating scale category and $k=0$ identifies the lowest score category in the response set (this ensures that the denominator sums the category characteristic curves from 0 through m so that the probability for category x is properly normalized); β_n represents person n 's estimated standing, or location, on the latent meaning-making continuum within a given context; δ_i is the scenario difficulty parameter, indicating the location of scenario i along the same latent continuum; and τ_j is the threshold parameter associated with the transition from response category j to the next higher category within a scenario. In this framework, scenario difficulty reflects the extent to which a scenario tends to elicit higher-level responses, with higher level response categories corresponding to greater meaning-making capacity. The analysis was conducted using WINSTEPS (Linacre, 2024).

Results

For each context, the systematic Rasch/Guttman scenario construction process generates an *a priori* expected ordering of the scenarios. Construct (or variable) maps, are an essential component of Rasch instrument development and analysis procedures. They are graphical representations that simultaneously display the scenario difficulty estimates and the respondent level of meaning-making estimates, both as locations along a logit continuum (Ludlow & Haley, 1995). If our reformulation of constructive developmental theory is correct, and the scenarios have been well designed, and the respondents are appropriate for the task, then the empirical ordering of scenarios within a context should

agree with the theoretical expectations for that set. Furthermore, the ordered sequence should define a quantitative continuum along which a student may then be located. For each context, students' locations on the corresponding scale thus provide a basis for a specific, actionable interpretation of their meaning-making capacities.

In addition to the construct maps, goodness-of-fit statistics evaluate how well the observed responses conform to the responses predicted under the Rasch model. Residuals are the differences between the observed and predicted responses. Variance weighted mean squared residuals ("Infit MNSQ") and unweighted mean squared residuals ("Outfit MNSQ") above 1.4 indicate potential model misfit (Linacre, 2019).

Figure 1 and Figure 2 present the construct maps for the school and family contexts, respectively. To the right of the vertical lines are the names of the scenarios in descending order of their difficulty estimates from hardest/most complex (John5 and Maria5) to say "Someday I might approach this kind of situation similarly to X, but I don't now" (top of the continuum) to easiest/least complex (John1 and Maria1) to say "In the past I might have approached this kind of situation similarly to X, but I don't now" (bottom of the continuum). The full text of each scenario is presented roughly adjacent to each label in order to qualitatively define each meaning-making position.

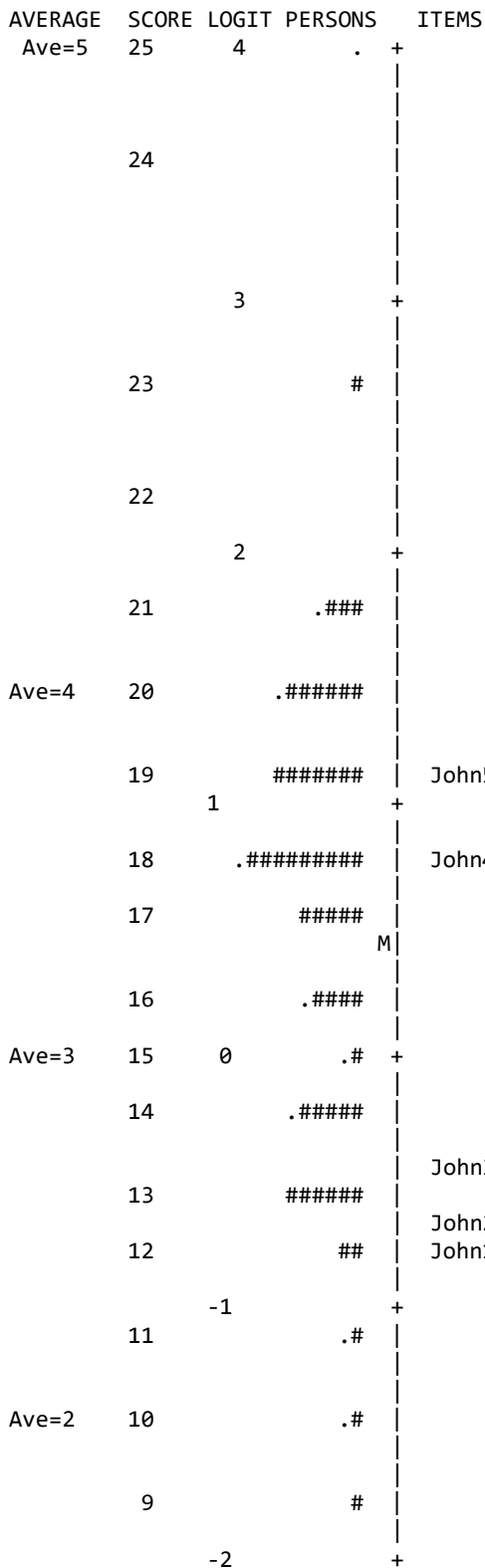
On the left sides of the vertical lines are the students meaning-making location estimates based on their raw score responses to the scenarios. The "#" symbols represent either 1 or 2 students at a given location. The "M"s represent the mean student location estimates. The "AVERAGE" column indicates where the raw score average responses of "2", "3" and "4" are located. The "SCORE" column corresponds to where the students are located in terms of their actual raw score. Furthermore, the Score column allows a reader to answer the scenario prompts, add up their responses, and locate where their responses put them on the respective continua. The "MEASURE" column contains the corresponding student and item locations in logits. Overall, both the scenario estimates and students' estimates display an appropriately broad range from lower to higher levels of complexity in each context.

Below Figure 1 and Figure 2 we provide a narrative to help the reader understand the connection between what a student's score means in terms of the scenarios adjacent to their location. That is, we answer the question: how is a given score to be interpreted?

Based on the systematic development of the scenarios: (a) John5 and Maria5 were written to be the most challenging for a student to say "Someday I might approach this kind of situation similarly to X, but I don't now" and their locations at the top of each construct map confirm that the students did, in fact, find them to be the most challenging. Likewise, John4/Maria4 and John3/Maria3 and John2/Maria2 were written to be mid-level in their degree of complexity, and John1/Maria1 were constructed to be the least challenging to say "In the past I might have approached this kind of situation similarly to X, but I don't now". The locations of these scenarios operationalizing each context, displayed in Figures 1 and 2, are consistent with their intended ordered positions, as presented in Tables 2 and 3.

The text boxes in Figures 1 and 2 provide substantively grounded and nuanced descriptions for interpreting individuals' scores. These interpretations derive directly from the construction of the scenarios, based on the integrated facet levels. For example, in Figure 1, students with meaning-making scores of 10 or below don't like questions that contain multiple perspectives, they tend to think concretely, may be unable to take the

Figure 1. Construct map for the School context.



ITEM SCENARIOS FOR SCHOOL CONTEXT

(John5) There are multiple ways to answer this question, and some are better than others. If I have the space and time, I think I will lay out a few options, and argue the merits and limits of each. But in the end, it will be really important that I make a clear argument for which one I think is most effective. I like questions like this; they offer a chance to really engage with the material.

(John4) There are multiple ways to answer this question, and some are better than others. On this question, I really think differently than the professor does. It feels a little risky, but I think if I make a strong, well-supported argument using the course material, I should be alright.

(John3) There are multiple ways to answer this question, and I really think differently about this than the professor does. But there's no way in the world I'm going to offer anything different from what I think the professor wants.

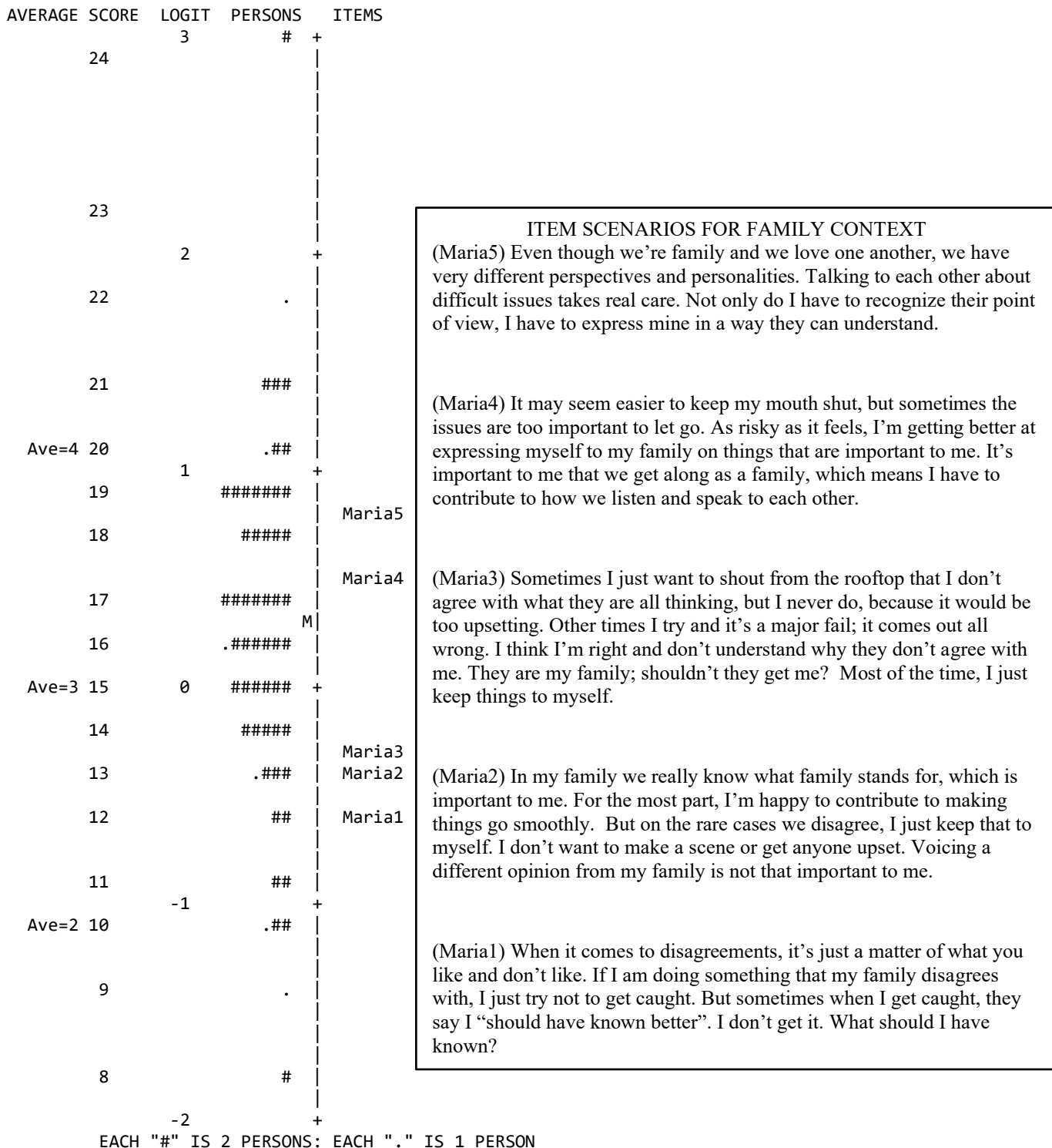
(John2) I really don't see how my opinion matters on this. There's got to be a right way to answer this question, whether it's naming the right steps or the right theorists. My job is to figure out what the right answer is.

(John1) I don't like questions like this. It feels like a trap. Do they want my opinion or the course material? Which is it? If they know what answer they want, they should just tell us what it is and not try to mess with us.

EACH "#" IS 2 PERSONS: EACH "." IS 1 PERSON

Legend: AVERAGE=the total scale score/5; SCORE=the sum of the 5 item responses; LOGIT=the logistic unit of measurement into which the raw scores are transformed (Ludlow & Haley, 1995).

Figure 2. Construct map for the Family context



Legend: AVERAGE=the total scale score/5; SCORE=the sum of the 5 item responses; LOGIT=the logistic unit of measurement into which the raw scores are transformed (Ludlow & Haley, 1995).

professor's point of view, focus on their own self-interest, and take no responsibility beyond their self-interest. At scores around 12, they still don't like this kind of question but will try to figure it out because they are beginning to recognize what the professor values, they are aware of the expectations for being an engaged member of the class, their self-interest begins to take into account the expectations of the class, and they begin to feel a need to respond appropriately within the class. At scores around 15 they don't necessarily dislike this kind of question as they begin to understand the professor's point of view and purpose and they do have a sense of belonging, but want to avoid conflict as they engage more fully with the task. At scores around 18, they begin to take the work more seriously and they are aware that they may think differently from the professor, but they are willing to take a risk and some responsibility to express their own position. Finally, at scores 20 and above, they recognize what the professor is asking of them, they are confident in their ability to express their point of view, and they are very aware of their relationship to the professor and how they can contribute to that relationship.

In Figure 2, a student with meaning-making scores of 10 or below tries to stay out of family conflicts, they don't understand why the family disagrees with them, they are unable to take the family's point of view, they focus on their own self-interest, and they take no responsibility beyond their self-interest. At scores around 12, they still try to avoid family disagreements even though they are aware the family disagrees with them and their self-interest keeps them quiet although they will engage when things are going smoothly. At scores around 15, they want to express their opinions but don't yet and they are conscious of the different positions of other members but are still not confident in their ability to express themselves appropriately. At scores around 18 and above, they begin to recognize the priorities and points of view of other family members and they are beginning to take some risks and responsibility for expressing and contributing to the family. Finally, at scores 20 and above, they accept that everyone has different experiences and perspectives and points of view about almost everything in family life, they have their own positions that they are confident in expressing, and they take a responsibility in contributing to the well-being of the family.

Appendix B and Appendix C contain the scenario characteristic statistics for the school and family contexts, respectively. The scenarios are displayed in the order in which they were presented. Both sets of scenario difficulty estimates (in their logit values and corresponding "Total scores") are monotonically increasing, as would be expected under the theoretical model. The "Model SE" is the standard error of the scenario estimates. Both sets of Infit and Outfit mean-squares suggest there were no strong patterns of unexpected responses. The item-total correlations for both sets of scenarios are all moderately positive (ranging from .45 to .66), as intended.

Appendix D and Appendix E present the category characteristic statistics for the school and family contexts, respectively. "Category Label" refers to the code for the response options. "Count" refers to how many times that response was provided. Both the "Observed Average" (in logits) for those who responded in each of the categories and the Andrich thresholds (the *transition* between adjacent ordered categories where a respondent becomes equally likely to choose category k instead of category $k-1$ (Andrich, 1978) monotonically increase, as intended.

The Infit and Outfit mean squares for the category responses show a slight degree of misfit for category one in both the school and family contexts but are otherwise

unremarkable. An inspection of the observed and expected responses for the entire sample revealed one student with an observed pattern of “1 5 4 4 4” for the five school scenarios. Their expected response to scenario one (John1) was 4.2 and their z-residual was -4. Two other students had -2 z-residuals on scenario one. For the family scenarios, there were four students with surprising responses to scenario one (Maria1) with z-residuals of -3, -2, -2, and -2.

A type of sensitivity analysis was performed, where the three students with the unusual John1 school responses and the four students with the unusual Maria1 family responses were removed from their respective data sets. The analyses were re-run and the item separation and reliability for the school scale increased from 6.63 and .98 to 7.02 and .98, respectively. The items maintained their same monotonic order, levels of acceptable fit statistics, and the category 1 fits decreased from 1.34 and 1.33 to 1.16 and 1.14 for Infit and Outfit, respectively.

The item separation and reliability for the family scale improved from 4.92 and .96 to 5.34 and .97, respectively. The items maintained their same monotonic order, levels of acceptable fit statistics, and the category 1 fits dropped from 1.52 and 1.50 to 1.4 and 1.38 for Infit and Outfit, respectively. In contrast, the item difficulty standard errors increased in each scale due to the decrease in the sample sizes. Although it is unclear why these different students responded relatively unexpectedly to the first scenarios in the two different contexts, their surprising responses may reflect a “start-up effect” due to their unfamiliarity with this type of unique item.

Finally, Rasch model item separation and item reliability indices (Linacre, 2002; Wright & Masters, 1982; Wright & Stone, 1979) help evaluate the quality of the item set in distinguishing among different locations on the latent trait—in this case, meaning-making capacity across school and family contexts. Item separation indicates how well the items in the instrument are spread out along the measured construct (school = 6.63; family = 4.92). Item reliability reflects the replicability of the item difficulty hierarchy if these same items were administered to another sample of similar respondents (school = .98; family = .96).

High separation indices (e.g., 6.63 or 4.92) suggest there are marked difficulty location distinctions among the items across the latent continua. Reliability values >.90 suggest that the item hierarchies are stable and would likely replicate in other samples. Collectively, these indices provide supportive preliminary evidence that the LAMP-C instrument is both psychometrically robust and theoretically well-aligned for measuring developmental progression across contexts.

Discussion

The purpose of this study was to assess the feasibility and psychometric defensibility of assessing meaning-making using the Rasch/Guttman Scenario (RGS) scale methodology, grounded in a reformulation of constructive developmental theory (CDT). Measuring a developmental construct like meaning-making is difficult—it requires clarity not only about the nature of the construct and how it unfolds over time, but also how it can be represented in ways that respondents in the target population can meaningfully engage with. The RGS scale methodology provided a promising strategy for addressing these formidable challenges.

From the outset, we adopted a facet-based approach to construct modeling, drawing on CDT to identify four core facets of meaning-making: cognitive ability, relational awareness, conflict resolution, and sense of responsibility. These facets were theoretically defensible and we anticipated that they could be operationalized in a reasonable way, though we recognized this would need to be tested empirically. Our aim was to integrate these facets into a series of scenarios—set in the contexts of school and family—that would vary systematically in complexity in order to provide a developmental ladder of meaning-making within each context. We used a sentence-mapping strategy to scaffold the complexity of each scenario corresponding to increasing orders of consciousness as described in Kegan’s model.

As reflected in the construct maps and scenario calibration statistics, the instruments yielded measurement structures that were consistent with our developmental hypotheses. In both the school and family contexts, the scenario difficulty estimates increased monotonically (as displayed in Figures 1 and 2, and reported in Appendices B and C), and the student meaning-making estimates spanned a wide range of the latent trait continuum (also seen in Figures 1 and 2). Rasch fit statistics, separation and reliability indices, and item-total correlations supported the internal structure of each scale and the ordered response thresholds confirmed that the rating categories functioned as intended (as reported in Appendices D and E). These are promising results, particularly given the complexity of the construct and the nascent stage of instrument development.

The iterative process of defining the construct, building scenarios, testing, identifying misfit, revising, and retesting—what Braun and Kirsch (2025) refer to as establishing procedural validity—is typical of how Rasch models are used to support the measurement of developmental constructs. As Ludlow et al. (2020a) have emphasized, model misfit is not simply a problem to fix; it is valuable diagnostic information that invites theory refinement and deeper construct understanding. In this case, the feedback we received from students in early phases of our pilot testing confirmed that some of our early assumptions about developmental clarity, linguistic accessibility, and contextual nuance needed to be reconsidered.

Because the Rasch/Guttman Scenario (RGS) scale methodology produces calibrated scores linked to specific, interpretable scenario responses, the results are theoretically very actionable and could be used to inform advising, program design, and longitudinal evaluation. For example, the construct maps in Figures 1 and 2 show how student locations along the continua correspond to qualitatively different ways of making sense of conflict. A student at a score near 10 on either the school or family continuum views conflict through a lens of self-protection and rule-following, whereas a student near 20 appreciates competing perspectives and acts with confidence and responsibility.

These developmental interpretations can guide advising conversations. Advisors can use score locations to match their approach to a student’s demonstrated capacity, offering more explicit structure for those at earlier levels and encouraging autonomy and leadership for those at later levels. Similarly, program designers can use aggregated results to align courses and co-curricular experiences with institutional goals. Curricula may be intentionally sequenced to provide opportunities for growth across the four facets of meaning-making: cognitive ability, relational awareness, conflict resolution, and sense of responsibility. Further, institutions could track student growth over time to evaluate the effectiveness of interventions. In this way, LAMP-C creates a direct link between theory

and practice: a developmental framework grounded in constructive developmental theory that also informs the day-to-day work of higher education.

Limitations and Future Research Directions

Although this study provides credible, preliminary evidence of construct alignment and developmental coherence for the LAMP-C instruments, several limitations temper interpretation and indicate directions for future research. The single-site, convenience sample severely constrains external validity. Institutional culture, mission, and student supports likely shape how respondents interpret the scenarios and construe developmental challenges; replication across diverse institutional types—including public universities, community colleges, and Minority-Serving Institutions (MSIs), as well as residential versus commuter settings—is needed before generalizing item hierarchies or score distributions. For instance, community colleges tend to serve many nontraditional students who juggle academic demands alongside work and family responsibilities—circumstances that may shape how they approach relational and conflict-resolution challenges. Similarly, students at MSIs may engage with questions of identity and responsibility in culturally specific ways. These differences make it essential to ensure that both the scenarios and the response formats in LAMP-C are relevant and meaningful for a variety of student populations. Expansion to additional contexts will allow testing of measurement invariance and differential item functioning (DIF) across subgroups defined by institution type, class year, first-generation status, race/ethnicity, and language background.

Response-process and linguistic factors also constitute important validity considerations. Scenario comprehension depends on the clarity and cultural resonance of the language and idioms used. Differences in first-language background, disciplinary jargon exposure, or reading fluency could influence ratings independently of meaning-making capacity. Although cognitive interviews with a *Student Voices Advisory Group* informed wording revisions, this group was small and also from a single institution. More systematic response-process evidence across diverse types of students, including linguistic subgroups—can help to identify sources of construct-irrelevant variance. Future testing will, therefore, combine multi-site replication with qualitative response-process studies to strengthen both external and internal validity.

Because the Rasch/Guttman Scenario scale methodology uses comparative, self-referential response options, ratings may also be susceptible to self-presentation or social-desirability effects. Students might over- or under-identify with a character's stance based on perceived desirability rather than genuine reasoning. Future iterations of LAMP-C should include validity checks such as social-desirability short forms and multitrait-multimethod comparisons to signal potential bias. At the same time, examining predictive validity—for instance, how meaning-making scores relate to advising engagement, leadership participation, or retention—will provide evidence of applied utility and real-world relevance.

A further methodological limitation involves the “start-up effects” observed when respondents first encounter a novel assessment format. Unexpected response patterns on initial items (e.g., 1–5–4–4–4) likely reflect poor format acclimation rather than construct differences, inflating residual misfit and attenuating stability. Prior Rasch research (Ludlow, 1983, 1986; Linacre, 2002; van der Linden, 2009) demonstrates that introducing

warm-up items can mitigate these distortions. Future LAMP-C implementations will include randomized item order and brief warm-up scenarios to reduce this source of measurement error.

Another consideration is context dependency, which is inherent to scenario-based measurement. School and family situations were chosen for relevance to college students, yet the degree to which each context mirrors a respondent's lived experience may vary. This variation can influence scenario response choices and consequently, observed developmental order. Future work will consider incorporating site-specific adaptations through focus groups with students, faculty, and staff to ensure that scenarios authentically reflect the developmental challenges salient within each institutional and cultural context. For example, community-college scenarios may highlight work–school balance, while those for MSIs could emphasize identity negotiation or systemic inequities.

Building on these refinements, the next phase of research will pursue a longitudinal validation agenda to test the sensitivity of LAMP-C to developmental change. Administering pre–post and multi-wave assessments within academic courses or co-curricular programs will help determine whether scores capture expected growth patterns, such as movement from compliance or avoidance toward principled self-authorship, and whether interventions targeted to specific facets (e.g., relational awareness) yield differential gains.

In collaboration with partners such as student affairs and career counselors, we also plan to design and evaluate score-reporting tools that translate person measures into developmental bands, accompanied by plain-language narratives and facet-specific growth suggestions. These reports will be tested for clarity, utility, and actionability through advisor and student feedback, ensuring that LAMP-C not only functions as a valid measure but also as a practical developmental feedback tool. Finally, future development will expand the current two-context framework by introducing a third relational context—friend relationships—within the LAMP-C instrument. This expansion will allow cross-context modeling of meaning-making development and provide a more complete picture of how students navigate meaning across personal, academic, and social domains. Collectively, these steps aim to enhance both the psychometric robustness and the practical utility of the LAMP-C framework, with the hope of advancing our understanding of meaning-making as an important component of student development across diverse higher-education settings.

Conclusion

This study focused on students from a single private university in the Northeast, offering an important starting point for our work. While the study findings provide a solid foundation, our broader goal is to develop a set of psychometrically integrated instruments that can capture how meaning-making evolves both within and across different contexts over time. By doing so, we aim to advance both theory and measurement, while gaining deeper insight into how meaning-making supports students' personal and educational growth throughout college and beyond.

The four core facets of meaning-making—cognitive ability, relational awareness, conflict resolution, and sense of responsibility—are not unique to any one institution or type of college. Rather, they reflect universal developmental processes that all students engage with as they navigate the challenges and opportunities of higher education and

beyond. Expanding this work to a wider range of settings and populations will allow us to create tools that are not only grounded in strong theory, but are also practical and adaptable for those working to foster student growth and transformation.

Appendices

[Appendix A: Descriptive Characteristics of Sample \(N=114\)](#)

[Appendix B: Scenario Characteristics for the School Context](#)

[Appendix C: Scenario characteristics for family context](#)

[Appendix D: Category characteristics for school context](#)

[Appendix E: Category characteristics for family context](#)

References

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573. <https://doi.org/10.1007/BF02293814>
- Baxter Magolda, M. B. (2001). *Making their own way: Narratives for transforming higher education to promote self-development*. Stylus.
- Baxter Magolda, M. B. (2009). The activity of meaning making: A holistic perspective on college student development. *Journal of College Student Development*, 50(6), 621–639. <https://doi.org/10.1353/csd.0.0106>
- Baxter Magolda, M. B. (2010). The interweaving of epistemological, intrapersonal, and interpersonal development in the evolution of self-authorship. In M. B. Baxter Magolda, E. G. Creamer, & P. S. Meszaros (Eds.), *Development and assessment of self-authorship: Exploring the concept across cultures* (pp. 25–43). Stylus.
- Boston College (2023). "The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Boston College (protocol code: 24.076.01e; approval date: 9 August 2023)."
- Braun, H. & Kirsch, I. (2025). Large-scale assessments in international contexts. In L. Cook & M. Pitoniak (Eds.) *Educational Measurement* (5th ed.). Oxford U. Press.
- Creamer, E. G., Baxter Magolda, M. B., & Yue, J. (2010). Preliminary evidence of the reliability and validity of a quantitative measure of self-authorship. *Journal of College Student Development*, 51(5), 550–562.
- Dalton, J., & Crosby, P. (2010). How we teach character in college: a retrospective on some recent higher education initiatives that promote moral and civic learning. *Journal of College and Character*, 11, 13–17. <https://doi.org/10.2202/1940-1639.1264>
- Fallar, R., Hanss, B., Sefcik, R., Goodson, L., Kase, N., & Katz, C. (2019). Investigating a quantitative measure of student self-authorship for undergraduate medical education. *Journal of Medical Education and Curricular Development*, 6, 2382120519896789. <https://doi.org/10.1177/2382120519896789>
- Glanzer, P. L., Hill, J. P. & Johnson, B.R. (2017). *The Quest for Purpose: The Collegiate Search for a Meaningful Life*. Albany: State University of New York Press.
- Helsing, D., & Howell, A. (2014). Understanding leadership from the inside out: Assessing leadership potential using constructive-developmental theory. *Journal of Management Inquiry*, 23(2), 186–204. <https://doi.org/10.1177/1056492613500717>
- Kegan, R. (1982). *The evolving self: Problem and process in human development*. Harvard University Press.
- Kegan, R. (1994). *In over our heads: The mental demands of modern life*. Harvard University Press.
- King, P. M. (2009). Principles of development and developmental change underlying theories of cognitive and moral development. *Journal of College Student Development*, 50(6), 597–620. <https://doi.org/10.1353/csd.0.0104>
- Lahey, L., Felix, S., Goodman, R., Kegan, R., & Souvaine, E. (2011). *A guide to the subject-object interview: Its administration and interpretation*. Minds at Work.
- Linacre, J. M. (2002). *Computer-adaptive testing: Memo 69*. Rasch Measurement SIG. <https://www.rasch.org/memo69.pdf>
- Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1), 85–106.

- Linacre, J. M. (2019). *Fit diagnosis: Infit outfit mean-square standardized*.
<https://www.winsteps.com/winman/misfitdiagnosis.htm>
- Linacre, J. M. (2024). *WINSTEPS Version 5.8.3.0*. Winsteps.com.
- Ludlow, L.H. (1983). The analysis of Rasch model residuals [Unpublished doctoral dissertation]. University of Chicago.
- Ludlow, L.H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement*, 10, 217-229.
- Ludlow, L.H. (2023). *Measuring meaning-making* [Unpublished IRB protocol]. Lynch School of Education and Human Development, Boston College.
- Ludlow, L.H. & Haley, S.M. (1995). Rasch model logits: Interpretation, use, and transformation. *Educational and Psychological Measurement*, 55, 967-975.
- Ludlow, L.H., Baez-Cruz, M., Chang, W-C C., & Reynolds, K. (2020a). Rasch/Guttman Scenario Scales: A Methodological Framework. *Journal of Applied Measurement*, 21(4), 361-378.
- Ludlow, L.H., Anghel, E., Szendey, O., O'Keefe, T., Howell, B., Matz-Costa, C., & Braun, H. (2020b). The Boston College Living a Life of Meaning and Purpose (BC-LAMP) Portfolio: An application of Rasch/Guttman Scenario methodology. *Journal of Applied Measurement*, 21(2), 134-153.
- McAdams, D. P. (2013). The psychological self as actor, agent, and author. *Perspectives on Psychological Science*, 8(3), 272–295. <https://doi.org/10.1177/1745691612464657>
- O'Keefe, T., Ludlow, L., Matz, C., Braun, H., Anghel, E., & Szendey, O. (2022, April 23). *Measuring growth in the development of meaning making: The analysis of Rasch model residuals* [Paper presentation]. American Educational Research Association (AERA) Annual Meeting, San Diego, CA, United States.
- O'Keefe, T. A., Warner, L., Matz, C., Ludlow, L. H., & Braun, H. I. (2025). A reframing of meaning-making and its measurement among emerging adults. *Religions*, 16(11), 1431. <https://doi.org/10.3390/rel16111431>
- Parks, S. (2000). *Big questions, worthy dreams: Mentoring young adults in their search for meaning, purpose, and faith*. Jossey-Bass.
- Piaget, J., & Cook, M. (1952). *The origins of intelligence in children* (Vol. 8, No. 5, pp. 18-1952). New York: International universities press.
- Pizzolato, J. E. (2003). Developing self-authorship: Exploring the experiences of high-risk college students. *Journal of College Student Development*, 44(6), 797–812. <https://doi.org/10.1353/csd.2003.0074>
- Stewart, C., & Wolodko, B. (2016). University Educator Mindsets: How Might Adult Constructive-Developmental Theory Support Design of Adaptive Learning?. *Mind, Brain, and Education*, 10(4), 247-255.
- van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46(3), 247–272. <https://doi.org/10.1111/j.1745-3984.2009.00080.x>
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.