Research & Practice in Assessment

Volume Twenty | Issue 2 | rpajournal.com | ISSN #2161-4120

A Publication of the Virginia Assessment Group

Research & Practice in Assessment

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 continous 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 a national readership. The journal continues to be published by VAG. A full history and timeline for the journal can be found at www.rpajournal.com.

Table of Contents

Editorial Staff

From the Editor

"Not everything that is faced can be changed, but nothing can be changed until it is faced."
— James Baldwin

As assessment continues to mature as both a scholarly field and a professional practice, we are increasingly called to confront its assumptions, methods, and consequences with greater clarity and intentionality. The articles in this issue of Research & Practice in Assessment reflect that call, offering work that faces longstanding challenges in assessment while advancing tools, frameworks, and perspectives designed to improve how evidence of learning is generated and used.

Medina and Holshouser introduce The Tri-Perspective Observation Tool (3-POT), presenting an approach to classroom observation that prioritizes descriptive evidence, transparency, and growth. Their study illustrates how observation tools can complement rubric-based systems while mitigating some of the constraints that accompany purely numerical judgments. In Peak Learning Moments, McGoey and colleagues analyze students' accounts of meaningful learning experiences, offering insight into how assessment can better capture learning as it is lived and perceived by students themselves.

In The Evidence of Learning and Impact Framework, Moreno and colleagues report findings from a Delphi study that advances a shared structure for linking evidence of learning to actionable impact. Their work provides a timely contribution for practitioners seeking coherence between assessment data, interpretation, and decision-making. Gibbons and colleagues then share Barriers and Enablers of Integrating Assessment with Curriculum and Instruction, highlighting the organizational and cultural conditions that shape whether assessment is experienced as an embedded practice or an external requirement.

Two articles in this issue push assessment thinking beyond traditional sequencing. In A New Approach to Learning Improvement, Lambert and Good challenge the assumption that assessment must always precede action, proposing an intervention-first model that reorients improvement work. Scully and colleagues, in Assessors' Engagement with Video-Recorded Performance Assessments in Nursing, explore how assessors interact with complex performance evidence, raising important considerations for validity, interpretation, and assessor cognition. Finally, Claydon and Bathgate share Leveraging Student Voices to center student perspectives to examine pathways, aspirations, and structural influences on doctoral participation. This work reinforces the value of assessment approaches that elevate student voice as a legitimate and necessary source of evidence.

Taken together, these contributions remind us that effective assessment requires more than technical precision. It requires reflection, responsiveness to context, and a willingness to reconsider how and why our practices endure. I hope this issue prompts both thoughtful dialogue and practical experimentation in your own assessment work.

Warmly,

Nicholas Curtis

Nicholas Curtis
Editor-in-Chief
Research & Practice in Assessment

The Tri-Perspective Observation Tool (3-POT): A Worthy Addition in a Comprehensive Teacher Observational System

Authors:

Adriana L. Medina
The University of North Carolina at Charlotte

Kaitlyn O. Holshouser
Gardner Webb University

ABSTRACT

The Tri-Perspective Observation Tool (3-POT) was developed with the purpose of lessening the constraints typically associated with rubric-based performance measures. This study examined how the 3-POT as an observation tool can support other observation instruments as part of a holistic teacher observational system prioritizing accountability and growth. Data was collected through interviews and focus groups from Student Teachers, Clinical Teachers, and University Supervisors. Student Teachers described how the 3-POT satisfied their desire for feedback. Supervisors reported how the 3-POT offered a new lens for conducting observations and allowed for transparency. The information captured in the 3-POT contextualized the score on the rubric-based performance measured instrument. The 3-POT captured what transpired, allowed for feedback focused on growth, was applicable in a variety of teaching contexts, and complemented a variety of measurement instruments. The 3-POT's validity, versatility, and variability make it a worthy tool for inclusion in a comprehensive teacher observational system.

Correspondence E-mail: AdrianaLMedina@Charlotte.edu
Keywords: Student Teaching, Teacher Observation, Teacher Education, Observation Tool
Funding: The Bank of America Faculty Research Fellowship Fund

Observation is a common and critical component of teacher evaluation systems (Ross & Walsh, 2019; Steinberg & Donaldson, 2016; Steinberg & Garrett, 2016). For both pre-service and in-service teachers, observations often serve one of two purposes: accountability or growth (Gabriel, 2018; Wise et al., 1985). While on some occasions, observations are used for the purpose of teacher development, on other occasions, observations aid in making high-stakes decisions, such as whether an in-service teacher's contract will be renewed or whether a student teacher will be recommended for licensure. The student teaching experience is a time for Student Teachers to grow in their teaching practice and develop the agency needed to successfully step into the role of a classroom teacher. The student teaching observation process is a critical component of most teacher education programs.

Classroom observations begin in teacher preparation programs and are typically thought of as necessary supports for Student Teachers as they transition into their roles as practitioners (Jonsson & Panadero, 2017). These observations are often conducted by a Clinical Teacher who is in the classroom, modeling teaching practices in real-time, and a University Supervisor who serves as the liaison between the University and the school system. These two supervisors are not the sole factors influencing the quality of feedback and the degree of support provided to Student Teachers. Caughlan and Jiang (2014) note the observation instrument is also an actor influencing the focus and feedback produced from an observation. They warn observation instruments "are not neutral but reflect the values of the programs that use them through particular (and sometimes contradictory) discourses of teacher learning and student learning" (p. 375). Observation instruments have been critiqued for their lack of depth, rigid criteria, relevant feedback, and user subjectivity (Bell et al., 2015; Cohen & Goldhaber, 2016; Gabriel, 2018). While according to Gabriel (2018) the goals of observations are typically accountability and growth, given rigid criteria and numerical values associated with many observation instruments in student teaching, emphasis is placed on accountability more so than growth. This is concerning as the primary focus of student teaching should be growth. Often the observation instruments focused on accountability are rubric-like and numerical in nature, that is, there is a number or score the Student Teacher receives for the competencies under observation. These instruments often do not allow much space for the observer to take and make notes regarding what was observed, nor are they accompanied by an observation tool to support the score. Without notes regarding what was observed, the numerical feedback from these instruments might seem to be more for accountability purposes than for fostering growth in the Student Teacher's ability to demonstrate proficiency in teacher competencies. Therefore, there exists a need for an observational tool that can complement these rubric and numerical types of observation instruments. There is a need for a tool that captures what transpired during an observation, allows for feedback focusing on growth, is applicable in a variety of teaching contexts, and can complement a variety of measurement instruments.

Literature Review

Rubrics are a primary instrument employed for observations in teacher preparation programs. When the observation instruments have criteria and numeric performance levels serving as a scoring guide to measure the proficiency levels of Student Teachers, these instruments can be classified as rubrics. Rubrics, in and of themselves and

independent of the observers, primarily focus on accountability and not growth. Gabriel (2018) posits that instruments cannot serve the purpose of accountability and growth simultaneously, therefore, one must ask which purpose do rubrics prioritize? Given the quantitative nature of rubrics, the goal is to determine whether Student Teachers meet a certain criterion demonstrating proficiency; therefore, it can be argued their central purpose is one of accountability. Shortcomings of these rubrics fall into the categories of lack of depth, rigid criteria, and relevant feedback.

Lack of Depth

A generic rubric might be used when observing a Student Teacher's classroom management which is a broad category and has several large facets which could take on a multitude of forms. Additionally, a generic classroom management rubric could transcend subject areas and grade levels. Although generic rubrics may not encourage criteria compliance to the same extent as specific instruments, they do not come without critique (Gabriel, 2018; Gabriel & Woulfin, 2015; McAbee, 2016). While highly specific observation instruments may provide narrow definitions of quality teaching, generic rubrics may fail to capture the complexity inherent in teaching, especially if the evaluator is not attuned to the best practices within the disciplines they are observing (McAbee, 2016). Tierney and Simon (2004) point out that although generic rubrics may afford the user more versatility, there exists a "tradeoff" in terms of rater reliability (p. 2). This is due to the fact that generic rubrics do "not contain concrete or task-specific descriptions to guide interpretation" (Tierney & Simon, 2004, p. 2). Whereas task-specific rubrics may compensate to some extent for an untrained evaluator, generic rubrics do not make such affordances.

Rigid Criteria

Currently, teacher evaluation measures "are used to make hiring, promotion, tenure, and dismissal decisions;" therefore, it comes as no surprise that observation instruments exert influence on the type of teaching occurring in schools (Harris et al., 2014, p. 74; Williams & Hebert, 2020). While this may have positive outcomes in some educational settings, McAbee (2016) argues that in some cases "the instrument is creating a situation where the teacher has to sacrifice quality instruction" (p. 179). For Student Teachers, the student teaching observations hold power over candidates' ability to obtain recommendation for licensure. Consequently, Student Teachers may feel pressure to teach to an observation instrument. This criteria compliance (Torrance, 2007), over time, could result in decreased agency (Caughlan & Jiang, 2014).

Cohen and Goldhaber (2016) discuss the fact many "constructs" of good teaching practices have emerged through educational research over time. Observation instruments that are highly specific may neglect other research-based teaching strategies, thus limiting teachers' creative freedom. Observation instruments that highlight one form of teaching over another fail to realize the intricacy of teaching to meet a given context (Connor, 2013; Cohen & Goldhaber, 2016). Therefore, time and attention must be given to the selection of appropriate observation instruments, since they often set the tone for what is valued as good teaching (Caughlan & Jiang, 2014; Connor et al., 2014).

Relevant Feedback

Feedback is two-pronged - what the user of the instrument provides as well as what the Student Teacher receives during debriefing conversations (Moskal, 2000). The nature of the instrument can influence the feedback provided (Caughlan and Jiang, 2014). A well-designed, task-specific rubric can be used for growth if additional feedback is provided beyond the score number. The language of each criteria can be used as a goal for the Student Teacher to attain. However, this is only possible if the feedback from the observation describes and shows what the Student Teacher did so the Student Teachers and the observer/evaluator can have a discussion of how to do things differently next time - not just to attain a higher score, but to work towards being a more proficient teacher. The conversations stemming from observations are just as important as the observation instrument used (Helgevold et al., 2015). Sosibo (2013) conducted a study with pre-service teachers to examine their perspective on the Student Teacher observation process. A common theme emerging from Sosibo's research was Student Teachers' desire for feedback. Student Teachers were not satisfied with "evaluators who merely made checks on the forms" (Sosibo, 2013, p. 150). Student Teachers desired detailed feedback and felt this feedback was necessary as evidence of growth in teaching.

To fill these gaps, the Tri-Perspective Observation Tool (3-POT) was developed with the purpose to lessen the constraints typically associated with rubric-based performance measures. The purpose of this study was to examine how the 3-POT as an observation tool can support a rubric-based performance measurement observation instrument to complement a teacher observational system prioritizing both accountability and growth.

Conceptual Frameworks

Ethnographic Practice

Ethnography is the study of culture and focuses on the observation and analysis of social practices and interactions in order to better understand a culture (Bloom & Green, 2018). Ethnographers examine, through a cultural lens, interactions and events within a particular community environment. A classroom serves as a community, and within the boundary of a classroom there exists a culture. Thus, using an ethnographic perspective as a lens for classroom observations can be helpful in understanding what is occurring in a classroom (Frank, 1999).

One ethnographic practice is making cultural ways visible through description. "Ethnography can be used as a tool by classroom observers to make visible what members are doing and learning in classrooms and to record, analyze, and represent the particular kind of classroom culture that is being created" (Frank, 1999, p. 3). The 3-POT is developed with this in mind (see Appendix A). The "What I Saw" and "What I Heard" columns allow for descriptions of the environment and of interactions which can make transparent the connection between what was seen and heard and what was thought. From the notes in these columns, the observed chain of events can be reconstructed. A related ethnographic practice is withholding quick judgement. These descriptive notes can later be used as evidence for interpretive notes which is also an ethnographic practice. A related ethnographic practice is taking fieldnotes, and the "What I Thought" column allows for observer fieldnotes, both in the moment and soon afterwards.

Validity

Validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). Observations and how they are conducted align with Messick's clarification of the term "score" which is "any means of observing or documenting consistent behavior or attributes" (p. 13). Messick explains "validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use" and that "what is to be validated is…the inferences derived from test scores or other indicators" (p. 13). Thus, for an observation tool to be useful, it should provide "scores" or information that will be used for its intended purpose. Messick posits that key to validity are these five elements: "interpretation, relevance, and the utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use" (p. 13). When deciding upon instruments to use as part of a teacher observational system (Hill et al., 2012), oftentimes data is not provided regarding the validity of the instrument's uses and inferences, that is, its interpretation, relevance, utility of scores, implication of scores for action, and the social consequence of the instrument (Messick, 1989). This study helps to avoid this pitfall for the 3-POT.

Method

Tool Description

The first author based the 3-POT on the research by Frank (1999) on ethnographic classroom observations and the benefits of notetaking, a description of what is being observed, and notemaking, the observer's thinking and interpretation of what is observed. The 3-POT divides notetaking into two columns, What I Heard and What I Saw. The third column, What I Thought, is used for notemaking. During an observation, the observer attempts to capture as much of the talk and action as is feasible by writing it down in the first two columns. As the action is taking place, or afterwards, the observer can record their thoughts, questions, and comments in the third column. The notemaking becomes the basis for the post-observation conference and the notetaking serves as evidence. The notetaking and notemaking inherent in the 3-POT make it an ethnographic tool for classroom observations. Notetaking allows for descriptive notes and notemaking allows for interpretive notes. Thus, the tool and its use allow for an ethnographic perspective. The 3-POT does not yield a numerical score. The observer uses the 3-POT to provide feedback related to what transpired during the observation based on what they saw, heard, and thought. The feedback information is used to guide the post-observation conversation.

Research Questions

To examine how the 3-POT as an observation tool can support a rubric-based performance measurement instrument, evidence was gathered to address the following questions:

- What information is gained through the 3-POT and what is missed?
- How do Student Teachers experience being observed with the 3-POT?
- How is the 3-POT used by observers to assess Student Teacher performance?

Setting and Participants

This study took place at a public urban research university in the southeastern part of the United States. The student teaching experience occurs during the last semester of a teacher candidate's program. The Student Teachers are assigned placement by the Office of Field Experience and their Clinical Teachers are designated by the school administration. The University Supervisors are designated by the Director of the Office of Field Experiences at the University and assigned to a school. The sample came from a pre-established process and therefore would be deemed a convenience sample. The study had a total of 21 participants - three University Supervisors, who each supervised three Student Teachers, who were each supervised by a Clinical Teacher. One University Supervisor, three Student Teacher, and three Clinical Teachers were at a high school. Two elementary schools each had one University Supervisor, three Student Teachers, and three Clinical Teachers. The participants mainly self-identified as females except for one male high school Student Teacher.

Data Collection

There were four stages to the research design. Initially, all of the participants were interviewed individually or in a focus group, depending on their teaching and class schedules. Student Teachers were asked about their expectations of the student teaching experience in reference to the students, classroom management, delivery of content, instruction, and being observed and evaluated.

The supervisors, that is, the Clinical Teachers and University Supervisors, were asked about their evaluation practices, philosophy of teacher education, and their use of the current (rubric-based performance measurement) observation instrument, the Student Teaching Assessment Rubric ([STAR] Jaus et al., 2007) which was comprised of ten standards[(a)] on which PSTs were observed four times during their student teaching semester. Next, the supervisors were trained on how to use the 3-POT. An explanation of the tool and a demonstration were provided. Then, Clinical Teachers and University Supervisors were asked to conduct one observation using the 3-POT.

Finally, participants were interviewed in focus groups again. The supervisors were asked to describe the barriers and facilitators to using the 3-POT. The semi-structured interview questions were: Did you find it easy or difficult to use the 3-POT? What information did you gain? How did you find the 3-POT as a tool for observing the instructional practices of Student Teachers? The Student Teachers were asked how their expectations matched their experience and about their feelings of being observed with the 3-POT. Specifically, they were asked: What information did you expect to gain from the observations of your teaching? What are your post-observation reactions to being observed with the 3-POT? Was there any difference in the information you received when observed with the 3-POT than when observed with other instruments?

[(a)] The 10 Standards were: Content Pedagogy, Student Development, Diverse Learners, Instructional Strategies, Motivation and Management, Communication and Technology, Planning, Assessment, Professional Growth, School and Community Relationships

Table 1. *Categories, Themes, Descriptions, and Codes*

Categories	Themes	Description	Codes (Frequency)
Positive Aspects	General Positives	General positive aspects of the 3-POT	Positive Aspects (11)
	Reflection	Note-taking on what was seen and heard allowed for deeper reflection and note-making afterwards	Allows for Reflection (17)
	Transparency	The 3-POT allows for transparency	Transcript (21) Depth of Observation (20) New Lens (6) Visual (6)
Limitations	Tool Limitation	3-POT limitations or limitations arising from observer influence over the 3-POT	Writing is Time Intensive (34) Multiple Interactions (5) Length of Observation (4) Learning a New Method (4)
Uses	Teaching Points	Using the 3-POT feedback for post-observation discussions and as evidence of demonstrated teaching skills and competence	Specific Issues (34) Snapshot (11) Teaching (9)
	Documentation	Using the 3-POT to document concerns	Document issues (11)
	Suggested Modifications	Suggestions for modifications to further the 3-POT's use as an observation tool	Timing (19) Modify Layout (10) Supplemental (7)
	Comparisons	Comparing/contrasting the 3-POT with other instruments	Broad vs Specific (24) Words vs Numbers (15) Filling it Out (8) Helpful/Useful (7) Deficit-Focused vs Asset-Focused (5) Level of Intrusiveness (1)

Data Analysis

Interviews were recorded and transcribed and de-identified of individual information. NVivo (QSR International, 2019) was used for data management and organization during the qualitative analysis process. The authors coded all the data

together using a constant-comparison method (Strauss & Corbin, 1998). Inductive coding was accomplished first through iterative coding of the transcriptions and deductive coding was employed to pull out the emerging themes within the codes (see Table 1). Data saturation was achieved through the transcripts. The supervisors' observation documents were collected and examined. They confirmed codes but did not add new information.

Results

Findings are presented with regards to 1) potential uses of the 3-POT, 2) validity elements, namely, interpretability, relevance, the utility of results, the implications of the results as a basis for action, and the functional worth of the results with regards to the social consequences of the use of the results, 3) positive aspects of the 3-POT, 4) limitations of the tool, and 5) suggested modifications. Direct quotes are italicized and codes are in bold.

Potential Uses

One Clinical Teacher suggested using the 3-POT to point out specific issues, "if a [Student Teacher] was having difficulties" with some aspect of teaching. One Student Teacher suggested the focus on specific issues as well, however with the idea that the Student Teacher could indicate what issue they "would like [the supervisor] to pay attention to" when observing them.

It was suggested the 3-POT could be used to capture a snapshot of the teaching. A University Supervisor said, "I do think it's nice to have a snapshot. I think it would be great to… transcribe 20 minutes of all your classes. By the time you got to the end of the semester, you would have a nice snapshot of what was going on in that teacher's classroom. …It could be 20 minutes from the beginning one time, 20 minutes in the middle the next time, etc." It was also suggested the 3-POT made for a "better teaching tool for [the Student Teacher] to see how she could handle things." Additionally, using the tool as a point from which to teach "could spark a discussion" during the debrief about what went well and what could be done differently next time.

Another suggested use of the 3-POT was as a method to document issues. One Clinical Teacher felt the transparency afforded by the 3-POT could be used for a Student Teacher "who maybe could not take criticism well." She felt the transparency might have the Student Teacher saying, "Oh yeah, [my Clinical Teacher] is not just giving me a difficult time. That actually did happen. I see that now." A University Supervisor indicated she could use the 3-POT observation as data for a Student Teacher who is struggling "to let [him/her] know what issues [he/she is] having."

On the other hand, one Student Teacher felt the observation with the 3-POT could help her justify her actions and show evidence of her being a competent teacher. She gave the example of when her Clinical Teacher wrote: "I re-write the problem after the student reads it and, it's just that I like repeating things." The Student Teacher defended herself by pointing out how "it's a lengthy word problem. I don't think it's a bad thing repeating because when you have a kid talking right here (points next to her), I think it's good that I repeat the problem for other kids to hear it."

Several stakeholders felt the 3-POT would be useful as a supplement to other existing tools. Many suggestions were related to timing. Some supervisors preferred to use the 3-POT early on in the student teaching experience. A University Supervisor agreed it

could serve as a "baseline observation." Other supervisors thought the 3-POT might be useful at mid-term. Another thought it could be used over time to "add to the big picture."

Interpretability

The stakeholders interpreted the 3-POT to be used as intended, for observation. The supervisors indicated this tool reminded them of other observation tools they had used. For example, one University Supervisor said, "…when I was Principal…I used the same format [scripting]." The Student Teachers interpreted the 3-POT to be used for observation. One Student Teacher noted, "on this one, you can see exactly where [the supervisors] saw those things; what you were doing when this happened."

Relevance

The stakeholders indicated the 3-POT was relevant within the student teaching context. The supervisors indicated the tool allowed them to capture the lesson and gave them the opportunity to provide specific, constructive feedback. One Clinical Teacher said, "if I would have done the observation just in my nature, … I probably would have not written down some things." Using the 3-POT, this Clinical Teacher identified relevant aspects of teaching for her Student Teacher to focus on she might not have found otherwise with another tool. She summarized, "I think there were a couple of small things to work on that probably would have not come out if I was doing it the other way." A University Supervisor felt the 3-POT was relevant in helping her remember more about the observations. "I remembered a lot more about it. … I had a lot more detailed."

The Student Teachers felt the 3-POT was relevant in providing them growth-oriented feedback. One Student Teacher noted, "… I could really see exactly what [my University Supervisor] was saying when she told me 'You need to do more of this' and I could see how I progressed." Another agreed, "It's not so cut and dry/black and white…. They can elaborate on some things." Another Student Teacher indicated the 3-POT helped the supervisor's comments be more relevant. "I like how you can see what she thought, based on what she heard and she saw how it corresponds with each other in rows." Another agreed, "I felt this only captured the lesson. It didn't go into all the community stuff as the STAR does. You only saw and heard what was in the classroom at that time."

Utility of Results

The stakeholders noted the usefulness of the results of the 3-POT. One Clinical Teacher felt the 3-POT allowed her to "give more anecdotal kind of support than [she] would probably put in a general observation." A University Supervisor agreed. She said, "I think I would focus using this instrument so I can go back and show the student teacher exactly what they were saying, what I was seeing and what the children were doing." Another University Supervisor stressed the 3-POT was useful because of its potential for transparency: "What it was, there it was. This is what I thought, this is what the children were doing, this is what you were doing, this is what was going on in the classroom that I saw. There aren't negatives or positives. It's not negatively or positively stated. It's what the observation was. This is what I saw."

The Student Teachers felt it was helpful to know what the supervisors were seeing and thinking: "it helps to point out [they've] been thinking this when you we were doing this." The Student Teachers felt the 3-POT provided more relevant information than other

tools used in the program. For example, one Student Teacher stated, "[it's] more detailed feedback on some of the things that's not on the InTASC , like, on actual teaching, instead of just concepts, and [it's] actually showing different strategies we used, … actually how we did it."

Implications for Action

The stakeholders were well aware of the importance of observation results for successful completion of the student teaching experience. One Clinical Teacher reflected, "I would have preferred to use [the 3-POT] for an earlier observation because this is very, very specific as to what I'm seeing and the things that I wrote down that I saw and I thought would be helpful in the next observation." Another agreed, "I think it would have been helpful to have it in the beginning." A third Clinical Teacher noted, "If you do it earlier, maybe on the 2nd or 3rd observation, it gives [the Student Teachers] a chance to use [the information] to try to improve areas where they might need a little more work to do." A University Supervisor thought the 3-POT might have more valuable implications for action for when "you have a struggling student and they don't see it. When you have a student to whom you say 'this is what my impression is of what you are doing' and they go (makes a blank face), it's like they don't get it, they don't see that, they can't step outside of themselves and see how they are perceived." One Clinical Teacher found the 3-POT to be so useful and valuable, she mentioned she'd like to use it as a mentor teacher at her school to observe a new teacher.

Worth and Social Consequence of Use

Student Teachers agreed the 3-POT results were worthwhile because it gave them information "that's in their [Supervisors'] head." One Student Teacher indicated "the most important column for [her] was the one about what [her] teacher thought about the actions [she] was doing." For some Student Teachers, "it's most important just getting that [supervisor's] feedback and the suggestions and the criticisms." One Clinical Teacher noted the worth of the 3-POT was "you can't argue with it, …, this is what you said and this is what the student did or, this is what you did. This is where you walked; this is where you stood most of the time. It is un-debatable; it's unquestionable."

Positive Aspects

The 3-POT allowed for depth of observation. One University Supervisor indicated that the column, "'What are you seeing?' made [her] more cognizant of looking." A Clinical Teacher indicated "this tool had me analyze some things more deeply. Taking a look at what I heard and then what I saw and then what I thought. I normally do those things … but probably not to the same depth." The 3-POT served as a new lens for seeing things not seen before. As one Clinical Teacher said, "there were a couple of small things to work on that probably would have not come out if I was [observing] the other way [with the other tools]."

Another positive aspect noted was the 3-POT allowed for transparency. As one Student Teacher remarked, "I don't think it was as subjective. It was there. It was all written down for me." Another positive aspect was the 3-POT functioned as a transcript and provided "word by word by word" what transpired during the observed time. All stakeholders indicated the transcript helped them create a mental picture of the

observation where they could "visualize this is what she said, this is what the kids are doing." While to some it felt the observation "was almost like you took a videotape and you wrote the videotape," to others it felt like "it kind of gives you a transcript of what you did without being on video, because on video you change." Another positive aspect was the 3-POT helped eliminate forgetting. One Clinical Teacher said: "what I found myself doing was trying to get as much of what I heard and saw because I felt that would be parts I would easily forget." This was important because Student Teachers indicated they felt there was a disconnect with the University Supervisor who was not always in the classroom. One Student Teacher said it this way: "when the University Supervisor is observing me, she doesn't see that I do give feedback to each child individually, but my Clinical Teacher sees it. And, so [the University Supervisor] will point out: 'you should have done more individual feedback' and I've done it before in another setting, it's just that in this certain setting that you are observing me on, it's not…I feel like my ratings are lower because she only sees a sip of what I've been doing; she doesn't see my full day or exactly everything I do. So, the University Supervisor was saying she didn't see this, she didn't see that, yet my Clinical Teacher was saying 'I see it.'" The transcript nature of the 3-POT could serve as counterpoint evidence for this type of discrepancy.

The stakeholders noted another positive aspect of the 3-POT – it allowed for reflection. Clinical Teachers indicated the 3-POT gave them a "chance to reflect back on what I heard and saw" and "a second chance to analyze."

Limitations of the Tool

The stakeholders felt the main limitation of the 3-POT was the writing. All of the supervisors wrote their observations by hand. The focus on "transcribing" and "getting everything down" was seen as a limitation. Some Clinical Teachers didn't mind the writing, but "did mind the typing it up part" afterwards. One Clinical Teacher stated, "This is 14 pages of typing. And it's really hard to type Math so, it took forever." Another agreed, "it did take much longer to transcribe and type." This concern is valid as the utility of the 3-POT could be compromised over time if observers reduce what they capture in writing. Capturing multiple interactions such as small groups and "all the chatter and questions of preschoolers" in a Kindergarten classroom was "difficult because there's so much going on." Naturally, it is not expected or possible for one observer to capture all that transpires during an observation; thus, there is selectivity and subjectivity inherent in observations when using any tool. A third limitation noted was this was a new method to learn and they already had methods they preferred. One University Supervisor confessed, "the first one I did I kept having this urge to write down numbers next to each of my comments. … That's a 2.4."

Suggested Modifications

Several stakeholders suggested modifying the layout of the 3-POT. One stakeholder suggested adding columns for the Student Teachers to weigh in on the observation. One Student Teacher felt this would have been useful for her observation where her Clinical Teacher noted: "You gave such-and-such the eye." During the interview, the Student Teacher countered: "That's what [the Clinical Teacher] saw. [She] didn't hear me say anything to go with it…" The Student Teacher did not feel comfortable confronting

her Clinical Teacher. With an additional column, she felt she could have responded and explained: "Well, they were tapping another kid."

One Clinical Teacher suggested changing the order of the columns and having "what I saw" to be the first column, then "what I heard" and "what I thought." A couple of Clinical Teachers wanted to change the orientation of the page from portrait to landscape. Another suggested adding specific categories of focus, for example, "How did the teacher open the class?" And then, what you heard, what you saw, what you thought about the opening of the class."

Limitations to the Study

It might be seen as a limitation that the supervisors' level of observation experience was not captured. Yet, two of the University Supervisors were part of an author team that developed observation instruments (some of which were used at this institution), and the Clinical Teachers were chosen by the Office of Field Experiences which works with district school administrators in the identification of qualified teachers to serve as Clinical Teachers. Secondly, the use of focus groups might be a limitation with regards to threats of social desirability. While the first author was not involved in the student teaching program, the Student Teachers were aware of her role as a faculty member and this might have influenced their critiques of the system they were currently participating in towards recommendation for licensure. Thirdly, it could also be seen as a limitation that the first author developed the 3-POT and she conducted the interviews. Future studies could employ an independent interviewer; however, the second author joined the study at the analysis stage and was beneficial in counteracting this limitation. Fourth, it should be noted that the instruments and tools influence the activity, so it is possible the supervisors saw differences in them because of what the instruments and tools ask them to do. Future studies might compare similar types of instruments or tools. The 3-POT was compared to a low inference instrument; the findings might be different if the 3-POT were compared to another ethnographic tool. Finally, the fact that the 3-POT was not used with middle school Student Teachers and was only used with pre-service teachers could be seen as a limitation. Future studies should include a more varied sample population.

Discussion

The 3-POT was developed to lessen the constraints typically associated with rubric-based performance measures and complement those rubrics by providing qualitative feedback. The absence of numeric proficiency levels of the 3-POT positions Student Teachers to receive feedback supporting their growth. A number score "tells" while the 3-POT allows for the observer to "show" what transpired during an observation to support the score earned. In essence, the 3-POT is simply the detailed notetaking and notemaking of one's person's observation. Yet, the information captured in the 3-POT is highly valuable because it contextualizes the score on the measurement instrument. There are many observation protocols that allow for observation notes, but it was beyond the scope of this study to identify those or compare them to the 3-POT. However, for scenarios where this is not the case, the 3-POT could be a simple and useful complementary tool. Measurement tools would benefit by being paired with the 3-POT as its validity, versatility, and variability make it a worthy tool for inclusion in a comprehensive teacher observational system.

Validity

The 3-POT was interpreted as an observation tool by all stakeholders. The tool had relevance within the student teaching observation experience with regards to the qualitative nature of the feedback. Unlike a score rubric with a specific criteria, the open-ended and ethnographic nature of the 3-POT allowed for growth-oriented and context-specific feedback which spurred conversation among supervisors and student teachers proving the utility of the 3-POT as an observation tool within the student teaching observation process.. This highlights the use of assessment as and for learning rather than solely using assessment of learning. Supervisors felt the 3-POT had implications for action because they could see themselves using it at various points throughout the student teaching experience and for various purposes within the observational system. The results indicated the 3-POT had worth for student teacher observation. Yet "validation is a continuing process" (Messick, 1989, p. 13); therefore, further research on variations of the 3-POT and its uses in different contexts will be informative. Generally speaking, observation instruments demand research attention because the consequential validity of many observation instruments is not documented in the literature.

Versatility

The 3-POT is versatile enough to stand alone or to be used in conjunction with other methods, instruments, and tools. The 3-POT can be used for observing a specific area in which a teacher needs support (e.g., classroom management). Additionally, the 3-POT is applicable to a variety of observation formats, subject areas, and grade levels. It can be used for remote observations, virtual teaching observations, and as a method of observing teachers on video (e.g., via YouTube) or videos uploaded to a video capture platform (e.g., GoReact). In light of the findings which suggested the 3POT was writing intensive, an appropriate modification could be converting it into a digital format allowing the observer to type thereby enhancing the versatility of the instrument.

The 3-POT could also be used by preservice teachers during field placements throughout their program of study and up to student teaching. Using the 3-POT as a classroom observation tool would allow preservice teachers to become familiar with the 3-POT before they are observed with it. The 3-POT could help preservice teachers understand the field placement classroom culture they are observing and entering.

The 3-POTs versatility has the potential to complement other tools used as part of a teacher observational system to help with teacher growth and accountability. While observational systems are in need of more tools that are growth-oriented rather than focused solely on accountability, it should be noted that the observer is the sieve through which the observation is captured. If the observer is not aware of their biases and does not approach the observation with a growth mindset, then the 3-POT will inadvertently serve the purpose of accountability. Thus reflection is needed when utilizing the 3-POT or any observation tool.

Variety

The 3-POT can be used within an observational system. Given the range of placements Student Teachers may find themselves in, there exists a need for a set of observation instruments for a variety of teaching contexts, without losing specificity as this aids in providing detailed feedback. While the 3-POT is versatile, as with many other

instruments, one tool cannot provide a well-rounded picture of teacher quality. Different instruments should be used for different purposes and in combination with each other in order to create a high-functioning teacher observational system (Hill et al., 2012). In the larger in-service teacher observational system, the 3-POT could be used informally, for peer-teacher observations, and formally within the teacher evaluation process. While the 3-POT was studied within the context of pre-service teacher education, its uses could extend to a variety of other settings.

Appendices

Appendix A: Tri-Perspective Observation Tool (3-POT)

References

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2015). Improving observational score quality: Challenges in observer thinking. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 50–97). Jossey-Bass. https://doi.org/10.1002/9781119210856.ch3

Bloom, D., & Green, J. L. (2018). Ethnography. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation*. SAGE Publications. https://doi.org/10.4135/9781506326139.n239

Caughlan, S., & Jiang, H. (2014). Observation and teacher quality: Critical analysis of observational instruments in preservice teacher performance assessments. *Journal of Teacher Education, 65*(5), 375–388. https://doi.org/10.1177/0022487114541546

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378–387. https://doi.org/10.3102/0013189X16659442

Connor, C. (2013). Commentary on two classroom observation systems: Moving toward a shared understanding of effective teaching. *School Psychology Quarterly, 28*(4), 342–346. https://doi.org/10.1037/spq0000045

Connor, C. M., Spencer, M., Day, S. L., Giuliani, S., Ingebrand, S. W., McLean, L., & Morrison, F. J. (2014). Capturing the complexity: Content, type, and amount of instruction and quality of the classroom learning environment synergistically predict third graders' vocabulary and reading comprehension outcomes. *Journal of Educational Psychology, 106*(3), 762–778. https://doi.org/10.1037/a0035921

Frank, C. (1999). *Ethnographic eyes: A teacher's guide to classroom observation*. Heinemann.

Gabriel, R. (2018). Reframing observation. *The Learning Professional, 39*(4), 46–49. https://learningforward.org/wp-content/uploads/2018/08/reframing-observation.pdf

Gabriel, R. E., & Woulfin, S. (2015). Evaluating the structure and content of observation instruments. In R. Gabriel & R. Allington (Eds.), *Evaluating literacy instruction: Principles and promising practices* (pp. xx–xx). Routledge.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal, 51*(1), 73–112. https://doi.org/10.3102/0002831213517130

Helgevold, N., Naeshiem-Bjørkvik, G., & Østrem, S. (2015). Key focus areas and use of tools in mentoring conversations during internship in initial teacher education. *Teaching and Teacher Education, 49*, 128–137. https://doi.org/10.1016/j.tate.2015.03.005

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Jaus, V. P., Cockman, N. R., Frazier, J. W., Hopper, C. J., & Rebich, S. K. (2007). *Student teaching assessment rubric*. Kendall/Hunt.

Jonsson, A., & Panadero, E. (2017). The use and design of rubrics to support assessment for learning. In D. Carless et al. (Eds.), *Scaling up assessment for learning in higher education* (pp. xx–xx). https://doi.org/10.1007/978-981-10-3045-1_7

McAbee, S. T. (2016). When leadership skills are not enough: The role of principals in high-stakes observations. In R. E. Gabriel & R. L. Allington (Eds.), *Evaluating literacy instruction: Principles and promising practices* (pp. 176–187). Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research, and Evaluation, 7*(3). https://doi.org/10.7275/a5vq-7q66

QSR International Pty Ltd. (2019). *NVivo* (Version 12) [Computer software]. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Ross, E., & Walsh, K. (2019). *State of the states 2019: Teacher and principal evaluation policy*. National Council on Teacher Quality.

Sosibo, L. (2013). Views from below: Students' perceptions of teaching practice evaluations and stakeholder roles. *Perspectives in Education, 31*(4), 141–154.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). SAGE.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340–359. https://doi.org/10.1162/edfp_a_00186

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis, 38*(2), 293–317. https://doi.org/10.3102/0162373715616249

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Research & Evaluation, 9*(2). https://doi.org/10.7275/jtvt-wg68

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria, and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education, 14*(3), 281–294. https://doi.org/10.1080/09695940701591867

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1985). Teacher evaluation: A study of effective practices. *The Elementary School Journal, 86*(1), 60–121. https://www.jstor.org/stable/1001217

Williams, K., & Hebert, D. (2020). Teacher evaluation systems: A literature review on issues and impact. *Research Issues in Contemporary Education, 5*(1), 42–50.

Peak Learning Moments: A Thematic Analysis of Student Experiences in Higher Education

Authors:

Kendall M. McGoey, Ph.D.
Auburn University

Kathleen B. Boyd, Ph.D.
Auburn University

C. Ben Farrow, Ph.D.
Auburn University

Tom Leathem, Ph.D.
Auburn University

Eric M. Wetzel, Ph.D.
Auburn University

ABSTRACT

Peak learning moments are meaningful events that change an individual's perspective. Indeed, collecting students' accounts of peak learning can inform university-wide improvements. Researchers, at a large Southeastern university, employed NVivo in a two-part study to identify and compare themes within students' peak learning moments. Researchers identified 14 peak learning moment themes, with responses being categorized as class (22%) or faculty (15%) most frequently. A second study then compared the frequency of responses into the same 14 themes across two colleges with similar missions. As expected, similar top themes emerged in the two-college sample. However, study abroad emerged as a top theme in one college, while internship surged in the other, likely due to different co-curricular requirements. These studies highlight the value of qualitative data at various organizational levels and its ability to deepen understanding of student learning when combined with quantitative research.

Correspondence E-mail: kmmcgoey@gmail.com

Keywords: Peak Learning Moments; Transformative Learning; Higher Education Assessment; Qualitative Research; Student Experiences

Chip and Dan Heath's (2017) work, "The Power of Moments," examines the meaningful and impactful events that change an individual's perspective (what they called peak learning moments). In this book, the concept of peak learning moments was based on four fundamental ideas. First, individuals reflecting on events in their lives tend to focus on key moments within that event. Second, the transformative moment itself is meaningful and memorable. Third, these moments are created through one or more distinct elements: (a) elevation, (b) insight, (c) pride, and (d) connection. And fourth, individuals must "elevate the ordinary" by intentionally crafting "peak moments" throughout their lives. This involves not only identifying and commemorating significant events but also establishing clear expectations and rituals around transitions. By doing so, individuals can transform ordinary moments into powerful experiences that shape their identity and narrative.

There is evidence that teacher-educators reflect on their peak learning moments with other colleagues and graduate students through written methods like prose and poetry, finding deep gratitude and connectivity in this experience (Waterhouse et al., 2020). Additionally, these teacher-educators found the alignment between their peak learning moments and professional practice helped to sustain the commitment to their roles. While these peak learning moments were explored with a unique sample of teacher-educators, one could argue that this same type of reflection can be just as meaningful for students. A reflection on peak learning moments can allow students the opportunity to respond to about any type of event that they deem impactful, allowing for diversity in perspective and the ability to capture experiences that are beyond the "typical" transformative learning experiences in a quantitative data format for institutions. To capture a wide array of these impactful events, the researchers of the present studies suggest rethinking how institutions frame questions in data collection to account for a multitude of experiences. The approach in the following studies was to prompt students to consider peak learning moments, allowing them the opportunity to reflect on their undergraduate studies in an open-ended response format.

Those within the field of assessment in higher education are tasked to evaluate student learning that occurs both inside and outside of the classroom, while also considering how this learning will transfer to postsecondary contexts. The assessment process typically includes quantitative approaches to measure student learning. For example, faculty may use rubrics to evaluate performance on written assignments or internship supervisors may score student performance on a rating scale. Importantly, higher education continues to evolve, and faculty must seek meaningful ways to capture student learning from multiple perspectives. Researchers and practitioners alike wish to understand student learning at a deeper level, and this understanding may be better reached with the use of qualitative approaches alongside quantitative ones. As mentioned by Newhart (2015), "since qualitative assessment may allow for more depth…we can begin to answer the calls for more accountability for the work we are doing in more detail—as well as respect the diverse student experiences that occur at our institutions" (p. 7).

Because of numerous misconceptions regarding qualitative research in assessment, the use of this approach tends to be underappreciated and underutilized (Qualitative Research Methods in Program Evaluation: Considerations for Federal Staff, 2016). Fortunately, researchers have taken the time to address certain misconceptions and provide clarity on the role this approach can play in assessment. For example, while

researchers typically want to identify data that is generalizable to inform intervention and practice, qualitative research allows for the embracing of differences in students' experiences as opposed to the similarities; these differing perspectives are crucial in informing best practices or making improvements (Harper & Kuh, 2007). Additionally, qualitative data may contextualize quantitative metrics, lending to more meaningful discussions about student learning and perhaps more valid approaches to improvement (Patton, 2002). Fortunately, these approaches are not limited to interviewing or focus groups and can capture large samples of data using a method like open-ended survey questions.

There is existing research conducted with qualitative approaches that touch on very important topics in higher education and specifically aim to enhance understanding of student learning. High Impact Educational Practices (HIPs) are specific practices intended to engage students in deep learning and promote active engagement with knowledge and skills both inside and outside of the classroom, with a few examples being service learning, writing-intensive courses, and common intellectual experiences (Kuh, 2008). Research has shown that students who participate in HIPs better retain information, have improved grades, incorporate the knowledge gained with their overall education, and engage with people that are different than themselves (Nelson et al.; Kramer et al., 2007; Peck et al., 2010). In 2014, Blaney and colleagues qualitatively explored HIPs that were discussed in students' written reflections during an experiential learning program at their institution. They identified themes regarding peer and faculty collaborations, applied coursework, and personal growth. They reasoned that these HIPs could be used to improve or share these types of experiences across institutions.

The present research includes two studies, the first conducted in the fall of 2022 and the second in the summer of 2023, that incorporate a qualitative approach to the assessment and analysis of peak learning moments in higher education. The purpose of the first study is to answer the following research questions.

- How do graduating students describe transformative peak learning moments? And, what types of experiences do graduating students identify as being transformative peak learning moments?
- What overarching, institutional themes exist within the dataset categorizing these students' responses?
- Does gender impact the types of learning moments that graduating students identify as transformative?

In identifying these themes, the researchers can determine which themes were mentioned with the highest frequency, thus informing future student experience-focused interventions and considerations of what is working well at the University. The findings from Study 1 are foundational for this process, with Study 2 addressing additional research questions that expand the thematic analyses.

- Does the field of study impact the types of learning moments that graduating students identify as transformative?
- Can collecting qualitative data about graduating students' transformative peak learning moments provide valuable context to data collected via more direct methods of assessment?

Method

Students at the University are required to complete a zero-credit graduation class. As part of that class, data is gathered via Qualtrics on various elements of their academic experience. Referred to as the Campus Experience and Engagement Survey (CEES), completion of the survey is required for students to meet graduation requirements. The CEES has approximately 40 questions that measure a variety of domains including demographic information, graduation term, perceptions of class experiences, HIP participation, peak learning moments, and expectations as an alum. In addition to HIP participation, this survey collects information about student experiences with HIPs, specifically assessing internships, co-ops, ePortfolios, undergraduate research, and study abroad. The students are asked to respond to questions aligned with the described eight key elements of HIPs, like significant investment of time and effort, diverse experiences, and public demonstration of one's competence (Kuh & O'Donnell, 2013). This study focuses on one question regarding peak learning moments asked as part of the CEES that allow a free response by the student:

> Describe a transformative learning experience, while a student at the University, that helped shape the person you are today (a short experience that was both memorable and meaningful). Please be descriptive and note that the moment could take place anywhere (classroom, internship, study abroad, work, athletics, fraternity/sorority, student government, etc.).

An Introduction to NVivo

NVivo is a software based on the work of Lyn and Tom Richards that provides numerous types of qualitative data via coding, searching for patterns, reporting/exporting data, complex searches and queries, and different modes of output (numeric, visual, and textual) (Jackson & Bazeley, 2019). The software can be used to conduct open-ended searches, queries, and attribute classifications with text data, the latter allowing demographic information to be explored within the created theme cases. The researchers have chosen NVivo for these reasons, and the software itself is helpful in working with large datasets like the ones used in the present studies. The approach to the extraction of themes at the institution level included a combination of NVivo's word and text search query functions and a manual coding review process. As mentioned by Welsh (2002), "NVivo can add rigor to the analysis process…[but] this searching needs to be married with manual scrutiny techniques so that the data are in fact thoroughly interrogated" (p. 5).

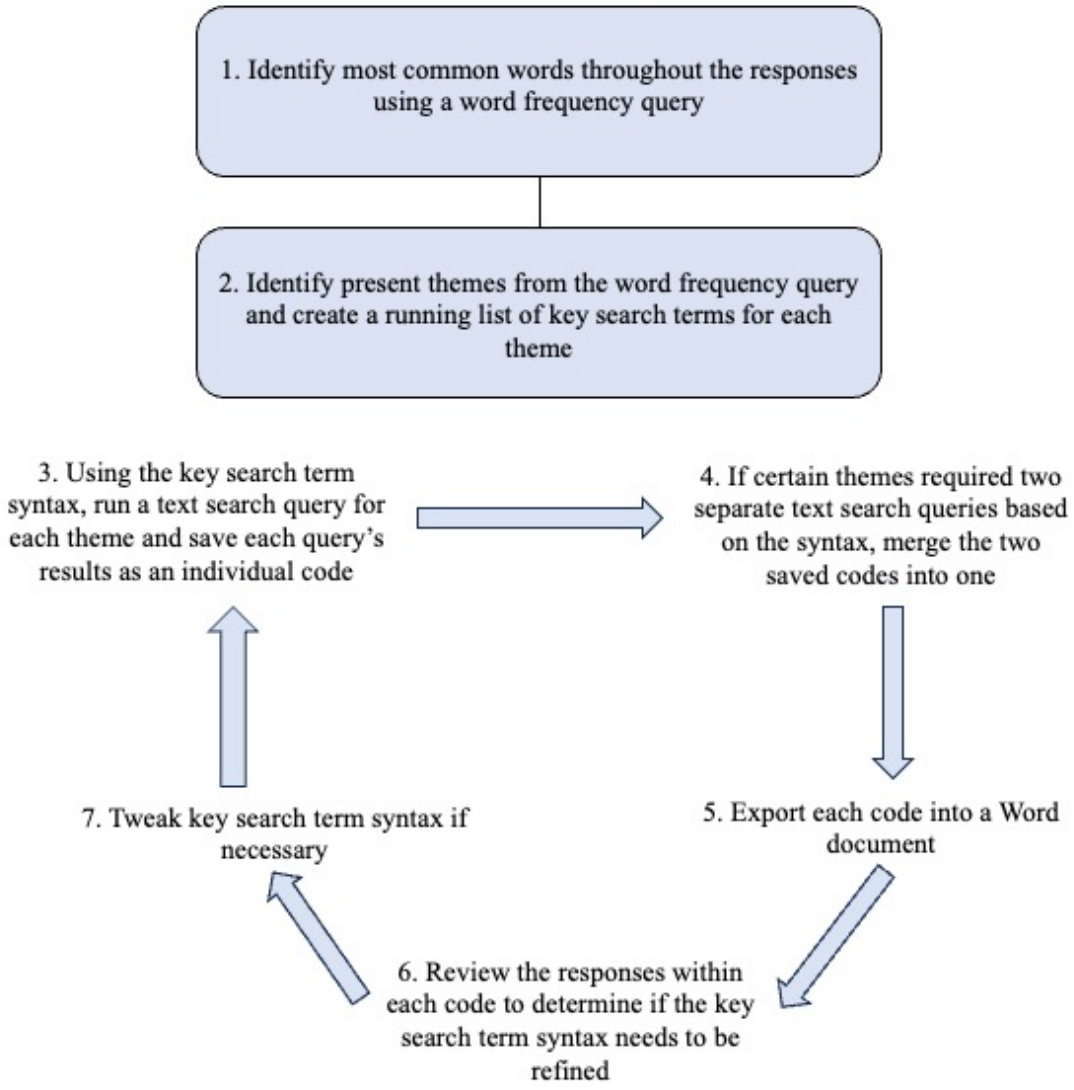
Study 1: Institution-Level Theme Extraction

Participants

The data collected in this study includes student responses from graduating seniors across ten academic semesters (Spring 2020-Spring 2023). In total, open-ended response data from N = 18,267 students about their peak learning moment at the University was collected. After removing duplicate and abstentions (i.e., "N/a", "Prefer not to answer"), the total response count for analysis was N = 17,867. The sample identified as 48.24% male, 82.82% White, 18.93% transfer student, 12.98% first-generation student, and 17.38% eligible for financial aid (Pell Grant eligible).

Procedures

To fulfill the aim of this research - the collection of student transformative experiences - the peak learning moment question allowed for students to freely respond based on any type of experience, not limited to a "traditional" HIP, or educational experience. Again, the research team used queries (word frequency and text search) within NVivo, along with manual coding techniques, to identify the top themes emerging from the data. Please refer to Figure 1 for a framework of the queries used to finalize the themes and the following details provided for each query's purpose.



Note: Steps 3 through 7 are repeated until the text search queries are refined.

Figure 1. *A framework of word frequency and text search queries, along with manual aspects, to code the top themes using NVivo*

A word frequency query, which "catalogues the words used most often in the data," was run first, and the research team manually created a list of present themes and possible key search terms within the response data from this query output (Jackson & Bazeley, 2019). Fourteen overarching themes were identified in the peak learning moment responses:

- Class (response mentioned a class-related experience such as an assignment, project, or exam)
- Co-op (response mentioned full-time, career-related, paid work experience that takes place over three semesters or more with the same organization)
- Diverse Experiences (response mentioned an experience with individuals the student perceived as not common to them)
- ePortfolio (response mentioned a personal website that communicates a professional identity and experiences)
- Faculty (response mentioned a meaningful interaction with faculty that left a lasting impression)
- Greek Life (response mentioned participation in a Greek fraternity or sorority)
- Internship (response mentioned a paid or unpaid professional learning experience that is related to a specific field of study or career)
- Leadership Positions (response mentioned an experience holding a leadership position at the University)
- On-Campus Organizations and Clubs (response mentioned participation in non-Greek organizations and clubs affiliated with the University)
- Personal Relationships (response mentioned a meaningful connection with peer(s), such as friends or significant others, at the University)
- Study Abroad (response mentioned a chance to study in a foreign country)
- Team-Oriented Experiences (response mentioned engaging in groups with a common goal, including athletic and competition-based teams)
- Undergraduate Research (response mentioned an opportunity to participate in research either as a member of a lab team or through classwork)
- Volunteer Experiences (response mentioned volunteering within the community as a part of various personal or organizational endeavors)

A text search query was then used to search for the specific words and phrases within the dataset that were identified during the word frequency query. Text search queries use purposeful search terms to help categorize responses into cases, referred to in this research as themes (Jackson & Bazeley, 2019). The researchers used their generated list of key search terms from the word frequency query, in conjunction with a list of similar words the team identified as important to include (e.g., specific campus unit names associated with diversity in the Diverse Experiences theme). The text search queries were narrowed in focus during the coding process by adding the word "NOT" in addition to "OR," in attempts to exclude responses that may better represent other themes. Of importance, the researchers use the language of "frequency" to report on the responses within each theme because students, while only asked to respond with one peak learning moment, may have mentioned multiple experiences in their responses. This speaks to the nature of qualitative data, but there is richness in identifying how frequently these particular experiences are mentioned throughout the student population. An example

Table 1. *Example Theme, Syntax, and Response*

Theme	Query	Example Response
Class	"class project" OR "class projects" OR "service learning" OR "service-learning" OR "group project" OR "team project"; class OR project OR lab OR laboratory OR test OR assignment OR presentation OR exam OR lecture OR read OR capstone OR technology OR homework OR coursework OR study OR studies OR studio OR preceptorship OR portfolio OR journal OR test OR academic OR course OR instruction OR curriculum OR syllabus NOT abroad NOT sorority NOT fraternity NOT greek NOT internship NOT intern NOT interned NOT interning NOT research NOT professor NOT teacher NOT co-op NOT leadership NOT "Dr." NOT friend NOT friends NOT classmate NOT classmates	"I would say that the most transformative learning experience was my Capstone project for my apparel design degree. We were asked to work in a group and create a collection for a brand of our choosing that could help that brand branch out to new customers. We created an androgynous line of clothing for the brand XX. We were required to create a manufacturer packet called a spec pack. We had to be so thorough with the project that we could send off what we made to a manufacturer they could understand it and start creating our fashion line if we wanted to. I learned so much and applied a lot of my learning and experience to the entire project."

Note: The ";" signifies that there were two separate text search queries that were merged to identify the final count of responses within the theme. The separate searches helped determine that the first set of phrases were captured accurately as they required exact matches.

Table 2. *Using NVivo To Conduct a Text Search Query*

Step 1	Highlight the file of interest and click "Query," then "Text Search Query"
Step 2	Enter in the "Search for" box a list of your key search terms. Separate the terms by "OR," "AND," or "NOT."
Step 3	When using a phrase, defined by quotation marks (i.e, "peer mentor"), or when an exact match is needed, make sure the "Exact Matches" option is clicked. When conducting any other search where it is helpful to see stemmed words (i.e., internship, internships), make sure the "With stemmed words" option is clicked.
Step 4	Choose the "Broad Context" option underneath the "Spread to" dropdown list. This allows NVivo to remove duplicate responses in the case. For example, if a student mentions one of the key search terms more than one time in their response, NVivo will count that as multiple responses. To be able to identify counts for how many students answered within a certain theme, how many female students talked about internships, etc., the researchers wanted to have an accurate number of responses per theme.
Step 5	When the Query is run, save this as a Case with the name of the theme. Of note, the researchers used Cases instead of Codes for saving the themes. This is because Cases are easily merged into other Cases if needed, and the Case Classification option was helpful with our attribution analysis.

theme and query of key search terms can be found in Table 1. In Table 2, there is a step-by-step process of a text search query, which highlights the importance of multiple rounds of coding to refine the themes.

Results

Fourteen overall peak learning moment themes were coded at the institution level, with the top three themes being "class", "faculty", and "internship". The frequencies and percentages of responses per top themes in the data can be found in Figure 2.

Qualitative comments by students reveal additional depth that helps one better understand some of the themed responses above. Comments on "class" included the following:

> One environment that I believed help shaped me into the person I am today, both personally and academically, was my classroom environment. I believe that my cohort, professors, and course work combined help me evolve throughout my experience at the University. With my cohort I learned teamwork and effective communication, allowing me to assess and discuss any obstacle that may have been hard to understand or overcome. With my professors, I believe I learned my value as a student and future employee. Time and time again I did not believe in myself and my potential, which is especially hard within my degree because a lot of our work is based on personal skills and talent. However, my professors always knew the right time to tell me what makes me personally valuable, and I will never forget that.

> I went through a rigorous process to get into my major called summer op. This summer intensive was 10 weeks long and we were in class from 8am-5pm MWF. It was a really long summer and incredibly hard. But I had to complete it to get into my program. It shaped me as a designer and as a student. It instilled in me my work ethic and my time management skills.

Comments on "internship" included the following:

> My child life internship at East Tennessee Children's Hospital and my child life practicum at East Alabama Health were transformative experiences for me because they allowed me to apply all of the learning I had done in my courses to real-world situations in the hospital environment. Although these experiences were different (in practicum, I was in an observer role, and in the internship, I was in a leadership role providing independent interventions), they both allowed me to see the goodness of fit for myself and this career path.

> During my senior internship at Storybook Farm, I was able to put all of my coursework, leadership experience, and skills into practice. Working with children facing adverse childhoods is so fulfilling. You can visibly see how much these children love Storybook and want to be there. There were a few

times I would interact with a child who I know is facing a lot of adversity, but there would be the look of pure joy on their face. That's why I chose the University for HDFS. For moments like these.

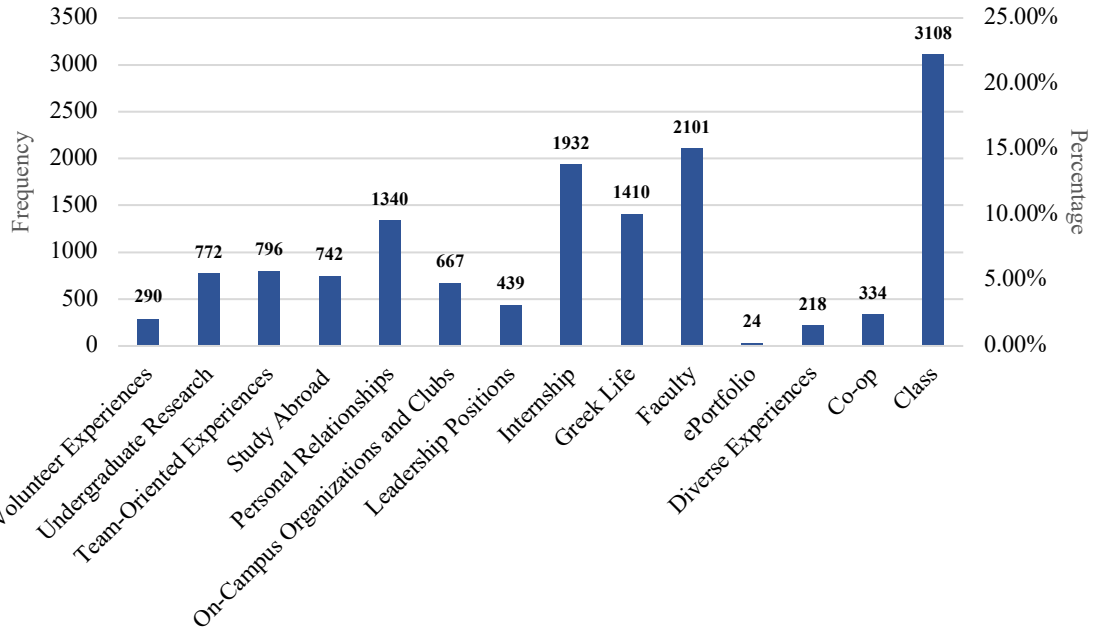


Figure 2. *Frequency and Percentage of Peak Learning Moment Responses per the Top 14 Themes at the Institution Level*

Crosstab Query: Analysis of Themes across Gender

Because students are not only diverse in their experiences but also in their individual backgrounds, it is important to consider how certain demographics may impact such experiences. In NVivo, a crosstab query can be used to provide an overview of response patterns, displaying counts for how each theme is distributed across certain attributes like term, college, gender, and so on (Jackson & Bazeley, 2019). For Study 1, the researchers were interested in conducting a crosstab query analyzing the themes across gender. This can be valuable information to help programs address demographic challenges with enrollment. For example, one of the programs focused on in this study is predominately male. If the crosstab query shows that females identify peak learning experiences different from males, the program can work to emphasize inclusion of experiences that are more appealing to females. The results of this query are shown below in Table 3. It is evident that "class" is the most frequently mentioned theme across both male (29%) and female (21%) students. Interestingly, when considering how this institution's HIPs are represented in the table, it appears that females are mentioning internships (17%), study abroad experiences (9%), and undergraduate research (7%) at a higher frequency as opposed to males mentioning co-ops (5%) at a higher frequency. The researchers found this information to be meaningful, yet it sparked additional questions. To elaborate, looking at gender is important, while it also might not give credence to how

students are represented in programs or colleges across campus as many colleges at this Southeastern university differ in their ratio of males to females. It may not necessarily be a gender effect shown below but a curricular effect. It is important to delve further into these differences as the University promotes these experiences for their students and hopes for them to be impactful across the student population.

Table 3. *Frequency of Responses per Theme Across Gender*

Themes	Gender = M	Gender = F	Response Count in Theme
Class	878 (29%)	746 (21%)	1624
Faculty	514 (17%)	620 (17%)	1134
Internship	395 (13%)	619 (17%)	1014
Greek Life	285 (9%)	485 (13%)	770
Personal Relationships	358 (12%)	360 (10%)	718
Study Abroad	117 (4%)	336 (9%)	453
Undergraduate Research	137 (5%)	234 (7%)	371
On-Campus Clubs & Organizations	166 (5%)	214 (6%)	380
Team-Oriented Experiences	218 (7%)	198 (6%)	416
Leadership Positions	80 (3%)	153 (4%)	233
Volunteer Experiences	37 (1%)	131 (4%)	168
Diverse Experiences	39 (1%)	83 (2%)	122
Co-op	156 (5%)	52 (1%)	208
ePortfolio	6 (.2%)	9 (.3%)	15
Total (unique)	3021	3597	6618

Note. A heatmap is displayed to highlight the frequency of responses per theme based on lowest (red) to highest (green).

Study 1 was useful in identifying themes at the institution level as it relates to students' peak learning moments, along with exploring how the themes are represented across genders. While within Study 1 the researchers began exploring other areas of interest, like themes across gender, there was a desire to delimit the sample population to examine themes at lower levels, like across departments, colleges, or programs. The purpose of this was twofold: (1) to identify if such differences in frequency of responses in themes across gender was more curricular in nature, and (2) to create more generalizability to other institutions hoping to conduct similar research but not having university-wide capacity.

Method - Study 2

Participants and Procedures

The purpose of Study 2 was to compare two colleges, a College of Architecture, Design, & Construction (CADC), and a College of Human Sciences (CHS), to see how their students responded to the peak learning moment prompt. The two colleges have similar levels of hands-on learning and design-focused mission. Being able to see similarities and differences in student responses and the frequency of responses within the existing 14 themes could inform practices within those colleges. Additionally, while the two have

similar missions, they differ in their gender distribution within the college and employ some different curricular structures, requirements, and opportunities for students. Because this study expanded upon Study 1, the researchers used the same sample of data from the campus engagement and experience survey, along with the same 14 themes. Again, N = 17,867 responses were included in the NVivo file and then delimited to student responses only from the two colleges.

Crosstab Query: College Comparison and College by Gender Comparison Analyses

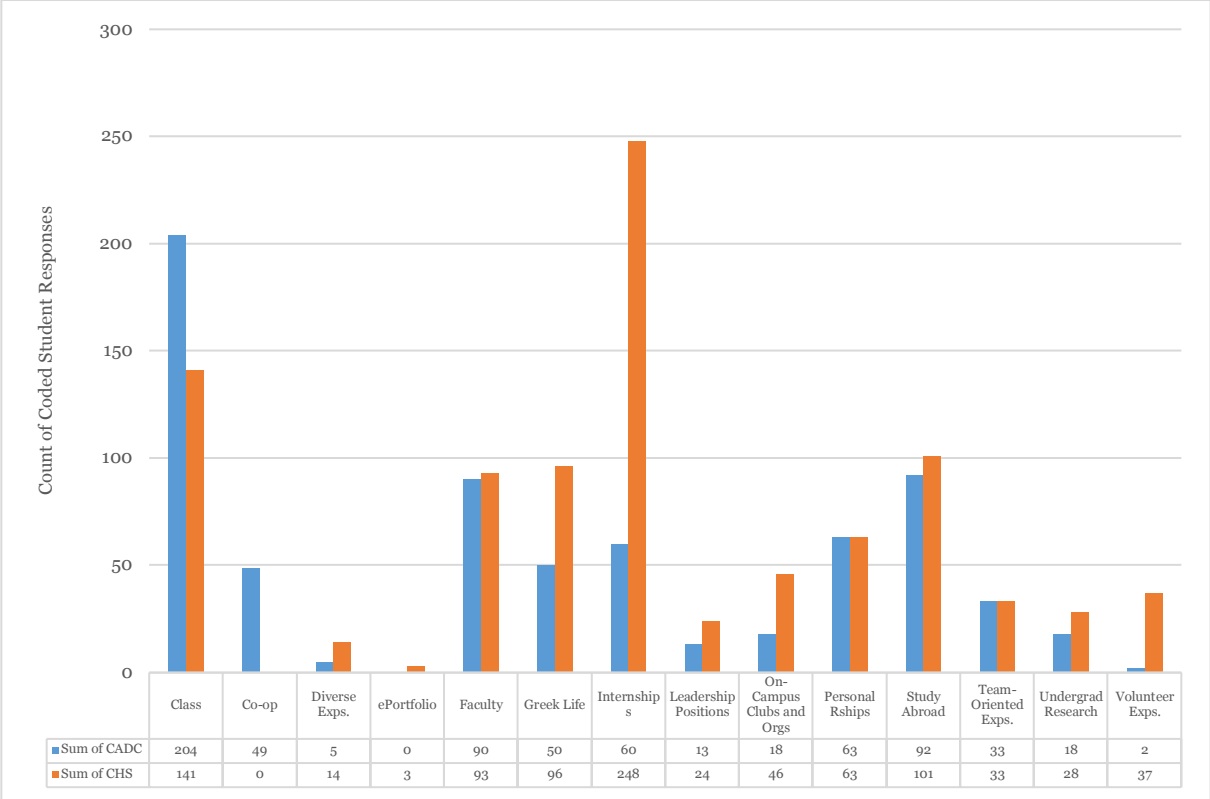
Two crosstab queries were run to address the research questions for Study 2, identifying the frequency of responses per theme between the two colleges, then exploring the differences by gender.

Results - Study 2

The initial results of Study 2 can be found in Figure 3, displaying the frequency of themes by college. Of note, 615 graduates of a CHS were represented in this data, with 93% of that total being female. With this gender distribution, the researchers did find similar trends between the top responses for females in a CHS when compared to the institution level across gender, with many female students mentioning internship experiences. In a CADC, there were 525 graduates represented in this data, with 34% of that total being female. The three themes mentioned with the highest frequency by both colleges are "class", "faculty", and "study abroad". When considering the ingrained experiences in each college, it makes sense that students in a CHS, who do not have co-op opportunities available to them, did not identify those experiences as peak learning moments. On the contrary, because of their required internship program, internships are mentioned most frequently by these students. Interestingly, students in a CADC have many studio-based classes that may heighten the mention of experiences by students about their transformative learning in the classroom.

The qualitative comments do provide additional depth to what quantitative data may have revealed regarding peak learning moments. One of the quotes about class specifically noted the importance of the "cohort" in a CADC. This approach puts a group of students in the same set of classes for a minimum of two years. Similarly, the quotes on internships demonstrate the importance of connecting classroom material to the real world in CHS.

The degree to which the co-op in a CADC and internships in a CHS was referenced is interesting. Co-ops in a CADC are available to only 7.5% of graduates. However, 7% of peak learning moment responses in a CADC noted that they had their peak experience through a co-op. This appears to imply that for students that did a co-op, it was typically their peak learning moment. The same does not hold true for internships, where less than 8% reported it as a peak experience in a CADC while it was available for 100% of the students at multiple times during their studies. In comparison, internships in a CHS were reported as peak experiences for 27% of their respondents. Internships in a CHS are required for multiple majors, and these are formalized, structured programs. Further study is warranted to understand why peak learning moments related to work experiences seem to occur more frequently in some settings than others.



	Class	Co-op	Diverse Exps.	ePortfolio	Faculty	Greek Life	Internships	Leadership Positions	On-Campus Clubs and Orgs	Personal Rships	Study Abroad	Team-Oriented Exps.	Undergrad Research	Volunteer Exps.
Sum of CADC	204	49	5	0	90	50	60	13	18	63	92	33	18	2
Sum of CHS	141	0	14	3	93	96	248	24	46	63	101	33	28	37

Notes. The trendlines represent the moving averages for both colleges across themes. CADC= College of Architecture, Design & Construction. CHS = College of Human Sciences.

Figure 3. *Frequency of Top 14 Themes and Comparison of Colleges*

The researchers hoped to further examine if the gender differences found in Study 1 at the institution level were in fact attributed to gender or instead because of curricular differences. References to class experiences differ by college and by gender which influences the institutional breakdown. Importantly, within a CADC, there were consistently high percentages of mentions about class experiences for both genders (29% of males and 35% of females). This leads the researchers to believe that the results of the institutional analyses by gender may have been impacted more by the college the students are in, which highlights curricular differences.

Discussion

The results of Studies 1 and 2 provide insight as to what types of transformative peak learning moments are taking place during college (within a single sample). The 14 identified themes appear to encompass a wide range of student experiences, including more nuanced ones like relationships with faculty or peers and team-oriented learning. According to the Wabash National Study, there are "three good practices that promote student learning: academic challenge and high expectations, diversity experiences, and good teaching/high-quality interactions with educators" (Goodman et al., 2011, p. 4). These practices seem to align well with the top two themes of "class" and "faculty", which supports the University's goal to foster experiences that researchers identify as best practices for learning.

The top two peak learning moments of "class" and "faculty" emphasize the importance of the educational experience in developing peak learning moments. Students who mentioned key words related to "class" highlighted challenging projects that developed time management and design skills, knowledge gained in the classroom, and having the chance to apply their expertise in studio settings. Students who mentioned key words related to "faculty" highlighted how their professors supported them, provided guidance, and gave helpful critiques that improved their skillsets. As institutions focus on the "student experience" and direct funding to initiatives that support that experience, they would be well-served to realize that peak learning moment data indicates the most common ones are associated with education and not with other out-of-class activities. Engaging classrooms, passionate professors, and mentors challenge students deeply and challenge them to think critically and develop skills necessary for lifelong learning. Connections forged with faculty and the wisdom gained from their expertise continue to shape and influence young minds positively. These interactions create transformative experiences for students and prepare them for impactful lives.

Regarding the crosstab query analyses for student attributes, the researchers found that the differences in frequency of responses by theme may be less affected by gender and more by college. This is beneficial information to have, as there are more ways institutions can improve curricular experiences for their students. For example, at this Southeastern university, it would be fruitful for the colleges to learn from one another in making curricular changes. A CHS could learn more from a CADC about what they do to make such impactful "class" experiences. A CADC might be able to learn more from a CHS about how their internships are so impactful.

Limitations

There are limitations worth noting. As previously mentioned, the researchers prompted students to respond with a transformative peak learning moment. While the prompt was expected to engage students to think about meaningful, memorable experiences, students had to define this for themselves. This could present slight error in the data as they may have not interpreted it as the researchers intended. And in relation to the subjective nature of this question, using self-reported data also lends itself to potential error. As is the case with most survey-based research, one limitation to this research is the bias that comes from collecting self-reported data. Indeed, this study relies on a graduation survey where students are asked to report their peak learning, at a time point that may be long after the experience has concluded. The nature of this survey is also low-stakes, which may not provide a lot of buy-in from students to respond with the best insights. However, the research benefits from having near census-level reporting from graduating seniors and allows for an analysis that explores trends and provides value through the review of themes rather than specific details of individual responses. In addition, the researchers expect that this data adds value by being jointly considered with other institutional and learning outcome data. And finally, the purpose of collecting these responses was to collect student perceptions, without locking them into responding about moments that were suggested to them.

It is important to note that because the data includes semesters of students who were impacted by the COVID-19 pandemic, certain experiences for student engagement may have been limited throughout this period of data collection (i.e., study abroad). The

lasting effects of the pandemic are expected to influence the educational landscape for the foreseeable future. The researchers believe that while certain experiences may have been limited or decreased during some of these academic semesters, the students were still prompted to consider their peak learning moments across their entire undergraduate journey. Therefore, the students can still recall other impactful experiences they may have had the opportunity to participate in, and their responses reflect the current state of higher education. Additionally, there was no pre-COVID data to compare to this, which would be worth exploring in a future study now being multiple years out of the height of the pandemic. The researchers do not expect that the themes would change drastically, but that an increase in certain experiences (e.g. study abroad) may shift the frequency or prevalence of certain themes.

Generalizability

Despite these limitations, this research has considerable generalizability to other institutions. The University has an ongoing and well-established data collection process that explores student experiences and success outcomes which other institutions have the opportunity to do as well, even if on a smaller scale. Institutions can collect qualitative data, whether through course-level surveying, broader surveys upon graduation, or focus groups with samples of students. And yet, institutions often collect this data only to let it remain underutilized because of how unwieldly it can be. While the researchers had a paid-for qualitative analysis software, there are free tools and AI-based software that make qualitative data analysis increasingly more attainable. Of note, a manual coding process is entirely possible depending on the size of the dataset and clearly used in this research as well to refine the analyses as best as possible. This research expands our understanding of how students' perceptions about their learning can be captured and analyzed in concert with more direct and quantitative metrics.

The qualitative data can and should be paired with other data to tell a more complete story of how, when, and where students are learning. And the 14 themes identified in this research appear to be generalizable across other research institutions that have very applied experiences built into their curriculum and degree offerings. Generally, these themes may not be a perfect fit for other institutions and their student body with different majors, required experiences, etc. The purpose of the present research study is not to state that these are the top themes across all institutions, but to provide a roadmap of how an institution explored peak learning moments in their student population of graduates, aiming to use this data to inform institutional practices and approaches to further enhance student learning and engagement.

Future Research

This research can inform both longitudinal studies and broader institutional practices. First, in relation to longitudinal research, it is worth identifying peak learning moments for both incoming students and alumni. The incoming students could provide peak learning moments from their formative years before college, and alumni could provide insight into whether their originally stated peak learning moments remained post-graduation. The latter would provide insight into how their peak learning impacted their success in their post-college endeavors.

There is opportunity to introduce the concept of "peak learning moments" earlier in a student's undergraduate journey. This would help students recognize, seek out, and reflect on these experiences throughout their education in ways that align with the University's desired approach. A key aspect of this University's mission is to not only provide accessibility to these experiences, but also to prepare students to articulate their learning for future success. It would be valuable to research whether additional preparation during peak learning experiences—specifically focused on articulating these experiences—enhances students' ability to achieve successful outcomes compared to those who do not receive this targeted support.

Conclusion

Understanding student learning experiences through qualitative data, like the peak learning moment question studied in this research, offers complementary insights that are crucial for assessment of the educational process. In addition to depth and nuance of data, the individual perspectives provided here offer a window into the world of each student allowing an institution to tailor instruction and support individual student needs. In addition, the comments provide information on how and why students learn and shed light on the emotional and motivational factors that add to learning. This in and of itself demonstrates significant value added by exploring qualitative data. The findings of this research emphasize the value of understanding student learning and respective experiences with approaches that may provide greater depth of understanding when viewed in conjunction with quantitative research.

References

Blaney, J., Filer, K., & Lyon, J. (2014). Assessing high impact practices using NVivo: An automated approach to analyzing student reflections for program improvement. *Research & Practice in Assessment, 9*(1), 97-100. https://eric.ed.gov/?id=EJ1062704

Goodman, K. M., Magolda, M. B., Seifert, T. A., & King, P. M. (2011). Good practices for student learning: Mixed-method evidence from the Wabash National Study. *American College Personnel Association and Wiley Periodicals, Inc., 16*(1), 2-9. https://doi.org/10.1002/abc.20048

Harper, S. R., & Kuh, G. D. (2007). Myths and misconceptions about using qualitative methods in assessment. *New Directions for Institutional Research, 2007*(136), 5-14. https://doi.org/10.1002/ir.227

Heath, C., & Heath, D. (2017). The power of moments: Why certain experiences have extraordinary impact. Simon & Shuster.

Jackson, K., Bazeley, P., & Bazeley, P. (2019). *Qualitative data analysis with NVivo*. Sage.

Kramer, P., Ideishi, R., Kearney, P., Cohen, M., Ames, J., Shea, G., Schemm, R., & Blumberg, P. (2007). Achieving curricular themes through learner-centered teaching. *Occupational Therapy in Health Care, 21*, 185–198. https://doi.org/10.1080/J003v21n01_14

Kuh, G. D. (2008). High-impact educational practices: What they are, who has access to them, and why they matter. Washington D.C. AAC&U.

Kuh, G. D., & O'Donnell, K. (2013). Ensuring quality and taking high-impact practices to scale. *Peer Review, 15*(2), 32. https://link.gale.com/apps/doc/A339018909/AONE?u=anon~eb93e534&sid=googleScholar&xid=c45ecfdf

Nelson, T. E., Shoup, R., Kuh, G. D., & Schwartz, M. J. (2008). The effects of discipline on deep approaches to Student learning and college outcomes. *Research in Higher Education, 49*(6), 469–494. https://doi.org/10.1007/s11162-008-9088-5

Newhart, D. W. (2015). To learn more about learning: The value-added role of qualitative approaches to assessment. *Research & Practice in Assessment, 10*, 5-11. https://www.rpajournal.com/dev/wp-content/uploads/2015/06/A1.pdf

Patton, M. Q. (2002). Qualitative Research and Evaluation Methods (3rd ed.) Thousand Oaks, CA: Sage Publications.

Peck, K., Furze, J., Black, L., Flecky, K., & Nebel, A. (2010). Interprofessional collaboration and social responsibility: Utilizing community engagement to assess faculty and student perception. *International Journal of Interdisciplinary Social Sciences, 5*(8), 205–221. https://doi.org/10.18848/1833-1882/CGP/v05i08/51853

Qualitative Research Methods in Program Evaluation: Considerations for Federal Staff. (May 2016). Office of Data, Analysis, Research & Evaluation Administration on Children, Youth & Families. https://acf.gov/sites/default/files/documents/acyf/qualitative_research_methods_in_program_evaluation.pdf

Waterhouse, P., Creely, E., & Southcott, J. (2021). Peak moments: A teacher-educator reflects (with colleagues) on the importance of heightened moments of teaching and learning. *Teaching and Teacher Education, 99*, 1-10. https://doi.org/10.1016/j.tate.2020.103275

Welsh, E. (2002). Dealing with data: Using NVivo in the qualitative data analysis process. In *Forum qualitative sozialforschung/Forum: qualitative social research, 3*(2). http://nbn-resolving.de/urn:nbn:de:0114-fqs0202260.

The Evidence of Learning and Impact Framework: A Delphi Study

Authors:

Kirstin Moreno, M.S.Ed., Ph.D.
Oregon Health & Science University

Sarah Jacobs, M.Ed.
Clark College

Constance Tucker, M.A., Ph.D.
Oregon Health & Science University

ABSTRACT

Learning outcomes frameworks are useful in course, program, and institutional assessment as well as continuing education or professional development contexts and help ensure that different aspects of learning are addressed. This article describes a Delphi study conducted to iterate and improve on the authors' novel Evidence of Learning and Impact Framework using assessment experts' feedback. This new framework is useful broadly within adult and higher education and uniquely incorporates an emphasis on attending to the impact of the learning on the learner, the impact of the learning on others, and encourages the use of equity lenses when examining learning. Niederberger and Spranger (2020) encourage more transparency about Delphi techniques used in scholarship, so the authors also provide many details about the Delphi process used. We hope that the Evidence Framework will challenge educators to think differently, more broadly, and more deeply about the kinds of learning they foster and assess.

Correspondence E-mail: kirstin@morenos.name
Keywords: Learning Outcomes Framework; Delphi Technique; Educational Assessment; Equity-Centered Assessment; Higher Education Assessment

Orienting to Our Previous Work

Since 2020, the authors have been engaged in a multi-step process to design a learning outcomes framework that better fits our institution's needs, and that also aligns with the assessment culture we want to promote in light of calls for incorporating antiracist and equity-focused approaches into teaching and learning (Alegría et al., 2024; Green & Malcolm, 2023; Henning et al., 2023; LaFever, 2016; Twyman-Ghoshal & Carkin Lacorazza, 2021). Through a scoping review and an analysis of existing frameworks, we created a crosswalk focused on the aspects of learning the various frameworks addressed (Tucker et al., 2024). Because we are assessment professionals in an academic health center, we had been using a combination of Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001) and Moore's Expanded CME Framework (Moore et al., 2009), which is typically used in continuing medical education and includes the following 7 levels: Participation, Satisfaction, Declarative Knowledge, Procedural Knowledge, Competence, Performance, Patient Health, and Community Health. It became clear that Bloom's is not really a learning outcomes framework, and while Moore's excels at helping us attend to the different levels of learning that can occur in clinical education specifically, it was not a good fit for some schools and programs in our institution such as public health, basic sciences, and health care management.

Our assessment frameworks crosswalk (Tucker et al., 2024), paired with our lived experiences using Bloom's Taxonomy and Moore's CME Framework, helped us determine what was useful from the existing frameworks and what might be missing. From this reflection and analysis, we created the initial *Evidence of Learning and Impact Framework*. The *Evidence Framework* expands beyond Bloom's narrow focus on knowledge and application to consider learning more broadly. From Moore's CME Framework, we adapted the highest levels (Patient Health and Community Health) into our Impact on Others level, making them relevant across disciplines. We also reoriented Moore's lowest levels (Participation and Satisfaction), which felt inconsistent with higher-order outcomes, into equity lenses that can be applied across all levels. This innovation is critical because as faculty plan assessments, they must also examine whether different groups of learners experience inequities, and the framework's design ensures equity is considered early and often. Additionally, our crosswalk with other frameworks revealed overlooked domains such as the "human dimension," "caring," "learning how to learn," "professional identity," and "empathy." We consolidated these into Impact of Learning on Self, capturing personal, relational, emotional, and reflective aspects of learning. Throughout the framework's development, we emphasized assessing "the impact of learners' learning," which is reflected in its final structure. We suggest that the development of this framework is novel because it is a synthesis of many frameworks and is able to be used across many disciplines and contexts.

In addition, one of the more unique features of our *Framework* are the framing questions we developed for each level, outlining concepts both for faculty and for students to consider. These framing questions encompass broader questions such as learners' shifting and growing as they learn, the extent to which they can combine knowledge and skills from different domains, and ways someone who is at the beginning stages of being an expert in their field can think about their potential for impact on others. Generally, we want to encourage thought about the end goals of learning in deep and multifaceted ways, but also present a framework that is streamlined and easy to use.

We found in our scoping review (Tucker et al., 2024) that some existing assessment frameworks are limited in usefulness because they are discipline-specific or only applicable to a specific learning environment which we attempted to avoid. Because the framework was created in the context of work in academic health, we wanted to bring in a group of assessment experts to provide feedback so our framework has wider applicability to multiple higher education settings. For easy reference, Figure 1 shows the graphic of the *Evidence Framework* as we initially conceived it. This paper outlines how we used a Delphi method to iteratively improve our *Framework*.

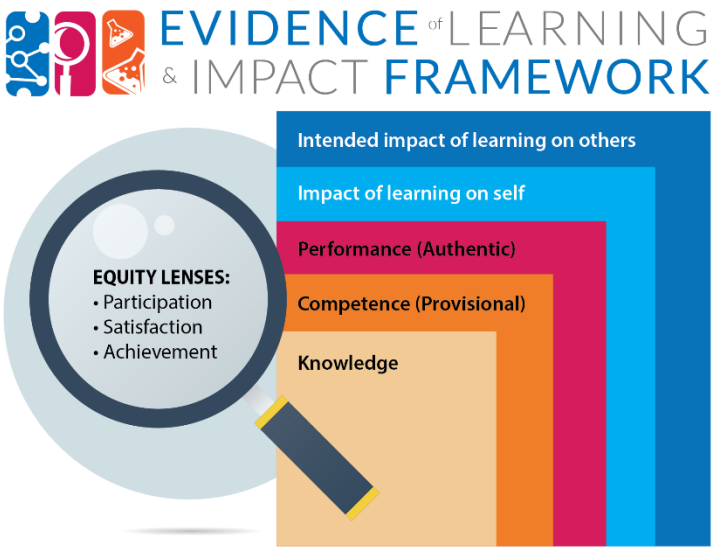


Figure 1. *Initial Version of Evidence of Learning and Impact Framework*

Delphi Approach to Refining Our Framework

Delphi techniques, which are typically used in the technical and natural sciences, are also used in health research to arrive at consensus (Niederberger & Spranger, 2020) and in educational research for a variety of uses including curriculum planning and setting university-wide educational objectives and learning goals (Green, 2014). The assessment experts who participated in our Delphi study provided anonymous feedback using a standardized questionnaire that was adapted over three iterative rounds, allowing us to revise our framework in between each round and reach consensus, which are hallmarks of studies using Delphi techniques (Barrett & Heale, 2020 Diamond et al., 2014; Nasa et al., 2021).

Niederberger and Spranger's 2020 meta-analysis of studies using the Delphi techniques found that it can be difficult to accurately judge the validity of outcomes from a Delphi study because precise details about the relatively complex Delphi technique are not always disclosed. The aim of this article is to share a new higher education learning outcomes framework we have developed while providing more nuance than the norm about the Delphi process we used to refine our framework. Drawing on Niederberger and Spranger (2020), we outline the following key areas for improvement:

1. Clear definition for consensus, along with a discussion of possible factors that may have influenced whether consensus was reached.
2. Inclusion of specific information around the modifications made to the method to help elucidate epistemic objectives and authors' thought processes.
3. Specific characteristics used to determine "who counts" as an expert and the impact of including "lifeworldly" expertise on the Delphi process.
4. Careful thought about the implications of having a small number of experts (in the "low double digits") on the consensus that is reached.
5. Attention to conveying the process for developing and monitoring the questions used in the Delphi surveys which collect experts' thoughts and scores.

We hope that our efforts to share more transparently the decisions we made about our framework revisions will demonstrate the rigor with which our framework was developed and also help higher education colleagues feel more confident in using Delphi techniques themselves.

Delphi Study Team

Before jumping into the details of both our new *Evidence Framework* and the related Delphi study, it is important to understand the study team's positionality. We are three women, two white and one Black, who work in higher education administration in the western United States. We are not health care professionals, but have backgrounds in different aspects of Education (educational psychology, student affairs, K-12, faculty development, and educational linguistics), and are responsible for guiding institutional assessment processes. We have experience in both qualitative and quantitative social science research, but none of us had undertaken a Delphi study before this project. We have strong orientations toward assessment as a tool to shift educational culture/practice and also toward qualitative, social constructivist approaches to thinking about assessment. Among us, we have some previous experience teaching for-credit courses. The institution we work for has publicly and repeatedly set a goal of us working toward becoming an antiracist organization and we have one of the largest groups of Native American students and faculty among academic health centers, though the overall student and employee population is predominantly white.

Preparing for the Delphi Study

The initial tasks we completed before the first round of Delphi are shown in Table 1. A few of these initial tasks merit some more description: determining our criteria for expertise; deciding what our survey questions would be; and our initial attempt at setting a standard for what would count as consensus.

Determining Expertise

Niederberger and Spranger (2020) discuss the ways the concept of expertise has been broadened in the selection of Delphi panels, including placing more value on "lifeworldy" experience (p.3). They cite studies which demonstrate that a more inclusive understanding of expertise leads to more creative and innovative results. In our study, we wanted to be inclusive of the many ways assessment experts gain expertise beyond a formal educational experience, as recommended by Niederberger and Spranger (2020). To that end, in the email we sent out to recruit higher education faculty and staff as

participants we only indicated they needed to be "engaged in assessment in higher education" to participate and did not select for specific titles or roles. Then in the

Table 1. *Tasks Completed Before Delphi Study Began*

Logistics Tasks	Content Tasks
Created list of contacts to invite to participate in the study, including personal contacts and large, international distribution lists related to assessment	Submitted and obtained IRB approval (IRB#: STUDY00024481; found to not be human subject research)
Estimated timeline for Delphi study	Created initial framework graphic in graphic design software
Determined amount and type of incentives for participating, to be distributed ASAP after each round closed	Developed the questions for experts to score, scoring scale, and qualitative, open-ended questions in Qualtrics (Provo, UT) survey
Determined criteria to show participant 'expertise' and developed demographic portion of survey for Delphi participants	Determined how to communicate to the experts about the motivations behind creating a new learning outcomes framework, the background, the framework itself, examples, and framing questions
Decided criteria for reaching 'consensus' and on which question(s) (once we knew what our survey questions would be)	After considering options including recording videos for the experts, we decided to create a slide deck with the narrative directly on the slides along with graphics. Examples of these slides are shared below for each Round

demographics section of the survey itself, we used responses to these questions shown in Table 2 to gauge expertise.

We had only one respondent who indicated they had been working in assessment less than one year. When we looked at the quality of their qualitative feedback in the survey, it was clear they were still novice-level in assessment and their comments lacked useful specificity, so we did not incorporate their feedback nor scores into our analysis. The wording of the second question around expertise was an effort to allow for a more equity-informed way of conceiving of "expertise". We received all "Yes" responses on this question. We received a significant amount of qualitative feedback throughout our Delphi rounds, and the remaining, included participants were able to speak to the questions we asked with insight and helpfully push us where they disagreed. We did not exclude any other participants from the Delphi study. Potential biases in the demographics of our participants are discussed in the Limitations section.

Table 2. *Expertise Questions*

Question	Options
How long have you been engaged in assessment work?	• Less than one year • 1–4 years • 5–9 years • 10+ years
Criteria to give feedback on the Evidence of Learning and Impact framework	• You work/have worked in higher education assessment. • You have credibility with the assessment community through either traditional roles in assessment (such as national and regional organizations, publications, research presentations at national conferences, etc.) and/or non-traditional roles (such as being the "go-to" assessment expert at your institution, have implemented assessment innovations, etc.)
Do you meet these criteria? (Answering "No" will skip you to the end of the survey and no further feedback will be needed.)	• Yes • No

Survey Questions to Collect Expert Feedback

Throughout the three Delphi rounds, we asked our participants to respond to both qualitative and quantitative questions in several categories: reactions to the framework as a whole, reactions to the graphic that represents the framework, and reactions to the individual levels of the framework. Being new to this process, we weren't sure at first if we would use the exact same questions for each round or not, or how much leeway we had to change the questions if we found they weren't working well. We decided to approach the Delphi consensus process with a constructivist research perspective. If our survey questions did not elicit the data we needed, we determined not to hesitate to revise them, of course making sure to note and disclose the details of that change. This allowed us to better hone in on the information we needed each round to further improve our *Evidence Framework*. All survey questions for the three rounds are in Appendices A, B, and C, and you will note that we pivot to some extent in each round to different topics, and also tweaked the wording of some of our quantitative questions.

Initial Consensus Setting

Throughout the Delphi study process, we were working from the assumption that the quantitative Likert scale questions was where we would calculate our official "consensus" to know when we had reached the end and our framework was sufficiently revised, and that the qualitative responses would provide the specific insight needed to make those revisions. Before starting Round 1, we agreed on the following to indicate consensus had been reached: on a Likert scale of 1-6, 90% of respondents would rate the final question for each of five *Evidence Framework* levels, which was "This level requires no editing/updating, or is good as written", as either Strongly Agree or Agree. We also

provided three other statements from the experts to rate about each level which we found very helpful in thinking about consensus: "This level is distinct from the other levels"; "the framing question helped me understand this level"; and "I feel confident I could map assessments appropriately to this level", but we did not initially include those three additional scores for each level in what we thought would be our formal determination of consensus.

The Iterative Delphi Study Rounds

Materials that we sent to the participants each round included a robust slide deck with graphics and explanations of our goals and ways of thinking about our framework, a link to a survey for them to respond to and a list of the survey questions, an information sheet summarizing the study, and to assure participant anonymity, a link to a separate survey to request their gift card incentive. We followed up with one reminder after a week and gave the participants about two weeks to provide feedback. Below, for each of the three rounds we will share examples of the materials we sent out, how many respondents we had, a snapshot of quantitative and qualitative feedback received and status of consensus, insight into what we were thinking as researchers, and a summary of the changes made in response to the experts' feedback.

Round 1

Demographics of Initial Set of Experts

For Round 1, from our recruitment list of contacts and national listservs, we had 39 expert participants provide feedback on our *Evidence Framework*. Additional demographic information on our experts is in Figure 2. Note that our gender data was corrupted, so we excluded it from the information. It is worth noting that only one Delphi participant was affiliated with a health professions institution, and the respondents represented many types of higher education institutions.

Table 3. *Expert Participant Demographics*

Formal Training in Assessment	
Yes	74%
No	26%

Years of Assessment Experience	
1- 4 years	8%
5-9 years	26%
10+ years	66%

Race/Origin	
White	80%
Asian	8%
Prefer not to answer	3%
Hispanic, Latino, or Spanish origin	3%
Mixed	3%
Black or African American	3%

Types of Institutions Represented	Count
Community, Junior, and Technical Colleges	13
Faith-Based Colleges	8
For-Profit Institution	4
Four-Year Colleges and Universities	39
Health Professions Institutions	1
International Colleges	1
Liberal Arts College/University	14
Minority-Serving Institutions (tribal colleges, HBCU, HSI, etc.)	8
Special Focus Institution (arts/design/music, military, law, etc.)	2

Age Range	
Under 30	5%
30-49	54%
50-69	41%
70 and above	0%

Materials Reviewed in Round 1

The slide deck the experts read through for Round 1 included study objectives and timeline; background information about why we created the Evidence of Learning and Impact Framework; an introduction to the new framework; practical examples of how it could be used; an explanation of the equity lenses and how we reframed those from Moore's Satisfaction and Participation levels; and references and resources.

The following two slides formed the primary material we were looking for feedback on. You can see in Figure 3 that we had adjusted the coloring of the Framework (but none of the wording) for the initial round. In Figure 4 we provided one example and 1-2 framing questions for each of the framework levels. You can also see the way we communicated with our experts in the wording on the slides. Appendix A contains the complete set of questions our experts responded to in Round 1 after reviewing the slides.

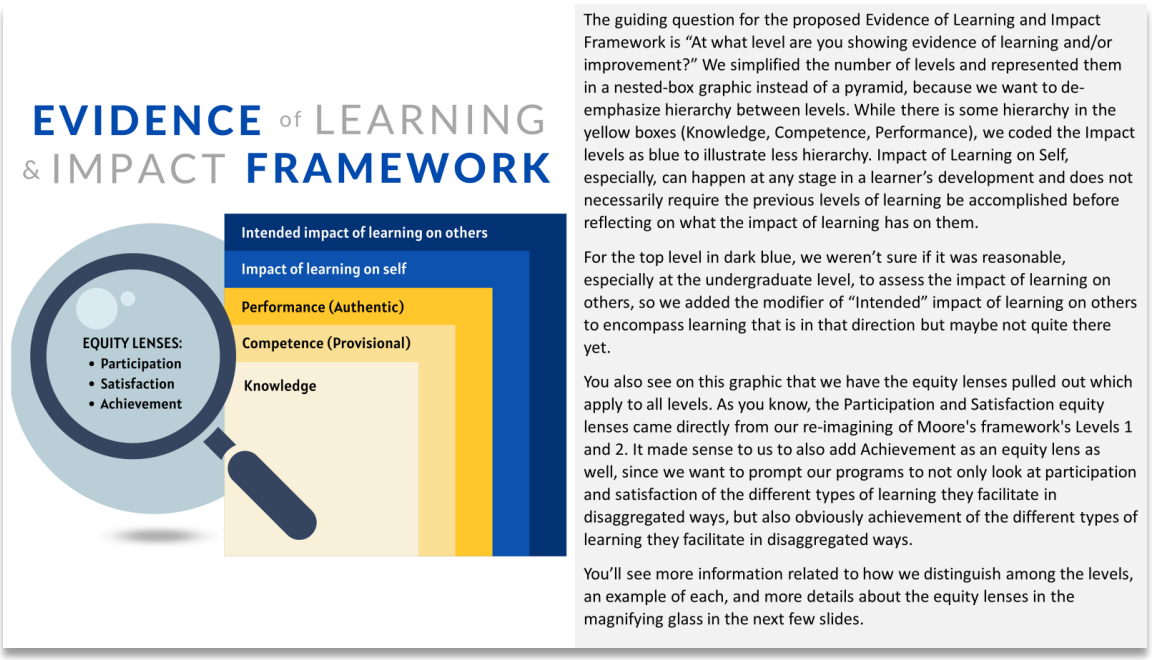


Figure 3. *Round 1 Evidence Framework Graphic*

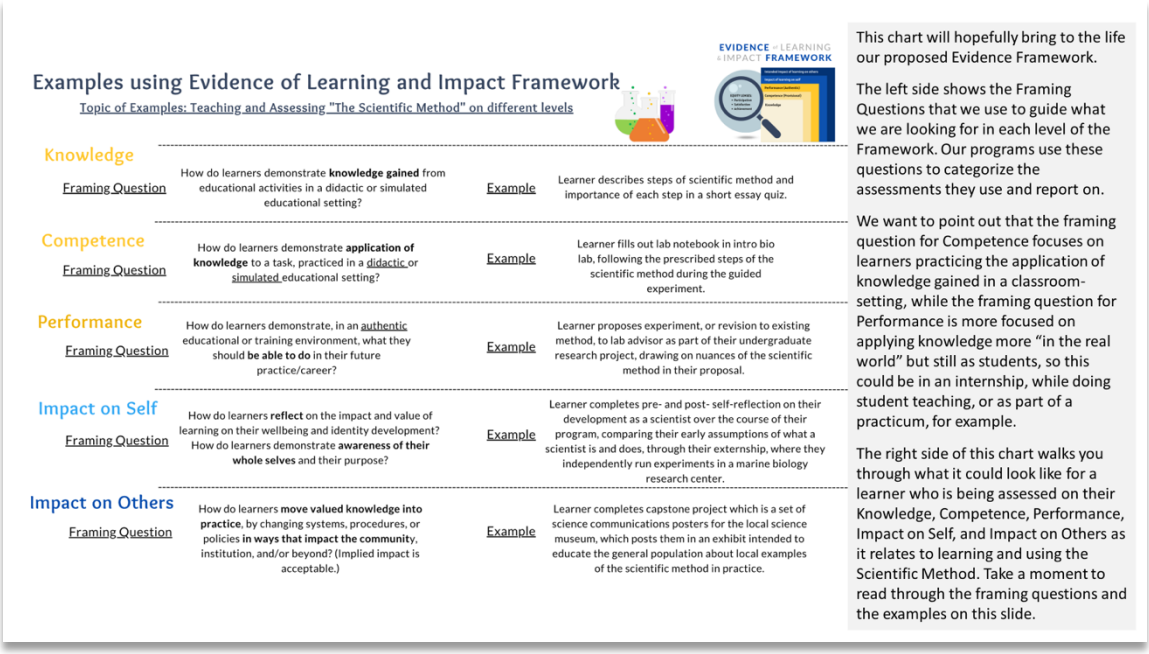


Figure 4. *Round 1 Examples and Framing Questions*

Table 4. *Quantitative Results from Round 1*

Question Text	Response Count (N)	High Agreement (%) (Strongly Agree, Agree)	Unclear (%) (Somewhat Agree, Somewhat Disagree)	Low Agreement (%) (Disagree, Strongly Disagree)
Framework as a Whole				
This framework would be easy to apply to a variety of disciplines.	39	89.74%	10.26%	0.00%
This framework would be easy to apply to interprofessional education (defined as learning with, from, and about each other)	39	94.87%	5.13%	0.00%
This framework makes space for diverse ways of knowing.	39	89.74%	10.26%	0.00%
This framework centers the learner in assessment (assessment for learning vs. of learning)	39	79.49%	20.51%	0.00%
Knowledge				
This level is distinct from the other levels.	35	85.71%	8.57%	5.71%
The framing question helped me understand this level.	35	80.00%	20.00%	0.00%
I feel confident I could map assessments appropriately to this level.	35	85.71%	11.43%	2.86%
This level requires no editing/updating, or is good as written.	35	48.57%	45.71%	5.71%
Competence (Provisional)				
This level is distinct from the other levels.	35	77.14%	17.14%	5.71%
The framing question helped me understand this level.	35	85.71%	11.43%	2.86%
I feel confident I could map assessments appropriately to this level.	35	80.00%	11.43%	8.57%
This level requires no editing/updating, or is good as written.	35	51.43%	31.43%	17.14%
Performance (Authentic)				
This level is distinct from the other levels.	35	74.29%	22.86%	2.86%
The framing question helped me understand this level.	35	77.14%	20.00%	2.86%
I feel confident I could map assessments appropriately to this level.	35	62.86%	34.29%	2.86%
This level requires no editing/updating, or is good as written.	35	31.43%	51.43%	17.14%
Impact of Learning on Self				
This level is distinct from the other levels.	35	91.43%	2.86%	5.71%
The framing question helped me understand this level.	35	71.43%	22.86%	5.71%
I feel confident I could map assessments appropriately to this level.	35	77.14%	14.29%	8.57%
This level requires no editing/updating, or is good as written.	35	42.86%	34.29%	22.86%
Intended Impact of Learning on Others				
This level is distinct from the other levels.	35	91.43%	5.71%	2.86%
The framing question helped me understand this level.	35	77.14%	20.00%	2.86%
I feel confident I could map assessments appropriately to this level.	35	57.14%	40.00%	2.86%
This level requires no editing/updating, or is good as written.	35	34.29%	45.71%	20.00%

After Round 1, three of our overall Likert-scale questions had achieved consensus: "This framework would be easy to apply to interprofessional education", "This framework would be easy to apply to a variety of disciplines", and "This framework makes space for diverse ways of knowing". Instead of asking those questions again in Round 2, we added two new overall questions to further dig into improving the framework: "The *Evidence Framework* graphic is an effective way to visualize the framework" and "Do you agree with the decision to move participation and satisfaction assessment methods out of the framework in order to use them as equity lenses?"

Qualitative Results from Round 1

Two of the authors took the qualitative input from the experts and, using Word, organized the data into "representative quotes about strengths" for each level, "quotes about weaknesses/possible action items", and "recurring themes" for each level. We paid attention to the quotes about strengths because wanted to make sure that we knew when to protect portions of our *Framework* from significant changes. Because we were primarily looking for specific action items that came out of the data that we could consider and use to make concrete changes to the framework, the examples, and the framing questions in a short timeframe, we did not more formally code the data in software. We used the Word doc as a comprehensive potential to-do list which we organized thematically and which framed our discussions as we met over the next couple of weeks to prepare for Round 2 and we documented our decisions about the feedback in the Word document itself to be sure we were considering all of the comments and on the same page as a research team. During these conversations we discussed the contradictions we saw in the experts' responses, the ways they aligned, and tried to reconcile these tensions in our revisions. We re-thought the boundaries between levels, the wording of framing questions and examples, and documented rationale behind our changes to incorporate into the next Round's slide deck. In total for the Delphi study we analyzed approximately 45 single-spaced pages of comments from our experts, which led to a substantial understanding of our participants' perspectives on our work.

During Round 1, qualitative data on the Knowledge level indicated that Knowledge was similar to other frameworks and fairly easily understood, basic, foundational, and clear. However, we needed to include more examples from other disciplines. We were also encouraged to provide a more encompassing use of "knowledge" and determine whether or not skills were included in Knowledge or in another level. The phrase "knowledge of, about, how, and why", which we added to the framework graphic after this round, came directly from one of our expert participants. Some other frameworks separate knowledge into smaller buckets, distinguishing among declarative and procedural knowledge, for example, or knowledge you remember vs. knowledge you apply. We decided we wanted faculty to put more focus on the other levels and not worry so much about precisely which kind of knowledge they are fostering. Knowledge of has to do with general awareness of a topic, knowledge about has to do with cataloging facts, knowledge how is remembering the steps to doing something, and knowledge why speaks to understanding the context, when different knowledge is applicable, and impacts of using that knowledge. This isn't meant to be thorough conceptualization of knowledge, and there may be other parts of knowledge that align to this level as well.

Qualitative data on the Competence (Provisional) level indicated there was confusion regarding terms like "provisional", "didactic", "simulated", and not everyone was clear on how Competence was different from Performance.

> "Depending on the content or task, I can imagine it being difficult to distinguish between Competence and Performance levels. Both require application of knowledge to a task."

Qualitative data on the Performance (Authentic) level indicated the experts had mixed feelings about whether Performance is useful or too close to Competence as noted above. The experts also expressed that it might be hard for some (perhaps less experiential-education focused) disciplines to identify what is Performance in their realm.

Qualitative data on the Impact of Learning on Self level indicated fairly strong support for incorporating a focus on the impact of learning on the learner in our framework. We also heard that the experts didn't understand how well-being fit into this level, that the examples need to be altered to focus on the student and not the task, and that we should expand our framing questions to be broader.

> "I wish Impact on Self more explicitly recognized our students as whole people and more explicitly referenced a growth mindset in learning."

Qualitative data on the Intended Learning of Impact on Others level indicated mixed feedback on the use of the word "Intended" in the name of the level, and perhaps too much overlap among this level, Performance (Authentic), and Impact on Self.

> "Very few learners will ever reach Impact on Others with all content they learn… especially not in bachelor's level programs."

Qualitative data on the graphic itself led us to revise it for Round 2.

> "For the graphic: Less emphasis on hierarchy… thus more of a perception that the learning taking place at all levels is valued."

Insight into Authors' Conversations

As we read through the qualitative data, we discussed whether faculty of any individual course should feel compelled to align with all five levels of the *Evidence Framework*, which was an assumption that some of the experts' feedback seemed to make. We decided that, theoretically, any program should be able to align to all five levels, but that an instructor of a specific course should feel free to align to whichever ones made sense for their content and course goals.

We also decided that we needed to provide more examples and framing questions in Round 2 for experts to review and hopefully get a better understanding of what we meant by each level as we honed in on the differences among them. And we wanted to provide more explanation and examples related to our Equity Lenses in Round 2 because we realized we did not receive specific enough feedback on that part of our framework

from our first Delphi round. Part of this conversation was us also determining that there were perhaps more components we needed feedback on than we had initially realized.

Consensus After Round 1

A slide (Figure 4) was shared in Round 2's slide deck and summarizes the state of consensus and timeline.

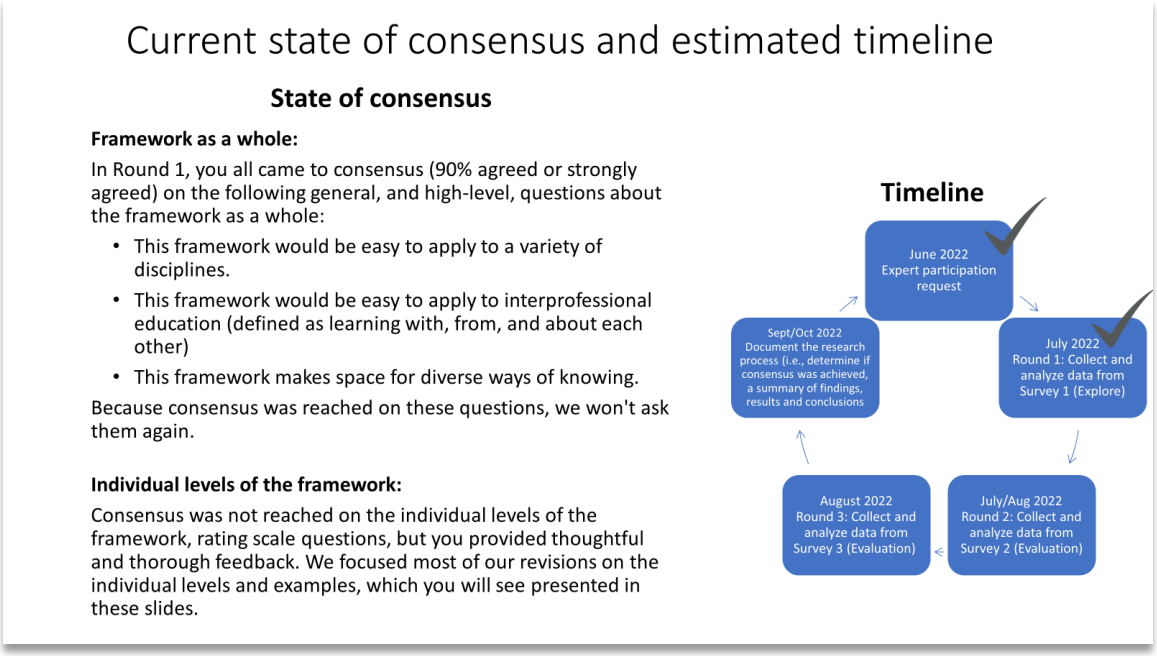


Figure 4. *State of Consensus after Round 1*

Round 2

Expert Participants Round 2

For Round 2, from the original set of 39 experts, 28 (72%) of them provided feedback on the revised *Evidence Framework* materials. No new experts were added to Round 2.

Materials Reviewed in Round 2

The slide deck the experts read through included a reminder of the crosswalk and other influences we were basing our framework on: a summary of the main changes we made, more in-depth exploration of the Equity Lenses, and for each individual level, a pair of slides that outlined additional framing questions, examples across the disciplines, and potential aligned assessment approaches. Examples of these are provided below.

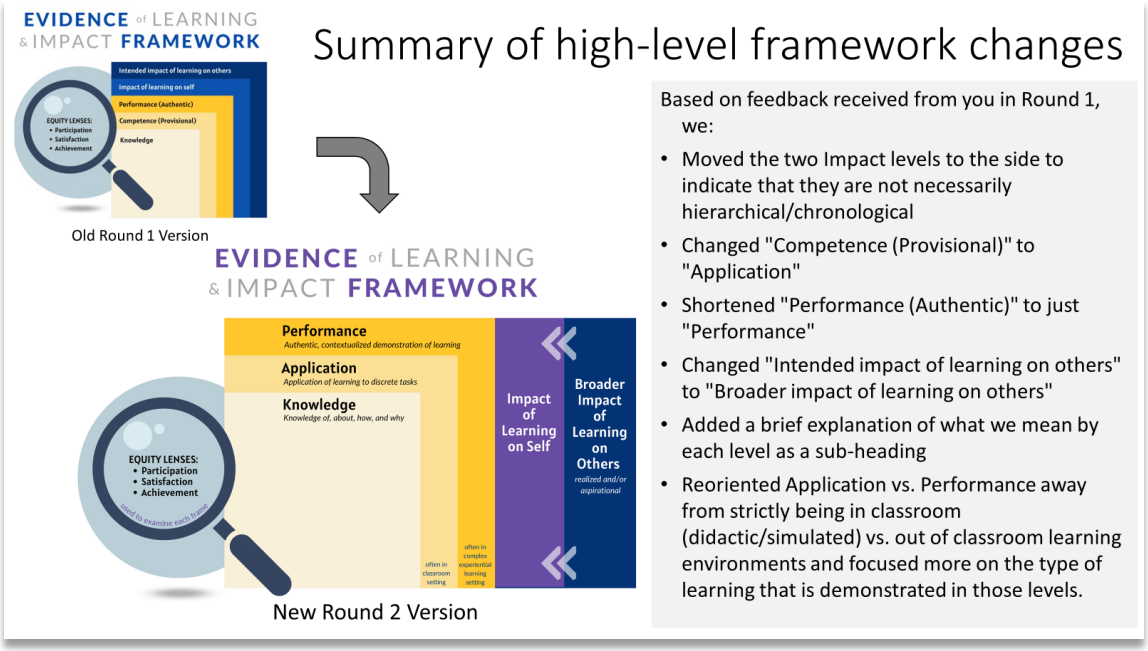


Figure 5. *Changes made after Round 1 to the Evidence Framework graphic*

Experts indicated in Round 1 that we needed to provide more clarity around why we moved Participation and Satisfaction out of the main structure of our previous (Moore's Outcomes Levels) framework, added Achievement, and re-imagined them all as lenses through which we examine equity. As we were creating the crosswalk (Tucker, Jacobs, and Moreno, 2024), we realized Participation and Satisfaction aren't evidence that learning is occurring, but rather they tell us who is there and whether they feel welcome/comfortable in the learning setting. The following two slides (Figures 6 and 7) were used to show how we defined Participation, Satisfaction, and Achievement Equity Lenses, the process by which our programs demonstrate the Equity Lenses portion of the framework, and a good example of what one program at our institution submitted for the equity lens question.

Finally, in the Round 2 materials, based on the qualitative feedback we received in Round 1, we provided some additional framing for reviewers to keep in mind when reading the set of slides on the individual levels (see figure 8).

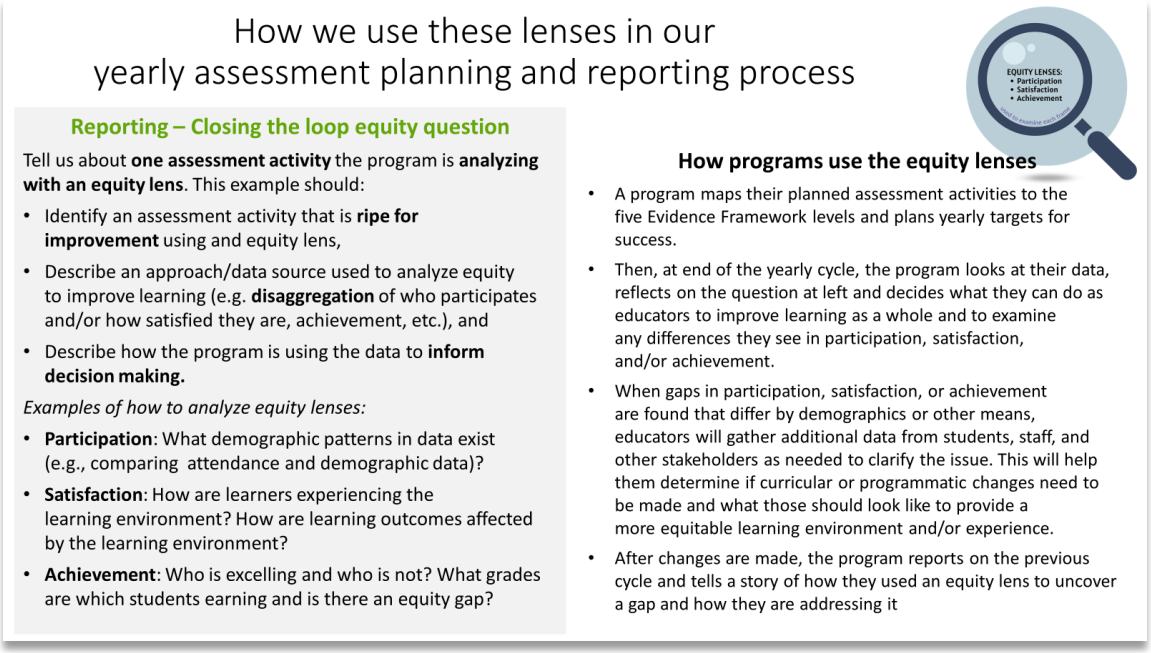


Figure 6. *Round 2 Equity Lenses*

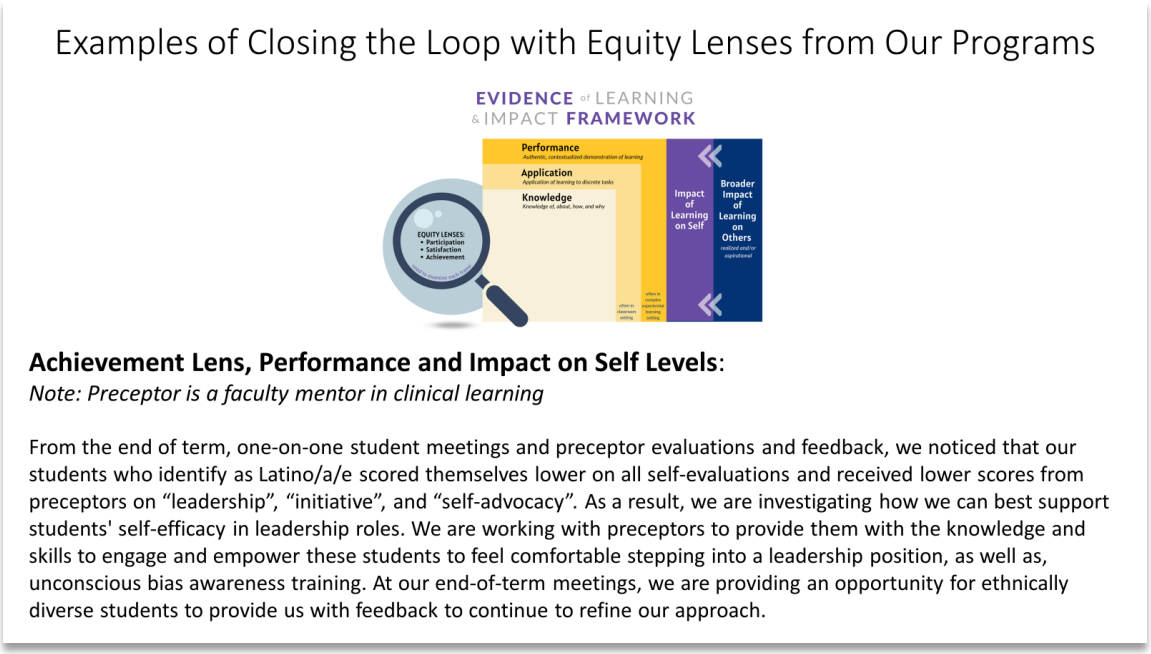


Figure 7. *Round 2 Example of Equity Lens in Use*

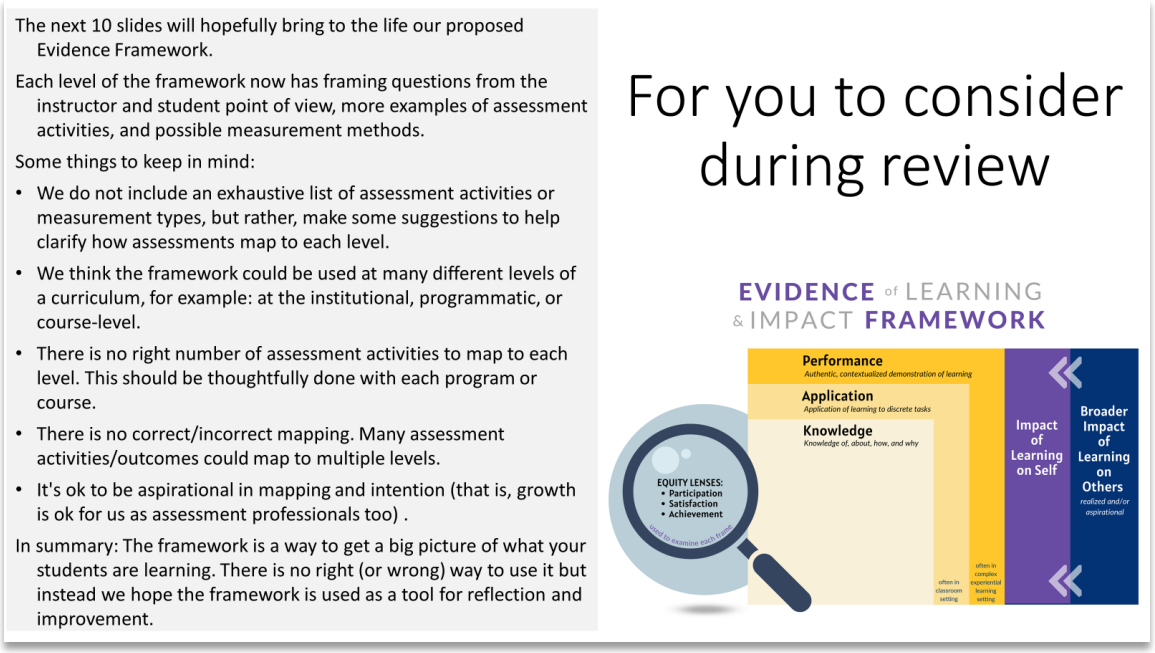


Figure 8. *Round 2 To Keep in Mind When Scoring Individual Levels*

Figure 9 and 10 illustrate two slides for the Performance Level that are representative of the kind of information we shared on each level during Round 2. You can see we shared changes made to the label of the level, revised framing questions both for students and for educators, and provided more examples. Appendix B contains the complete set of questions our experts responded to in Round 2.

Quantitative Results from Round 2

After Round 2, our overall Likert-scale question "This framework centers the learner in assessment" had achieved consensus (at least 90% Agree or Strongly Agree) and almost 93% of the participants responded "Yes" that they agreed with moving Participation and Satisfaction into Equity Lenses. For the specific levels, significant progress was made in getting closer to consensus on the Knowledge, Impact on Self, and Impact on Others levels. We also calculated whether we had reached consensus if we were to expand it to include Slightly Agree and it did not make much difference, so we decided to keep consensus as at least 90% that chose Agree or Strongly Agree. You can see from Table 5 that some of the individual questions about the levels were starting to reach consensus after Round 2.

Performance*

*This level was formerly called "Performance (Authentic)". Based on feedback, and in the interest in clarity, we have renamed it Performance.

Framing question(s) for educators:

- *To what extent do learners integrate their learning and demonstrate their depth and breadth of knowledge and skills as they negotiate a complex task?*

Framing question(s) for students:

- *How am I showing I am drawing on my knowledge and skills to navigate complex tasks?*
- *How am I adapting to new or conflicting information as I apply my knowledge to real-life tasks?*
- *How am I using what I've learned to make decisions based on multiple inputs and with various audiences and goals in mind?*
- *How do I unite my technical skill and creative expression in an authentic setting?*

Potential aligned assessment approaches:

- Comprehensive/Holistic Review: Program Committee/Panel Review - TAC/DAC reviews, theses, dissertations, capstones, portfolio, proposal defense
- External Review of Student Work(s): Manuscript feedback, external national assessments, IRB approval, grant review, peer reviewed blogs and presentations
- Internal Performance Observation: Simulation/Clinical Exercise, Research lab notebooks, portfolio, dissertation/thesis/capstone, peer evaluation
- Self-Assessment: Written, oral or other, portfolio development

Figure 9. *Round 2 Performance Framing Questions*

Performance - Examples

- ***Public Health***: Students draft a policy memo relevant to their internship audience, which is then evaluated by their preceptor and/or other internship stakeholders.
- ***Computer Science***: Students work collaboratively in their internship placements to develop a decision tree.
- ***Applied Linguistics:*** During student teaching, discuss the cognitive benefits of bilingualism with a parent who is upset that their Kindergartener still isn't reading in either language.
- ***English Literature***: Student publishes a poem they wrote for class in a local literary magazine, navigating revisions to the poem and professional expectations from the publisher, with coaching from their professor.
- ***Student Life*:** Student leads a team-building activity in the group they started and the student life trainer gives feedback.
- ***Career Services***: Student makes decisions about which student employment position and which clubs to participate in based on identified values.
- ***Undergrad science:*** Learner notices undergraduate research project is not resulting in usable data and proposes a revision to the existing method to lab advisor, drawing on knowledge of the scientific method and ethical research practices.
- ***Clinical***: In a simulated or real clinical setting, students take a history and physical exam on a patient, and then create a differential diagnosis.

Figure 10. *Round 2 Performance Examples*

Table 5. *Quantitative Results from Round 2*

Question text	Response Count (N)	High Agreement (%) (Strongly Agree, Agree)	Unclear (%) (Somewhat Agree, Somewhat Disagree)	Low Agreement (%) (Disagree, Strongly Disagree)
Framework as a Whole				
This framework centers the learner in assessment	28	92.86%	7.14%	0.00%
The Evidence Framework graphic is an effective way to visualize the framework	28	85.71%	14.29%	0.00%
This level is distinct from the other levels.	28	96.43%	0.00%	3.57%
Knowledge				
The framing questions helped me understand this level.	28	82.14%	17.86%	0.00%
I feel confident I could map assessments appropriately to this level.	28	96.43%	3.57%	0.00%
This level requires no editing/updating, or is good as written.	28	57.14%	35.71%	7.14%
Application				
This level is distinct from the other levels.	28	82.14%	14.29%	3.57%
The framing questions helped me understand this level.	28	71.43%	25.00%	3.57%
I feel confident I could map assessments appropriately to this level.	28	85.71%	14.29%	0.00%
This level requires no editing/updating, or is good as written.	28	42.86%	46.43%	10.71%
Performance				
This level is distinct from the other levels.	28	64.29%	17.86%	17.86%
The framing questions helped me understand this level.	28	71.43%	17.86%	10.71%
I feel confident I could map assessments appropriately to this level.	28	67.86%	14.29%	17.86%
This level requires no editing/updating, or is good as written.	28	46.43%	25.00%	28.57%
Impact of Learning on Self				
This level is distinct from the other levels.	27	100.00%	0.00%	0.00%
The framing questions helped me understand this level.	28	85.71%	7.14%	7.14%
I feel confident I could map assessments appropriately to this level.	28	78.57%	17.86%	3.57%
This level requires no editing/updating, or is good as written.	28	64.29%	25.00%	10.71%
Broader Impact of Learning on Others				
This level is distinct from the other levels.	28	92.86%	7.14%	0.00%
The framing questions helped me understand this level.	28	75.00%	21.43%	3.57%
I feel confident I could map assessments appropriately to this level.	28	78.57%	17.86%	3.57%
This level requires no editing/updating, or is good as written.	28	60.71%	32.14%	7.14%

Qualitative Results from Round 2

We organized our qualitative data again in Word to determine what changes we wanted to make based on expert feedback. Here are the major takeaways we found at each of the framework levels:

Qualitative data showed there was some confusion as to whether demonstration of skills belonged at the Knowledge or Application level of the framework.

Qualitative data on the Application level indicated the experts like the choice of "Application" for this level more than "Competence". Much of the substantive feedback was on needing to further distinguish between Application and Performance, namely that the location of the assessment (in class vs. in experiential learning) is not perfectly mapped onto more discrete tasks vs. more complex ones. A couple quotes from participants demonstrate the kind of feedback we received on this topic:

> "I have questions about the knowledge and skills part of the explanation… where do skills become Application or Performance vs. Knowledge"

> "As much as you can, play up the complex or multi-faceted/varied circumstances or environment of Performance to differentiate it from the Application level."

Qualitative data on the Performance level indicated both that Performance is too much of a leap from Application and also that they aren't distinguishable enough. Clearly,

the experts were struggling with the way we presented two levels in the slide deck as well. We did get a few really positive comments about Performance that made us think we are on the right path with it:

> "Robust framing questions and potential aligned assessment approaches are very clear and useful; All of the framing, aligned assessments, and examples are really strong. I think this is my favorite level as it is currently written; The scaling up to complexity, authenticity, and real-life settings and tasks; It represents complex thinking/learning/doing for a variety of contexts and circumstances."

> "The holistic aspect of it invites a long-term, wide lens of examination to assess student progress, which is unique from the other two levels. I see Knowledge and Application as really course-based, held within specific assignments, while Performance is more program-based, with several possible points of data."

Qualitative data on the Impact of Learning on Self level was primarily centered on the scope and wording of the framing questions we provided. Experts pointed out that the thrust of the Impact on Self framing questions for educators weren't fully aligned with the framing questions for students, for example. We also received some positive comments about this level:

> "This is the part of learning that is often left out of assessment frameworks, so I am glad to see it called out here."

> "I love this level! We do not often ask students to reflect on how their learning is affecting them."

> "Productively student-centered and metacognitive."

Qualitative data on the Broader Impact of Learning on Others level indicated the experts were not sure they liked the addition of "Broader" to the title, but they didn't agree on using "Intended" either. Much of the rest of the feedback was on finer points of the wording in the framing questions. A few positive comments about this level were also shared:

> "It helps higher education to reestablish itself as a public good by centering the emphasis on building healthy communities as the grand goal of learning."

> "I like that it covers the recognition, awareness, and intentionality of learning in relation to others. the part about recognizing new ways to perceive the world around them is particularly compelling."

Insight into Authors' Conversations

Much of our discussion after Round 2 was centered on trying to decide what we really meant by Application vs. Performance. We firmly believe that there is something different about a nursing student *performing* learning such as drawing blood in front of a family, while a baby is crying, while trying to speak a second language with patients versus *applying* learning to doing the blood draw on a mannequin arm in a simulation center. Fundamentally, we decided it wasn't just about where the learning or assessment was occurring, but also about the requirement to draw on lots of disparate sets of knowledge to accomplish the task that distinguished Performance from Application. However, we know this is not a perfect distinction.

We also discussed whether we wanted to call the different parts of the framework "levels" or some other term (category, group, set, kind, measure, section, frame, type). We decided to add a question about this to Round 3 to see what the experts thought.

Consensus After Round 2

A slide (Figure 11) was shared in Round 2's slide deck and summarizes the state of consensus and timeline.

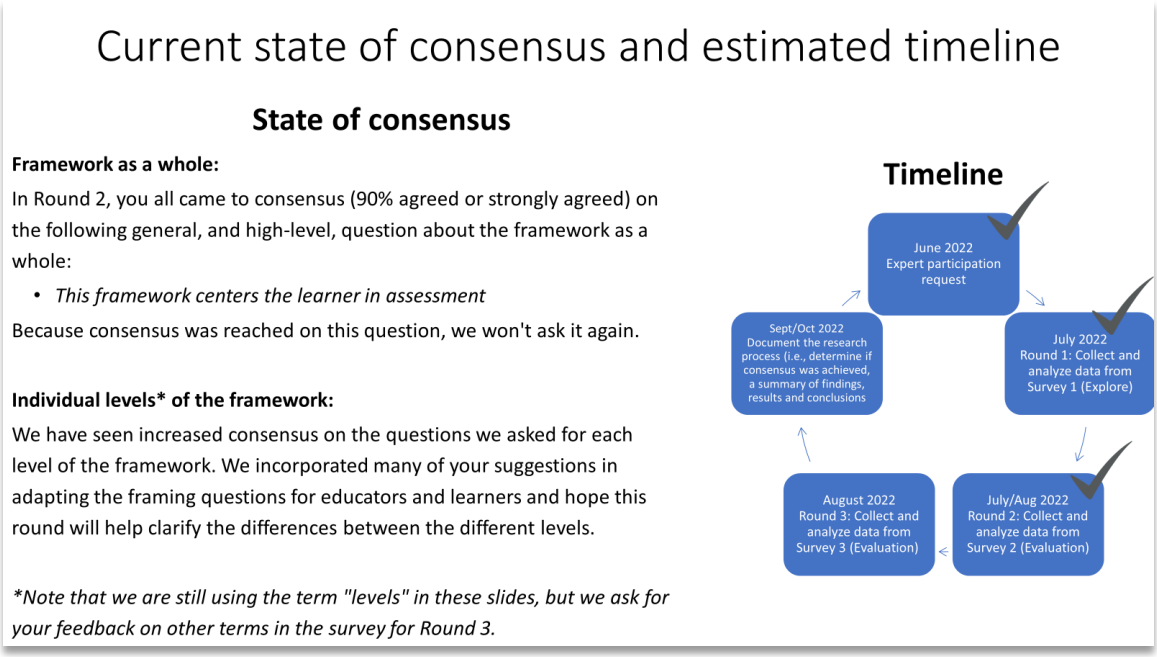


Figure 11. *State of Consensus after Round 2*

Round 3

Expert Participants Round 3

From the original set of 39 experts, we had 22 Round 3 expert respondents (22 participants is 78% of those who responded in Round 2, or 56% of those who responded in Round 1). No new experts were added to Round 3. We estimate that it took at least an hour for the experts to respond to each round of the Delphi study, which makes the numbers dropping each round unsurprising, and also consistent with literature (Boel et al., 2021). We think we were able to retain the number of experts that we did because we provided a relatively quick turnaround from round to round, we were responsive to their feedback, and we shared with them transparently what changes we made and why.

Niederberger and Spranger (2020) identified an area described for improvement we highlighted in the introduction to this paper: '4. Careful thought about the implications of having a small number of experts (in the low double digits) on the consensus that is reached.' Our Delta rounds, from 22-38 expert participants each, generated a lot of rich, insightful data in line with what you would expect from a rigorous qualitative study. If we had received the same amount of detailed feedback from 80 or 100 experts we would have been drowning in data. And because we maintained a relatively high bar in terms of reaching consensus on the quantitative data, we are confident that the revisions we made to the Evidence Framework were well-conceived.

Materials Reviewed in Round 3

The slide deck for Round 3 focused primarily on clarifying and differentiating the five levels of the framework and finalizing the framing questions. See figure 12 for a summary of changes made after round 2. From qualitative feedback in Round 2, we decided that the framing questions were a more powerful way to define the different levels, and pivoted to focus just on those during Round 3 instead of on refining our examples. The focus on framing questions is communicated in Figure 13. Figure 14 is an example of the revisions we made to each level's slide focused on framing questions. Performance is shown for continuity with the example provided above in Round 2 (see figures 9 and 10). Figure 15 shows how we conceived the differentiation between Application and Performance after two rounds of feedback telling us we needed to make that clearer. Appendix C contains the complete set of questions our experts responded to in Round 3.

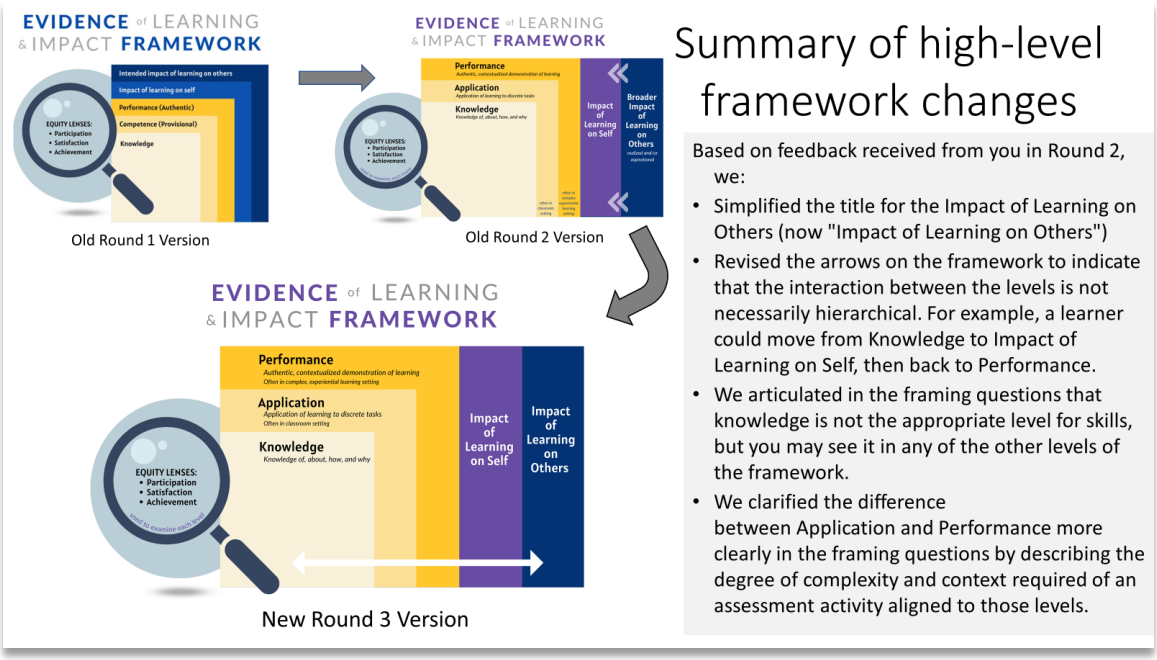


Figure 12. *High-Level Changes Made after Round 2 to the Evidence Framework*

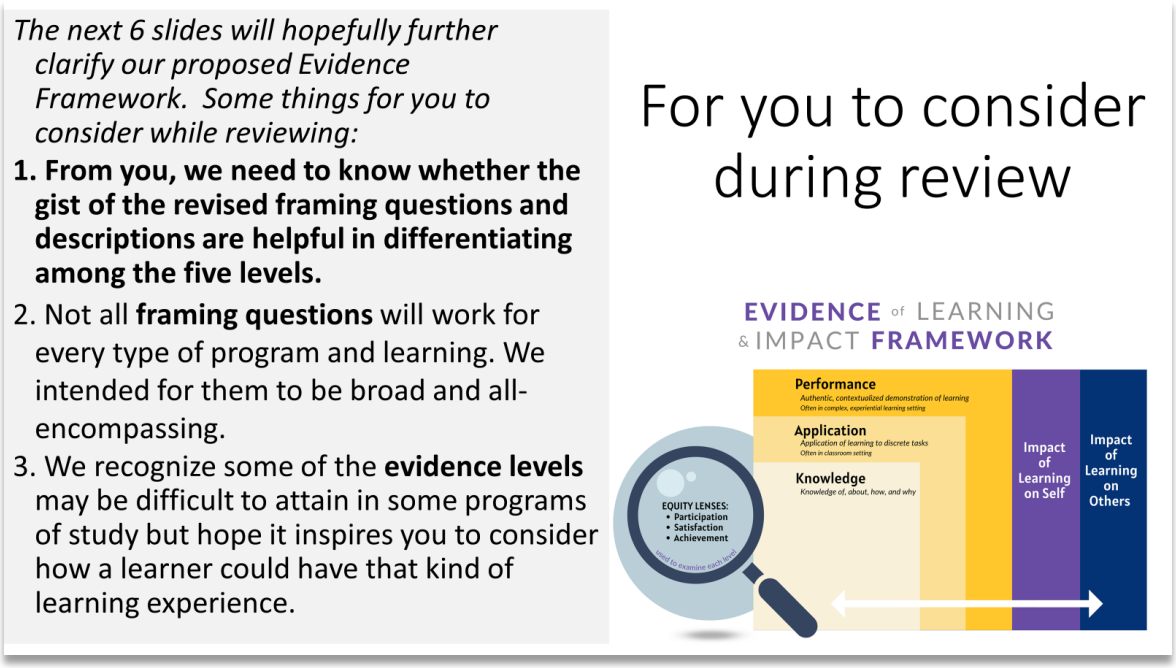
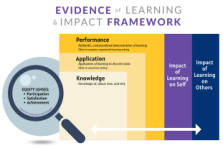


Figure 13. *To Consider While Reviewing Round 3*

Performance

Authentic, contextualized demonstration of learning; often in complex or novel contexts. Usually seen in experiential learning settings that require navigating the situation as a whole by using knowledge/skills from different domains.



Framing question(s) for educators:

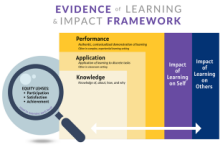
- To what extent do learners integrate their learning and demonstrate their depth and breadth of knowledge and skills as they navigate complex tasks and/or situations?
- To what extent do learners navigate unexpected challenges as they apply their learning? Do they transfer knowledge in messy situations?

Framing question(s) for students:

- How do I show that I am able to thoughtfully apply knowledge and skills in complex ways under real constraints and conditions? Am I able to move beyond the script or formula? How am I adapting to new or conflicting information?
- How do I draw on knowledge and skills from disparate subject areas?
- How am I using what I've learned to make decisions based on multiple inputs and with various audiences and goals in mind?
- How do I integrate technical skill with creative expression to demonstrate learning and accomplish objectives?

Figure 14. *Round 3 Performance Description and Framing Questions*

Application versus Performance



Application is:

Defined by application of learning to discrete tasks. This often occurs in a classroom setting or in controlled contexts and may rely on formulas or scripts. Learners practicing skills would fall under application.

Performance is:

Defined by an authentic, contextualized demonstration of learning. It often occurs in complex or novel contexts and is usually seen in experiential learning settings that require navigating the situation as a whole by using knowledge/skills from different domains.

Application Examples	Performance Examples
Computer Science: Using sample user interface data from a popular software application, students work with classmates to draft a list of suggested updates/changes to improve user experience with software.	**Computer Science**: Student identify a user interface design problem within the company where they are interning, research possible solutions, identify appropriate data to collect, and propose a solution to improve user experience. (This same assessment activity could be mapped to Impact on Others, if the project were actually completed, and user experience data was gathered and analyzed for further improvement).
English Literature: Students practice writing poems that focus on elements such as repetition, symbolism, and figurative language.	**English Literature:** Student publishes a poem they wrote for class in a local literary magazine, navigating revisions to the poem and professional expectations from the publisher, with coaching from their professor.
Undergrad Science: Learner follows steps of scientific method during intro bio lab and writes up experiment in lab notebook.	**Undergrad Science:** Learner notices undergraduate research project is not resulting in usable data and proposes a revision to the existing method to lab advisor, drawing on knowledge of the scientific method and ethical research practices.

Figure 15. *Round 3 Application vs. Performance*

Quantitative Results from Round 3

After Round 3, we determined that we had reached consensus, though not exactly in the way we had originally identified when we started the Delphi study. In looking at Round 2 data, we worried that wording our main consensus question for each question "This level requires no editing/updating, or is good as written", and requiring Strongly Agree or Agree at 90% was perhaps setting a rather perfectionistic bar for ourselves. We decided to slightly revise the wording of that question in Round 3 to "This level requires minor editing/updating, or is good as written." We still wanted to keep consensus at 90% Strongly Agree or Agree.

Table 6. *Quantitative Results from Round 3*

Question Text	Response Count (N)	High Agreement (%) (Strongly Agree, Agree)	Unclear (%) (Somewhat Agree, Somewhat Disagree)	Low Agreement (%) (Disagree, Strongly Disagree)
Knowledge				
This level is distinct from the other levels.	22	100.00%	0.00%	0.00%
The framing questions helped me understand this level.	22	95.45%	4.55%	0.00%
I feel confident I could map assessments appropriately to this level.	22	95.45%	4.55%	0.00%
This level requires minor editing/updating, or is good as written.	22	72.73%	9.09%	18.18%
Application				
This level is distinct from the other levels.	22	90.91%	9.09%	0.00%
The framing questions helped me understand this level.	22	100.00%	0.00%	0.00%
I feel confident I could map assessments appropriately to this level.	22	100.00%	0.00%	0.00%
This level requires minor editing/updating, or is good as written.	22	72.73%	13.64%	13.64%
Performance				
This level is distinct from the other levels.	22	90.91%	9.09%	0.00%
The framing questions helped me understand this level.	22	95.45%	4.55%	0.00%
I feel confident I could map assessments appropriately to this level.	22	100.00%	0.00%	0.00%
This level requires minor editing/updating, or is good as written.	22	77.27%	4.55%	18.18%
Impact of Learning on Self				
This level is distinct from the other levels.	22	100.00%	0.00%	0.00%
The framing questions helped me understand this level.	22	100.00%	0.00%	0.00%
I feel confident I could map assessments appropriately to this level.	22	90.91%	9.09%	0.00%
This level requires minor editing/updating, or is good as written.	22	81.82%	9.09%	9.09%
Impact of Learning on Others				
This level is distinct from the other levels.	22	100.00%	0.00%	0.00%
The framing questions helped me understand this level.	22	95.45%	4.55%	0.00%
I feel confident I could map assessments appropriately to this level.	22	90.91%	9.09%	0.00%
This level requires minor editing/updating, or is good as written.	22	86.36%	0.00%	13.64%
Framework as a Whole				
The Evidence Framework graphic is an effective way to visualize the framework	22	100.00%	0.00%	0.00%
Assuming minor edits based on feedback, the evidence framework, as a whole, works for me.	22	95.45%	4.55%	0.00%

Round 3 data showed that with our new, slightly modified question, in combination with all the revision made to the *Evidence Framework*, that we now had between 72% and 86% of experts Agreeing or Strongly Agreeing that each level was done. However, when we looked at the three sub-questions for each level ("this level is distinct", "framing questions help me understand this level", and "I could map to this level confidently"), each one of those had reached consensus of at least 90%. We decided to declare consensus reached in this way and ended the Delphi study.

Additional quantitative data collected in Round 3 was whether we should use a different term than "level" when discussing the *Evidence Framework*. 52% of the experts said we should keep "level", and the term with the second most votes (33%) was "category". We chose to retain the use of "level" for now.

Qualitative Results from Round 3

In Round 3, feedback we received included wording tweaks for the Application framing questions; that we had more adequately addressed the difference between Application and Performance; and that we needed to reconsider the use of "scholarship" in our framing of the Impact of Learning on Others. We received many comments similar to "Solid, clear to understand", "You nailed it", "Framing questions are fantastic", and some really lovely comments at the end of the survey.

> "THANK YOU! Your courage and practice in the development of this framework are so impressive. As I apply your very framework to the work you have done, I think you check all of the boxes. I'd like to be able to use some of your resources in my work conducting faculty development on my campus, and I hope you consider licensing this for others to use with an attribution. Thank you."

> "This has been a great process and the framework is better because of it. Thank you for your contribution to the field."

An analysis of the qualitative data also showed that the degree of participant confusion had decreased, as did the scope of the changes that were being suggested as we progressed through the three rounds, which helped to confirm the consensus calculations that came out of the quantitative data.

Summary of Final Changes Made

We integrated many of the Round 3 suggestions we received into the final version of the framing questions and added some additional text to the graphic.

Discussion

Final Version of Evidence of Learning and Impact Framework

The final version of the *Evidence of Learning and Impact Framework* is shown in Figure 16. We invite faculty, programs, schools, and universities to consider adopting and, importantly, adapting it with attribution for their own use.

We hope that the *Evidence of Learning and Impact Framework* will challenge educators to think differently, more broadly, and more deeply about the kinds of learning they foster and assess. We also hope the framework pushes higher education, and in our case, especially STEM, toward a space where multiple narratives and multiple interpretations of information are valued, and where students are given opportunity to reflect meaningfully on the holistic impact of their learning on themselves and those around them.

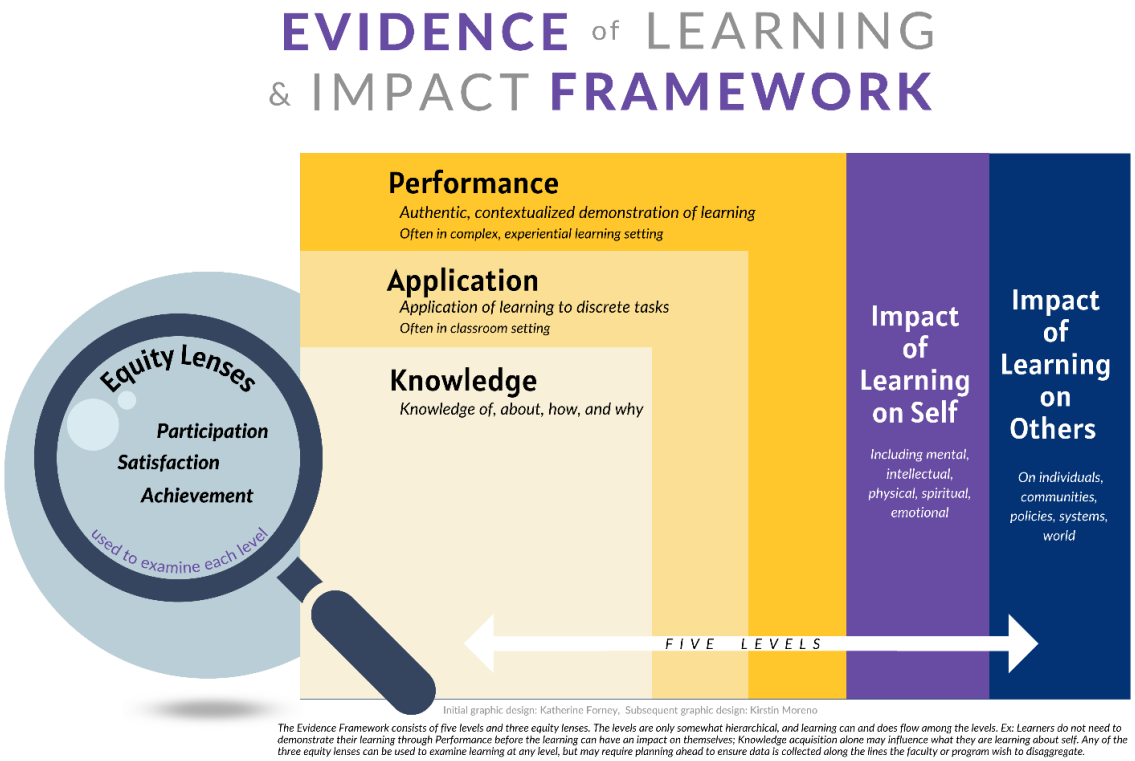


Figure 16. *Final Version of Evidence of Learning and Impact Framework (or "Evidence Framework")*

Implementation Challenges

We have been using our *Framework* since 2020 and have noted a few implementation challenges. At a high level, it can be challenging to shift a group of faculty from a known paradigm such as Bloom's Taxonomy to a new way of thinking about assessment. We re-visited the distinctions between Bloom's, Moore's and our *Evidence Framework* with assessment staff and faculty over time to help them fully understand the aims behind the new framework. In addition, a couple of the levels are a bit challenging to learn how to assess or measure. For example, incorporating Impact of Learning on Self has not been as difficult as we expected, possibly because reflective practice is common in the health professions, but Impact on Others has taken more work to orient to. Most of our programs have maybe only one or two Impact on Others-aligned activities, as this level is a big lift and may apply well to capstone projects. To support faculty in learning a new framework, it was important to reassure them that they don't need to master every aspect at once and acknowledging this has been helpful. Finally, the boundaries between Application and Performance, and also between Performance and Impact on Others can be a bit of a grey area, but we have found that looking at the intentions of the assessment activity and the method used to assess the learning helps to clarify the correct level to assign. In some cases, it's appropriate to assign more than one level, so we've built in flexibility to allow for that. Overall, we've found that the *Evidence of Learning and Impact Framework* aligns well with the needs of our academic and student affairs departments, and we've encountered fewer barriers compared to the previous frameworks we used.

Additional Resources to Assist with Implementation

To support readers in effectively applying the *Evidence of Learning and Impact Framework*, we have included a set of supplementary resources. Specifically, the final list of framing questions—revised in response to feedback from Round 3—can be found in Appendix D. These questions are intended to facilitate immediate implementation of the framework, particularly for those who require minimal additional guidance.

In addition, Appendix E contains a list of theoretical example activities that align to the *Evidence Framework* from various disciplines that was revised based on feedback from Round 2 of the Delphi study. The most significant revisions we made based on the Delphi study to these example activities in Appendix E are:

- The addition of History examples to capture different aspects of Impact on Self and how it can be connected to the Knowledge level and to affect.
- The addition of phrasing to the Performance level examples to emphasize the importance of bringing together multidisciplinary knowledge and skills, based on feedback such as "As much as you can, play up the "complex" or multi-faceted/varied circumstances or environment of this learning demonstration to differentiate it from the "Application" level."

Considerations for Adopting the Evidence of Learning and Impact Framework

Based on our experience having faculty and staff implement the Evidence Framework at OHSU, we recommend the following:

1. Begin by having programs align their assessments to the framework levels, supporting this work with charts, trainings, and/or consultation.
2. Consider allowing programs to map to both an older framework (e.g. Bloom's Taxonomy, etc.) and the Evidence Framework at the same time in order to scaffold understanding about the Evidence Framework.
3. Build in checks on quality alignment to Evidence levels, such as trained peer reviewers, giving feedback and suggesting realignment/changes where needed.
4. Programs often struggle most with the two Impact levels, so we recommend additional training and exercises to reinforce the types of measurement which work best for these two levels (e.g. journals and self-knowledge tests for Impact on Self, or community impact data for Impact on Others). These two levels draw on practices from Impact Evaluation, so we recommend you familiarize yourself.
5. Once initial alignment and instruction has happened, introduce the equity lenses and how they might be applied in context, for example, by asking learners their perspectives on the learning environment, or by disaggregating student outcome performance data and examining the results for possible inequities.
6. Starting small is key, whatever that looks like at your institution. We used pilot groups, double (or triple) mapping, and trained framework champions to help with implementation.

Please adapt the framework levels and equity lenses to fit your institutional context.

Limitations

Our Delphi study and subsequent *Evidence Framework* would be more robust if we had been able to recruit more expert participants from places outside of North America, especially because we imagine conceptualizations around ideas of Impact on Self and Impact on Others may be quite different in cultures that are more collectivist in nature, or that perhaps have less centering of whiteness, colonization, capitalism, and similar orientations. We have mitigated this somewhat by ensuring that the scoping review (Tucker, Jacobs, and Moreno, 2024) we conducted included research and thinking about learning and assessment from countries in the global south or indigenous perspectives. We drew on these studies to help us construct some of our framing questions for each level and to refine our equity lenses, but more geographically dispersed Delphi experts would have been useful partners in this work.

Future Work

In future work, we hope to share more details about how this framework could be used (in assessment planning and reporting at the institutional level, in program- and course-level assessment practices), more examples of the equity lenses in action, and a discussion of assessment methods that pair well with the framework levels. The final list of framing questions, revised based on Round 3 feedback, are in Appendix D so the framework can be implemented now for those who need less guidance.

Additional future work could examine whether the Evidence of Learning and Impact Framework is indeed applicable to different sectors of higher education including career technical education and the humanities given that we have only been using it in an academic health center. Future scholarship could also expand upon specific ways to approach the equity lenses in an assessment framework of this scope. Finally, the Impact on Self and Impact on Others levels of the Evidence Framework are especially rich spaces to think differently and more broadly about the impact of learning on learners and those around them. We would love to see more assessment scholarship that dives into these levels.

Using Delphi Techniques for Improving Educational Innovations:

Delphi techniques used in this study were a very powerful way to significantly improve the rigor of the framework we had already put a lot of time and thought into developing. As this was the first time we employed a Delphi process, we did not know what to expect in terms of the quality or amount of feedback we would receive, or whether our participants would choose to continue to stick with the study. Happily, the experts provided an abundance of input, very generously sharing their thoughts with us, and prompted many nuanced conversations as we figured out how to synthesize their feedback and be responsive to their insights. The Delphi process forced us to really think about the different components of our framework and what we needed to address with each. We fully endorse employing Delphi techniques if you are looking to refine something that has broad appeal and have the time to put in the work to run a Delphi study well. We did not realize how long it would take to pull together all of the materials for each round, to clean the data to complete the consensus calculations, and to sort through the qualitative feedback. Estimate more time than you think!

There are various things about Delphi techniques that are not intuitive. Those include determining exactly how to determine consensus, for example: which questions(s), at which level of agreement, from which percentage of respondents, and whether you should average scores in doing so, etc. Determining whether the qualitative feedback plays a role in consensus-determination is also not something that occurred to us intuitively. On these things you'll need to trust your gut in addition to looking at how other Delphi studies have handled determining consensus, and then just be transparent about what you did (Niederberger and Spranger, 2020). There are many ways to approach running a Delphi study.

Tips and Crucial Mindsets for Using Delphi Techniques

- Anonymity is key. A small portion of the (anonymous) feedback the experts had to give us was quite critical, but necessary for us to hear.
- Take time to really determine what the different, concrete aspects of your project are that you want expert feedback on. Determine if you will ask for that all at once in the first round, or if you will space it out, prioritizing different aspects for different rounds of the Delphi process.
- It is critical to carefully document your process throughout so you can provide detailed transparency in your own scholarship that draws on Delphi techniques.
- Be transparently flexible with consensus setting. Check that you're not being perfectionistic, but maintain high standards. Determining when to claim consensus is an art, not a science.
- Be responsive and take input you receive in good faith, and share back what you changed based on that input. It really makes a difference to the experts who take time to participate in the study.

Appendices

Appendix A: Round 1 Survey Questions
Appendix B: Round 2 Survey Questions
Appendix C: Round 3 Survey Questions
Appendix D: Framing Questions for the *Evidence of Learning and Impact Framework*
Appendix E: Evidence of Learning and Impact Framework Example Activities

Acknowledgements: The authors wish to thank the OHSU Assessment Council for their collaboration on initial versions of the Evidence Framework; our Delphi Study participants who truly provided an abundance of useful and insightful feedback as we went through our three rounds; and Kathie Forney, for her initial graphic design of our Evidence Framework.

Previous Presentations: Some details from our Delphi study and revised Evidence of Learning and Impact Framework were presented by the authors at the International Association for Medical Education (AMEE) in Lyon, France, Aug 27-31, 2022, at the IUPUI Assessment Institute in Indianapolis, Indiana, Oct 9-11, 2022, and at AALHE, Jun 3-6, 2024 in Portland, Oregon.

References

Alegría, M., Thurston, I. B., Cheng, M., Herrera, C., Markle, S. L., O'Malley, I. S., Porter, D., Estrada, R., & Giraldo-Santiago, N. (2024). A learning assessment to increase diversity in academic health sciences. *JAMA Health Forum*, *5*(2), e235412. https://doi.org/10.1001/jamahealthforum.2023.5412

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn & Bacon. Boston, MA. ISBN: 978-0801319037

Barrett, D., & Heale, R. (2020). What are Delphi Studies? *Evidence-Based Nursing*, *23*(3), 68–69. https://doi.org/10.1136/ebnurs-2020-103303

Boel, A., Navarro-Compán, V., Lawendé, R., & van der Heijde, D. (2021). Two different invitation approaches for consecutive rounds of a Delphi survey led to comparable final outcome. *Journal of Clinical Epidemiology, 129*, p. 31–39. https://doi.org/10.1016/j.jclinepi.2020.09.034

Green, R. A. (2014). The Delphi technique in educational research. *SAGE Open, 4*(2). https://doi.org/10.1177/2158244014529773

Green, M., & Malcolm, C. (2023). Degrees of change: the promise of anti-racist assessment. *Frontiers in Sociology, 8*. https://doi.org/10.3389/fsoc.2023.972036

Henning, G., Rice, A., Heiser, C., & Lundquist, A. (2023). Equity-centered assessment practices: Survey findings and recommendations. *Research & Practice in Assessment, 18*(2), p. 20–30. https://www.rpajournal.com/dev/wp-content/uploads/2024/03/Equity-centered-Assessment-Practices-RPA.pdf

LaFever, M. (2016). Switching from Bloom to the Medicine Wheel: Creating learning outcomes that support Indigenous ways of knowing in post-secondary education. *Intercultural Education*, *27*(5), 409–424. https://doi.org/10.1080/14675986.2016.1240496

Moore, D. E., Jr, Green, J. S., & Gallis, H. A. (2009). Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. *The Journal of Continuing Education in the Health Professions*, *29*(1), 1–15. https://doi.org/10.1002/chp.20001

Nasa, P., Jain, R., & Juneja, D. (2021). Delphi methodology in healthcare research: How to decide its appropriateness. *World Journal of Methodology*, *11*(4), 116–129. https://doi.org/10.5662/wjm.v11.i4.116

Niederberger, M., & Spranger, J. (2020). Delphi technique in health sciences: A map. *Frontiers in Public Health*, *8*, 457. https://doi.org/10.3389/fpubh.2020.00457

Tucker, C., Jacobs, S., & Moreno, K. (2024). Branches from the same tree: A scoping review of learning outcomes frameworks. *Intersection: A Journal at the Intersection of Assessment and Learning*, *5*(4), 53–82. https://doi.org/10.61669/001c.123960

Twyman-Ghoshal, A., & Carkin Lacorazza, D. (2021, March 31). Strategies for antiracist and decolonized teaching. *Faculty Focus*. https://www.facultyfocus.com/articles/equality-inclusion-and-diversity/strategies-for-antiracist-and-decolonized-teaching/ retrieved April 30, 2024.

Barriers and Enablers of Integrating Assessment with Curriculum and Instruction

Authors:

Rebecca E. Gibbons, Ph.D.
Embry-Riddle Aeronautical University

Arthur Hernández, Ph.D.
University of the Incarnate Word

Teresa Flateby, Ph.D.
Independent Consultant

Karla Hardesty, Ed.D.
Colorado Mountain College

Yuerong Sweetland, Ph.D.
Franklin University

ABSTRACT

The extent to which three essential processes—assessment, curriculum, and instruction—are systematically integrated provides a window into the efficacy of assessment in improving higher education. This paper explores findings from open-ended responses to a survey of assessment stakeholders to understand this integration. The authors synthesize respondents' perceptions of how the factors outlined in the Assessment Integration Model derived from a previous interview study (Flateby et al., 2023; 2024): Leadership Support, Assessment Processes, Assessment Purposes, Recognition & Reward, Faculty Support/Development, Collaboration, and Accreditation, contribute to integration. Respondents emphasized the critical impact of leadership, leadership changes, and siloing (even at small institutions) on integration. In addition, responses from assessment stakeholders suggest a shared sense of urgency and a desire to establish integrated systems despite the numerous challenges. The findings indicate that some institutions have well-developed assessment processes; however, the processes are not often integrated with curriculum and instruction.

Correspondence E-mail: Rebecca.gibbons@erau.edu

Keywords: Higher Education, Assessment, Integration, Curriculum, Instruction

Faculty and staff at institutions of higher education (IHEs) seek insight and strategies to increase the relevance and quality of instructional programs. Data derived from the assessment process (specifically, assessing student learning outcomes) is commonly used as a source of information to achieve that end (Kuh & Ewell, 2010). Assessment scholars have posited (Flateby et al., 2023; 2024) that institutional success in meeting the instructional/educational mission requires integrating assessment as an essential component of curricular and instructional planning to foster student learning outcomes effectively. Faculty regularly use formative assessment processes (Wiliam & Black, 1996) to gauge and (in more advanced practice) enhance student learning in the classroom (Webber & Tschepikow, 2011). However, outside the assessment profession, the assessment process is not often used as a tool in curricular and instructional implementation beyond the course level (Banta & Blaich, 2010; Peterson & Augustine, 2000). We assert that integrating three specific components of IHE operations—assessment, curriculum, and instruction—can serve as an enhancement lever.

The assessment process was designed to integrate with curriculum and instruction (National Institute of Education, 1984). The primary role of assessment, to improve student learning, has been supported since the assessment movement began in the mid-1980s (Ewell, 1991). After the movement began, assessment was formalized by institutional accreditors within standards (Gaston, 2018). More IHEs completed assessment activities, most often operationalized as the completion of assessment reports for academic programs, after they were included in accreditation standards (Jankowski et al., 2018). Two paradigms for assessment grew from this environment: accountability and improvement, which can be in tension (Ewell, 2009). Ewell (2009) characterized the improvement paradigm as internally focused on enhancing teaching and learning. The improvement paradigm, then, is similar to the manufacturing/industrial concept of continuous quality improvement widely embraced in the medical context (Colton, 2000). On the other hand, the accountability paradigm has a compliance and external reporting focus.

Under the accountability/accreditation standards paradigm, institutions have generated assessment processes within a positivist mindset. The faculty perception of assessment through the lens of the accountability paradigm can be characterized as unwelcome, rooted in the perception that assessment could be used punitively toward underperforming instructors (Pepin, 2014). In the improvement paradigm, assessment is embedded in curricula and goals are "continually map{ped} to, and reinforced by, the teaching and learning process throughout the curriculum…" (Ewell, 2009). This assessment can be transformative, enhancing student learning and equity at institutions (Henning et al., 2022).

We conceptualize the improvement paradigm through the lens of integration to provide a framework to approach the ideal of assessment. The inquiry described in this manuscript explores assessment stakeholders' perceptions of the factors that influence the integration of assessment with curriculum and instruction (Figure 1). An integrated system emphasizes the importance of faculty and staff in engaging with assessment data to effect change in their programs. We posit that the benefits of creating an integrated system span alignment and clarity, evidence-based decision-making, improvement for learning's sake, and accountability/accreditation.

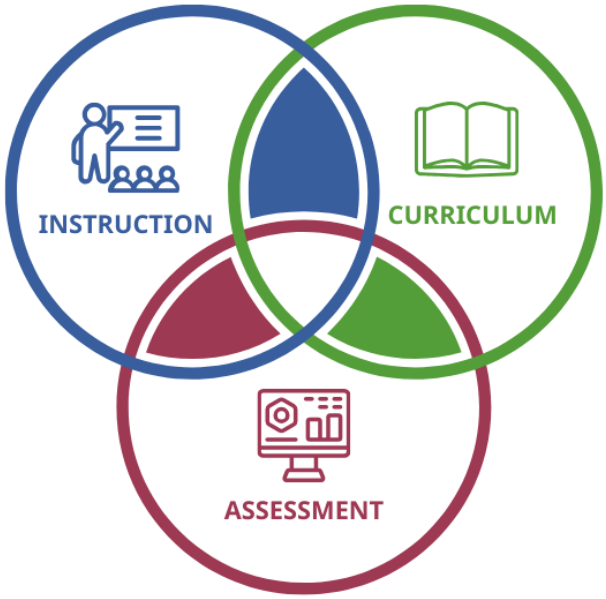


Figure 1. *Assessment Integration*

In practice, "curriculum" refers to the structured educational experiences provided to learners within a specific program or course. Effective instruction, informed by educational theory and tailored to diverse student needs, is how the curriculum's goals are realized. Assessment, within this integrated model, is more than measuring student achievement; it is an ongoing process of inquiry that explores and enhances the effectiveness of the curriculum and instruction.

We consider the challenges and opportunities facing those implementing assessment at IHEs. We leverage qualitative data sources from a United States-scale survey, enabling us to demonstrate the factors that assessment stakeholders (i.e., practitioners and faculty) articulate are influencing the dynamics between the three integrated components. Our analysis is anchored in the Assessment Integration Model (AIM, described below).

Rationale for Integration

Assessment can make gains for learning and improvement in IHEs when integrated with curriculum and instruction (Kinzie, 2019). Pending a large body of literature on the nature of integration, we propose, through the researchers' lived experiences, that integration can "move the needle" on the perceptions of faculty as an "add-on" or unnecessary practice (Flateby et al., 2023; 2024).

The definition of integration as articulated here emerges from the foundational works in assessment, although no extant literature specifically describes integrating curriculum, instruction, and assessment into a single system. Scholarly interpretation of the former American Association for Higher Education (AAHE) Principles of Good Practice (Astin et al., 1992) nods to integration (Hutchings et al., 2012). For instance, Principle 2 states, "Assessment is most effective when it reflects an understanding of learning as multi-dimensional, integrated, and revealed in performance over time."

Further, in Principle 4, the authors mention attending to outcomes and also the "experiences along the way – about the curricula, teaching, and kind of student effort that led to particular outcomes."

Our lived experiences as assessment professionals and faculty within IHEs indicate that when faculty are focused on the curriculum, integrated assessment changes their perceptions about assessment. For example, well-implemented integration can help faculty to realize that assessment is essential for answering important questions for themselves and their programs. One of the authors' institutions, a doctoral-granting regional university with approximately 25,000 students, developed a university-wide voluntary initiative to improve students' written communication skills and specific critical thinking skills within their majors, intending to transfer these skills to their careers. Because the initiative director was an Institutional Effectiveness administrator, it was initially perceived primarily for accreditation. After the initiative was implemented for several semesters and centered on faculty development, integration emerged organically when faculty viewed assessment as tied to instruction (Flateby et al., 2023; 2024). Specific faculty participating in the program saw such benefits in this integration that they voluntarily began leading curricular revisions in their programs.

Integration positively reinforces aligning curriculum with program learning outcomes, as neither curriculum nor assessment can deviate from the other in an integrated system. One of the goals of assessment processes is to create a context in which decisions about instruction and curriculum are based on data and evidence. In most IHEs, these decisions are made ambiguously. Where integration is successful and the curriculum is tightly aligned with learning outcomes, decisions to impact student learning are directly informed. This is especially relevant in the case of updating the curriculum; in some cases, curricular decision-making is limited to those aspects of disciplinary knowledge that appear current to faculty experts' perspectives. While this input is essential, the information provided by assessment can enhance curricular innovation born of disciplinary trends by identifying those current curricular areas that might require more emphasis across the courses and experiences in a program.

Integration also fosters a cultural drive toward continuous improvement, as it promotes environments where reflection on instructional practice serves as a key decision-making tool, creating opportunities to utilize assessment results for enhanced learning. This culture cultivates a more agile IHE, adaptable to changing student learning needs and encouraging instructional innovation to meet those needs.

Thus, the integrated and reciprocal relationship between curriculum, instruction, and assessment in IHEs forms one interdependent system where each element continuously informs and shapes the others.

Assessment Integration Model: Development and Components

In a previous qualitative investigation, the perspectives of assessment practitioners at IHEs recognized by the assessment community for high-quality assessment work (Flateby et al., 2023; 2024) provided unique insights. The Assessment Integration Model (AIM, Figure 2) was developed after this interview-based study with eight representatives from mostly large, research-oriented institutions. These institutions were identified by a group of assessment experts in the Association for the Assessment of Learning in Higher Education (AALHE's) Knowledge Development Task Force via reputation and a review of

websites as institutions that demonstrate highly successful assessment practice. The interviews with representatives from these institutions were qualitatively analyzed to identify the representatives' stated factors that impact their integration work. Upon conclusion of this analysis, the AIM identifies seven influential factors: Leadership Support, Assessment Purposes, Assessment Processes, Collaboration, Recognition & Reward, Faculty Development/Support, and Accreditation. All factors worked together to cultivate an environment for integration (although not always fully realized) in the IHEs that were identified by the Task Force as having effective assessment practices. The comprehensive approach of AIM emphasizes the synergy between the seven factors for improving student learning. AIM is predicated on the understanding that effective integration requires more than just the implementation of assessment tools; it necessitates a holistic and strategic approach that engages all stakeholders, is driven by faculty in collaboration with assessment practitioners, and is aligned with the institution's broader educational objectives.

Leadership Support

The first factor considered in AIM is Leadership Support, which is foundational in navigating the challenges associated with integration, including resource allocation, faculty buy-in, and alignment of assessment with institutional goals and accreditation standards. All assessment professionals in the original study indicated that effective leadership is crucial for championing integration implementation. Leaders can also articulate a vision and mobilize financial and human resources to support assessment initiatives. Institutional leaders' (e.g., presidents, provosts, deans) commitment and active involvement are critical for setting the tone and priorities of assessment initiatives and creating an institutional culture that values and supports assessment as the principal means to enhance student learning and success. Effective leadership support also involves advocating for assessment initiatives and addressing barriers to implementation by provisioning resources through strategic and operational planning.

Assessment Purposes

The next factor is the Assessment Purposes. The purpose and driving motivation for completing assessment work is one of the most influential aspects of a successful integration environment. As detailed in the literature reviewed above, the focus on assessment for accountability via accreditation serves one route, and the focus on assessment for continuous quality improvement opens the door to successful integration initiatives. While IHEs must conduct the assessment process to demonstrate accountability through accreditation requirements, the purpose of assessment moving beyond this requirement and towards improvement is key for integration.

Assessment Processes

Assessment Purposes are considered distinct from Assessment Processes in AIM. Transparent and sustainable assessment processes are vital for ensuring the relevance and efficacy of integration. An institution cannot progress toward successful integration without an existing process that can seamlessly integrate assessment, curricular, and instructional work.

Recognition & Reward

AIM highlights the importance of Recognition & Reward mechanisms to foster faculty and staff engagement in integration. Acknowledging the efforts and contributions of individuals and teams participating in assessment activities can significantly enhance motivation and commitment. The prior study revealed such strategies as formal recognition in performance evaluations, professional development opportunities, monetary rewards, or public acknowledgment of individual contributions to enhancing educational quality.

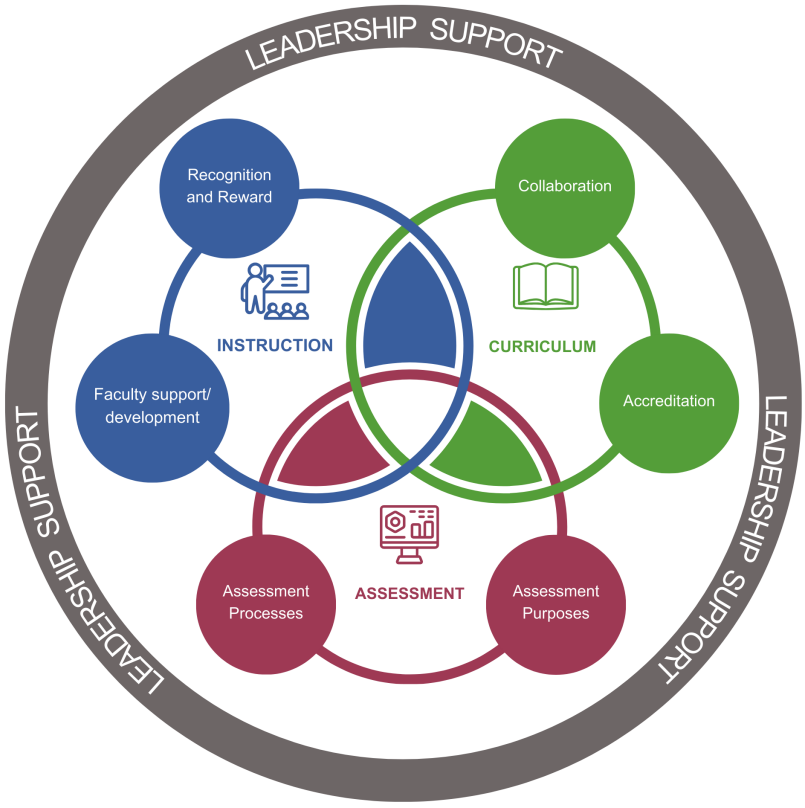


Figure 2. *Assessment Integration Model*

Faculty Support/Development

Faculty Support/Development is a critical component of AIM, recognizing that faculty members play the central role in curriculum design and delivery and, consequently, in the implementation of integration. This factor includes professional development opportunities focused on assessment methodologies, data analysis, curriculum alignment, and using assessment results for continuous improvement.

Collaboration

No individual assessment professional can create integrated processes alone at an IHE. AIM incorporates Collaboration because collaboration across departments and units, such as the assessment office, institutional research, faculty governance bodies, and

curriculum-decision-making bodies and disciplines, can facilitate the sharing of best practices and promote a cohesive approach to integration. Another unit hosted by many institutions, a center for teaching and learning, or a similarly-named unit, is a centralized space for faculty development in pedagogy that can be another key collaborator in integrated assessment. The importance of collaboration across the institution is echoed in reflections on the AAHE's Principles (Astin et al., 1992; Hutchings et al., 2012). The previous interview-based study revealed that collaboration will vary from institution to institution.

Accreditation

The final factor in AIM is Accreditation. The impact of accreditation on integration is an important driver in that it provides the environment in which assessment processes occur within IHEs. However, when narrowly perceived as the single objective of assessment, accreditation becomes a double-edged sword for integration. Accreditation standards generally go beyond recommending implementing isolated assessment processes by emphasizing informing improvements with assessment.

Note that AIM does not assume the directionality of impacts. So, the impact of work on the development of curricula on Collaboration is considered in the same way that Collaboration impacts the work done on the development and revision of curricula. The interview study demonstrated that these factors can all work as facilitators and hindrances to implementing integration; for example, a lack of collaboration between a centralized assessment office and a faculty committee working on general education might reduce the effectiveness of the impact of assessment on general education.

Research Aim

Following the previous in-depth interviews, this survey inquiry's goal was to gather further information about the integration of the assessment process with curriculum and instruction from a wide body of IHEs. In this manuscript, we detail additional qualitative information to contextualize the details generated in the interview study and will explore quantitative information in further writing.

Method

Data Sources and Methodology

The insights presented in this paper are derived from qualitative comments in response to open-ended questions from an instrument designed to measure the respondents' impressions of the impact of each AIM factor on their institutions' integration status. While no quantitative data collected by the closed-response items to the instrument will be discussed in this manuscript, the instrument was iteratively designed, beginning with the original research team generating a large pool of potential items. These items were reviewed by a team of 5 external assessment experts to identify the extent to which the content of the constructs being identified (i.e., the factors in AIM) were captured by the items. Based on the response from these experts, some items were deemed unnecessary and therefore removed from the pilot instrument. For example, in the "Assessment Purposes" category, the original items included "Student learning assessment informs instruction and curriculum," "Faculty use student learning assessment information to make evidence-based changes to improve their instruction,"

and “Faculty use student learning assessment information to make evidence-based changes to improve curriculum.” These were deemed redundant as a group of three, and the first was removed from the final instrument. A pilot administration of the instrument was then conducted to establish response processes, resulting in changes to the scales used to measure individual quantitative items. The final instrument consisted of a series of closed-ended, Likert-type items about specific aspects of each AIM factor, with open-ended, qualitative items included for each factor. A final question on the survey asked respondents to reflect not only on integration but also on their professional development needs in terms of integration.

The survey was distributed from 5/17/2023 to 7/10/2023 via Qualtrics software. The survey was administered via the ASSESS listserv, a United States-based group of assessment professionals, faculty, administrators, and staff, hosted by AALHE and the University of Kentucky. The survey was shared with weekly reminders sent to the entire listserv each week after the initial posting until the survey closed. Ultimately, 150 responses were collected. Fully null responses (N = 59) were removed, resulting in 91 analyzed responses. These responses offer a panoramic view of the current state, challenges, and potential of integration in U.S. higher education because they were collected from diverse academic stakeholders.

Respondents’ experiences with integration were broadly represented by a wide range of IHEs (specifically targeted, although not limited, to IHEs in the United States). Respondents had an average of 9.5 years of experience, with a range of 0 - 25 years, a mode of five, and a median of eight. They represented administrators, faculty, staff, and those in faculty/administrator roles. Respondents represented institutions from below 1,000 full-time equivalency (FTE) students to above 20,000, with the largest group representing institutions between 1,000 and 4,999 FTE. Public and private, 2-year, 4-year, and for-profit institutions were all represented by respondents, with the majority being 4-year and above. All of the institutional (formerly regional) accreditors were represented, with the majority of respondents coming from the Southern Association of Colleges and Schools Commission on Colleges (SACSCOC) and the Higher Learning Commission (HLC) regions, aligned with the national distribution of institutions in the U.S. at the time of survey administration. The institutions’ Carnegie classifications spanned associate’s, associate’s/bachelor's, bachelor's, professional, master’s, doctoral, and special focus. Therefore, the sample represented a more diverse body of institutions than the original interview study, which nearly exclusively included large public 4-year institutions; however, there is still a smaller representation from the 2-year sector in these data.

The Present Investigation

Analysis Methods

The researchers engaged in inductive coding to identify the key themes within responses to the open-ended items affiliated with each factor of AIM (Creswell & Poth, 2018). Responses to each item were coded by two independent reviewers, who listed codes in shared codebooks. Discrepancies between reviewers were discussed to a consensus. These codes were refined into themes. Simultaneously, a generative AI tool (ChatGPT 3.5, OpenAI, 2023) was engaged to perform a thematic analysis. The results from both independent reviewer notes and AI notes were compared and evaluated to synthesize insights gained from the responses. To identify any threats to the trustworthiness via

confirmability of the analysis as the result of reviewer bias (Shenton, 2004), member-checking in the traditional sense (Hoffart, 1991) could not be completed as the respondents were anonymous. Because all members of the research team represented respondents' professional roles as assessment professionals and/or faculty, they engaged in a checking procedure, wherein final insights from the two-reviewer team assigned each question were reviewed with other members of the research team.

ChatGPT Data Analysis

Given the novelty of applying an AI to data analysis in the scholarly literature at the time of this publication, it is important to describe the process used for this study. Incorporating ChatGPT as a tool for thematic analysis necessitated a careful balance between the exploratory nature of qualitative research and the structured rigor expected in scholarly analysis. For our purposes, this integration was guided by principles rooted in the Critical Reflective Practitioner (Thompson, S. & Thompson, N., 2023) approach, which prioritizes empathy, reflexivity, and cultural responsiveness while adhering to established methodological standards. The core objective was to direct GPT in a manner that facilitated meaningful engagement with qualitative data while systematically uncovering patterns and themes without prematurely imposing rigid analytical frameworks. By leveraging thoughtful prompts, ChatGPT iteratively refined its analysis, providing nuanced and credible interpretations aligned with recognized qualitative research practices (e.g., Maxwell, 2013; Creswell & Poth, 2018).

To achieve this, the design of the prompting process focused on key aspects of qualitative data analysis such as data familiarization, initial coding, theme development, and reflection on bias and credibility. Each analysis stage was explicitly framed through task-specific prompts to ensure that ChatGPT's responses aligned with rigorous qualitative standards. These prompts guided GPT step-by-step through coding, identifying themes, and refining them, allowing for both structure and flexibility within the interpretive process. A notable feature of this approach was the deliberate avoidance of rigid frameworks at the outset, instead fostering an iterative analytical process to enhance depth and reliability.

One of the central strategies involved instructing GPT to adopt the persona of a Critical Reflective Practitioner. This persona emphasized holistic, human-centered analysis, balancing empathetic engagement with critical inquiry and reflection. Such an approach was essential for uncovering insights that respected the complexity of the informants' lived experiences while remaining grounded in the practical objectives of the research. Within this framework, ChatGPT was encouraged to reflect on its interpretations, explore alternative perspectives, and account for possible contradictions in the data, fostering a more nuanced and comprehensive analysis.

Providing clear context and background information was another essential component of the prompting process. Detailed descriptions of the research objectives, background institutional and informant characteristics, and data type (i.e., survey) ensured that GPT had the necessary contextual understanding to prioritize and interpret information accurately. This clarity enabled GPT to focus on relevant aspects of the data, avoiding misinterpretation or tangential responses. Structured prompts further enhanced this focus by specifying the exact task at each stage of the thematic analysis, such as initial coding or synthesizing themes, thereby preventing vague or unfocused outputs.

Reflection was embedded as a critical part of the process, with prompts encouraging the GPT to consider potential biases and alternative interpretations. These sub-prompts guided the GPT to critically evaluate its outputs, question its assumptions, and explore contradictory evidence. By integrating reflective practice into the analysis, the GPT provided deeper, more robust insights while ensuring no significant perspectives were overlooked. Culturally responsive framing further enriched this reflective component, wherein prompts were designed to acknowledge diverse perspectives and power dynamics. Although the data pertained to institutional characteristics rather than individual experiences, culturally responsive instructions were included as a matter of thoroughness.

The iterative nature of the thematic analysis was another key factor in ensuring rigor and depth. Prompts were designed to encourage ChatGPT to revisit and refine its outputs continuously. This iterative revision process allowed for ongoing evaluation of initial codes and themes, with attention to internal coherence, interrelationships among themes, and gaps in the analysis. Through this recursive approach, ChatGPT made necessary adjustments based on emerging insights or contradictory evidence, thereby enhancing the overall quality and credibility of the analysis.

The design and execution of this approach reflect an intentional effort to integrate artificial intelligence into qualitative research in a way that honors the complexity and rigor of human-centered inquiry by thoughtfully structuring prompts, fostering reflexivity, and embedding culturally responsive practices.

Results

Respondents identified the challenges and successes they had experienced concerning each of the identified AIM factors. What follows is a key-point summary of the themes and codes identified in the analysis process described above under each AIM factor, presented in the same order as the factors of AIM are presented above and as they were presented to the respondents in the survey. The evidence collected in this study demonstrates only the respondents' self-described experiences rather than a ranking of the importance of each factor in the respondents' assessment experience. This survey did not ask respondents to rank their responses or the importance of the factors in their experiences with integration.

Leadership Support

Open-ended responses offered by survey respondents (N = 20) related to leadership indicated that respondents believed that their institutional leaders thought the purpose of assessment and its value was for accountability and accreditation. In general, there was little evidence of institutional leaders understanding any value of assessment beyond accreditation. Several respondents reported that associate provosts (i.e., not senior, but high-level leaders) in charge of curriculum understand the value of integrating curriculum, instruction, and assessment practices more than other institutional leaders. Although a limited number of respondents mentioned receiving support for the integrated process, typically they suggested that institutional leaders had little understanding of or support for an integrated process or its potential. Furthermore, respondents indicated that leaders offered little assessment resource allocation short of a basic administrative structure for assessment, with one respondent stating, "Outward support. Little resource

allocation has been my experience." However, a few respondents indicated that they were provided adequate staffing.

More than one assessment respondent mentioned political issues on campus that excluded them, with one even stating, "No, I, as the Assessment Director, am literally barred from even being in the room" when conversations or meetings related to curriculum and instruction arose. Although not all respondents were pessimistic, one suggested that a greater understanding of the assessment's value may occur with the forthcoming leadership change.

The respondents' articulated Leadership Support challenges were conceptualized by the authors as a beautifully packaged gift that is revealed to be empty inside. Leaders may speak of support for assessment, but few follow through with resource allocation aligned with this spoken support, as represented in Figure 3.

Note. Image conceptualized and described by the authors and generated by CoPilot (Microsoft, 2025).

Figure 3. *Leadership Support challenges articulated by respondents, as conceptualized by the authors.*

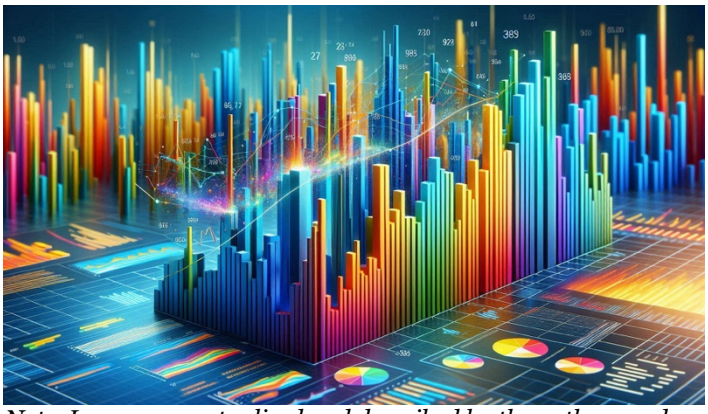
Assessment Purposes

A recurring theme among respondents was that the purpose of assessment at their institutions remains accountability, with a primary focus on data collection or a "box-checking" approach grounded in a positivist mindset. Some respondents assumed that integration of assessment was taking place and that sometimes data is used to make improvements in specific departments or programs. However, these respondents were housed in institutions without a robust information-sharing system or a central assessment office for information about improvements to be delivered. Many did not think their institutions had established a broad culture of integrating assessment with curriculum and instruction. Only one response commented on having a strong culture of assessment (N = 24).

Respondents indicated that the purpose of assessment varied widely across departments and programs within their institutions, leading to inconsistent practices. The extent to which faculty and programs integrated assessment into curriculum and instruction also differed: some were actively participating, and others either lacked involvement or did not fully understand the purpose of assessment. One respondent

indicated that some faculty misunderstood the concept of "closing the loop," focusing on changing assessment methods rather than using data to improve curriculum. Most respondents suggested that while curricular improvements were relatively more feasible, instructional changes posed greater challenges.

The authors conceptualized the Assessment Purposes challenges articulated by the respondents as inconsistency, with variations within and between institutions. This is conceptualized in Figure 4 as bars of varying heights.



Note. Image conceptualized and described by the authors and generated by CoPilot (Microsoft, 2025).

Figure 4. *Assessment Purposes challenges articulated by the respondents, as conceptualized by the authors.*

Assessment Processes

Respondents commented on their institutional experiences with assessment processes' impacts on integration (N = 16). The responses provided a general insight that assessment processes and integration happen at most institutions, but not in concert.

The distinct themes that emerged from the item about processes were: the uneven distribution of information about processes among levels of the institution; senior leadership changes impacting process engagement; and, the involvement of the assessment office and center for teaching and learning (or similarly-named units) as providing professional development as well as facilitating the processes.

The most prevalent theme was the uneven distribution of adherence to assessment and integration processes within institutions. One respondent exemplified this by indicating that the impacts of processes were "definitely true for some faculty/programs, but not all." Another shared additional context: "…programs are asked to share and discuss relevant assessment findings with the appropriate stakeholders, and their (sic) is some evidence in the assessment reports that programs are doing this, but it's difficult to determine if this is happening systematically and with which stakeholders…" This sentiment addresses one of the key needs of integrating assessment with curriculum and instruction: communication between all groups involved. Some responses detailed that the role of leading assessment was distributed between the colleges and assessment offices.

While one theme revolved around the consistency of process implementation, another challenge faced was the role of the centralized assessment office in accessing information about process implementation or integration in general. For example, one

respondent indicated that "the work is being done, but it is not across and transparent at the institutional level," others discussed review processes in which reading and evaluating assessment reports is a distributed model across colleges and individual assessment offices, without centralization. On the other hand, centralized assessment offices might suffer from a lack of engagement with embedded units: "There is a one-person office that facilitates program learning outcome assessment, including organization and facilitation of the use of the data, but has no role in *actual* use of the data." This kind of "siloing" is noted across the results for all items, particularly in this theme. Based on these responses, the "right hand" of institutional assessment does not necessarily know what the "left hand" embedded within curricula and classrooms is doing.

Another theme that emerged was that assessment processes, even if well-established, might not be resilient to changes in senior leadership. While integration is impacted by assessment processes, assessment processes are similarly impacted by various contextual factors. The respondents indicated that a change in leadership has an impact, sometimes detrimental, on the progress of integration at their institutions.

Overall, Assessment Processes challenges articulated by the respondents can be conceptualized in the common metaphor of the silos found at most IHEs, demonstrated in Figure 5.

Note. Image conceptualized and described by the authors and generated by CoPilot (Microsoft, 2025).

Figure 5. *Assessment Processes challenges articulated by the respondents, as conceptualized by the authors.*

Recognition & Reward

Of the responses about Recognition & Reward ($N = 14$), the consensus was that the institution does not provide these for assessment. However, assessment may be recognized as a portion of other rewards given to faculty members. For example, one respondent stated, "Department-level assessment coordinators are given a full or partial course release for their assessment work, and many of them do work with their departments' curriculum committees." In certain cases, "non-monetary incentives" have either been used in the past or included in some units but not all. There was also a plurality of respondents looking toward the future for recognition & reward. For example,

"resources are not dedicated in this way, but the desire is there to change that." Moreover, one reported, "We do not do these things currently, although we should do more," and another stated, "I feel like faculty at my institution want to see students learn and see assessment as important. It would help if it was incentivized a little more."

Overall, the Recognition & Reward challenges articulated by the respondents can be described like a child writing a letter to Santa (Figure 6); these respondents have hopes, and many indicate having plans, of future work in recognition & reward or fond memories of past recognition & reward systems without demonstrable recognition & reward at present.

Note. Image conceptualized and described by the authors and generated by CoPilot (Microsoft, 2025).

Figure 6. *Recognition & Reward challenges articulated by the respondents, as conceptualized by the authors.*

Collaboration

While participants (N = 12) seemed to agree on the need for collaboration between assessment and other institutional offices or stakeholders, the current situation, as reported by participants, varied across institutions. One respondent highlighted the collaborative efforts that started during COVID-19 and continued afterward at their institution, involving the assessment area, center for teaching and learning, student affairs, advising, library, and technology departments, resulting in beneficial outcomes such as resources for faculty. Another respondent reported being "a part of everyone's team," indicating widespread collaboration between assessment and other functional areas at their institution. However, other respondents reported a lack of collaboration at their institutions, leading to siloed operations by individual units such as offices, departments, councils, and committees. Several respondents reported that assessment units could be excluded from discussions on curriculum or program reviews. Various factors could have contributed to these siloes, including the "decentralized and territorial" nature of some institutions or a lack of infrastructure for facilitating and supporting such collaborations, particularly in smaller, underfunded community colleges, as pointed out by one respondent. However, despite the historical lack at some institutions, there was hope that collaboration would occur in the future, as expressed by some respondents.

Respondents reported positive movement toward institutional collaboration, occasionally faced with roadblocks (Figure 7).

Note. Image conceptualized and described by the authors and generated by CoPilot (Microsoft, 2025).

Figure 7. *Collaboration challenges articulated by respondents, as conceptualized by the authors.*

Faculty Support and Development

Faculty support and ongoing professional development (PD) are necessary for integration; however, responses (N = 15) indicate that while most institutions offer PD, many have room for improvement and are growing in this area, especially with PD focused on assessment. The level of faculty support and PD varied across responses, from a couple reporting a lack of professional development at their institutions, with most others indicating some PD is available. Some participants commented that voluntary faculty participation in PD is low and that events are not well attended. Some institutions offer PD for integration more formally through offices of teaching and learning, and some through the assessment office. Many participants reported that faculty support and development were continually being improved at their institution, especially after seeing a decline in PD due to the pandemic.

The respondents' experiences with Faculty Support and Development are nascent (just beginning/experiencing low attendance) or reminiscent (recognizing that PD was more widely attended in the first years of COVID-19). The authors conceptualize this as institutions prepared and ready at the starting line for providing holistic PD related to assessment and integration (Figure 8).

Note. Image conceptualized and described by the authors and generated by CoPilot (Microsoft, 2025).

Figure 8. *Faculty Support challenges articulated by the respondents, as conceptualized by the authors.*

Accreditation

The responses to the open-ended item about accreditation (N = 11) showcased the respondents' complicated relationships with the accreditation enterprise as assessment stakeholders. There was a nearly even split of respondents providing a generally positive view of impact and negative, for both professional/specialized and institutional accreditors. The respondents indicated that they believed that both levels of accreditation contribute to integration in the sense that they mandate assessment activities to happen, as articulated by one respondent: "People don't want to do it, but still, without it, there wouldn't be assessment happening." This sensation of mandating assessment did not seem to extend into fully formed models of integration: "The relationship between accreditation requirements and assessment is clear, but less so with the integration of assessment with curriculum and instruction." Therefore, accreditation serves as a foundational aspect that creates assessment activities, which then can be integrated when other institutional environmental characteristics are fit. This item also reiterated that assessment, as an ongoing practice, is not integrated per se with curriculum and instruction, but that integration would be another level of assessment implementation. Respondents did not report this perception (and reality) of assessment activities as a "requirement" of accreditation bodies as a positive feeling among their institutional assessment stakeholders, but begrudging at best.

Overall, the respondents' sentiments related to accreditation were that assessment (non-integrated) would not occur without the influence of accreditors at both the institutional and individual program level. Respondents see assessment as a positive influence, but unfortunately, it is brought out only through the pressure of accreditation expectations, conceptualized as the pressure of a hand squeezing a stress ball to produce assessment processes (Figure 9). The respondents did not feel this pressure was leading to effective integration of assessment with curriculum and instruction.

Note. Image conceptualized and described by the authors and generated by CoPilot (Microsoft, 2025).

Figure 9. *Accreditation challenges articulated by the respondents, as conceptualized by the authors.*

Discussion

In the previous study, all factors within the Assessment Integration Model (AIM) were well-developed and contributed positively at the identified IHEs to creating a context conducive to successful integration. The findings of this study are more representative of all involved stakeholders, such as both large and small, 2- and 4-year institutions from all regions of the U.S. The findings suggest that most institutions have well-established assessment processes, consistent with trends observed by other scholars since the 1980s in examining the evolution of assessment practices (Peterson & Augustine, 2000). However, while these institutions possess well-developed assessment processes, our respondents indicate that the processes are often siloed rather than fully integrated at the institutional level.

This lack of integration can be attributed to variations in how factors of AIM, particularly Collaboration and Assessment Purposes, impact the process. Respondents indicated that barriers to collaboration, stemming from institutional roadblocks and internal politics, hinder integration efforts. Specifically, limited collaboration with curriculum-defining bodies, such as faculty committees or individual faculty members completing assessment reports, at both institutional and college levels, has reduced awareness of integration activities among assessment practitioners. The extant processes for assessment are only enacted due to accreditor requirements, per the respondents, rather than for the purpose of continuous quality improvement.

Leadership Support emerged as the most frequently mentioned critical factor from AIM that has stagnated progress toward fully integrated systems. Respondents highlighted leadership changes as a significant challenge to sustainability. Institutions appear to benefit most from context-specific strategies to initiate integration, with respondents consistently emphasizing the importance of sustained leadership support to ensure that assessment processes are durable and widely implemented enough to achieve integration.

Based on the findings from this study, assessment stakeholders agree that the factors outlined in AIM have provided a foundational structure for integration. However, respondents indicated that integration remains in its early stages. They stated that

integration is not merely a best practice but an essential strategy for realizing the full potential of assessment in enhancing student success.

While many professionals are actively working to promote the value of integration within their institutions, few perceive these efforts as currently successful. Institutional structures and politics present significant barriers to advancing this conversation. For example, siloed reporting practices, where programs submit a single, decontextualized report to a central assessment office, limit opportunities for collaboration across different offices, teams, and committees. This, in turn, hinders the ability to create and implement faculty development initiatives effectively.

Fluctuating impacts on integration come from factors such as Recognition & Reward and Faculty Development. For instance, during the recent pandemic, the shift to remote communication led to increased faculty engagement, but this engagement diminished as institutions returned to a "new normal." Similarly, while some leadership teams have promoted incentives such as rewards or grants for successfully integrated assessment work, leadership changes often shift the institutional political climate, reducing support for resource-based efforts toward integration. Although some departments and faculty are actively using assessment to improve instruction and curriculum, most institutions lack a consistent, cross-disciplinary collaboration structure necessary to expand this practice. Despite these challenges, there is a shared sense of urgency and a strong desire among assessment practitioners to establish systems that effectively integrate assessment with curriculum and instruction. A visual interpretation of these findings in the context of the overall AIM can be found in Figure 10.

While the advantages of integration are clear, implementation is not without challenges. Institutions must navigate issues of stakeholder buy-in, resource constraints, and the need for the skills and knowledge to successfully interpret data into useful information. General resistance to change is one characteristic noted by the respondents, both institutional assessment professionals and faculty members. Institutions' existing practices, such as siloing decision-making around assessment and curriculum and isolating faculty to their instructional context only, can prohibit successful integration. There can be limited faculty engagement in existing assessment structures and processes. The needs of both institutional and specialized accreditors must be interpreted for minimum compliance and balanced with the contextual needs of faculty within the IHE environment. Overcoming these challenges requires a concerted effort to build institutional capacity for assessment.

Based on the perspectives shared by the respondents, we argue that the compelling value proposition, combined with clear cost benefits, underscores the necessity of integration in enhancing educational quality, ensuring institutional effectiveness, and promoting student success. Here, we offer strategic recommendations to leverage the factors of AIM to best cultivate an environment where integration can thrive.

Note. Sub-images conceptualized and described by the authors and generated by CoPilot (Microsoft, 2025).
Figure 10. *Assessment Integration Model with visual representations of respondents' challenges within each factor.*

Integration: Value Added

Value Proposition of Integration

Integrating assessment with curriculum and instruction offers institutions a tangible means to demonstrate the benefits of their academic programs. Specifically, by delineating expected student learning outcomes, faculty and IHEs now possess a tangible framework for evaluating educational effectiveness. The value proposition of integration lies in its ability to facilitate evidence-based decision-making processes. While also creating an environment for making decisions rooted in data, systematically collecting and analyzing data allows IHEs to showcase the concrete learning outcomes achieved by students. This focus on demonstrable student learning ensures graduates are prepared for their chosen fields and life's multifaceted challenges. Furthermore, well-integrated assessment, curriculum, and instruction systems enable institutions to remain agile and align their educational offerings more closely with industry standards, societal needs, and, most importantly, student needs. Integration can be transformative for IHEs in how curriculum and instruction are delivered to center students and their learning experiences with cultural responsiveness (Montenegro & Jankowski, 2017).

Cost Benefits and Institutional Advantages

Beyond the educational merit, integration presents significant potential in cost benefits. In the responses provided by survey respondents analyzed here, it is clear that some institutional stakeholders (e.g., faculty, assessment professionals) perceive and often have assessment work as an "add-on" to existing responsibilities. In addition, at most institutions, curricular decisions happen via committees and other gathered bodies of faculty and staff. By leveraging assessment data for curricular and instructional decisions, institutions can optimize resource allocation in the time and talents of those participating in such assessment work and committee sessions, ensuring financial prudence. Despite the initial investment required for developing and implementing assessment processes, the long-term benefits include improved educational quality, increased institutional accountability, and enhanced student learning experiences and satisfaction. Moreover, these processes can enhance an institution's accreditation standing, potentially unlocking additional funding avenues.

Recommendations

We now provide strategies for effective implementation, resource allocation, leadership involvement, and continuity to address these challenges. These recommendations are rooted in IHEs' responsibility for their students' learning and success rather than accountability to external accreditors or regulatory bodies.

For example, we recommend that assessment leaders and practitioners engage in the following activities as part of daily practice to help promote integration:

- Educate institutional leaders about the benefits of integrating planning and implementation processes as described in AIM, which may require building partnerships with others who have more direct contact with the leaders. Build knowledge about the benefits of the integration in academic chairs' training.
- Build collaborations with offices and units, such as student affairs, centers for teaching and learning, and the library, to inform them about integration and to highlight benefits to student learning, academic programs, and the institution.
- Build curricular integration into assessment reporting structures. When faculty members and staff are required to submit reports, respondents indicate that they perceive these reports to be decontextualized and complete the reports due to the requirement, so incorporating the aspects of curriculum and instruction improvement into the required report can start to build a culture of considering these other components with assessment.
- Build assessment infrastructure into the institutional fabric and culture so the integrated system survives leadership change. Design succession activities for integration to be beneficial and sustainable when assessment leadership within the IHE changes.
- Design and maintain professional development and encourage/promote sustainable recognition and reward structures within the institutional context.
- Engage in a collaborative and guidance approach to the assessment leadership role.

The work of reinvigorating IHE assessment with the opportunity to create integrated environments cannot be and is not solely the responsibility of individual assessment practitioners, many of whom, as indicated in this study, operate in relatively low-resourced and often siloed offices. It is clear from the literature and the responses here that external accrediting bodies' influence is the most influential on the progress of integration at their institutions, as it informs the Assessment Purposes and Leadership Support. Leaders and representatives (i.e., peer reviewers) from the bodies through which regulation is enacted in higher education, the accreditors themselves, can create an environment where there is strong encouragement for institutions to have an infrastructure for sustainable assessment practice. Expectations from accreditor staff and peer reviewers can promote the resiliency of institutions' assessment systems that can survive changing leadership expectations. Individuals affiliated with the external accrediting bodies are positioned to provide strong support for assessment professionals in their quest to develop an integrated system via recommendations for enhancement, especially in cases where leadership is frequently changing or does not strongly emphasize assessment.

Conclusion

The Assessment Integration Model offers a holistic framework emphasizing the interdependence of Leadership Support, Assessment Purposes, Assessment Processes, Recognition & Reward, Faculty Development, Collaboration, and Accreditation. By optimizing the support of and reducing the impact of obstacles provided by each aspect of the model, IHEs can create a culture that values and prioritizes assessment processes as a tool for enhancing student learning via classroom instruction, improving educational quality via curricular enhancement, and achieving strategic institutional goals by creating a culture of continuous quality improvement. As U.S. higher education continues to evolve, the role of assessment as a cornerstone of academic excellence and institutional integrity becomes increasingly evident. Integrating assessment into the fabric of curriculum and instruction brings an IHE to a level beyond meeting accreditation requirements or administrative checkboxes, genuinely understanding and enhancing student learning. This establishes a framework for longer-term learning improvement success and offers a promising pathway to enhance quality and accountability. This qualitative data analysis, utilizing data from assessment stakeholders, demonstrated a high degree of urgency and desire to create integrated systems within their home institutions, driving initiatives despite challenges and in collaboration with leadership. Through a concerted focus on these strategic areas, higher education institutions can achieve comprehensive and sustainable integration. The future of higher education, marked by accountability and a relentless pursuit of excellence, will undoubtedly see integration as a pivotal instrument in the symphony of academic achievement.

References

Astin, A. W., Banta, T. W., Cross, K. P., El-Khawas, E., Ewell, P. T., Hutchings, P., Marchese, T. J., McClenney, K. M., Mentkowski, M., Miller, M. A., Moran, E. T., & Wright, B. D. (1992). Principles of good practice for assessing student learning. American Association for Higher Education (AAHE) Assessment Forum.

Banta, T. W., & Blaich, C. (2010). Closing the Assessment Loop. Change: The Magazine of Higher Learning, 43(1), 22–27. https://doi.org/10.1080/00091383.2011.538642

Colton, D. (2000). Quality improvement in health care: Conceptual and historical foundations. Evaluation & the Health Professions, 23(1). https://doi.org/10.1177/01632780022034462

Creswell, J. W., & Poth, C. N. (2018). Qualitative inquiry and research design: Choosing among five approaches (4th ed.). Sage. ISBN: 978-1544398396.

Ewell, P. T. (1991). Assessment and public accountability: Back to the future. Change, 23(6), 12–17. https://doi.org/10.1080/00091383.1991.9940589

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

Flateby, T., Sweetland, Y., Ghasemi, A., Gibbons, R. E., Hardesty, K., & Hernández, A. (2023, June 5-8). Are you integrating assessment, curriculum and instructional practices and processes? [Conference presentation]. Association for the Assessment of Learning in Higher Education Conference 2023, New Orleans, LA.

Flateby, T., Sweetland, Y., Hardesty, K., Hernández, A., & Gibbons, R. E. (2024, June 3-6). Assessment integration with curriculum and instruction: Results of a national survey. [Conference presentation]. Association for the Assessment of Learning in Higher Education Conference 2024, Portland, OR.

Gaston, P. (2018). Assessment and accreditation: An imperiled symbiosis. (Occasional Paper No. 33). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Henning, G., Baker, G. R., Jankowski, N. A., Lundquist, A. E., & Montenegro, E. (Eds.). (2022). Reframing assessment to center equity: Theories, models, and practices (1st ed.). Stylus Publishing, LLC.

Hoffart, N. (1991). A member check procedure to enhance rigor in naturalistic research. Western Journal of Nursing Research, 13(4), 522–534. https://doi.org/10.1177/0193945991013004

Hutchings, P., Ewell, P., & Banta, T. (2012). AAHE principles of good practice: Aging nicely. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Jankowski, N. A., Timmer, J. D., Kinzie, J., & Kuh, G. D. (2018). Assessment that matters: Trending toward practices that document authentic student learning. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Kinzie, J. (2019). Taking stock of initiatives to improve learning quality in American higher education through assessment. Higher Education Policy, 32, 577–595. https://doi.org/10.1057/s41307-019-00148-y

Kuh, G., & Ewell, P. (2010). The state of learning outcomes assessment in the United States. Higher Education Management and Policy, 22(1). https://doi.org/10.1787/hemp-22-5ks5dlhqbfr1

Maxwell, J. A. (2013). Qualitative research design: An interactive approach. Thousand Oaks, CA: SAGE Publications.
Microsoft Copilot. (2025). Microsoft visual creator. https://m365.cloud.microsoft/
Montenegro, E., & Jankowski, N. A. (2017). Equity & assessment: Moving towards culturally responsive assessment. (Occasional Paper No. 29). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).
National Institute of Education, Study Group on the Conditions of Excellence in American Higher Education. (1984). Involvement in learning: Realizing the potential of American higher education (ED publication 246833). Washington, DC: Department of Education.
OpenAI. (2023). ChatGPT (Feb 13 version) [Large language model]. https://chat.openai.com
Pepin, C. K. (2014). The dilemma of assessment in the US. In Q. Li & C. Gerstl-Pepin (Eds.), Survival of the fittest (pp. 85–100). Springer. https://doi.org/10.1007/978-3642-39813-1_6
Peterson, M. W., & Augustine, C. H. (2000). Organizational practices enhancing the influence of student assessment information in academic decisions. Research in Higher Education, 41, 21–52. https://doi.org/10.1023/A:1007038212131
Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. Education for Information, 22, 63–75. https://doi.org/10.3233/EFI-2004-22201
Thompson, S., & Thompson, N. (2023). The critically reflective practitioner. United Kingdom: Bloomsbury Publishing.
Webber, K. L., & Tschepikow, K. (2011, May 21-25). Learner-centered assessment: A comparison of faculty practices in US colleges and universities 1993 to 2004. [Conference presentation]. AIR Forum, Toronto, Canada.
Wiliam, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? British Educational Research Journal, 22, 537–548. https://doi.org/10.1080/0141192960220502

A New Approach to Learning Improvement: Starting with an Intervention

Authors:

Laura A. Lambert
James Madison University

Megan R. Good
James Madison University

ABSTRACT

In higher education, institutions routinely assess student learning in degree programs driven by institutional accreditation requirements. Assessment is intended to provide an avenue for improvement with collected data inspiring curricular change and future assessments providing evidence of the efficacy of those changes (i.e., learning improvement). However, evidence of actual learning improvement resulting from assessment is rare, partly because learning improvement projects are resource-intensive, requiring significant faculty time and potentially departmental funding. In this article, we explore a novel approach to learning improvement by using existing data in a STEM undergraduate program. We identified 11 student learning objectives aligned with a previously implemented curricular change. Trained faculty used a common rubric to evaluate student work submitted both before and after the curricular change was implemented. We found evidence of improved student learning. Our findings contribute to the definition of learning improvement generally and in STEM, noting the importance of resource considerations.

Correspondence E-mail: laycocla@jmu.edu

Keywords: Learning Improvement; Curricular Change Assessment; Historical Data Analysis; STEM Education Outcomes; Program-Level Student Learning Objectives

The higher education assessment cycle articulated by Erwin (1991) includes the articulation of student learning objectives, formulating an assessment design, collecting and analyzing data, and using results. The primary goal of assessment is to use results to improve student learning. In the mid-1990s and early 2000s, states and institutional accreditors began requiring assessment of student learning as a means of accountability (Ewell, 2009). These dual purposes (i.e., improvement and accountability) seemed to be in tension, even though the accreditation standards required *improvement* (Smith et al., 2015).

Fast-forward to today and assessment for accountability is evident. For example, among institutions accredited by the Southern Association of Colleges and Schools – Commission on Colleges (SACSCOC) in 2013, 21% of institutions were found non-compliant with the degree program assessment standard; in 2023, this figure fell to 6% (Matveev, 2014, 2024). In addition to institutional accreditation requirements, disciplinary accreditors require assessment. As an example, ABET, a disciplinary accreditor for engineering and applied natural science programs, has Criterion 4: Continuous Improvement (ABET, 2021b).

Evidence of improved student learning resulting from assessment, however, is lacking. Indeed, Banta and Blaich (2011) found very few examples of learning improvement – that is where programs assessed student learning, made targeted changes, and monitored learning over time to determine if the change was indeed an improvement. How could this be!? One explanation is that the definition of improvement was unclear.

In 2014, Fulcher et al. presented a clear definition of learning improvement through their seemingly 'simple' model: one must assess student learning at the degree program level (Assess), make a change (Intervene), and then re-assess learning to determine if the change was indeed an improvement (Re-assess). However, learning improvement as defined by Fulcher et al. (2014) is resource intensive.

Specifically, learning improvement projects require energy from multiple faculty members given the focus is on programmatic outcomes. One faculty member – a champion – is necessary to provide momentum and stamina (Fulcher & Prendergast, 2021). Often, assessment practitioners and educational developers from outside the program are needed to support this work (Smith et al., 2018). Last, learning improvement projects often take years to implement. It's a big lift! The resources needed to initiate a learning improvement project are problematic given that faculty capacity following the COVID-19 pandemic is increasingly limited (McClure et al., 2023). And yet, care and attention to increasing educational quality is important work that must be done.

In this paper, we explore an approach to learning improvement that challenges Fulcher et al.'s (2014) model breakdown where a program lacks baseline data. Indeed, if a program has consistently gathered student work (i.e., data) without assessing it, there is an opportunity to retroactively assess the efficacy of a curricular change with significantly fewer resources than beginning a new learning improvement project.

Background

Theoretical Framework – Learning Improvement

As noted, Fulcher et al. (2014) provided a 'simple' model for learning improvement: 1) Assess student learning at the degree program level, 2) Intervene, or do something differently in the curriculum, and 3) Re-assess future cohorts of students in the

program to determine if the changes were indeed improvements. While conceptually simple, it is tough to pull off. Often, programs have one or two of the three components, but rarely all three in coordination.

Highlighting this, Fulcher et al. (2014) describe three ways the simple model can break down:

1. *"~~Assess~~, intervene, re-assess*. For this breakdown, faculty make a coordinated curricular change. Following implementation, they assess student learning. While significant, Fulcher et al. note that this is not true learning improvement given the lack of baseline learning.
2. *Assess, ~~intervene~~, re-assess*. Here, a program dutifully assesses student learning annually without making curricular changes. Essentially, the focus is on the measurement of learning rather than changes to the curriculum.
3. *Assess, intervene, ~~re-assess~~*. In this scenario, a program assesses student learning and makes intentional curricular changes. However, they do not assess cohorts of students who experienced the curricular change to determine if the change was an improvement. Given the years it often takes to implement changes that span a curriculum, the assessment component can get lost" (p.6).

Completing all the steps of the simple model is challenging on multiple fronts, not only to initiate, but to sustain and see to completion. Indeed, successful projects require significant human resources: multiple faculty members from the program, a faculty champion who leads the charge, assessment practitioners to refine measures (e.g., ensure measures are sensitive enough to capture change) and educational developers to collaborate on scaffolded curricular change. Specifically, the faculty champion is one who not only 'lead the charge', but often contributes sustaining energy to the effort as well as coordinating other needed experts (Fulcher & Prendergast, 2021). Additionally, because creation and implementation of changes to a curriculum can take years to implement, getting results to determine if the changes were improvements is a long-term time investment.

Since the publication of the simple model in 2014, a few programs have successfully engaged in a learning improvement project. A Learning Improvement Community was formed in 2017 (*Learning Improvement Community: About Us*, 2024) and includes nine learning improvement stories (*Learning Improvement Community: Stories About Learning Improvement*, 2024). One of the first examples of learning improvement was James Madison University's Computer Information Systems, where the program dramatically improved students' ability to elicit requirements during consultations with clients (Lending et al., 2018). The project took several years to complete, included a core team of four faculty members, an assessment practitioner, an educational developer, and was supported by the program's administrative team (i.e., department head). This example illustrates the amount of time and resources needed for a successful learning improvement project. Considering that many faculty are still feeling the effects of the disruptions of the COVID-19 pandemic (i.e., rapid and sudden shifts in teaching modality, general uncertainty, etc.), faculty capacity for such a resource-intensive effort is often very low.

Given resource limitations, we challenge Fulcher et al (2014)'s first model breakdown (~~*Assess*~~*, intervene, re-assess*). In this scenario, a program has no baseline assessment data, but has successfully implemented a major curricular change. This is the situation in which the lead author found herself. However, this was not considered a dead-end for learning improvement, given a host of positive situational factors. First, capstone student work had been retained from before, during, and after the curricular change. Second, the academic unit head expressed interest in engaging in learning improvement. Third, the faculty noted they were curious about the impact of the prior curricular change, and also indicated they had limited capacity to engage with anything new. The purpose of this paper is to explore learning improvement with existing data to determine if a prior curricular change was effective. If successful, this approach could provide evidence of learning improvement with minimal faculty resources.

Program of Interest: Integrated Science and Technology

The Integrated Science and Technology (ISAT), BS program is situated in an R2 institution in the mid-Atlantic region. The unit head expressed interest in engaging in a learning improvement project. Before initiating the project, the authors used a *Readiness for Learning Improvement* tool – a resource created by the aforementioned Learning Improvement Community (Lambert & Good, 2024). This tool was used to gather perspectives from faculty members about their readiness to engage in a resource-intensive project by considering the capacity, culture, and commitment of the program.

Ultimately, the tool highlighted that the program had a strong culture for learning improvement and a high level of commitment from department leadership – two of the three key ingredients for learning improvement. Unfortunately, faculty clearly expressed that they did **not** have the capacity to engage in a project. At the same time, they expressed an interest in better understanding the impact of a prior curricular change that was implemented over the past several years. This interest inspired the current project, where we consider a new avenue to learning improvement: the use of existing data. This approach could embrace the positive culture and supportive leadership while limiting resource needs.

Prior Curricular Change

In 2010, ISAT faculty initiated a review of their upper-level curriculum to define and implement elements that would make ISAT graduates uniquely valuable. This process led to the development of 'habits of mind' that would characterize ISAT graduates working in the field years after graduation. The result was a list of objectives and a proposed series of courses dubbed the "Holistic Problem-Solving Spine" (hereafter "Spine"), which focused on complex problem solving and systems thinking. In 2017, this work was integrated into student outcomes, prompting the creation of new courses. During the readiness evaluation process, faculty expressed interest in understanding the impact of this curriculum change, which had been based on prior assessment data.

The Spine was designed to allow ISAT majors to practice a holistic approach to problem solving – a valued way of thinking in the discipline. Importantly, the Spine was designed to encourage students to consider systems as the underlying object of study, to evaluate culture, institutions, nature, and technology as intertwined systems, and to allow for collaborative learning communities. It accomplishes this by giving a structured

exposure to complex problems as well as an introduction to a broader interdisciplinary community. As students advance through the Spine courses, they encounter progressively more open-ended and ill-defined problems, learning methodologies and skills to understand these challenges and define specific areas of inquiry within them.

During the Spine's creation, the faculty created new program-level student learning objectives and aligned the scaffolded courses to these outcomes. Given the intensity of work needed from the faculty to redesign the curriculum, shift credit hours, and design new courses, the process took years. Courses were designed and implemented sequentially, starting with the first-year level classes (100-level courses). The following year, the first-year level courses were revised as needed and the sophomore level classes were implemented, and so on. The first students to have experienced the entire Spine series of courses graduated in 2019.

Faculty noted that while it had been nearly five years since the first graduates experienced the full Spine series of courses, there had been no evaluation of the objectives. Was student learning better after implementation of the Spine? They further noted that in the intervening time, one of the courses had been altered from how it was originally imagined. While acknowledging the major impact the COVID-19 pandemic had on the ability to gather accurate representations of student work, faculty still felt as though this major curricular change should be assessed.

Data situation

Since its inception, the ISAT program has had students research, write, and present a senior capstone project. Students work independently or in groups during their senior year to complete these projects, and they have been a much-celebrated hallmark of the program. We were able to determine that the written capstone papers have been saved electronically since the program's start (1996). While examples of learning improvement projects are rare in the literature, what examples do exist (e.g., Good, 2015; Learning Improvement Community, 2022; Lending et al., 2018; Smith, 2017) have used data gathered during the course of the project. These stories follow the Fulcher et al. (2014) format for documenting learning improvement (i.e., Assess-Intervene-Re-assess).

To evaluate the efficacy of an intervention, it is important that consistent student products are available both before and after the intervention. In our case, we had just that: historical data (i.e., student capstone projects) from before, during, and after the creation and implementation of the Spine. The purpose of this paper is to engage in a learning improvement project using the available historical data. To date, learning improvement projects have been future-oriented, and the authors are unaware of existing literature on learning improvement projects using historical data. Focus on the impact of a prior curricular change – if successful – can save time and provide meaningful evidence to stakeholders.

Method

Student work

Since the beginning of the program in 1996, capstone research has been designed such that students in the ISAT program work with a faculty advisor to complete their research activities. Given the integrated nature of the program, as well as the different concentration options, student projects are varied in scope and topic. Additionally, the

faculty advisors ultimately grade the capstone papers, each coming from diverse content backgrounds. This diversity is reflected in the papers: topics can range from social science to bench science, with formats ranging from more typical prose to a paper that could be published in a scientific journal.

We gathered capstone papers from both before and after the creation of the Spine, focusing on the time periods that were definitively "pre" and "post". We defined pre-Spine as 2010 and earlier, since the creation and implementation of the Spine was a multi-year process, with courses being added in a staggered fashion. Additionally, we defined post-Spine as 2019 and later, with 2020 being excluded from the analysis due to the major interruptions of the pandemic. Capstone papers from 2019 represent the first class to have fully participated in the Spine sequence of classes, and all future cohorts will have fully participated. In addition to papers from 2019 being included in the analysis, papers from 2021 and 2022 were also included. Papers from 2023 were not yet available at the time of rating.

Papers were deidentified to minimize any indication of whether the paper was pre-Spine or post-Spine. Specifically, student names were redacted. Additionally, if an "Acknowledgements" section was present in the paper, that was also redacted to remove references to advisors. That said, citation dates, tools, and software referenced throughout the papers likely provided hints as to when a paper may have been written. Raters were instructed to ignore any such information and to strictly rate the paper on the objectives being examined and the rubric being used.

Measure

A rubric was created to evaluate a subset of the Spine learning objectives (see Appendix A); only objectives aligned with the capstone paper were included in the rubric. A previous rubric used for program assessment served as the baseline for the rubric used in this study, which has tighter alignment with the Spine objectives. We acknowledge that not all Spine objectives can be assessed from capstone papers. When training raters, we asked that they only evaluate the elements on the rubric (e.g., do not evaluate writing quality). Several suggestions from the faculty raters were incorporated into the final version of the rubric used. Given that our interest was in the mean difference between pre-Spine vs. post-Spine scores, exact rater agreement on each item was less of a concern than was rater consistency.

Rater recruitment

Faculty from the department were invited to serve as capstone paper raters during a summer rating session. Raters were in-person for one full day of rater training followed by a time of asynchronous rating. A total of five faculty members agreed to serve as raters and participated in rater training. During the rater training, raters had the opportunity to examine the rubric and the objectives being examined, to have any questions clarified, and to suggest changes or improvements to the rubric. After faculty were acquainted with the rubric and clear on the interpretation of each prompt, they used it to rate two different capstone papers, with group discussion happening between each rating session.

After training, the capstone rating sessions were asynchronous, and raters managed their time within the agreed-upon rating window. Faculty were paid an hourly rate for their time in recognition of work performed over the summer, outside of the

Table 1. *Objectives Evaluated During the Capstone Rating Process*

	Objective
A	Describe the systems out of which a defined complex problem emerges, including interactions between the social, cultural, natural, and technological forces.
B	Evaluate the structure of a system to explain how the system and its components affect the complex problem.
C	Chronicle the history, evolution, and current manifestations of a complex problem.
D	Relate historical events and illustrate how these influence the trajectory of the current system.
E	Use an appropriate temporal scale to evaluate solutions.
F	Account for the governance contexts and constraints that impact a complex problem's stakeholders.
G	Identify avenues and processes for change within a given governance context.
H	Account for stakeholder perspectives, including cultural and institutional power differentials, to inform decision-making.
I	Apply relevant and credible sources of information to address a complex problem and its dimensions.
J	Select an appropriate method to define, analyze, and solve a complex problem.
K	Evaluate and integrate multidisciplinary sources of information to analyze a complex problem.

Note. Each objective starts with "ISAT graduates will…"; this was removed for the sake of brevity.

typically contracted academic year. Of note, some raters agreed to rate a total of 40 papers while others agreed to rate 80 papers – this was due to personal time constraints of each individual. Capstone papers were delivered to the faculty via a shared folder on the institutional OneDrive network along with a rating spreadsheet for data entry.

Rating

Overall, the five raters evaluated 193 capstone papers in total. All post-Spine papers were rated, given that they only spanned three years. For the pre-Spine papers, a random number generator within R (R Core Team, 2025) was used to select which pre-Spine papers to rate. There were three common papers embedded within each of the faculty members' folders to permit checking of inter-rater reliability if desired in the future.

Each paper was examined with the eleven Spine objectives in mind. Faculty raters rated each objective as a 1 (low), 2 (medium), or 3 (high), resulting in a paper having a possible maximum score of 33 and a minimum score of 11. In addition to a numerical rating, raters had the option to leave qualitative feedback either for the overall process or for specific capstone papers. This qualitative feedback was not analyzed explicitly for this

study, though it was discussed in the rater debrief session, and will likely lead to future projects.

Statistical Methods

Our main research question was whether historical data could be used to provide evidence of learning improvement. Our data was generated from the ratings assigned on the rubric by raters. Given that there were two groups being compared - capstone papers written prior to and after the implementation of the Spine curricular change - an independent samples t-test was used within R (R Core Team, 2025). This was not a traditional pre/post repeated measures design, as the papers written in the pre-Spine group did not share authors with papers written in the post-Spine group. After evaluating the assumption of homogeneity of variance via a Levene's test, it was determined that equal variances were not present, and the Welch two-sample t-test was used to compare the means. Additionally, Cohen's d was calculated to determine the practical significance of any differences.

In addition to the overall difference of pre-Spine vs. post-Spine, we had information broken down by objective. We further examined the effect of the Spine broken down by objective by using an independent samples t-test with a Bonferroni adjustment to account for the multiple tests being performed. This allowed us to examine which objectives changed, and which, if any, did not change after the implementation of the Spine. Effect sizes were also calculated to determine the practical significance of any differences.

Reliability

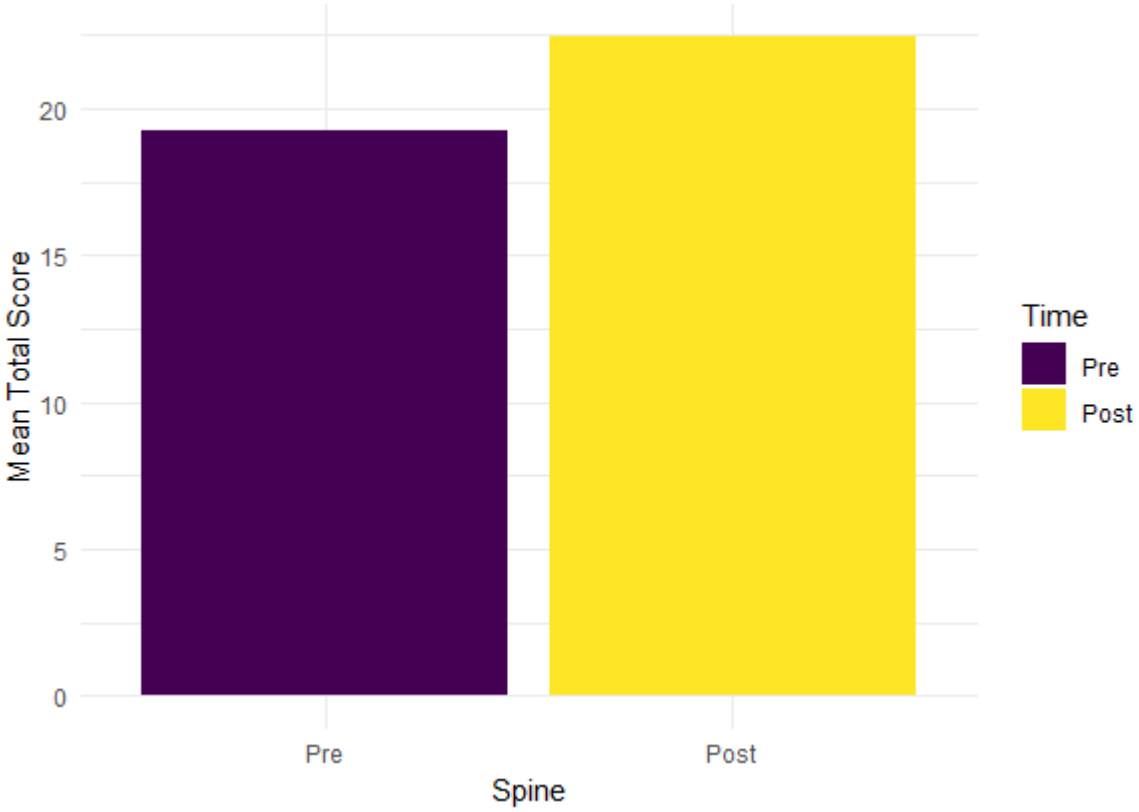
We elected to have a design where each capstone paper was only read by one rater, rather than by two raters. While this did not allow for a calculation of inter-rater reliability, this did allow us to rate twice as many papers. Further, given our parameter of interest was the difference between group means and each rater was scoring an even mix of pre-Spine and post-Spine papers, having one rater per capstone paper would result in a smaller error variance. This design resulted in raters being fully crossed with group, with each rater scoring an equal ratio of pre and post capstone papers. Therefore, the mean square error (MSE) was equal to $\frac{\sigma^2(s:g)}{n_s} + \frac{\sigma^2(rg)}{n_t} + \frac{\sigma^2((s:g)*r)}{n_s n_r}$, where n_t is the total number of raters, n_s is the harmonic mean, and n_r is the number of raters who score each student.

Ultimately, we decided it was better to have more capstone papers rated with a smaller error variance but larger individual student error variance than fewer papers rated and smaller individual student error variance. Another way to consider the MSE is that the *(s:g)*r* component will be twice as large with one rater per paper, but when computing the group mean there will be twice as many papers to average. Therefore, *(s:g)*r* will contribute the same error to the group mean regardless of whether we had 1X papers each rated by 2 raters or 2X papers each rated by 1 rater (C. DeMars, personal communication, 8 May 2023).

Results

Overall

The average total score for papers completed prior to the Spine (N = 110) was a 19.2 while the average total score for capstone papers completed after the Spine (N = 83) was 22.4, resulting in a significant difference ($t(157.87) = 4.66$, $p < .001$), with a 95% confidence interval for the mean difference of (1.84, 4.55). This indicated that for the objectives evaluated using our rubric on student capstone papers, students performed significantly better after the introduction of the Spine series of courses as compared to before. From a learning improvement lens, this indicated that student learning improvement was able to be captured, and the curricular change of the Spine resulted in improved student performance on the objectives we evaluated.



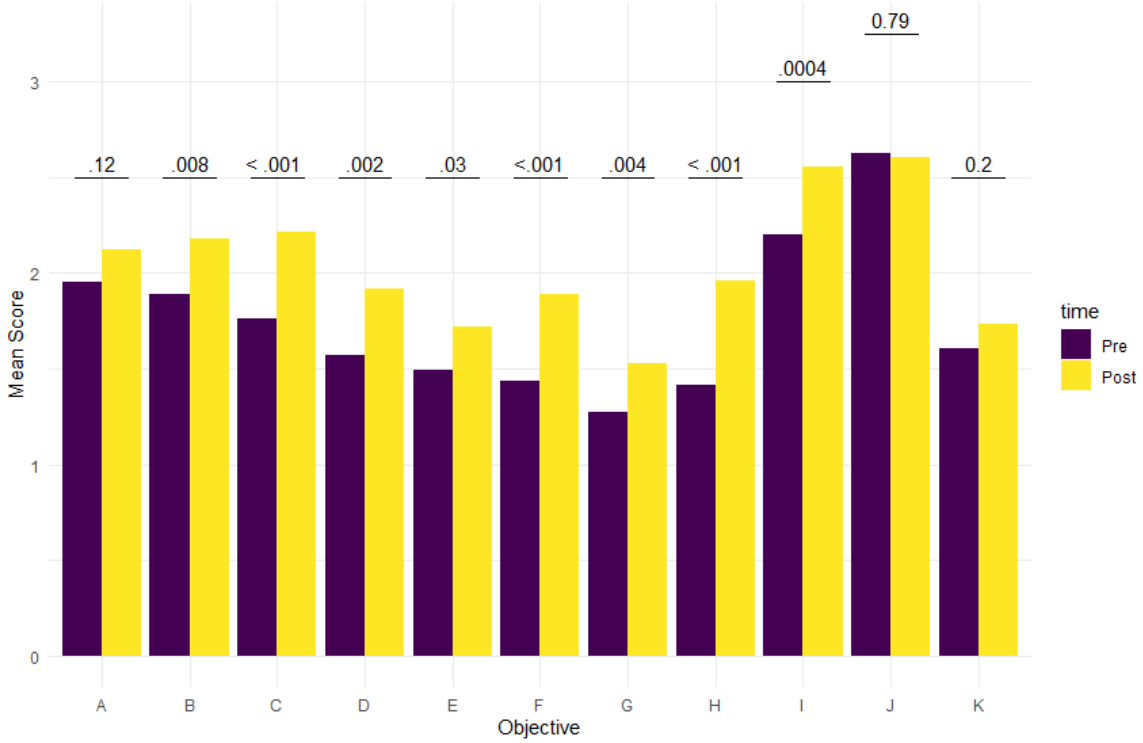
Note. Pre = pre-Spine; Post = post-Spine. The mean difference was statistically significant (t(157.87) = 4.66, p < .001).

Figure 1. *Group Mean Comparison of Student Performance*

We also determined if this was a practically significant effect. Cohen's d was calculated to be 0.69, which, by established rules of thumb, is a moderate to large practical effect size (Cohen, 1988). This can be interpreted to mean that not only is there a statistically significant difference in overall student performance on the assessed learning objectives after the curricular change, but there is also a practical, 'real-world' difference.

Objective breakdown

In addition to examining overall results, we also examined results by objective. Given that multiple pairwise tests were being run, a Bonferroni adjustment to the alpha level was used, resulting in an alpha of 0.005.



Note. Pre = pre-Spine; Post = post-Spine. P-values are indicated above each objective.
Figure 2. Group Mean Differences Broken Down by Objective

Examining each objective (See Table 1 for a complete list), it was found that 7 out of 11 tested showed significant improvement. Of those that did not show improvement, one was likely due to a ceiling effect: objective J was already at 2.6/3, leaving little room for improvement. This objective addressed student ability to select an appropriate method to solve a complex problem. Objective K also did not show significant improvement. This objective was evaluating student ability to integrate multidisciplinary sources of information. Unlike objective J, this lack of difference is not likely due to a ceiling effect (mean of 1.7/3 in the post-Spine group), and instead pointing out an opportunity for further student improvement in the future. Objectives A and E were also not significant and are again likely opportunities for student improvement in the future, and perhaps a potential area of focus for future learning improvement projects. These objectives deal with a student's ability to describe the systems out of which a defined complex problem emerges (A) and use an appropriate temporal scale to evaluate solutions (E). Objective E specifically was referenced in the faculty debrief session as one that may need to be revisited in the future, as its interpretation seemed to be vague.

While objective G did show significant improvement from pre-Spine to post-Spine, overall mean scores were the lowest of all the objectives (1.53/3 in post-Spine). Objective G had been pointed out during the faculty debrief session as one of two objectives that

seemed redundant, dealing with identifying avenues for change within governance contexts. It is possible that this objective and objective F, which dealt with accounting for governance contexts that may impact stakeholders, overlapped too much. Alternatively, it may be that faculty struggled to differentiate the two within the capstone paper, resulting in lower scores on objective G. Regardless of the reason for the low scores, faculty have indicated a desire to revisit objectives F and G specifically to better determine what is unique between them and if both are needed.

Discussion

Learning improvement – that is, true evidence that curricular changes were indeed improvements – is worthwhile. However, learning improvement projects are complex and resource-intensive. Often, programs engage in robust curricular change without assessment. This 'breakdown' in the simple model (Fulcher et al., 2014) can also be seen as an opportunity when student work has been retained from both before and after a significant change.

Undertaking a post-hoc evaluation of learning improvement following a curricular change can be viewed as an easier approach to documenting real learning improvement, as well as contributing to more rigorous evidence to regional and disciplinary accreditors (e.g., ABET). This approach still takes the approach of the 'simple model' as proposed by Fulcher et al. (2014) while adding a viable alternative approach practitioners can take if parts of the model have broken down. In particular, this approach offers an alternative when programs have an intervention, but no assess or re-assess (e.g., when it is ~~assess~~ – intervene – ~~re-assess~~) (Fulcher et al., 2014). Given that programmatic curricular changes happen regularly, yet assessment of these changes is not common, we propose that this approach will be of value to many programs. For those programs that have a disciplinary accreditor, in particular ABET, this approach can aid them in capturing improved student learning and continuous improvement (ABET, 2021a).

Additionally, in this situation, the curricular change has already occurred – what is potentially the hardest part of the simple model is done. Recognizing the intensity and heavy lift of that work, assessing how the curricular change impacted student learning allows faculty to see if all that work 'paid off'. Further, assessing after the intervention is markedly easier than starting a new learning improvement project from scratch – there is much less new work to be performed. This approach honors the hard work already done while still capturing any change in student learning.

Further, learning improvement projects require careful consideration of cultural norms, available resources, and the willingness to adapt to change within academic programs. Namely, it is important to consider the human resources alongside the more commonly considered financial resources. In the case of ISAT, the faculty were exhibiting a culture and commitment to assessment, but were lacking in capacity. A unique aspect of our study was the use of historical data to assess the impact of a prior curricular change—the introduction of the "Spine" series of courses. Most learning improvement projects focus on future-oriented assessments, making our approach distinctive. Historical data spanning a curricular change provided a rare opportunity to evaluate the long-term effects of a programmatic shift. The use of historical data allowed us to examine the tangible effects of the curricular change, offering valuable insights, while respecting the lack of

capacity expressed by faculty with respect to starting a learning improvement project "from scratch", as described by Fulcher et al. (2014).

Generalizability

This study captures the ability of a program to 'look back' at curricular changes – this could be a series of courses, as illustrated here, or other changes such as the addition/removal of single courses, capstone or research requirements, or a reworking of Student Learning Outcomes. Programs are dynamic, with courses, outcomes, and requirements being evaluated in light of changing student needs. Additionally, it has been well documented that Learning Improvement projects, while providing valuable insights into student learning, are very resource-intensive (Fulcher & Prendergast 2021; Fulcher et al. 2014). A backwards-looking approach to learning improvement makes it possible for programs to evaluate prior curricular changes, as well as alleviating some of the resource intensity of building a learning improvement project from the ground up.

While this project was in a STEM program accredited by ABET, the process is not exclusive to such a program. Any program that has continuity of student artifacts – portfolios, final exams, course projects, etc. – over a curricular change could take this approach. The key piece was the availability of student work. For this study, it was programmatic data collection (i.e., the program collected and stored student capstone reports). However, if a history program implemented a curricular change and Professor Jones had a record of student papers she assigned each semester, those papers could potentially serve as artifacts spanning the curricular change. Each program will have to consider the context of their change, and then discuss with faculty what student work is available to 'look back' over the change. From there, they will be able to determine if they have adequate coverage, both before and after, of the curricular change to be able to assess the impact on student learning. The defining feature of student work in this application is continuity and relevance to the outcome(s) being examined rather than a specific assignment type.

Additionally, many institutions consider the assessment timeline to be on an annual cycle – programs are often asked to submit a report annually. Given this short institutional focus, a longer learning improvement project may not be prioritized. It will take longer than a single year to conceive, design, implement, and assess. This approach can be beneficial for programs accredited by ABET. The Continuous Improvement criteria of ABET (*ABET*, 2021a) offers a longer view of the program's improvement history, and this type of work could be productive in showcasing the efficacy of changes made. Recognizing that many programs implement curricular change on more anecdotal evidence (e.g., "Our students seem to be struggling with xyz. Let's implement these curricular changes to support them."), taking the approach detailed in this paper will allow programs to capture the results of their efforts. This approach also allows programs to capture the effects of curricular changes for external accreditors.

Faculty Benefits

An unanticipated, yet positive, secondary outcome of this study was that the rating experience itself was a form of professional development for the faculty engaging in the rating. During the debrief session held after faculty had completed rating the capstone projects, it was found that they had deeper engagement with the Spine student learning

objectives as well as with the capstone project. Faculty observed that there were objectives that may need to be revisited, and others that seemed redundant. This has spurred further faculty discussion around these objectives, feeding into the paradigm of gathering data and using that data for curricular change.

Additionally, faculty were able to identify a seeming deficiency in student writing ability and a perceived stylistic change from capstone papers written pre-Spine to those written post-Spine. This has also spurred further conversation around the topic, including where in the curriculum students are exposed to writing tasks, where in the General Education curriculum they are exposed to writing, and if the program needs to revisit how writing is taught. The stylistic change has also prompted conversation, jointly with the writing skills, surrounding the creation of a style guide or manual that both advisors and students could reference. The acknowledged challenge surrounding the style and quality of writing in the capstone paper stems from academic freedom of the capstone advisors; these advisors are the ones who grade and provide feedback on the paper. Understandably, faculty have different views as to length, content, and what constitutes quality. Regardless, this experience has spurred a desire for further conversation around this topic.

Perhaps an important lesson from the debrief is that engagement should be kept going so that when faculty have the capacity to engage with larger learning improvement projects, the culture and commitment will be well maintained to support such an effort. It is likely that further curricular adjustment will occur as a result of this project, as well as the potential for a larger learning improvement project.

Limitations

One major limitation in this study is the variability of the capstone papers themselves. The exact contents, length, and rigor are determined by the individual capstone advisors, leading to variability in final products. Additionally, each concentration area varies in the approach to scientific writing. Some lean much more towards the approach taken in primary literature of biology and chemistry, while others approach writing more like the social sciences. This variety in approaches also leads to a variety in final products. We are hopeful that the impact of this variety can be minimized, as all concentrations are represented in both the pre-Spine and post-Spine groups.

Conclusion and Future Work

Our study emphasizes the multifaceted nature of learning improvement projects in higher education. It underscores the importance of considering situational factors, such as capacity and historical data, when embarking on such endeavors. By using historical data to assess the impact of a curricular change, we both illustrate the feasibility and additionally provide a pathway for other programs interested in documenting learning improvement to examine existing data in a new light.

The implications of this study are potentially very impactful – programs wishing to engage in learning improvement projects yet that lack the resources to do so 'from the ground up' may find themselves able to make use of historical departmental data to evidence learning improvement. A program interested in taking this approach needs only to look for student work that has spanned curricular changes that can map to objectives impacted by those changes. Perhaps that comes in the form of a senior capstone report, as presented in this study. However, it may also come in the form of a final project for a

course, a homework assignment that has remained largely unchanged for a number of years, or presentations given during courses. While it is recognized that historical data will not exist for every situation, there are situations where it may exist. This will not only allow programs to evaluate curricular changes through a learning improvement lens, but it will also reduce the resource intensity of a learning improvement project, as well as looking at improvement over a larger time scale than the yearly assessment cycle seen in most institutions.

Appendices

Appendix A: Spine learning objectives rubric

Acknowledgements: We would like to thank the Integrated Science and Technology department for allowing this project to take place and for giving us access to the historical capstone reports. Additionally, we would like to thank the faculty raters who took time out of the start of their summer to rate student reports. We would also like to thank Dr. Christine DeMars for her expertise in reliability in aiding study design.

References

ABET. (2021a). abet.org

ABET. (2021b). *Assessment Planning*. https://www.abet.org/accreditation/get-accredited/assessment-planning/

Banta, T. W., & Blaich, C. (2011). Closing the assessment loop. *Change (New Rochelle, N.Y.)*, *43*(1), 22–27. https://doi.org/10.1080/00091383.2011.538642

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Erwin, T. D. (1991). *Assessing student learning and development: A guide to the principles, goals and methods of determining college outcomes*. Jossey-Bass Inc.

Ewell, P. T. (2009). Assessment, accountability, and improvement: Revisiting the tension. *National Institute for Learning Outcomes Assessment, Occasional Paper #1*. https://www.learningoutcomeassessment.org/documents/PeterEwell_005.pdf

Fulcher, K. H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014). *A simple model for learning improvement: Weigh pig, feed pig, weigh pig*. https://in.ewu.edu/facultycommons/wp-content/uploads/sites/129/2016/12/A-Simple-Model-for-Learning-Improvement_Weigh-Pig-Feed-Pig-Re-Weigh-Pig.pdf

Fulcher, K. H., & Prendergast, C. (2021). *Improving student learning at scale: a how-to guide for higher education* (1st ed.). Stylus Publishing, LLC.

Good, M. R. (2015). *Improving student learning in higher education: a mixed methods study* [ProQuest Dissertations Publishing]. https://commons.lib.jmu.edu/diss201019/18

Lambert, L., & Good, M. R. (2024). An intentional approach to starting a learning improvement project. *Intersection: A Journal at the Intersection of Assessment and Learning*, *5*(4), 25–34.

Learning Improvement Community. (2022, March). [Online document]. *Readiness for learning improvement tool*. https://learning-improvement.org

Learning improvement community: About us. (2024). [Organization website]. https://learning-improvement.org/about-us

Lending, D., Fulcher, K., Ezell, J. D., May, J. L., & Dillon, T. W. (2018). Example of a program-level learning improvement report. *Research & Practice in Assessment*, *13*(2), 34.

Matveev, A. (2014). *Top 10 most frequently cited principles in reaffirmation reviews: 2013 reaffirmation class institutions*. SACSCOC. https://sacscoc.org/app/uploads/2019/09/Most-Frequently-Cited-Principles_2013_Track_A_B-rev.pdf

Matveev, A. (2024). *Most frequently cited principles of accreditation in decennial reaffirmation reviews: Class of 2023*. SACSCOC. https://sacscoc.org/app/uploads/2024/03/2024_Most-Frequently-Cited-Principles_Class-of-2023_web.pdf

McClure, K. R., Sallee, M. W., Ford, J., Gonzales, L. D., Griffin, K., Jenkins, T. S., Kiyama, J. M., Martinez, M., Martinez-Aleman, A. M., Miao, S., Perez, R. J., & Porter, C. J. (2023). *The impact of covid-19 on faculty, staff, and students: Using research to help higher education heal through the pandemic and beyond*. Association for the Study of Higher Education. https://www.ashe.ws/positiontaking

R Core Team. (2025). *R: a language and environment for statistical computing* (Version 4.5.1) [Computer software]. R Foundation for Statistical Computing. www.R-project.org

Smith, K. L. (2017). *Integrating implementation fidelity and learning improvement to enhance students' ethical reasoning abilities* [James Madison University]. https://commons.lib.jmu.edu/diss201019/153

Smith, K. L., Good, M. R., Sanchez, E. H., & Fulcher, K. H. (2015). Communication is key: "Unpacking" use of assessment results to improve student learning. *Research & Practice in Assessment*, *10*, 15–29.

Smith, K. L., Good, M. R., & Jankowski, N. (2018). Considerations and resources for the learning improvement facilitator. *Research & Practice in Assessment*, *13*, 20-26.

Appendices:

Appendix A: Spine learning objectives rubric

Assessors' Engagement with Video-Recorded Performance Assessments in Nursing: A Qualitative Study

Authors:

Conor Scully, PhD.
Dublin City University

Prof. Michael O'Leary, PhD.
Dublin City University

Zita Lysaght, EdD.
Dublin City University

Mary Kelly, PhD.
Dublin City University

ABSTRACT

The use of video in performance assessments has accelerated as a result of the COVID-19 pandemic. Research has tended to focus on the administrative and cost implications of setting up remote assessments: however, few studies have explored how the process of assessing a performance is altered when the assessment is conducted through video. This qualitative paper reports on findings from a simulated video assessment in nursing. As part of the study, 12 nursing assessors watched the same four videos of undergraduate nurses performing either a blood pressure measurement or a naso-gastric tube insertion. They were asked to "think aloud" while doing so, and were also subject to an interview about their assessment practices. Findings revealed that the use of video allows assessors to reduce guesswork when assessing, yet it also limits their field of vision, and in some cases harms the perceived validity of the assessment.

Correspondence E-mail: conor.scully@dcu.ie

Keywords: Video-based performance assessment; Objective Structured Clinical Examination (OSCE); Assessor cognition; Inter-rater reliability in nursing education; Remote clinical skills evaluation

This paper explores how assessors engage with the task of evaluating student performance in a simulated video Objective Structured Clinical Examination (OSCE), in which student nurses were recorded—and graded—ompleting two tasks: blood pressure measurement and naso-gastric tube insertion. The OSCE is a type of performance assessment, an assessment modality in which test-takers have to "construct an answer, produce a product, or perform an activity" (Darling-Hammond & Adamson, 2010, p. 7). In contrast with other assessments such as multiple-choice examinations, performance assessments are distinguished by their closeness to real-world situations that test-takers may encounter (Darling-Hammond & Adamson, 2010).

In nursing, as well as the health sciences more broadly, performance assessments such as the OSCE have been used to evaluate student performances at a range of practical skills that they will go on to use in the "real world" of clinical practice (Rushforth, 2007; Khan et al. 2013). The OSCE was designed due to deficiencies in more traditional assessment formats—notably the short and long cases—that were popular in medicine, particularly the perceived autonomy of assessors to grade students according to personalised or subjective criteria (Harden et al., 1975; Khan et al. 2013). The use of standardised marking tools in the OSCE, as well as the fact that all students who complete the assessment have to perform the same skills, were implemented to ensure that students are graded in a manner that is consistent (Khan et al. 2013).

In common with all assessments, those administering OSCEs need to build a validity argument about the inferences that are to be made on the basis of OSCE scores and provide evidence that these inferences are justified (Fraenkel & Wallen, 2006; Khan et al., 2013; AERA et al. 2014). Two sources of evidence are usually considered important for the validity argument associated with OSCE scores:

1. Fidelity to the real world: OSCEs entail a simulation of a situation that test-takers can be expected to face when they enter the world of clinical practice. As such, they are perceived to entail a high level of real-world fidelity. In general, the closer the approximation between the environment of the test and the real-world situation it is simulating, the stronger the associated validity argument regarding the test scores will be (Darling-Hammond & Adamson, 2010; Eva & Hodges, 2012; Hodges, 2013).
2. Inter-rater reliability (IRR): OSCEs usually require the use of a range of assessors to grade student performances. It is important that these assessors interpret the performances in the same way, such that students receive the same score regardless of who is assessing them (Khan et al. 2013; Gwet, 2014). If there was a notable variation within a sample of assessors, it would be less defensible to make an inference about a test-taker on the basis of their score, as they may have received a different score had they been assessed by a different assessor. The question of whether OSCE scores do in fact demonstrate high levels of IRR has received significant attention in research literature in both medicine (e.g., Brannick et al., 2011) and nursing (Rushforth, 2007; Navas-Ferrar et al., 2017; Goh et al., 2019).

In part due to these two factors, OSCEs have become a popular assessment format across the Western world, have expanded beyond their original implementation in medicine, and are commonly used in nursing (Rushforth, 2007; Patrício et al., 2013).

In the wake of the COVID-19 pandemic, universities across the world were forced to suspend in-person teaching and implement new or altered assessment modalities. In a traditional OSCE, test-takers, assessors, and—in some cases—Standardised Patients are required to be in the same room (Khan et al. 2013). As such, the onset of the pandemic meant that educators were forced to come up with alternative ways of conducting assessments of students' clinical skills (Lara et al., 2020). Some researchers have described the administration of fully remote OSCEs, in which the OSCE is conducted "live", via an online platform such as Zoom or Microsoft Teams (e.g., Hopwood et al., 2020; Lara et al., 2020). Others have discussed the possibility of allowing students to record their own videos and upload them to an online platform for grading by assessors, a scenario that had been described elsewhere in the literature prior to the pandemic (Purpora & Prion, 2018).

In all cases, the role of the assessor shifts from one where they are able to observe a student's performance in-person, and in real time, to one where their grading of a performance takes place through video. The issue of how this affects the process of judging and grading performances is central to this paper, which addresses the following research question: *How does the use of video in performance assessments in nursing affect assessors' engagement with the task of assessment?*

This issue has implications for the validity argument associated with test scores, as evidence that assessors use meaningfully different strategies when assessing through video compared with in-person assessments may affect the IRR of test scores and may reduce the comparability of test scores across different administrations of the assessment.

Literature Review

Assessor cognition

The issue of what specifically happens when examiners are given the task of assessing student performance at clinical tasks is one that has received significant attention in the medical assessment literature, particularly in the last decade and a half. This emergent field, known as "assessor cognition" or "rater cognition", broadly seeks to understand how assessors "interpret and construct their own personal reality of the assessment context" (Govaerts & van der Vleuten, 2013, p. 1169). This area of inquiry is underpinned by the recognition that because all assessors are individuals, it is likely that they bring with them their own ideas and perspectives that may influence how they engage with the task of assessment (Eva & Hodges, 2012).

Researchers in the field of assessor cognition have used methods such as interviews and "think alouds" in order to develop models that map assessors' thoughts as they watch and grade student performances (e.g., Kogan et al., 2011; Yeates et al., 2013; Roberts et al., 2020). These methods have revealed numerous ways in which assessors may diverge in their approach to the assessment task. For instance, some assessors may be prone to making inferences about a student beyond what is directly visible to them when the student is completing an assigned task, or might compare a student's performance with performances graded immediately before (Kogan et al., 2011; Yeates et al., 2015; St-Onge et al., 2016). Gauthier et al. (2016) conducted a review of assessor cognition literature and developed a model for understanding what happens when assessors watch and interpret student performances: their model conceptualises judgement formation as a three-stage process of observation, processing, and integration.

Perhaps unsurprisingly given the focus inherent in assessor cognition research on divergences in judgement formation, studies on assessor cognition have often taken an inter-rater reliability (IRR) perspective. Seen this way, differences in how assessors approach the task of assessment may result in divergences in the scores they award, even when they are watching the same performance (Gingerich et al., 2014a). This poses a threat to the IRR of assessment scores and may ultimately undermine the validity of decisions made on the basis of these scores (AERA et al., 2014). Multiple studies have attempted to measure the amount of score variance that can be attributed to divergences in assessors' patterns of judgement (e.g., Gingerich et al., 2014b; Roberts et al. 2020). For example, a study by Gingerich et al. (2017) found that assessors' varied cognitive processes accounted for between 21% and 53% of score variance in scores awarded to students in a recorded performance assessment.

It is notable that the vast majority of published studies on the subject of assessor cognition have taken place in the field of medicine, rather than other healthcare disciplines. This is in spite of the documented popularity of performance assessments (especially the OSCE) in other fields, particularly nursing (Rushforth, 2007; Navas-Ferrar et al., 2017; Goh et al., 2019). An exception to this is a 2014 study by East et al., in which nursing assessors were interviewed about their assessment practices, and a recent study (Scully et al., 2024) which mapped nursing assessors' cognitive processes. The issue of how nursing assessors specifically engage with the task of assessment is one that is under-researched.

Use of Video in Performance Assessments

Video has been incorporated into performance assessments for many years in various different ways. For example, numerous studies have described the creation of video exemplars, the purpose of which is to demonstrate optimum performance at an assessed task (or series of tasks) that test-takers will be instructed to complete (e.g., Barratt, 2010; Massey et al., 2017). Researchers have used the recording of performances as a means of providing formative feedback to test-takers, who are able to understand comments from examiners by looking at a video of their own performance (e.g., Paul, 2010). Researchers have also described studies in which students had recorded themselves completing specific tasks and uploaded these videos to an online platform for grading by assessors, an approach deemed to be beneficial for those participating in distance learning (Purpora & Prion, 2018).

Since the COVID-19 pandemic, numerous studies have described remote performance assessments which take place "live" over Zoom or other online platforms. (e.g., Hopwood et al., 2020; Lara et al., 2020; Major et al., 2020). These assessments were usually developed and administered as a result of the prohibition on in-person teaching and learning that was implemented during the pandemic (Felthun et al., 2021). As of 2024, enough studies describing the development of these remote assessments (usually, though not always, classified as OSCEs) have been published so as to allow for several review pieces. These papers have the aim of drawing conclusions about the feasibility of conducting live performance assessments, and the potential barriers to their effective implementation.

Felthun et al. (2021) reviewed 11 studies of video-based performance assessments (which they broadly label "teleOSCEs") administered during the pandemic. Of these 11,

seven entailed a "live" examination, while four involved the submission of student-produced videos, which were uploaded to a portal for grading by assessors. More recently, Giri and Stewart (2023) reviewed 28 studies describing the use of video in performance assessments in medicine, nursing, and dentistry, again finding a mix of studies which described "live" assessments and those involving student-produced videos. Both reviews found that performance assessments conducted through video were perceived by stakeholders as being viable, particularly when there is a lack of in-person assessment possibilities.

However, in both reviews, the authors described a lack of insight into how assessors engage with the task of assessment when an examination is conducted through video as opposed to in-person. As noted by Felthun et al. (2021, p. 4), their review "revealed little about whether examiners extract different information about student performance from teleOSCEs and in-person assessments". They mention that further research into the use of video in performance assessments should "focus on how the online platform impacts… examiners' judgments" (p. 4). This lack of information might have implications for the reliability of assessment scores, especially if test administrators need to compare scores from a remote administration of the assessment with previously determined scores from an in-person administration of the same assessment. As discussed by Giri and Stewart (2023, p. 14): "remote assessment of practical skills should be interpreted with caution because of a lack of correlation between the assessment scores of the face-to-face examiner and remote examiner". As such, while researchers have published many studies describing the administration of remote performance assessments, these studies have generally failed to examine in detail the potential changes in assessors' cognition that may take place when an assessment is conducted through video, and the resultant reliability implications.

The present study, therefore, sits at the intersection of two areas of inquiry: research into assessors' cognitive processes and research into the use of video in performance assessments. This paper is among the first to investigate the specifics of how assessors approach the task of assessment when the assessment takes place through video, and whether there are observable differences in their approach vis-a-vis in-person assessment.

Methods

This qualitative paper reports on findings from a larger study that had the aim of exploring assessors' cognitive processes as they watched and discussed the same OSCE performances (Scully et al., 2024) As part of the study, the first and fourth authors filmed six videos of three students each completing two OSCEs: blood pressure measurement (BP) and naso-gastric tube insertion (NG). The students were at different stages in completing their General Nursing programme (one student from years 1, 2 and 3, respectively) which offered the opportunity for some divergence in performance levels to be assessed. Four of these videos recorded were used in the study: the first-year student completing the NG OSCE (P01NG), the second-year student completing the BP OSCE (P02BP) and both third-year student videos (P03BP and P03NG). The BP OSCE was performed on a real person, while the NG OSCE was performed on a mannequin. The videos were recorded using UniCam, with a camera and microphone embedded into the ceiling that the researchers could operate with their phones.

Having completed the video recording, 12 assessors were recruited —using convenience sampling methods —to participate in the study. All 12 were employed in the same university nursing department, either as lecturers or clinical skills nurses (who have the responsibility of teaching clinical skills to nursing students). All participants had experience of assessing undergraduate nursing OSCEs, with nine of the 12 having over five years of experience. When asked to rate their proficiency as assessors, one selected *Advanced Beginner,* three selected *Competent,* six selected *Proficient*, and two selected *Expert* (Benner, 1982).

The 12 participants engaged in a one-to-one, semi-structured interview with the first author, to discuss how they perceived their roles as assessors, and the processes they employed when making judgements about students' OSCE performances (East et al., 2014). Additionally, each assessor participated in a cognitive interview (Ericsson & Simon, 1980; Willis, 2015), which provided insight into their interpretations of the two OSCE marking guides used in the study.

In the final stage of the process, each assessor watched the four recorded videos and, using a think-aloud protocol, shared their opinions as to how well or badly the students were performing. "Thinking aloud" is a technique that has been used in numerous studies of assessors' cognitive processes (e.g., Kogan et al., 2011; Yeates et al., 2013; St-Onge et al., 2016), and provides a rich insight into how assessors engage with the task of assessment. During this section of the study, assessors were given autonomy to pause and rewind the video as they wished. When this happened, the first author probed them by asking why they had done so.

Qualitative data from the study were analysed according to the principles of thematic analysis, a six-step process of *familiarisation, coding, theme search, theme review, theme refinement* and *write-up* (Braun & Clarke, 2006; 2021). Coding largely focused on the semantic meaning of the words, however in some instances latent codes were identified that indicated participants' views about video assessment. The data allowed for a range of insights into the cognitive processes of the assessors who participated in the study (reported more widely in Scully et al., 2024). This paper reports on aspects of the data that relate to how assessors engage with the task of assessment in ways that are specific to the video format, and how these processes of engagement differ compared to when the OSCE is administered in-person.

Results

Analysis of the data revealed three ways in which the use of video affects assessors' engagement with the assessment task: *reducing guesswork, obstructed vision,* and *reduction in perceived validity*. These factors are discussed in turn below, augmented with illustrative quotes from participants.

Reducing guesswork

When assessments happen in real life, events unfurl quite quickly and assessors have to make snap decisions as to what took place in front of them. When video is used — as in the present study —assessors have the opportunity to pause or rewind the video as needed, in order to be sure that they are correctly observing a student's performance and, therefore, making the correct decision about how well that student executed the required tasks. Ten of the 12 assessors in the sample spoke about this phenomenon as being a key

benefit of video OSCEs. These assessors noted that the use of video allowed them to make decisions with more confidence, and therefore rely less on instinct or guessing, which they admitted had happened occasionally when OSCEs were administered and assessed live:

> This just reminded me of when I'm doing it, that when you're doing so many of them in real life, sometimes you're going "Did she do that?" and that's where the video is helpful because you can actually stop it and look back, and you can't do that in real life.
> - Assessor 10

> I would probably watch the whole thing through and then I'd go back. I might highlight something that might pop up, that I need to look a little bit more carefully. And then I would go back and I might stop it a few times if I'm not sure, especially with the first few until I become familiar with exactly what the student is doing.
> - Assessor 1

For these assessors, the use of video technology within the OSCE allows them to view the same performance multiple times to ensure they have not missed anything and can confidently award a grade to the student. As such, the use of technology affected their process of coming to a judgement about a student. Seen this way, using video technology should improve score reliability, as assessors would be less reliant on guesswork or short-term memory when judging a student.

Shielded vision

When assessment takes place through video, assessors are unable to move around in order to see something with more clarity, as they would in real life. As a result, their vision is bounded by what the students have filmed, and they may be unable to see specific parts of the procedure that the student is performing. Although students are usually given specific instructions to make all aspects of their performance visible on the video, the reality expressed by participants in this study is that this was not always the case. Assessors mentioned that, when assessing OSCEs in the past they have had to guess whether a student had completed a specific step on the guide:

> Sometimes it can be due to a camera angle as well, and the quality of the video sometimes isn't that good. So you are kind of guessing, “did they or didn't they?”
> - Assessor 11

This phenomenon was reported to be more pronounced for minute or intricate tasks such as —in the case of this study —locating the brachial artery. Assessors discussed how it was impossible to tell through video if the student had correctly located the artery on the patient. As a result, they had to choose whether or not to either give the benefit or the doubt to students regarding their completion of this step on the marking guide. In this sample, different assessors reported different strategies for what to do when something

was not clearly visible to them: some would fail to award a mark for an item that they could not see clearly, while others would award the mark. The implications in terms of IRR are perhaps obvious, with the clear possibility that the same performance would be graded differently depending on the assessor. Indeed, for one of the recorded performances in this study, there was widespread disagreement within the sample of 12 assessors as to whether the student had correctly located the brachial artery during a blood pressure measurement.

Some assessor participants noted that in spite of the potential limitations of the video format in terms of allowing them to observe all aspects of a performance, the students could make up for this by narrating what they were doing. In this way, students could still communicate to assessors what they were doing, even if the video could not show it in detail:

> With the patients [in the NG OSCE], sometimes you can get curling of the tubes at the back of the throat and so forth. You might not see that all the time with the mannequin, it might just have one way down and it just goes down. So that can be a little bit difficult to see. But if they're vocalising that to you, they may say, "well, when I'm sliding it in, I want it to go upwards and backwards, inwards and backwards", and then they know the basis of it.
>
> - Assessor 9

However, the potential for narration to make up for the lack of detail in the recorded performance is not one that was discussed by all assessors in the sample, which again indicates differences in how assessors engage with the task of assessment and has potential implications for IRR.

Reduction in perceived validity

As noted at the beginning of this paper, performance assessments are perceived to be effective in part because of fidelity to the real world. Three of the assessors in the current study discussed how the use of self-recorded videos affected their perception of the assessment. These assessors emphasised that when OSCEs are administered through video, students can record their attempt at the skill multiple times; as such, the disconnect between the "real world" of clinical practice, where a nurse may only have one attempt at performing a procedure, and the testing environment, is increased. For these assessor participants, the use of video may decrease the objectivity of the OSCE:

> But of course, if it's a physical assessment, you will probably have gotten more information… And I would have been able to assess more objectively because it's real time. And whatever mistakes they're making, they're making it in real time… Whatever thing they're doing, it's real time. And it gives you more objectivity for sure.
>
> - Assessor 2

A reduction in the perceived validity of the assessment on the part of the assessors has implications for how they engage with the task of assessment. The assessors who

discussed how they preferred in-person assessment were more likely to note that they intentionally deviated from the marking scheme in order to reward students that they perceived to be competent.

Discussion

The present study sought to determine whether there are differences in assessors' cognitive processes when an assessment takes place through video, rather than in-person. In order to address this question, 12 assessors of undergraduate nursing OSCEs were interviewed about their assessment practices and participated in a "think aloud" during which they vocalised their thought processes while watching and grading four videos of students completing OSCE stations. Analysis of the data resulted in three themes being identified: *reducing guesswork, obstructed vision,* and *reduction in perceived validity*. These themes speak to the specific ways that the use of video in performance assessments affects how assessors engage with the task of assessment.

As noted in several review articles about the use of video in performance assessments (Felthun et al., 2021; Giri & Stewart, 2023), such studies have lacked a focus on assessors and how the use of video affects their cognitive processes. In spite of this, several findings in the present study have been noted elsewhere. Specifically, Chen et al. (2018) discussed how assessors may be limited by the use of video, as they are unable to walk around the room to improve their ability to see certain aspects of a student's performance. Relatedly, Chan et al. (2014) found that while videos in general could be used for assessment of clinical performance, caution should be exercised when using video for intricate physical tasks, as these were much less likely to be visible on a video. As determined in the present study, this has IRR implications, as some assessors may resort to guessing whether a student completed a task correctly or allowing them to compensate by narrating what they're doing.

A vocal minority of assessors in the present study expressed that the use of video reduced the objectivity of the assessment, as it decreased the fidelity of the OSCE in relation to the real world of clinical practice. The issue of whether OSCEs, and other performance assessments, are perceived as valid by assessors is one that has recurred in the literature on such assessments (e.g., Roberts et al., 2020; Hyde et al., 2022). Indeed, Hyde et al. (2022) found that, driven by a lack of belief in the utility of the OSCE, experienced assessors were more likely to intentionally deviate from marking guide designed for an OSCE station, and judge students according to what they personally believe to be important. As such, any adjustment to the OSCE, such as the incorporation of video, is likely to increase the risk of assessors intentionally deviating from the marking guide, which would affect the IRR of assessment scores.

There are two notable implications of the present study for those using video-based performance assessments. The first is to ensure that there is a robust procedure in place for conducting reliability checks within a pool of examiners. Calculating the IRR of awarded scores should be a step in any assessment, particularly one that is high-stakes (Khan et al., 2013; AERA et al., 2014). The present study indicates that the use of video can bring about specific threats to IRR, and as such the need for reliability checks may be amplified. Secondly, this study has indicated that certain tasks, especially those which are physically intricate, such as the location of a brachial artery, may be difficult or impossible to assess consistently through video. As a result, test administrators should consider the

type of skills they wish to assess, and whether it is feasible to do so through video (Chan et al., 2014; Giri & Stewart, 2023). In spite of the specificity of the present study, these general principles will apply in any domain where practical skills are assessed through video: someone carrying out a video assessment in dentistry should also have a strict procedure in place to ensure that IRR levels are sufficient and would also need to think about what skills are feasible to be assessed through video.

In terms of theoretical implications, the present paper adds to the assessment literature by contributing to a growing body of research on assessors' cognitive processes (e.g., Kogan et al., 2011; Yeates et al., 2013; Roberts et al., 2020), and is one of the few studies to focus on nursing assessors specifically (e.g., East et al., 2014; Scully et al., 2024), as well as one of the first to focus on how the use of video affects assessors' judgements. As noted elsewhere, incipient research on video assessment has tended to focus on the cost and feasibility of setting up such assessments (e.g., Felthun et al., 2021; Giri & Stewart, 2023). This paper refocuses the literature by examining the assessors themselves, and whether they extract different information from a performance when it takes place through video.

A clear next step for researchers is to determine the comparability of assessors' grades when an assessment is conducted through video, as opposed to in-person. As noted by Giri and Stewart (2023, p. 14), there may be "a lack of correlation between the assessment scores of the face-to-face examiner and remote examiner". A study by Dagnaes-Hansen et al. (2018) is one of the few pieces of research to measure this specifically (using assessment scores of a cystoscopy exam), and their research should be extended into other contexts where possible.

The content of this study should be noted in light of its limitations. Firstly, a sample of 12 assessors working at the same institution necessarily limits the generalisability of the findings beyond this context. Ideally, researchers wishing to apply these findings in other contexts (especially outside the domain of nursing and the health sciences more broadly) should exercise caution, and conduct a comparable study where possible. Secondly, the assessments used in the study were simple, first- and second-year nursing OSCEs. It is possible that the effects of the use of video may be different if assessors were tasked with the interpretation of more complex tasks. As noted above, research has indicated that specific skills (such as communication) may be better suited to assessment through video. This study does not allow for generalisability beyond the specific, technical competencies assessed during a blood pressure measurement or a gastric tube insertion. Finally, the lack of score data in the current study prevents measurable links being made between assessors' cognitive processes and the scores they award. In other words, it was not possible to determine whether the changes in assessors' engagement as a result of the incorporation of video led to measurable effects regarding their awarded scores (e.g., Gingerich et al., 2014b; Chahine et al., 2016). Ideally, score data would be obtained when an assessment is conducted in-person, and compared when the same assessment is conducted through video (e.g., Dagnaes-Hansen et al., 2018). Such an approach would allow for a quantifiable measure of the effect of an assessment taking place through video. In spite of these limitations, this study will be of interest to researchers in nursing assessment, as well as those pursuing remote assessments more generally.

References

American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for Educational and Psychological Testing*.

Barratt, J. (2010). A focus group study of the use of video-recorded simulated objective structured clinical examinations in nurse practitioner education. *Nurse Education in Practice, 10*(3), 170–175. https://doi.org/10.1016/j.nepr.2009.06.004

Benner, P. (1982). From Novice to Expert. *American Journal of Nursing, 82*, 402–07.

Brannick, M. T., Erol-Korkmaz, H. T., & Prewett, M. (2011). A systematic review of the reliability of objective structured clinical examination scores. *Medical Education, 45*(12), 1181–1189. https://doi.org/10.1111/j.1365-2923.2011.04075.x

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–01. https://doi.org/10.1191/1478088706qp0630a

Braun, V. & Clarke, V. (2021). *Thematic Analysis: A Practical Guide*. Thousand Oaks, CA: Sage Publications.

Chahine, S., Holmes, B., & Kowalewski, Z. (2016). In the minds of OSCE examiners: uncovering hidden assumptions. *Advances in Health Sciences Education, 21*(3), 609–625. https://doi.org/10.1007/s10459-015-9655-4

Chan, J., Humphrey-Murto, S., Pugh, D. M., Su, C., & Wood, T. (2014). The objective structured clinical examination: can physician-examiners participate from a distance?. *Medical Education, 48*(4), 441–450. https://doi.org/10.1111/medu.12326

Chen, T. C., Lin, M. C., Chiang, Y. C., Monrouxe, L., & Chien, S. J. (2018). Remote and onsite scoring of OSCEs using generalisability theory: A three-year cohort study. *Medical Teacher, 41*(5), 578–583. https://doi.org/10.1080/0142159X.2018.1508828

Dagnaes-Hansen J., Mahmood O., Bube S., Bjerrum, F., Subhi, y., Rohrsted, M. & Konge, L. (2018). Direct observation vs. video-based assessment in flexible cystoscopy. *Journal of Surgical Education, 75*(3), 671–677. https://doi.org/10.1016/j.jsurg.2017.10.005

Darling-Hammond, L. & Adamson, F. (2010). Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning. Stanford Center for Opportunity Policy in Education. Available at: https://globaled.gse.harvard.edu/files/geii/files/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning-report_0.pdf

East, L., Peters, K., Halcomb, E., Raymond, D., & Salamonson, Y. (2014). Evaluating Objective Structured Clinical Assessment (OSCA) in undergraduate nursing. *Nurse Education in Practice, 14*(5), 461–467. https://doi.org/10.1016/j.nepr.2014.03.005

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251. https://psycnet.apa.org/doi/10.1037/0033-295X.87.3.215

Eva, K., & D Hodges, B. (2012). Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Medical Education, 46*(9), 914–919. https://doi.org/10.1111/j.1365-2923.2012.04310.x

Felthun, J. Z., Taylor, S., Shulruf, B., & Allen, D. W. (2021). Assessment methods and the validity and reliability of measurement tools in online objective structured clinical examinations: a systematic scoping review. *Journal of Educational Evaluation for Health Professions, 18*, 11. https://doi.org/10.3352/jeehp.2021.18.11

Fraenkel, J.R. & Wallen, N.E. (2006). *How to Design and Evaluate Research in Education*. New York, USA: McGraw Hill.

Gauthier, G., St-Onge, C., & Tavares, W. (2016). Rater cognition: Review and integration of research findings. *Medical Education, 50*(5), 511–522. https://doi.org/10.1111/medu.12973

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014a). Seeing the "black box" differently: Assessor cognition from three research perspectives. *Medical Education, 48*(11), 1055–1068. https://doi.org/10.1111/medu.12546

Gingerich, A., Van Der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2014b). More consensus than idiosyncrasy: Categorizing social judgments to examine variability in Mini-CEX ratings. *Academic Medicine, 89*(11), 1510–1519. https://doi.org/10.1097/acm.0000000000000486

Gingerich, A., Ramlo, S. E., van der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2017). Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. *Advances in Health Sciences Education, 22*(4), 819–838. https://doi.org/10.1007/s10459-016-9711-8

Giri, J., & Stewart, C. (2023). Innovations in assessment in health professions education during the COVID-19 pandemic: A scoping review. *The Clinical Teacher, 20*(5), e13634. https://doi.org/10.1111/tct.13634

Goh, H. S., Zhang, H., Lee, C. N., Wu, X. V., & Wang, W. (2019). Value of nursing objective structured clinical examinations: A scoping review. *Nurse Educator, 44*(5), E1–E6. https://doi.org/10.1097/nne.0000000000000620

Govaerts, M., & van der Vleuten, C. P. (2013). Validity in work-based assessment: Expanding our horizons. *Medical Education, 47*(12), 1164–1174. https://doi.org/10.1111/medu.12289

Gwet, K.L. (2014). *Handbook on Inter-rater Reliability*. Gaithersburg, MD, USA: Advanced Analytics.

Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal, 1*(5955), 447–451. https://doi.org/10.1136%2Fbmj.1.5955.447

Hodges, B. (2013). Assessment in the post-psychometric era: learning to love the subjective and collective. *Medical Teacher, 35*(7), 564–568. https://doi.org/10.3109/0142159X.2013.789134

Hopwood, J., Myers, G., & Sturrock, A. (2020). Twelve tips for conducting a virtual OSCE. *Medical Teacher, 43*(6), 633–636. https://doi.org/10.1080/0142159X.2020.1830961

Hyde, S., Fessey, C., Boursicot, K., MacKenzie, R. & McGrath, D. (2022). OSCE rater cognition – an international multi-centre qualitative study. *BMC Medical Education, 22*(6). https://doi.org/10.1186/s12909-021-03077-w

Khan, K. Z., Gaunt, K., Ramachandran, S., & Pushkar, P. (2013). The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. *Medical Teacher, 35*(9), 1437-46. https://doi.org/10.3109/0142159x.2013.818634

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education, 45*(10), 1048–1060. https://doi.org/10.1111/j.1365-2923.2011.04025.x

Lara, S., Foster, C., Hawks, M. and Montgomery, M. (2020). Remote assessment of clinical skills during COVID-19: A virtual, high-stakes, summative paediatric objective structured clinical examination. *Academic Paediatrics, 20*(6), 760–761. https://doi.org/10.1016%2Fj.acap.2020.05.029

Major, S., Sawan, L., Vognsen, J., & Jabre, M. (2020). COVID-19 pandemic prompts the development of a Web-OSCE using Zoom teleconferencing to resume medical students' clinical skills training at Weill Cornell Medicine-Qatar. *BMJ Simulation & Technology Enhanced Learning*, *6*(6), 376–377. https://doi.org/10.1136/bmjstel-2020-000629

Massey, D., Byrne, J., Higgins, N., Weeks, B., Shuker, M. A., Coyne, E., Mitchell, M., & Johnston, A. N. B. (2017). Enhancing OSCE preparedness with video exemplars in undergraduate nursing students. A mixed method study. *Nurse Education Today, 54*, 56–61. https://doi.org/10.1016/j.nedt.2017.02.024

Navas-Ferrer, C., Urcola-Pardo, F., Subiron-Valera, A.B., & German-Bes, C. (2017). Validity and reliability of Objective Structured Clinical Evaluation in nursing. *Clinical Simulation in Nursing, 13*(11), 531–543. https://doi.org/10.1016/j.ecns.2017.07.003

Patrício, M. F., Julião, M., Fareleira, F., & Carneiro, A. V. (2013). Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Medical Teacher, 35*, 503–514. https://doi.org/10.3109/0142159X.2013.774330

Paul, F. (2010). An exploration of student nurses' thoughts and experiences of using a video-recording to assess their performance of cardiopulmonary resuscitation (CPR) during a mock objective structured clinical examination (OSCE). *Nurse Education in Practice, 10*(5), 285–290. https://doi.org/10.1016/j.nepr.2010.01.004

Purpora, C., & Prion, S. (2018). Using student-produced video to validate head-to-toe assessment performance. *Journal of Nursing Education, 57*(3), 154–158. https://doi.org/10.3928/01484834-20180221-05

Roberts, R., Cook, M., & Chao, I. (2020) Exploring assessor cognition as a source of score variability in a performance assessment of practice-based competencies. *BMC Medical Education, 20*(1), 168. https://doi.org/10.1186/s12909-020-02077-6

Rushforth, H. E. (2007). Objective structured clinical examination (OSCE): Review of literature and implications for nursing education. *Nurse Education Today, 27*(5), 481–490. https://doi.org/10.1016/j.nedt.2006.08.009

Scully, C., Kelly, M., Lysaght, Z., & O'Leary, M. (2024). The cognitive processes employed by undergraduate nursing OSCE assessors: A qualitative research study. *Nurse Education Today, 134*, 106083. https://doi.org/10.1016/j.nedt.2023.106083

St-Onge, C., Chamberland, M., Lévesque, A., & Varpio, L. (2016). Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. *Advances in Health Sciences Education, 21*(3), 627–642. https://doi.org/10.1007/s10459-015-9656-3

Willis, G.B. (2015). *Analysis of the Cognitive Interview in Questionnaire Design*. Oxford: Oxford University Press.

Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013). Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Advances in Health Sciences Education: Theory and Practice, 18*(3), 325–341. https://doi.org/10.1007/s10459-012-9372-1

Yeates, P., Moreau, M., & Eva, K. (2015). Are Examiners' Judgments in OSCE-Style Assessments Influenced by Contrast Effects? *Academic Medicine: Journal of the Association of American Medical Colleges, 90*(7), 975–980. https://doi.org/10.1097/acm.0000000000000650

Leveraging Student Voices to Explore Career Interest in STEM PhD Programs

Authors:

Jennifer Claydon, PhD
Yale University

Meghan Bathgate, PhD
Yale University

ABSTRACT

As career landscapes within and outside of academia shift, higher education STEM programs increasingly must navigate assessing student experience and career goals with an eye towards actionable improvements in career preparation for myriad roles. In this study, we aimed to better identify and respond to career interests across four years of graduate students (N=364) as they considered their post-graduation plans at different points in their training. These results revealed students are most interested in pharmaceuticals/biotechnology careers, postdoctoral positions, research-heavy faculty positions, and non-faculty academic roles, though the most notable results are the reductions in degree of career interest between the first and last year of PhD training. Specifically, first year students express significantly higher levels of interest (depth) across more career areas (breadth) compared to graduating students. This work highlights the changing pattern of career interests alongside how a collaborative assessment approach can inform choices in how to best support students.

Correspondence E-mail: Jennifer.Claydon@yale.edu
Keywords: Assessment, STEM education, graduate school, career development

Many students enter graduate education with an eye towards securing a position in a tenured faculty role and the traditional apprenticeship model of training these students follows that desire (Anderson, et al., 2011; Walker, et al., 2008). Unfortunately, the availability of those highly sought after tenure track roles has been in decline over the past decade (National Academies of Sciences, Engineering, and Medicine, 2018). In 2008, 25.9% of STEM graduates had secured a tenure track faculty position within 5 years of graduating, and only 17.7% percent of STEM Ph.D.'s secured these positions by 2015 (NSB, 2018). Studies by Larson and colleagues explored whether there was a dearth in positions, or a burst of additional PhD graduates that contributed to this trend (Larson et al., 2014; Xue & Larson, 2015). They found that although the overall number of STEM PhDs had been climbing steadily, the number of tenure track positions remained nearly constant in most fields.

The lack of new tenured positions for graduates has shifted the landscape towards non-traditional roles in fields such as biotechnology, consulting, and finance (Garrison, 2024) with a majority of doctoral graduates pursuing positions outside of academia (Larson et al., 2014). As data collection and data accessibility have grown, administrators are seeking ways to assess the student experience that align with actionable ways to apply findings (Montenegro & Jankowski, 2020; Cubarrubia, 2019). Implementing improvements in higher education training first requires institutions to understand the successes and challenges of their graduate student population, making the path to best support new scholars both a disciplinary and assessment challenge. Despite evaluation of the graduate student experience in STEM PhD programs being a way to better identify and remedy potential gaps in career development during training, these assessments have been poorly documented at a national level (Reeves et al., 2022).

The challenges to designing and sharing common models of how student experience is assessed are varied, including where accreditation requirements are standardized. Even when students share their experiences, the follow-through to reviewing and responding to this feedback is often incomplete (Blaich & Wise, 2011; Banta & Blaich, 2010; Baghramian & Roberts, 2023). The gaps in knowledge create missed opportunities for departments to highlight strengths and allow higher education to remake the same mistakes cohort after cohort (Alberts et al., 2014).

The Current Study

Our work here describes how an interdisciplinary program in the biomedical and biological sciences used collaborative assessment to reflect on and apply student perspectives to the training graduate students received. Specifically, our research questions are:

- R1: What career areas are graduate students in the Biological and Biomedical Sciences Umbrella Program (BBS) interested in, and do those interests shift over the course of graduate training?
- R2: How can using collaborative assessment contribute to designing measures that collect and use data to center student experiences in STEM PhD programs?

Funding agencies for graduate training have been shifting their language to support approaches that modify graduate training to align with job market demands (Blume-Kohout, 2007; Denecke et al., 2017; Fuhrmann et al., 2011). Finding viable and efficient avenues for graduate programs to implement career preparedness across various fields has proven challenging (Subramanian et al.2022; Bixenmann et al., 2020). Anecdotally, our program knew students were consistently exploring non-traditional career options outside of mainstream tenure-track faculty positions but had not consistently examined student experiences about career interests. Ganapati and Ritchie explored student perceptions of how their graduate programs prepared them for positions outside of the traditional academic pathways and found that students consistently requested tailored professional development for various career trajectories (2021). As we explored ways to assess student experience, and subsequently how to initiate evidence-based decisions in PhD career development, we found variability in how students had been engaged in the assessment process and how training programs addressed career preparation (Gibbs & Griffin, 2013). Golde and Dore discussed the concerns that PhD graduates "often struggle to make the transition out of the academy and into the workforce," perhaps because of a lack of specific career preparedness programming for a variety of career trajectories (2011). In order to best understand our population of student needs and interests, we aimed to design an assessment survey built on input from our student population. Towards this end, we embedded collaborative assessment techniques into the entire process of designing questions, collecting and analyzing data. The details of these procedures are outlined in the Methods. Additionally, once data were collected and analyzed, we partnered with students and departmental leadership to instigate action at the programmatic level.

Context of Collecting Program-Specific Data in Higher Education

There has been an increasing demand for assessment expertise within educational departments, in part driven by evaluation criteria for research and training grants. Despite a recognition of the importance of assessment, there are challenges in using assessment data. First, as assessment professionals, those collecting the data in higher education tend to have a lack of direct authority to implement any of the changes that might be warranted or recommended. This includes institutional research offices, teaching centers, student affair offices, and the like. These hubs of data are often well-informed, though removed from the direct training of students. Second, once an assessment report has been created and sent along with recommendations, it is not uncommon for there to be minimal follow-up from stakeholders. At times, our colleagues have informally described feeling as if their reports "float into the ether." A variety of explanations might exist for why administrators struggle to navigate the connection between data on student experience and how to shift or improve programming to influence that data in future assessments. At the crux of these explanations, communication among all involved individuals plays the largest role in supporting stakeholders as they digest data and balance administrative pressures within programs.

Furthermore, understanding how data are actively used can often be muddied by the data themselves. Sometimes in higher education, assessments can be scaled too large, with stakeholders attempting to collect too many variables at one time. As survey fatigue continues to climb, low response rates can call into question what to do with the data that

are collected (Fass-Holmes, 2022). Additional research has demonstrated that it is atypical within higher education to use survey feedback in decision-making to improve educational programs (Jonson et al., 2014). Finally, poor communication of the findings to key stakeholders—including to the participants themselves—can dampen enthusiasm for future participation and reduce the value placed on providing feedback for educational programs.

In order to use student feedback in actionable ways, the design of any assessment tool needs to incorporate deep consideration of the participant experience coupled with collaborative stakeholder engagement. The relationships built during the entire assessment process are the foundation for effective evaluation (Chouinard & Cousins, 2009); Wilson, 2008). A primary goal of our assessment team is understanding how to engage students, faculty, and administrators in the discussion, purpose, and use of evaluation. In designing surveys of the graduate student experience in the BBS program, we employed the Collaborative Assessment Model (CAM), first presented by Bathgate and Claydon in 2021. CAM advocates that effective and successful assessment requires four broad principles to guide the work and elevate interpersonal relationships throughout the assessment cycle. Specifically, program assessment must be aligned to program goals, actionable when data are collected, sustainable for programs to maintain over time, and contextual in their conception, design, collection, analysis, and reporting (see Figure 1).

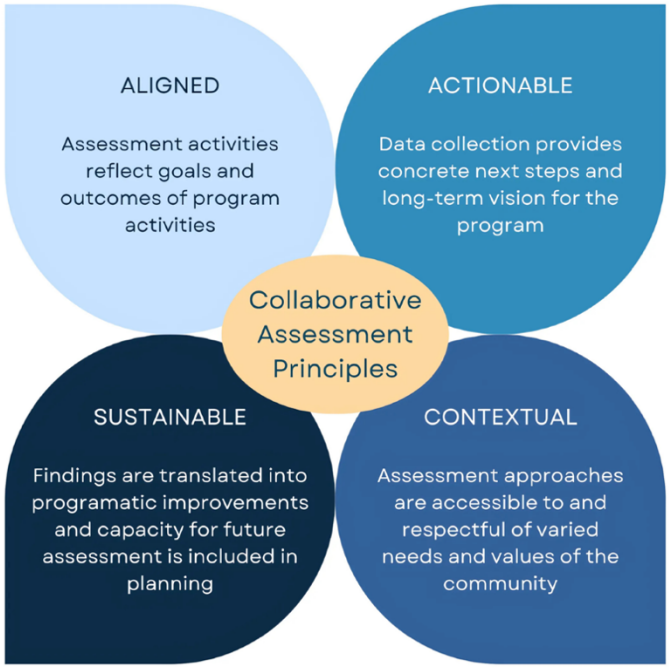


Figure 1. *The Four Principles of the Collaborative Assessment Model.*

Before reviewing our specific methods for this study, we first outline ways in which we applied CAM to this work. First, the authors met to articulate where our assessment activities would align with program goals, which were largely shaped by the program's

objectives for meeting PhD milestones and by the grant-specific objectives towards training highly skilled scientists. These discussions included meeting with faculty Principal Investigators on training grants (NIH T32s) to ask them about how they assess whether students are meeting their department specific goals. Throughout the survey development, we met with these individuals a couple of times, and traded drafted language via email, to ensure our questions for students aligned with the goals of the program. Questions were also discussed twice with the Yale BBS Development & Involvement Community (YBDIC) student group of more than 20 individuals to ensure they were interpreting the questions the way we were intending, a technique known as cognitive interviewing (Beatty & Willis, 2007). Cognitive interviewing has emerged as a prominent way to create ideal question formats and language that is understood by the target participants.

In making the findings as actionable as possible, we engaged stakeholders (students, faculty) in the BBS program in the roles of Directors of Graduate Studies (DGSs), student program leaders, specifically in the YBDIC, early and frequently, including attending departmental meetings across the BBS program, and articulating the goals behind the assessment work. We met with each department twice: once before the survey launched, and once to discuss the findings after everything was collected and analyzed. We worked collaboratively with faculty and students to show them why we were asking specific questions, how the answers to those questions could inform programmatic changes at the departmental level, and how those changes would benefit the student experience specifically with career preparedness. We also presented findings once to the larger BBS group of DGSs and departmental registrars to encourage them to review the findings and seek out ways to implement changes to improve the graduate student experience.

Sustainability of the assessment was a focus from initial discussions with faculty program leaders, as we recognized that longitudinal data from students and their mentors needed to be nimble enough to fit within existing structures with changing staff roles and robust enough to meaningfully inform the program. In this light, we decided not to survey all graduate students in the program, but targeted our surveys to the end of the first year when students are choosing a PhD lab and department, and at the end of their academic career in the BBS program when they are about to graduate, to obtain perspective on career interest at different parts of the program. We also used program alignment and student feedback to maintain a shorter survey so that students did not feel overly taxed with the time required to share their experiences.

Finally, we recognized that the context in which the assessment took place inherently shapes the framing of the data. For example, who we included to engage for feedback, the language we used in communications, and the timing, method, and tone in which data were shared back, all influence this work. As such, we reflected as a team and with program leaders on the selection of our methods and measures to best encompass and represent the perspectives of all students and discussed how to iteratively share results to allow for broad faculty access to these findings.

In this study, we share assessment findings describing patterns in career interest among four cohorts of first year and graduating students who complete their STEM PhD training in our interdisciplinary umbrella program. We also provide examples of how positioning, implementing, and reflecting on assessments within STEM PhD programs

can promote responsiveness to student feedback. Throughout, we share our assessment approach and how it allowed us to engage departmental programs at different strategic points. Researchers interested in STEM persistence may be most drawn to the findings of this work, while assessment professionals may be more interested in the process of gathering, reflecting, and acting on these data. With either lens, the context of the data collection and content of the findings reflect each other in ways that further spur conversations in the higher education assessment community.

Method

In line with the principles outlined in CAM, we walk through the specifics of our methodological decisions here. We prioritized equity-centered practices as we designed measures exploring multiple areas of the graduate student experience (Henning & Lundquist, 2022). A single survey, even if excellently designed, cannot fully capture the broad range of students' experiences nor fully represent the nuance for any single student. The complexity in the context and pressures of graduate school vary in ways that are challenging to capture. To help represent and respect these differences, we engaged a few strategies that allow for students to meaningfully shape and guide our approach and methods. Namely, we integrated data from multiple sources so as not to rely on a single timepoint or item type (Hohensinn & Kubinger, 2011), engaged students and program leaders in co-creating assessment measures, collaborated with students to verify our interpretations of the data reflected their experiences, used a series of norming conversations to reduce bias when coding qualitative data, and collected mixed-method data to better represent the nature and variety among students' perspectives and voices. We also met with faculty and graduate students to discuss and receive their reflections on the analyses, and subsequently sought multiple avenues for disseminating our findings to share back the findings with the community (Oliveri et al., 2019).

As we designed our measures, we started with a literature review of current assessment methods in graduate education and incorporated several existing measures (Anderson et al., 2016; Sinche et al., 2017). Additionally, we designed our own question of career interest across thirteen common sectors. These thirteen sectors were previously identified using BBS alumni data housed in the Institutional Training Grant hub (ITG hub) database of which career areas our graduates have been working over the past 25 years. We also included an "other" category where students could write in a topic or career area they did not see listed among the options.

Our quantitative career interests items asked students to rate the following sentence, "I am interested in pursuing a career in this field" for each of thirteen career areas (see Table 1) on a Likert scale of "1-strongly disagree", "2-moderately disagree", "3-disagree a little", "4-agree a little", "5-moderately agree", to "6-strongly agree." Demographic questions were also collected such as department tract, year of study, and gender. We also asked students the open-ended question, "Do you have any comments or suggestions for the BBS program regarding your graduate student experience?"

The surveys were finalized by working collaboratively with graduate student groups, such as the YBDIC, with Directors of Graduate Studies (DGSs) and training program directors, sharing our final survey drafts before distributing the survey to the target population. The current study is part of the larger programmatic assessment of the BBS department and its tracks, complementing its training grant funding, and

institutional goals. There were other measures in the broader survey that included questions about satisfaction with specific training activities and confidence across skill sets, for example. For our current research questions, we focus on career interests measures.

Table 1. The thirteen career areas students were asked to rate their interest in pursuing.

1. Postdoctoral Training	2. Finance or Law
3. Faculty in academia at research intensive institution	4. Government or Non-profit
5. Faculty in academia at teaching intensive institution	6. Business/Entrepreneurship
7. Academic, other job type	8. Publishing/Communications
9. Pharmaceuticals or Biotechnology	10. K-12 Education
11. Healthcare or Clinical	12. Library Science
13. Consulting	Other (please specify)

Procedure

To explore whether and how graduate student career interests may change over their course of study, career interest was measured at the end of the first year of graduate training when students are joining their PhD labs and again at the end of their training as they defend their dissertations and graduate (see Figure 2). Surveys were sent to students' emails via Qualtrics software in May of 2021, and have been administered annually in May since then (2022, 2023, and 2024). Students were identified using the ITG hub database to generate a list of all first-year students, and a list of all students who were graduating the respective year. Data were collected confidentially, but not anonymously. Being able to link student data over time would eventually allow us to match within subjects from the end of their first year to the year they graduated (another few years of data will be needed before within subjects' comparisons can be explored due to the years of study necessary for students to receive their degrees). Data were de-identified after exporting from Qualtrics before conducting any analyses. This study was approved by the Yale University Institutional Review Board (# 2000024769).

Analysis

Quantitative and Likert scale questions were processed and compared using SPSS software (version 24.0 for Mac). All eight surveys (two per year from 2021-2024, inclusive) were combined into one large dataset to explore all BBS student career interests and delineated by whether each response reflected first year students or graduating students. The open-ended question "Do you have any comments or suggestions for the BBS program regarding your graduate student experience?" was thematically coded according to the 6-

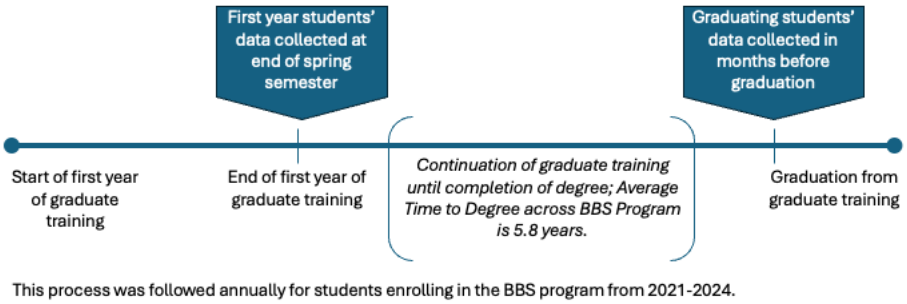


Figure 2. *Data collection process for surveying first year and graduating students.*

step process detailed by Braun and Clarke, using NVIVO qualitative analysis software by the authors (QSR International, 1999; Braun & Clarke, 2006 & 2013). Responses relating to career interest, career development, career self-efficacy, or professional development were identified and further coded. Two norming and discussion sessions were held between the authors to develop a shared coding structure and familiarize ourselves with the data, which used grounded theory to enable initial codes to emerge. Open coding resulted in several themes that were then refined further with axial coding to review potential themes in the responses. The authors then defined and named the themes in the responses and defined the themed results presented here. The sample size for the open-ended coding was smaller, relative to the overall sample, and codes, counts, and percentages are included in the results section to complement themes found in our quantitative analyses.

Results

Over the four years of survey data collection, a total of 364 students participated in the survey, representing a 47% response rate from the overall potential population (55% from first years and 39% from graduating students, on average). There was a relatively stable response rate of about 40-50 people per cohort (first year students, graduating students) per year (2021-24). Sample sizes across first year students and graduating students are included in Table 2, with about 60% of the sample being female.

Table 2. Sample of BBS participants across 8 surveys.

	Total Students	**Percent**
All Students	364	100%
1st Years	224	62%
Graduates	140	38%

Overall Interest

Students' interests varied across career categories, as Figure 3 shows.[a] On average, students were most strongly interested in a pharma/biotechnology career (M=4.53, SD=1.50), pursuing a postdoctoral position (M=4.10; SD=1.87), or becoming a faculty member at a research-intensive institution (M=3.89; SD=1.89). Average interest levels varied across the remaining careers, with lower average interest towards careers in finance or law (M=2.20; 1.45), K-12 education (M=2.06; SD=1.37), and library science (M=1.71; SD=1.08).

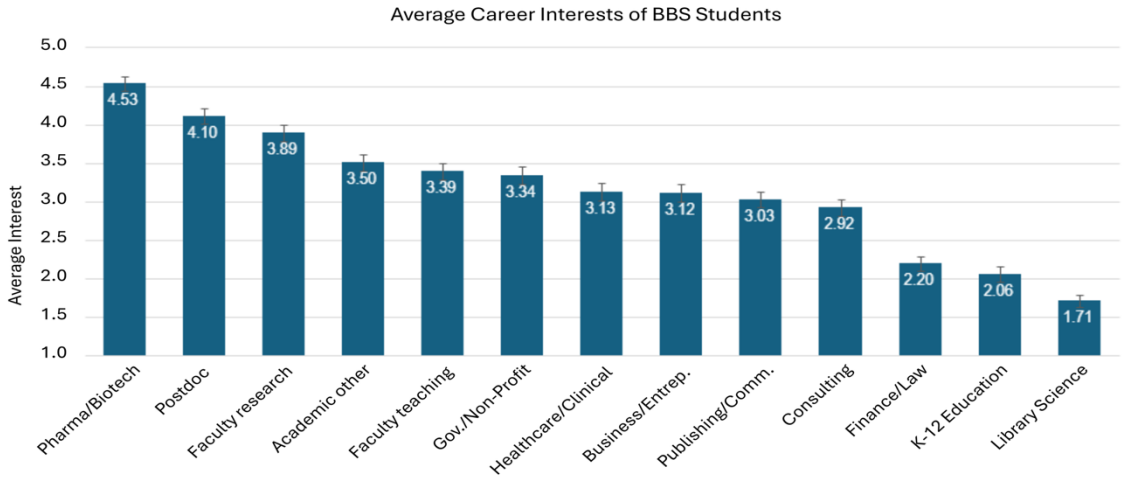


Figure 3. *Average Career Interest of BBS students from 2021-2024. For additional descriptive statistics for whole sample's average interest in disciplines, please see Appendix A.*

Interest Differences Between First Year and Graduating Students

To explore whether career interest differed between first-year students and graduating students, we ran a repeated measures ANOVA including cohort as a between-subjects variable and career category as the within-subject variable. Students' interest varied significantly by career category, which reflects the spread of means in Figure 3 ($F(1,6.17)=94.70$, $p<.001$, partial eta^2=.22)[b].

[a]While there are statistical differences across some categories, we do not report every pairwise comparison across categories, both for theoretical and statistical limitations (e.g., compounding error across more dozens of comparisons within the same sample). Additionally, a handful of students opted to write in a career choice in the open-text box, though there was little consistency in these areas and they are not included in these analyses. These write-in areas included: administrative positions, project management, science policy, data science, scientific art production, biotechnology, research at non-academic institutions, translational research, medical school, sustainability, artificial intelligence, psychology, community college professor, culinary arts, and further education.

[b]Greenhouse-Geisser correction applied, given Mauchly's Test for Sphericity was significant ($W=.029$, $p<.001$)

Further, we found a significant effect of cohort (F (1,6.17)=40.63, p<.001, partial eta^2 =.11) and a significant interaction between cohort and career categories (F (1,6.17)=7.70, p<.001, partial eta^2=.02). This means first year students' interest varies from graduating students, and that interest varies by career category for both first year students and graduating students, respectively.

While this overall difference is important to note, additional analyses are needed to explore specific contrasts between first year and graduating students interests in careers. To examine pair-wise differences, we conduced t-tests comparing first year students' average interest to graduating students' average interest across each career option. Bonferroni corrections were applied to establish a more rigorous significance threshold that better accounts for repeated comparisons (α=0.0038). There were five significant differences emerging with three additional differences that were significant at the p<.05 level but did not meet the more stringent Bonferroni corrected alpha. Figure 4 and Table 3 give details across each comparison. The largest differences were all within the academic industry: postdoctoral position, faculty research, faculty teaching, or other academic position. Graduating students' interests in each of these four areas were significantly lower than first year's interests. Also notable is that these are four of the top five career interests of first year students. Publishing/communication also showed lower interest for graduating students. The consulting, healthcare/clinical, and library science showed trending differences with lower graduating student interest compared to first years, though library science showed the lowest interest overall. These results show us that while there are trends of career interests within the sample, the overall average obscures notable variations in the depth and type of career interests from early to late graduate school.

In addition to the disciplinary differences in interest between cohorts, we also see a difference in the breadth of interest, as measured by the number of topics students selected they were at least "a little" interested in ($\geq$4 on the six-point scale). Specifically, students in the 1st year selected an average of 6.7 areas of interest (i.e., selecting that they are at least "a little" interested in a given career), with an average of 48% of students reporting interest in more than half the career areas. Graduating students selected an average of 5.2 areas of interest, with an average of 23% of students reporting interest in more than half the career areas. Taken together, we see a narrowing and deepening of career interest in graduating students compared to a broader interest in first year students.

The open-ended question about improving the BBS program had a total of 103 participants respond, from the 364 who took the survey (28% response rate to the open-ended question). These 103 responses were reviewed by both authors and 11 responses (11%) were selected for further coding for the current study as they mentioned a career-related improvement to the BBS program (Cohen's Kappa, k = 1.0). The other responses were unrelated to career interest and either focused on various other aspects of graduate training, such as stipend, length of lab rotations, and administrative suggestions, or were made up of 'no', 'N/A', or praise for the overall program. Upon initial coding, the first author determined five main themes out of the 11 responses which were discussed with the second author. Given the low number of total responses for this sample, the authors agreed 100% on the five themes below. None of the five themes had more than five responses total. The purpose of coding this question was to consider whether additional

nuance or depth about students' career interests and preparation could be offered, recognizing that most comments were unrelated to the current research questions.

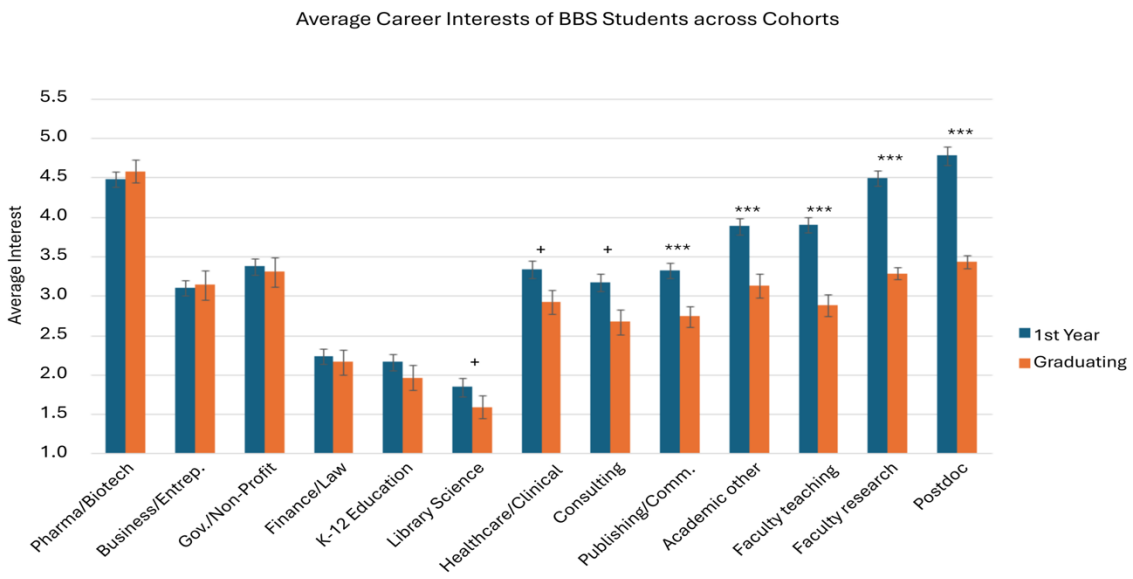


Figure 4. *Average Career Interest separated by first year and graduating BBS students from 2021-2024.*

Table 3. *Pairwise t-test comparisons of differences in career interest between first year and graduating students.*

	t	df	sig.	d
Postdoc	6.30	201.45	***	0.77
Faculty Research	5.67	220.05	***	0.67
Faculty Teaching	5.45	253.12	***	0.62
Academic Other	3.98	235.58	***	0.46
Publishing/Communications	3.58	339	***	0.40
Consulting	2.79	255.64	+	0.32
Healthcare/Clinical	2.21	338	+	0.25
Library Science	2.20	309.17	+	0.24
K-12 Education	1.30	339	n.s.	0.14
Finance/Law	0.47	337	n.s.	0.05
Gov/Non-Prof	0.38	339	n.s	0.04
Business/Entrep	-0.21	340	n.s	-0.02
Pharma/Biotech	-0.57	250.28	n.s.	-0.07

Significance is indicated by *** when the t-value exceeded the Bonferroni corrected alpha (α=.00384) and + when t-value exceeded alpha threshold of (α<.05). Cohen's d effect size is also included. Corrected degrees of freedom were used in cases where tests for homogeneity of variance were significant.

The 11 responses were coded into five themes, including the need for additional statistics courses (n = 5), better support for career exploration (n = 4), support for finding funding sources (n = 2), additional time for career planning (n = 1) and prioritizing teaching (n = 1). Selected deidentified representative quotes are presented below.

> "Also, I would appreciate if the program were more accepting and supportive of first years pursuing internships. The BBS advertises internships/extracurricular activities in the newsletter, but I received somewhat negative feedback from my program director about how the summer internship would set me back even though I am pursuing the internship to learn a skill and even though some of my colleagues and peers are still rotating over." [Time for Career Planning]

> "My only other suggestion would be to perhaps prioritize teaching a bit more, or at least have avenues for that. While we require TF semesters, most of the time the courses available do not require any independent teaching (at least that's been the experience I've observed). That's fine for some, but for others who want to teach, I think it poses a challenge." [Prioritize Teaching]

> "Data analysis courses could be offered in 1st year from BBS, and that would be extremely helpful for many BBS students." [Additional Statistics Classes]

> "I also don't feel that the program provided enough resources or support for people trying to enter fields other than academia or industry. Often there wasn't even acknowledgement that other options exist for people with science PhDs." [Support for Career Exploration]

> "Need more career support and scholarship support" [Finding Funding & Support for Career Exploration]

Disseminating Findings and Curricular Responses

We emphasized actionable and sustainable elements of the collaborative assessment model by sharing results not only with participants, but all possible stakeholders to help decision makers design future workshops, webinars, or courses that align with student needs based on equitably designed assessment measures. In our dissemination of results, we presented data in multiple ways and across multiple audiences. This included presentations emphasizing findings (e.g., PowerPoint slides of graphs and tables), written and verbal executive summaries of changes to consider, and communication strategies for how to share these findings within departments. Results were shared with current BBS students through summary reports in the program newsletter and to key student groups to share within graduate student networks. Students were encouraged to have discussions with their labs to bolster awareness of student career interests and support further career exploration throughout their PhD training. Formal presentations and slides were also given and shared with the entire 50-person faculty and

registrars within the BBS Executive Committee who were then encouraged to share broadly within their home departments. Presentations included not only the results from the survey, but also recommendations and possible action steps for departments and tracks to consider as they reviewed the student experience data on career interest. Our hope in sharing summary data was to develop a collaborative process to address the needs students were highlighting and to acknowledge the broader calls to revolutionize training programs based on evidence and student experience (Gammie & Gibbs, 2017).

Curricular & Programmatic Responses

Similar to O'Meara et al., (2014) who revealed departmental attitudes and support for non-traditional academic career paths can support graduate student awareness and interest in a variety of career trajectories, the BBS program made actionable shifts in curriculum and programs offered to the graduate student population based on these data and the subsequent discussions of the results. Below, please find several examples of how BBS programs have implemented changes based on the graduate student feedback from these surveys.

- Interdepartmental Neuroscience Program: incorporated student voice by adding students to their Executive Committee and making significant curricular revisions to focus on diverse skills development that prepare students for a variety of career sectors.
- Computational Molecular Quantitative Biology: added a required statistics course for all trainees to develop deeper knowledge of scientific research avenues and added events with diverse guest speakers to raise awareness of additional career avenues outside the traditional academic routes.
- Genetics Department: revamped their orientation programming to offer student-led "bootcamps" in various coding, statistics and data analysis skill sets.
- YBDIC, a group independently run by graduate students across all PhD programs in the BBS, hosted a day long panel in the spring of 2025 on diverse careers.

Discussion

Our exploration into career interests of STEM graduate students was conducted to explore firstly what career areas our students expressed interest in, and secondly how we could use collaborative assessment principles to design measures that collect and use data to center student experiences in STEM PhD programs. Results showed differing interests between first year and graduating students. Specifically, first year students showed greater depth of interest in many careers (measured by their average interest) and greater breadth of interest in careers (measured by nominal interest across careers). These findings suggest that students narrow their career interests during their course of study. To some extent, this is expected, given that students gain disciplinary expertise during their training, learn the ins and outs of their fields, and begin to more practically consider their next career step as they approach graduation. Yet, the strength of these differences and the overall attenuation of interest to lower averages across many disciplines was notable, particularly in academic careers. There is an increasing pattern of graduate students considering careers outside of academic track faculty, whether by choice or by a dearth in these positions (Bird & Rhoton, 2021; Cidlinska et al., 2023; Roache & Sauermann, 2017; Sinche, 2018). Our current work adds to this conversation by showing that students, at

least in this sample and program, start out with academic positions (e.g., postdoc, research faculty position) among their highest interests.

Similar to Sinche (2018), who found that "only 14% of PhD graduates in science occupy tenure-track positions five years after completing their degree," we noted elevated student interest for non-academic career trajectories. This finding was first noted on a smaller dataset exploring how career interests can change over the course of a semester in graduate school (Claydon et al., 2021). These researchers found that a semester long course, with each week focusing on different career avenues, was enough to shift student knowledge of and interest for various career trajectories. Building on that knowledge, our recent survey data demonstrates clear avenues to support student exploration of career interest across a variety of career sectors.

While differences between first year and graduating students are themselves notable, the dissemination of these results and the accompanying programmatic response is a major contribution to why this work matters and is worthy to share. To cultivate program-level impacts required to maintain a focus on actionable assessment, we intentionally shared back results with key stakeholders to build better awareness, transparency, and responsiveness to student experience. Meeting with departments to reflect on patterns of students' career interests provided an opportunity to bring data into curricular action, with four curricular and programmatic shifts made directly in response to these discussions within a year of their review. Without coordinating discussions to bring these findings to department and program meeting agendas, they would not have as readily been integrated into a strategic plan and the students may not have seen the impact of their feedback.

Our research demonstrates how assessment processes can be used to build community within and across programs, with the goal to best prepare trainees for future success across myriad career trajectories. How data are used can visibly shift the perspectives of the participants providing that data, and if students can see changes being implemented in their programs, this can inspire trust in the stewardship of their feedback and create an excitement to provide additional data in future endeavors.

Collaborations have been blossoming nationally to support these initiatives, such as the workshops on "Enhancing Dissemination of Evidence-Based Models for PhD Career Development" (Bixenmann et al., 2020). Raising awareness among more faculty to both understand the current market landscape for graduates, in addition to understanding the career interests of the students in their programs, can support broader training and skills development that would seamlessly integrate into a variety of positions. Our embedded assessment efforts built momentum for students to share their data and for faculty to receive the results in a way that allowed for actionable programmatic changes.

For other institutions, their student populations may be looking for avenues to share their career interests in the current career landscape for STEM PhDs. Understanding student perspectives can provide faculty with the information needed to examine their career development offerings and to ascertain whether changes need to be made to best prepare their future scientists. The collaborative evaluation approach presented here can be used as a framework by other institutions to help them conduct evaluations of their own programs, starting with identifying what assessment support exists at their institutions. If there is an Office of Career Strategy, or a place where graduates can go to learn about varied career avenues, faculty could partner with these

individuals to design survey questions (or adopt other methods, such as focus groups or a town hall), to first gather student perspectives on career development and career interest. Communicating with students about the purpose of a career interest assessment will help generate student excitement to share their perspectives, while communicating with training program leaders can support new initiatives to use collected data to adjust career development offerings. If a goal of STEM PhD education is to prepare the next generation of scholars adept at integrating across myriad career trajectories, then building supports to prepare students for those trajectories can be accomplished by engaging students, faculty, and staff to assess career interests and subsequently develop relevant content.

Limitations and Future Directions

Our work can be expanded on and bolstered given the boundaries of what can be concluded using our data. One limitation of the current work are the middling response rates, especially for graduating students, revealing the need for streamlined processes to capture when students are defending their dissertations, communicate expectations for and demonstrated use of students' feedback, and to potentially offer motivational incentives (monetary or not) for completing surveys within an expected timeline. We are navigating various administrative infrastructures to coordinate across a dozen departmental registrars such as building a simple but centralized database to capture key information that will drive timing of assessment collection methods. Future work with this project also includes assessing the impact of the program and departmental changes that were made in response to these data.

Additionally, we recognize that while the breadth of the BBS tracks encompasses a range of STEM disciplines, our sample draws students who may be seeking a particular trajectory based on the faculty areas of study or the university's association with medical school access, for example. The extent to which we can verify our current students' views represent the career interests of students across higher education is limited. As we build our dataset over time, share our work with colleagues exploring related questions, and look within and across the disciplines represented in our institution, we can speak to more robust patterns.

Our current analyses also do not explore patterns related to race, gender, or international status, though we acknowledge there are notable challenges of access and equity in graduate STEM education associated with such categories (Posselt et al., 2021). We opted not to include these variables in our analyses for a few key reasons. First, the completion rates paired with the limited percentage of members within given categories (whether self-described or gathered from admissions), particularly when intersectionality is considered, limits our ability to make patterned claims based on such demographic categories. Given this sample size, analyzing people based on such categories also limits our ability to adhere to confidentiality and lends towards overinterpreting the influence of one demographic factor over another. For example, there are a small set of people who are first year, female, Black, international students who have enrolled in the past few years. If we parse people along multiple categories, we risk pointing towards identifying members who contributed to this work and are likely overestimating which part of one's multi-dimensional identity may be associated with any pattern in the data. We recognize that people's individual and cultural identities are a factor in how they interact with the world around them, and want to respect the need that without much larger samples that allow

us to understand individual and contextual factors, we may be misattributing effects with limited data to help us explain such differences.

Similarly, membership within given labs may influence people's career interest over time, assuming that skill sets and career examples may be more central to some disciplinary tracks and lab staffing than others. For similar reasons related to confidentiality and limited identifiable contextual factors (mentorship, perception of principle mentor encouragement, resource support, teaching experience), we do not examine lab-by-lab effects in our current sample. We look across the averages of the sample to give us a broader direction of patterns in career interest across years of study. In relation to the open-ended data presented here, we recognize that a specific question related to career interest and preparedness would provide better results than the broad question used on these surveys. In future iterations, we would like to expand the section for open-ended feedback on graduate student experiences with career interest and development during graduate school.

Future directions for this research involve alumni research to explore how recent graduates navigate the job market, and to what degree their expressed interests while in graduate school translate to their post-graduation job positions. Additionally, given the timeline of our data collection and students' time to degree, we are in the first years of exploring within-sample data that follows the same set of students from enrollment through graduation, paired with career choices of alumni. A repeated measures approach will allow us to finetune our analyses and look at individual patterns or sets of patterns over time. The current dataset takes a between-sample approach using cross-sectional cohorts of students, which may be more sensitive to individual differences. We aim that the year over year analyses of this work will allow us and our colleagues to continuously improve the BBS program and best prepare the next generation of scientists.

Appendices:

Appendix A: Descriptive Statistics for Whole Sample's Average Interest in Disciplines

References

Alberts B., Kirschner M.W., Tilghman S., Varmus H. (2014). Rescuing US biomedical research from its systemic flaws. *Proceedings from the National Academy of Sciences U S A, 111*(16), 5773–5777. https://doi.org/10.1073/pnas.1404402111.PMID: 24733905; PMCID: PMC4000813

Anderson, W.A., Banerjee, U., Drennan, C.L., Elgin, S.C.R., Epstein, I. R., Handelsman, J., & Warner, I.M. (2011). Changing the culture of science education at research universities. *Science*, *331*(6014), 152-153. Doi: 10.1126/science.1198380

Anderson, C.B., Lee, H.Y., Byars- Winston, A., Baldwin, C.D., Cameron, C., & Chang, S. (2016). Assessment of scientific communication self- efficacy, interest, and outcome expectations for career development in academic medicine. *Journal of Career Assessment*, *24*(1), 182-196. doi:10.1177/10690 72714 565780.

Baghramian, A., & Roberts, L. (2023). Change is hard: Closing the loop in the assessment process, *Western State Law Review, 50(1),* 1-20.

Banta, T.W., & Blaich, C. (2010). Closing the assessment loop. *Change: The Magazine of Higher Learning, 43*(1), 22-27. https://www.slu.edu/provost/educational-program-development-review/assessment-student-learning/program-level/assessment-resources/closing_the_assessment_loop.pdf

Bathgate, M.E., & Claydon, J.L. (2021, October). Designing Collaborative Assessment that is Inclusive and Actionable, Assessment Institute in Indianapolis (IUPUI presentation).

Beatty, P.C., & Willis, G.B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*(2), *287-311. https://doi.org/10.1093/poq/nfm006*

Bird, S.R., & Rhoton, L.A. (2021). Seeing isn't always believing: Gender, academic STEM and women scientists' perceptions of career opportunities. *Gender & Society, 35*(5), 422-448. DOI:10.1177/08912432211008814

Bixenmann, R., Natalizio, B.J., Hussain, Y., Fuhrmann, C.N. (2020). *Enhancing Dissemination of Evidence-Based Models for STEM PhD Career Development; a Stakeholder Workshop Report*. Worcester, MA: Professional Development Hub, University of Massachusetts Medical School. https://doi.org/10.13028/79a5-ym66

Blaich, C. & Wise, K. (2011). *From gathering to using assessment results: Lessons from the Wabash National Study*. NILOA.

Blume-Kohout, M. (2017). "On what basis? Seeking effective practices in graduate STEM education." Washington, DC, National Academies of Sciences, Engineering and Medicine.http://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_186176.pdf

Braun, V. & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3, 77-101. http://dx.doi.org/10.1191/1478088706qp063oa

Braun.V. & Clark, V. (2013). Successful Qualitative Research: A Practical Guide for Beginners. https://www.researchgate.net/publication/256089360_Successful_Qualitative_Research_A_Practical_Guide_for_Beginners

Chouinard, J. A., & Cousins, J. B. (2009). A review and synthesis of current research on cross-cultural evaluation. *American Journal of Evaluation*, *30*(4), 457-494. https://doi.org/10.1177/1098214009349865

Cidlinska, K., Nyklova, B., Machovcova, K., Mudrak, J., & Zabroska, K. (2023). "Why don't I want to be an academic anymore?" When academic identity contributes to academic career attrition. *Higher Education, 85,* 141-156. https://doi.org/10.1007/s10734-022-00826-8

Claydon J, Farley-Barnes K, Baserga S. (2021). Building skill-sets, confidence, and interest for diverse scientific careers in the biological and biomedical sciences. *FASEB Bioadvances, 3*(12), 998-1010. doi:10.1096/fba.2021-00087. PMID: 34938961; PMCID: PMC8664047.

Cubarrubia, A.P. (2019, October 13). We all need to be data people. *The Chronicle of Higher Education*. https://www.chronicle.com/article/We-All-Need-to-Be-Data-People/247306

Denecke, D., Feaster, K., & Stone, K. (2017). Professional development: Shaping effective programs for STEM graduate students. Council of Graduate Schools, Washington, D.C.

Fass-Holmes, B. (2022). Survey Fatigue—What is its role in undergraduates' survey participation and response rates? *Journal of Interdisciplinary Studies in Education, 11*(1), 56-73. https://ojed.org/jise

Fuhrmann, C.N., Halme, D.G., O'Sullivan, P.S., & Lindstaedt, B. (2011). Improving graduate education to support a branching career pipeline: Recommendations based on a survey of doctoral students in the basic biomedical sciences. *CBE Life Sciences Education, 10,* 239-249. https://doi.org/10.1187/cbe.11-02-0013

Gammie, A., & Gibbs, K. (2017, October). Catalyzing the modernization of graduate biomedical training. Presentation at the Council of Graduate Schools.

Ganapati, S., Ritchie, T.S. (2021). Professional development and career-preparedness experiences of STEM Ph.D. students: Gaps and avenues for improvement. *PLoS ONE, 16*(12), e0260328. https://doi.org/10.1371/journal.pone.0260328

https://www.nap.edu/catalog/25038/graduate-stem-education-for-the-21st-century

Garrison, H. (2024). The changing employment distribution of life science doctorates. In: Markovac, J., Barrett, K.E., Garrison, H. (Eds.), Life Science Careers (Perspectives in Physiology). Springer, Cham. https://doi.org/10.1007/978-3-031-50694-9_1

Golde, C.M. & Dore, T.M. (2011). At cross purposes: what the experiences of doctoral students reveal about doctoral education. *The Pew Charitable Trusts; 2001*. www.phd-survey.org/report%20final.pdf

Gibbs, K.D. Jr, & Griffin, K.A. (2013). What do I want to be with my PhD? The roles of personal values and structural dynamics in shaping the career interests of recent biomedical science PhD graduates. *Cell Biology Education— Life Sciences Education*. *12*(4), 711-723.

Henning, G., & Lundquist, A. (2022). Using assessment to advance equity. *New Directions Student Services, 185-194*. DOI: 10.1002/ss.20439.

Hohensinn, C., & Kubinger, K.D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*(4), 732-746. DOI: 10.1177/0013164410390032.

Jonson, J.L., Guetterman, T., & Thompson, R.J. (2014). An integrated model of influence: Use of assessment data in higher education. *Research & Practice in Assessment*. 18-30.

Larson R.C., Ghaffarzadegan N., Xue Y. Too many PhD graduates or too few academic job openings: the basic reproductive number R_0 in academia. *Systems Research and Behavioral Science*. 31(6), 745-750.

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

National Academies of Sciences, Engineering, and Medicine. 2018. Graduate STEM Education for the 21st Century. *Washington, DC: The National Academies Press*. https://doi.org/10.17226/25038.

National Science Board. 2018. Science and Engineering Indicators, NSB-2018-1. Alexandria, VA: National Science Board. https://www.nsf.gov/statistics/2018/nsb20181/assets/901/tables/tt03-16.xlsx

Oliveri, M. E., Lawless, R., & Mislevy, R. J. (2019). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. *International Journal of Testing*, *19*(3), 270-300.

O'Meara, K., Jaeger, A., Eliason, J., Grantham, A., Cowdery, K., Mitchall, A., & Zhang, K. (2014). By design: How departments influence graduate student agency in career advancement., *International Journal of Doctoral Studies*, *9*, 155-179. http://ijds.org/Volume9/IJDSv9p155-179OMeara0518.pdf

Posselt, J., Baxter, K., & Tang, W. (2021). *Assessing the landscape for diversity, equity and inclusion efforts in U.S. STEM graduate education: A systematic literature review*. A report commissioned by the Alfred P. Sloan Foundation. https://pullias.usc.edu/wp-content/uploads/2023/05/USC-Rossier-DEI-literature-review.pdf

QSR International. NVivo Qualitative Data Analysis Software; 1999. Available from https://qsrinternational.com/nvivo/nvivo-products/

Reeves, P., Claydon, J., & Davenport, G. (2022). Program evaluation practices and the training of PhD students in STEM. *Studies in Graduate and Postdoctoral Education, 13*(2),109-131.

Roach, M., & Sauermann, H. (2017). The declining interest in an academic career. *PLoS ONE, 12*(9), e0184130. https://doi.org/10.1371/journal.pone.0184130

Sinche, M. (2018). *Next gen phd: a guide to career paths in science*. Harvard University Press. National Academy of Sciences (2011).

Sinche, M., Layton, R.L., Brandt, P.D., O'Connell, A.B., Hall, J.D., Freeman, A.M., Harrell, J.R., Cook, J.G. and Brennwald, P.J. (2017). "An evidence-based evaluation of transferrable skills and job satisfaction for science PhDs. PloS One, *12*(9), p. e0185023, doi: 10.1371/journal.pone.0185023.

Subramanian, S., Hutchins, J.A., & Lundsteen, N. (2022). Bridging the gap: increasing collaboration between research mentors and career development educators for PhD and postdoctoral training success. *Molecular Biology of the Cell*, 33, 1-4.

Walker, G.E., Golde, C.M., Jones, L., Bueschel, A.C., & Hutchings, P. (2008). *The formation of scholars: Rethinking doctoral education for the twenty-first century*. San Fransisco, CA: Jossey-Bass.

Wilson, S. (2008). *Research is Ceremony: Indigenous research methods*. Fernwood Publishing, Halifax and Winnipeg.

Xue, Y., and R. Larson. 2015. STEM crisis or stem surplus? Yes and yes. *Monthly Labor Review*. https://www.bls.gov/opub/mlr/2015/article/stem-crisis-or-stem-surplus-yes-and-yes.htm.