JAHE | The Journal of Assessment in Higher Education

Volume 1 | Number 1 April 2020

https://journals.flvc.org/assessment/ ISSN: 2642-0694

Linking Program Assessment to Institutional Goals

Liz Grauerholz*, Patrice Lancey, Kristen Schellhase & Cory Watkins

University of Central Florida

Abstract

Integration of institutional research-based planning and evaluation processes is a mechanism to improve institutional quality and effectiveness by focusing all university constituents on implementing and evaluating strategic initiatives. While educational program assessment to foster evidence-based improvements is strongly infused in the culture of many universities, drawing intentional connections between program assessment, which primarily focuses on student learning outcomes, and institutional strategic planning, can be challenging for faculty. This paper highlights the assessment work of three diverse disciplines in a large public research institution that have articulated connections between their program's student learning and outcomes to elements of the university strategic plan and other organizational requirements. This paper explores the benefits and challenges of explicitly linking outcomes or measures in program assessment to university planning.

Keywords: Program assessment, strategic plan, university goals, student learning outcomes

* Corresponding author: Liz Grauerholz, Department of Sociology, University of Central Florida, 400 Central Florida Blvd., Orlando, FL, 32816. Elizabeth.grauerholz@ucf.edu

Recommended citation*:* Grauerholz, L., Lancey, P., Schellhase, K., & Watkins, C. (2020). Linking Program Assessment to Institutional Goals. *Journal of Assessment in Higher Education, 1*(1), 1-29.

doi:10.32473/jahe.v1vi1i.117164

©Journal of Assessment in Higher Education. Published under Creative Commons License CC BY- 4.0 International.

Introduction

Virtually all colleges and universities engage in strategic planning and institutional assessment; indeed, program assessment is a requirement of regional accreditation. Similarly, strategic planning is critical to helping organizations lay out a path of growth and improvement, which often fuels fund-raising and helps satisfy the demands of stakeholders and oversight committees and boards. Because strategic planning occurs at the institutional level—envisioning strategies for strengthening and transforming the institution's prominence, efficiency, and culture to meet present and future challenges—very often the strategic plan feels disconnected from the concerns and challenges facing specific academic departments and programs, which typically focus on improving student learning outcomes. This perceived disconnect between the broader institutional goals and the specific educational programs that collectively carry out the mission of the university or college (student learning, research innovation) can generate a sense of isolation, a kind of "us-against-them" mentality, and disinvestment in the institutional assessment process that may be seen as out of touch with program-level concerns and challenges.

In this paper, we argue that intentionally connecting program assessment that primarily focuses on student learning outcomes, and institutional goals, is mutually beneficial. Programs are better able to articulate their mission and role within the larger structure, and the institution is enhanced by having multiple academic units working in conjunction with these broad goals. Ultimately, students are the primary beneficiaries, as programs and organizations work collaboratively to serve their needs.

Strategic Planning and Assessment in Higher Education

Both strategic planning and assessment of student learning and programs in higher education gained traction in the United States during the 1980s as demands for greater accountability from state and federal governments and accrediting commissions intensified (Hinton 2012). While some argue that the two (learning outcomes and institutional planning processes) should and must be closely connected (Hinton 2012; Serbin 2004), in reality, assessment and strategic planning serve different purposes and are frequently carried out by different personnel, and thus the overlap may be slight.

Strategic plans typically outline the organization's mission, values and goals, and strategies for achieving those goals, typically for the next 5-10 years. Calls for more "strategic" planning in higher education emerged as many colleges and universities were facing crises due to enrollment declines and shrinking financial support from governmental and business sources during the 1970s and early 1980s (Keller 1983). Keller (1983) proposed as a solution that strategies commonly employed in corporate and commercial settings, namely "strategic marketing planning," be adopted by institutions of higher education as a way for colleges and universities to plan for and survive the shifting landscape. As Kotler and Murphy (1981, p. 488) argued, "The future that appears to hold many threats for most colleges and universities should become less imposing with the judicious use of strategic planning." Strategic planning was a way to prioritize increasingly limited resources and promote greater focus within the institution (Hinton 2012).

Assessment's roots can be traced to the First National Conference on Assessment in

Higher Education held in 1985 following two publications—*Involvement in Learning* (National Institute of Education, 1984) and *Integrity in the College Curriculum* (Association of American Colleges, 1985). Assessment practitioners and policy shapers gathered to discuss the report recommendations which argued, based on scholarly research, that several conditions were needed to promote student achievement, including setting high expectations, involvement in active learning, and providing prompt and useful feedback. But the report also emphasized that higher education institutions could benefit from feedback about their own performance. This final recommendation was consistent with voices within higher education that were focused on curriculum and pedagogical improvement to create a cohesive experience guided by ongoing scholarly measurement of student learning.

Though a handful of colleges and universities initiated attempts to measure student competencies, it was the publication of a report from the National Governors Association that fostered early response from state governing boards (National Governors Association, 1986). States mandated the use of standardized tests to compare across institutions (e.g., Texas) or required higher education institutions to establish their own approach to articulate, measure and gather evidence about student learning (e.g., Colorado and Virginia). By the end of the decade, about two thirds of states had installed requirements that institutions assess student learning (Banta, 1993; Ewell, 2002).

Reauthorization of the Higher Education Act of 1988 along with tight state budgets spurred transfer of the public accountability agenda from state authorities to regional accreditors. Armed with new language in the act, many regional accreditors took up the charge to require by the early 1990s that all colleges and universities participate in assessment of student learning. Furthermore, colleges and universities were required to provide strategic plans. As Hinton (2012, p. 7) notes, “institutions began to find themselves under serious scrutiny during their reaccreditation processes if they did not have a working strategic plan and some form of assessment plan in place.”

Although strategic planning remains integral to institutions of higher education, Hinton (2012) notes that by the late 20th century, even those educational institutions that had forged successful plans and fruitful processes began to dismantle planning offices and focus instead on assessment initiatives. However, this shift from strategic planning to assessment came with its own challenges. Historically, assessment of student learning has been mired by a dual purpose—calls for accountability and calls for authentic study of teaching and learning to improve outcomes. Because the motivation to conduct assessment resulted from administrators’ efforts to meet compliance standards, assessment of student learning was considered by many at colleges and universities as an add-on activity rather than an integral part of the teaching and learning process. Faculty with this perspective propagated the use of summative standardized tests and surveys of students.

Perhaps even more distressing, one result of the compliance agenda was the erroneous faculty beliefs that assessment was divorced from their academic mission and scholarship and was merely something they did periodically to satisfy administrative requirements (Ewell, 2008; Astin & Antonio 2012). Unfortunately, this focus on compliance hindered the pursuit of *authentic assessment* which is systematic, ongoing, and formative as well as summative in nature, aimed at gaining understanding of and improving explicitly stated student learning outcomes about what students should know, be able to do and value. In this approach, multiple measures are embedded by faculty into student assignments in a curriculum of study so that student work, scored with a rubric or other scoring protocol, is the evidence of focus to improve curriculum design, pedagogy and learning over time.

More recently, a survey of provosts or chief academic officers in 2009 revealed the most common use of assessment data remained regional or discipline accreditation (Kuh & Ewell, 2010). However, perhaps this study captured evidence of a turning point for assessment in higher education. Provost responses also showed a commitment to use assessment data to improve learning through revising learning goals, informing strategic planning, modifying general education curriculum, and improving instructional performance (Kuh & Ewell, 2010).

Making this paradigm shift in purpose can be challenging for institutional leaders. Executive leadership must carefully weigh the benefits and costs of investing in building and sustaining an ongoing, systematic and effective assessment culture and planning process focused on evidence-based improvement. Additionally, integrating and making connections between institutional planning processes that matter to members of the community can be daunting: educational program assessment is usually focused on using results gleaned from annual assessments analyzed over time to improve student learning outcomes rather than broader institutional goals prevalent in strategic planning. Serban (2004) notes that comprehensive models that “coherently integrates all levels, from courses and programs to the overall institution” (p. 26) are lacking. In short, institutional student learning outcomes (ISLOs) are often not emphasized in strategic planning, especially at larger universities whose mission includes and typically prioritizes research funding, posing challenges to those charged with program assessment to link their efforts to institutional goals.

Structuring program assessment so that faculty make intentional connections between the outcomes or measures in their educational program assessment plans to elements of the strategic plan can help to bridge two planning levels by linking program assessment to broader institutional and organizational goals. Benefits include harnessing the energies and expertise of all constituents in the institution to achieve broad planning goals through ongoing practice of evidence-based decision making that promotes improvement in mission-driven institutional priorities. Here we describe the institutional effectiveness assessment model practiced at University of Central Florida (UCF), a large public research institution that aims to foster links between program assessment and strategic planning. We also highlight the assessment work of three diverse disciplines within the institution where faculty have articulated connections between their program’s student learning and program outcomes and elements of the university strategic plan or their professional requirements.

The Institutional Effectiveness Assessment Model at the University of Central Florida

The decision to deepen investment in institutional planning processes as a mechanism to foster quality, innovation and improvement was made by leaders at UCF in 2000. Consistent with its core mission and strategic plan, administrators at UCF implemented its own institutional effectiveness (IE) assessment policies and procedures. UCF faculty and staff members defined expected outcomes, assessed the extent to which these outcomes were achieved, and have modified and improved their academic programs and administrative units based on assessment results since 1994. By 1996, the faculty of each academic program and administrative unit had developed an assessment plan (mission, objectives, outcomes, and measures) and completed one cycle of reporting results and use of results. A three-year review cycle was instituted initially, followed by an annual review in 2000. This change was prompted by a memorandum by the president that restated the importance of assessment and established a new office, Operational Excellence and Assessment Support, to support assessment activities.

The UCF Institutional Effectiveness Assessment process is directly tended by Divisional Review Committees (DRCs) that are aligned to colleges and divisions. The UCF IE assessment model consists of two broad categories, academic programs and administrative units, and is used to guide assessment in both areas. Academic programs include undergraduate and graduate educational programs (with selected tracks) and certificates.

Assessment coordinators (faculty members) for each program work with program faculty to:

- develop a plan with student learning outcomes consistent with the mission using SMART[1] guidelines;
- select and implement measures using MATURE[2] guidelines; and
- analyze results and plan for improvements based on the results that are then assessed in the subsequent plan (that is, closing the loop). The results and plan for improvement are documented in an assessment report.

The components of the assessment report that is submitted annually are described as follows:

1. Results of the previous year's assessment plan (data and analysis).
2. A reflective statement about the results describing the implications of the findings and how the evidence can be used to make improvements. Reflections are based on a trend analysis of results for outcomes gleaned from annual assessment over time.
3. Implemented and planned strategies to bring about improvements to curriculum, pedagogy and academic processes based on these results.
4. An assessment plan for the current year, which includes measurement of the effect of improvements made. The plan consists of a mission statement, assessment process description, a description of how the program assessment outcomes or measures link to the university’s strategic plan, outcomes—at least six outcomes (for undergraduate programs) or at least three outcomes (for graduate programs and administrative units) that are central to their mission—and at least two

[1] SMART outcome guidelines include specific, measurable, aggressive and attainable, result-oriented and time bound.

[2] MATURE measure guidelines include matches the outcome, uses appropriate methods, sets performance targets, is useful to improve, is reliable and is effective and efficient.

measures (one of which is a direct measure) per outcome with performance criteria or targets that provide evidence about how well the outcomes are being achieved. Methodologically sound practices are employed by faculty to measure student learning and operational outcomes.

5. Results and plans are submitted to DRCs for reviews designed to promote excellence in assessment and improvement based on the results. A web application report and review system houses common structured templates for assessment coordinators, DRC chairs, and DRC members. Using the UCF IE Assessment Rubrics, DRC members provide feedback to the coordinators about the assessment results and plans.

Each DRC is charged with working collaboratively with its programs or units to mentor the members in their assessment team and to provide a review of the quality of the assessment reports based on established criteria. These criteria are defined in the UCF IE Assessment Rubrics, designed in 2009 and revised in 2013 by the University Assessment Committee as a tool for providing specific feedback on plans and results. Each program or unit is reviewed by multiple members of the Divisional Review Committee (DRC)—often one member and the chair. Assessment coordinators then address the feedback and resubmit the results and plans back to the DRC. The results and plans go through several review iterations prior to final approval by the DRC Chair.

Broad-based participation is the foundation of the UCF assessment model and is characterized by active involvement and contributions of faculty, staff, and administrators who are organized into DRCs that are aligned to the colleges and divisions. Each Division Review Committee has a chair who sits on the University Assessment Committee. The University Assessment Committee (UAC) was established by the UCF President to support a process of continual self-evaluation and improvement. The primary purpose of the UAC is to oversee and assist academic and administrative units in conducting ongoing assessment to improve student-learning and operations. The UAC ensures the quality of the reviews conducted by the DRCs through its oversight of the review process. The chairs of each of the 21 DRCs comprise the university-level committee. Annually, each member of the UAC presents a DRC report about the quality of the results and plans. It contains examples of how the programs or units used assessment results to make improvements.

The expectation that program assessment coordinators make intentional connections to strategic planning was introduced to the university community by the UAC beginning with the drafting of 2009-10 IE Assessment Plans. The strategic planning alignment criteria was included in the 2009 IE Assessment Plan Rubric. However, after several years of applying the 2009 IE Assessment Rubrics to academic program plans, DRC members observed that a more specific rubric criteria was needed to help faculty structure intentional connections between program student learning outcomes assessment and strategic planning. The IE Assessment Plan Rubric criteria related to strategic planning was revised in 2013 to foster deeper alignment between these institutional planning processes. The 2009 rubric criteria asked faculty members to “describe the relationship between the IE plan and the University’s Strategic Plan.” By contrast, in the revised 2013 IE Assessment Plan Rubric, the strategic planning criteria was redesigned to increase specificity by stating that, “the plan explicitly links one or more outcomes or measures to strategic planning.” An accompanying IE Assessment Plan Rubric narrative was also developed to provide additional rubric criteria guidance. Further, a dedicated area was created in the IE

Assessment Plan template for faculty to detail the strategic planning links. Finally, the IE Assessment Plan Rubric levels were adjusted in 2013 to increase rigor. The strategic planning criteria was one of the two criteria that could be satisfied to earn an IE Assessment Plan Rubric rating of “Accomplished” on a five-point scale where 1 is “Beginning,” 2 is “Emerging,” 3 is “Maturing,” 4 is “Accomplished,” and 5 is “Exemplary.” With the implementation of the most recent UCF strategic plan in 2017, the accompanying IE Assessment Plan Rubric narrative was revised to further clarify to align to the “promises” or “metrics” in the current strategic plan.

UCF’s Strategic Plan

Planning for the UCF Strategic Plan, or Collective Impact Statement, began in fall 2015 and was implemented in summer 2017. Like most university strategic plans, many of the objectives are designed to enhance reputation, prestige and funding of the institution (e.g., attract $100 million in new funding) and are seemingly separate from the student-learning mission. Of the five broad, overarching “promises” outlined in the plan, one does address students and faculty: “Attract and cultivate exceptional and diverse faculty, students, and staff whose collective contributions strengthen us” and one goal in the strategic plan relates directly to student success—“increasing student access, success, and prominence” (University of Central Florida, 2017). Yet, because the strategic plan is aimed at these higher-level goals, it follows that the metrics and strategies associated with these institutional-level goals are also broad (e.g., “enroll a student population whose family incomes reflect the distribution of the region”). As a result, there is a perceived disconnect between the day-to-day workings of individual faculty or departments/programs and the university’s goals and strategies.

Thus, an assessment challenge confronting faculty and program directors is linking improvements at the departmental or program level to institutional goals. These challenges may be confounded even further when programs are accredited by a professional governing board (e.g., Accreditation Committee for Education in Nursing or Association for Behavior Analysis International). In addition to university goals, these programs must demonstrate that they have satisfied criteria outlined by the accrediting body. Despite these challenges, some faculty and programs have attempted to intentionally link their program assessment to these larger institutional goals. Here we review three such examples: The first (Criminal Justice) highlights efforts to directly link assessment of student learning to the strategic plan’s goal of increasing students’ access and success; the second (Social Sciences) describes efforts to link program assessment (apart from student learning) to the strategic plan and its goals of increasing student success, diversity and inclusion; and, the third (Athletic Training Program) illustrates how one program linked assessment of student learning to both the university strategic plan’s call for greater student access, success and prominence, and the professional requirements dictated by discipline accreditation.

Illustrative Examples: Three Programs

Linking Assessment of Student Learning to the Strategic Plan: The Criminal Justice Bachelor of Arts and Bachelor of Science Program

Part of the mission of the Department of Criminal Justice at UCF is to serve the university's strategic goal of providing the best undergraduate criminal justice education to students coming from diverse backgrounds. The makeup of the more than 1400 students majoring in Criminal Justice (CJ) consist of a blend of first time in college (FTIC, 42%) and transfer students (58%) from local area state colleges who are admitted with Associates of Arts or articulated Associates of Science degrees through the DirectConnect to UCF program (DirectConnect to UCF guarantees admission to the population of transfer students from our partner colleges). The CJ majors are required to complete core courses in the areas of policing, courts, corrections, research methods, statistics, and a Capstone Experience.

Criminal Justice was one of three pilot programs selected in which faculty implemented the Student Success Collaborative (SSC), established in 2015-16, which is designed to enhance student success, retention, and timely graduation. This program stems from an emphasis at the state level and subsequently by top UCF administrators to improve student success in these areas, and its focus aligns directly with the university's strategic plan regarding student success. SSC uses a predictive analytics platform to aid program directors, coordinators, faculty and advisors in more effectively monitoring student success. SSC is used to pinpoint success or failure markers for struggling students in a timely manner to reduce or prevent course repeat, failure and negative trajectories. Reports are generated to inform program personnel about students who may be in jeopardy of missing a high-enough grade in an important success marker class or whose behaviors may show a pattern across multiple courses that could indicate a more serious problem.

Contextual knowledge combined with SSC pilot program analytics confirmed that two required courses—Research Methods and Data Analysis—historically inhibited student success, retention and timely graduation. Thus, in 2017, based on this contextual knowledge, combined with SSC pilot program analytics, annual assessment results dating back to 2013-14, and recommendations from the program's review in 2013, the faculty decided to restructure the curriculum with a close eye on these two courses to ensure improved performance on program outcomes aimed at meeting our learning compacts. The faculty established Research Methods and Data Analysis committees to review course structuring and curriculum, examined sibling programs at UCF (i.e., Psychology, Sociology, Political Science, Public Administration), and conducted a statewide analysis of institutions with CJ programs offering similar courses. The result was the addition of a one-hour weekly lab in both courses, making a "hands on experience" a vital component of the courses, and development of a fixed curriculum and datasets to be used by all students in these courses.

Two direct outcome measures in the program's assessment focus on learning in these two courses. As seen in Table 1, students were assessed on their ability to design a research project and consume CJ research, and their ability to understand national crime databases

(UCR, NCVS), and how crime data are collected and presented to the public. The direct measures supported improved student performance in these areas tied to programmatic changes. Additionally, indirect measures of both outcomes were used to gauge student perception of success using the Graduating Senior Survey conducted annually.

Table 1

University Goals and Corresponding Program Outcomes and Measures for Criminal Justice Program

Strategic Planning Goal: Increase student access and success (six-year graduation rate of 75% and transfer student graduation rate of 75%)

Program Assessment Outcome 1 (Research Methods): Criminal Justice students will demonstrate an ability to design a research project and intelligently consume the results of criminal justice research conducted and presented by others.

Measures:

1. Annually, panel of CJ faculty will evaluate research projects to determine both the students` ability to design a research project and the students` ability to intelligently consume the results of criminal justice research. All research projects from all sections of CCJ4701 are reviewed. At least 75% of students will score 75% or higher on their research project evaluation. (**Direct)**
2. Annually, all graduating majors are asked to rate their level of agreement with statement: "As a result of my Criminal Justice education at UCF, I am able to design a research project." Respondents will be able to respond with strongly agree, agree, neutral, disagree, strongly disagree. At least 80% of students will respond that they strongly agree or agree with the statement. **(Indirect)**

Program Assessment Outcome 2 (Data Analysis): Criminal Justice students will demonstrate knowledge of national crime and victimization databases and how crime data are collected and presented to the public.

Measures:

1. Annually papers and/or projects from all students enrolled in all sections of CCJ 4746 will be evaluated to determine if students demonstrate a knowledge of national crime and victimization databases/how crime data are collected and presented to the public. At least 75% of students sampled will score 75% or higher on the evaluation. **(Direct)**
2. Annually, all graduating majors are asked to rate their level of agreement with the statement: "The criminal justice program at UCF has provided me with the knowledge of national crime and victimization databases and how crime data are collected and presented to the public." Respondents will be able to respond with strongly agree, agree, neutral, disagree, strongly disagree. At least 80% of students will respond that they strongly agree or agree with the statement. **(Indirect)**

Note. Information about this program may be found at https://www.ucf.edu/degree/criminal-justice-bs/

These efforts to restructure the curriculum and assessment to align more clearly with university goals surrounding student success, retention and timely graduation, has been beneficial to students. Although it is too early to provide sound long-term data, some preliminary highlights and anecdotal data lend encouragement to the department's efforts. That is, program assessment results for 2018-19 revised curriculum, compared to 2016-17 results, before the program changes occurred, show signs of student learning gains in ability to design a research project and intelligently consume the results of criminal justice research conducted and presented by others. In 2018-19, 84.4% (n=260/308) of students in the population scored 75% or higher on the research project using a rubric, compared to 79.2% (n=415/524) of students in 2016-17 who scored similarly on this measure. In support of the direct measures, the results from the indirect measures indicate that 83% (n=329/395) of all graduating CJ majors in 2018-2019 agreed or strongly agreed that as a result of their Criminal Justice education at UCF, they are able to effectively design a research project, compared to 78% (289/370) in 2016-2017. Further, 95% (375/395) of all graduating CJ majors in 2018-2019 compared to 91.9% (350/381) in 2016-2017 agreed or strongly agreed that the criminal justice program at UCF has provided them with the knowledge of national crime and victimization databases and how crime data are collected and presented to the public.

Anecdotally, conversations with core faculty and instructors teaching Research Methods and Data Analysis indicate that students appear to be more effectively grasping methodological concepts and data analytic skills because of incorporating lab-based environments. Informal advising sessions and discussions with students also seem to indicate a more positive attitude toward these courses along with more positive communication streams among students, leading to higher retention. Early data also suggest students are less likely to repeat these courses, and thus are more likely to graduate on time.

Admittedly, it is early in the process as the Criminal Justice program has had one year of comparative data since restructuring the program with a pointed eye on research methods and data analysis courses. Nonetheless, the program and curriculum changes noted above, stemming from a triangulated approach (program analytics contextual knowledge, review of trends in previous annual assessment results, and recommendations from the previous program reviews), assisted with meeting departmental goals and aligned the program with the larger institutional mission. Other programs may explore the use of data analytics linked with other assessment and implementation approaches to accomplish departmental outcomes and link to larger institutional goals.

Linking Program Assessment to the Strategic Plan: The Social Sciences Bachelor of Science Program

The Social Sciences program at UCF is an interdisciplinary program composed of programs within the social sciences (sociology, psychology, political science, communication, anthropology, and women's and gender studies). The major is structured in such a way that students complete the requirements for the minor in three of the six programs, in addition to completing a course in basic statistics and a methodology course related to one of the student's area of concentration. Currently, there are approximately 120 majors in the program.

Until 2017, the sole focus of the Social Sciences program assessment was to assess cognitive learning outcomes. During the semester of graduation, students were required to take an “exit exam” in which they were quizzed on their statistical literacy, methodological knowledge (especially concerning ethics), and knowledge of basic principles and concepts related to their three areas of concentration. This exam assessed outcomes corresponding to basic knowledge, comprehension, and to a lesser extent, application of concepts, according to Bloom’s (1956) taxonomy. The exit exam was intended only for assessment purposes. That is, students were not required to pass the exam with any specific level of competency; they were simply required to complete the exam.

With the emergence of UCF’s new strategic planning goals, the faculty began exploring ways to further align the program (and its assessment) to the larger institutional mission. Because the UCF strategic plan, as discussed earlier, is geared toward institutional outcomes and not student learning per se, bridging the gap between institutional goals and this relatively small academic program was challenging. Faculty decided to focus on two strategic planning goals that pertained to undergraduate students: “student success” and “student diversity and inclusiveness.” It should be noted that the program continues to focus primarily on student learning outcomes, and that the new outcomes based on the strategic plan are in addition to the student learning outcomes that were previously created and are still in use.

As seen in Table 2, one UCF strategic planning goal concerning student success states that all students will participate in a positive, high impact student experience. Here, “positive, high impact student experience” includes research, internship, service-learning, or study abroad experiences. Thus, the program created a new assessment outcome and measures for the social sciences program. The desire was to align with the university goal and have Social Sciences majors participate in these high-impact experiences including research, internships, and study abroad. It should be noted that given the interdisciplinary nature of the Social Sciences program, the program director has no control over the departments or programs that students are minoring in. For instance, some departments have extensive internship or research opportunities available to students, others don’t. Given these challenges, the program created modest measures—10 percent increases in each category: research experiences, internships and study abroad.

The second approach to linking program goals to the broader university goals was through student diversity and inclusiveness (see Table 2). The university goal was to increase degree attainment of specific diverse student cohorts across all academic disciplines by 10%, and the program was also dedicated to attracting these students to the program and serving diverse students within the major. Thus, the measures developed to assess progress in this area were twofold: to increase by 10% the Social Sciences majors representing diverse and underrepresented groups, specifically students of color and transfer students; and to increase by 10% the percentage of majors who represent diverse groups. The data used to measure diversity and high-impact experiences are provided by the university.

Table 2

University Goals and Corresponding Program Outcomes and Measures for Social Sciences Program

Strategic Planning Goal: Student Success (100% of undergraduates participate in a positive, high impact student experience either on or off campus).

Program Assessment Outcome 1: Social Sciences majors will participate in positive, high-impact experiences including research, internships and study abroad.

Measures:

1. There will be a 10% increase in number of students who participate in research (Honors in the Major, Independent Directed Studies, Student Undergraduate Research Experience [SURE])
2. There will be a 10% increase in number of students who participate in internships (experiential learning).
3. There will be a 10% increase in number of students who participate in study abroad.

Strategic Planning Goal: Student Diversity and Inclusiveness (specifically, increase by 10%, degree attainment of specific diverse student cohorts across all academic disciplines)

Program Assessment Outcome 2: Social Sciences will attract and serve diverse students to/within the major

Measures:

1. The percentage of students representing diverse and underrepresented groups (students of color and transfer students) who major in Social Sciences will increase by 10%.
2. The percentage of graduates representing diverse groups will increase by 10%.

Note: Information about this program may be found at https://www.ucf.edu/degree/social-sciences-bs/

Although it is too early to document the Social Sciences' program success, by meaningfully and intentionally working not only to meet desired student learning outcomes but also working in synergy with the university to achieve its promise, the program is better positioned to contribute to the larger institutional mission. By intentionally linking the program outcomes to the broader university goals such as enrollment patterns and retention data by demographics, any program can evaluate how well it meshes with the path that the larger institution has laid out, and whether it is doing enough to facilitate student success and attract diverse students.

Linking Assessment of Student Learning to the Strategic Plan and Accreditation: The Athletic Training Program

The Athletic Training (AT) Program at UCF illustrates the linking of professional standards to strategic planning and program-specific student learning outcomes. Athletic trainers are healthcare providers who serve in a primary care role within secondary schools, colleges and universities, outpatient rehabilitation facilities, industry and military, as well as any other location where physically active people sustain injuries and illnesses. Approximately 56 students are enrolled in the program.

Unlike academic programs highlighted above, professional clinical programs such as the AT Program that allow graduates to sit for the certification examination are guided by a list of competencies. There are eight Professional Knowledge content areas found in the 5th Edition of the Educational Competencies (National Athletic Trainers' Association, 2011). The accreditor for athletic training programs requires that "There must be a comprehensive assessment plan to evaluate all aspects of the educational program." The AT Program interprets "comprehensive" to mean that the program needs to: 1) assess faculty and preceptors as teachers/mentors of the students; 2) assess clinical sites for their ability to provide appropriate experiences; 3) assess the curriculum to determine if the program is preparing students in all aspects of practice; and 4) assess some of the "soft skills" that graduates need to be good athletic trainers. The first-time pass rate on the Board of Certification (BOC) examination provides important evidence of student success.

Table 3 reveals the direct link between the professional standards—the 8 competency content areas—and the strategic plan for the university. All outcomes are student learning outcomes and include all 8 areas. Direct measures are practical skills, exams, essays, and projects. Indirect measures assess the graduates' perceived confidence in their abilities in the 8 areas.

These professional requirements and goals align with one of UCF's Strategic Plan Priority Metrics—increasing student access, success, and prominence. The way the AT Program increases success is by ensuring that all graduates are well qualified to become entry-level practitioners. The AT Program increases prominence by ensuring that graduates are well prepared to pass the BOC examination at a rate that establishes UCF as a leader across the country.

Using this same approach, any academic program can review the professional standards documents from their national professional association (e.g., the American Speech and Hearing Association's "Big 9" areas of practice and the Council on Social Work Education's "Nine Core Competencies and Behaviors") and use them to assess the curriculum in their program. Doing so creates an additional linkage between the program's student learning outcomes and institutional goals surrounding prominence and student success.

Table 3

Professional Goals and Corresponding Program Outcomes and Measures for the Athletic Training Program

Strategic Planning Goal: Increase student access, success and prominence

Professional Association Goal: Satisfy professional knowledge competencies

Program Assessment Outcomes 1-7: AT Program students will be competent with the knowledge, skills and abilities in seven Professional Knowledge content areas of prevention & health promotion, clinical examination & diagnosis, acute care of injury & illness, and therapeutic interventions, psychosocial strategies & referral, healthcare administration, and professional development & responsibility found in the Professional Education Council's 5th Edition of the Athletic Training Education Competencies.

Measures:

1. 90% of students will earn a grade of "B-" (80%) or better on the cumulative final competency examinations for each practicum course (ATR 3812L, 3822L, 4832L, 4842L). The first-time pass rate will meet or exceed the first-time pass rate for the prior year. Commonly missed questions will be identified and categorized so that an action plan to improve can be implemented during the subsequent cycle. (**Direct)**
2. 90% of all students in the AT Program will earn a "B-" (80%) or better on the Psychosocial Intervention essay in the Case Studies in Sports Medicine (ATR 4103 course). Scores will be adjusted for formatting errors (the rubric has 72 points related to content and 28 points related to structure and format – students can also lose 25% for a late grade). This measure assesses the psychosocial strategies & referral content area. **(Direct)**
3. 90% of students will earn a grade of "B-" (80%) or better on the cumulative final examination for the Organization & Administration in Athletic Training course (ATR 4512C). This measure assesses the healthcare administration (HA) and professional development & responsibility (PD) content areas. **(Direct)**
4. 90% of graduating seniors will report on the AT Program Exit Survey (prior to graduation), that they "agree" or "strongly agree" that they are confident regarding their knowledge and ability to perform in the seven Professional Knowledge content areas measured in this outcome. Each mean score will meet (within 1 standard deviation) or exceed the mean score from the prior year. **(Indirect)**

Program Assessment Outcome: AT Program students will demonstrate information fluency and critical thinking through proficiency with the 5 steps of evidence-based medicine (EBM - defining a clinically relevant question, searching for best evidence, appraising evidence quality, applying evidence to practice, and evaluating the process).

Table 3 (continued).

Professional Goals and Corresponding Program Outcomes and Measures for the Athletic Training Program

Measures:

1. 90% of students will earn a grade of "B-" (80%) or better on the Therapeutic Modalities in Athletic Training (ATR 4302C) EBM Project. (Direct)
2. 90% of students will earn a grade of “B- “(80%) or better on EBP examination questions given on the Advanced Rehabilitation in Athletic Training (ATR 4315C) final examination. (Direct)
3. 90% of graduating students will "strongly agree" or "agree" that the AT Program fostered critical thinking skills and that they are able to provide care that is evidence-based. The mean scores will meet (within 1 standard deviation) or exceed the scores from the prior year. (Indirect)
4. 90% of graduating seniors will report on the AT Program Exit Survey (prior to graduation), that they “agree” or “strongly agree” that they are confident regarding their knowledge and ability to perform in the Professional Knowledge content area of Evidence-Based Medicine (EBM). The mean score will meet (within 1 standard deviation) or exceed the mean score from the prior year. (Indirect)

Note: Information about this program may be found at https://healthprofessions.ucf.edu/kpt/athletictraining/

Conclusion

In this paper, we present three examples of programs from diverse disciplines that have confronted the challenge of linking program assessment to larger institutional goals, including those associated with professional accreditation. Admittedly, the linkages feel, at times, elusive, as the difficulties of bringing university and organizational level goals and metrics down to the department or program level persist. Such linkages would be more straightforward if institutions prioritized institutional student learning outcomes (ISLO) (Serban 2004) so that each program could link to some or all of these outcomes. Strategic plans outside teaching-oriented institutions often fail to include ISLOs or ones applicable to various programs. These disconnects not only reveal the different purposes of strategic planning and assessment, but also the tension between accountability and authentic assessment that have historically played out within the larger arena of assessment in higher education. Yet, as seen here, as faculty become more intentional in the way they develop assessments to align with larger institutional goals and strategies, the tension between these two approaches can be resolved. Programs can play a more direct role in influencing institutional goals and metrics related to student retention, diversity/inclusion or similar outcomes. By working together in this manner, students, faculty and administrators all stand to benefit from institutional assessment in higher education.

References

Association of American Colleges (1985). *Integrity in the college curriculum: A report to the academic community.* Washington, D.C.: Association of American Colleges.

Astin, A. W., & Antonio, A. L. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education* (2nd ed.). Lanham, MD: Rowman & Littlefield.

Banta, T. W. and Associates (1993). *Making a difference: Outcomes of a decade of assessment in higher education.* San Francisco, CA: Jossey-Bass.

Bloom, B. S. (1956). *Taxonomy of educational objectives. Handbook 1: Cognitive domain.* New York: McKay.

Ewell, P. (2002). An emerging scholarship: A Brief history of assessment. In T. W. Banta and Associates (Eds.), *Building a scholarship of assessment.* San Francisco: Jossey-Bass.

Ewell, P. T. (2008). Assessment and accountability in America today: Background and context. *New Directions for Institutional Research*, *S1*, 7-17. https://doi.org/10.1002/ir.258

Good, J. M., & Kochan, F. (2008, February). *Creating a quality program by linking strategic planning and assessment through collaboration.* Paper presented at the Annual Meeting of the American Association of Colleges of Teacher Education, New Orleans, LA.

Hinton, K. E. (2012). *A practical guide to strategic planning in higher education*. Society for College and University Planning.

Keller, G. (1983). *Academic strategy: The management revolution in American higher education*. Baltimore, Maryland: The Johns Hopkins University Press.

Kotler, P., & Murphy, P.E. Murphy. (1981). Strategic planning for higher education. *The Journal of Higher Education, 52*(5), 470-489.

Kuh, G. D., & Ewell, P. T. (2010). The state of learning outcomes assessment in the United States. *Higher Education Management and Policy, 22*(1), 1-20. https://doi.org/10.1787/hemp-22-5ks5dlhqbfr1

National Athletic Trainers' Association. (2011). *Athletic training educational competencies*. 5th ed. Dallas, TX: National Athletic Trainers' Association.

National Governor's Association (1986). *Time for results: The Governors' 1991 report on education.* Washington, D.C.: National Governor's Association.

National Institute of Education (1984). *Involvement in learning: Realizing the potential of American higher education.* Washington, D.C.: U.S. Government Printing Office.

Serban, A. M. (2004). Assessment of student learning at the institutional level. In A. M. Serban & J. Friedlander (Eds.), *Developing and implementing assessment of student learning outcomes* (New Directions for Community Colleges, No. 126, pp. 17-28). San Francisco, CA: Jossey-Bass.

University of Central Florida. (2017). *Creating our collective impact: An overview of our promises and strategies*. Retrieved from https://www.ucf.edu/strategic-plan/files/2017/07/Creating-Our-Collective-Impact-rev072017.pdf.

JAHE | *The* Journal of Assessment *in* Higher Education

Volume 1 | Number 1 April 2020

https://journals.flvc.org/assessment/ ISSN: 2642-0694

Creating Diagnostic Assessments: Automated Distractor Generation with Integrity

Darryl Chamberlain, Jr. & Russell Jeter

University of Florida * *Emory University*

Abstract

The goal of this paper is to propose a new method to generate multiple-choice items that can make creating quality assessments faster and more efficient, solving a practical issue that many instructors face. There are currently no systematic, efficient methods available to generate *quality distractors* (plausible but incorrect options that students choose), which are necessary for multiple-choice assessments that accurately assess students' knowledge. We propose two methods to use technology to generate quality multiple-choice assessments: (1) manipulating the mathematical problem to emulate common student misconceptions or errors and (2) disguising options to protect the integrity of multiple-choice tests. By linking options to common student misconceptions and errors, instructors can potentially use multiple-choice assessments as personalized diagnostic tools that can target and modify underlying misconceptions. Moreover, using technology to generate these quality distractors would allow for assessments to be developed efficiently, in terms of both time and resources. The method to disguise the options generated would have the added benefit of preventing students from working backwards from options to solution and thus would protect the integrity of the assessment. Preliminary results are included to exhibit the effectiveness of the proposed methods.

Keywords: Assessment Generation, Automated Item Generation, Diagnostic Tools

* Corresponding author: Darryl Chamberlain, Jr., Department of Mathematics, University of Florida, 368 Little Hall, P.O. Box 118105, Gainesville, FL, 32611. Dchamberlain31@ufl.edu

Recommended citation*:* Chamberlain, Jr., D. & Jeter, R. (2020). Creating diagnostic assessments: Automated distractor generation with integrity. *Journal of Assessment in Higher Education, 1*(1), 30-49.doi 10.32473/jahe.v1i1.116892

©Journal of Assessment in Higher Education. Published under Creative Commons License CC BY- 4.0 International.

Introduction

Assessment is a critical component of every course. There are two common types of assessment: summative and formative. S*ummative* assessments strive to record student achievement while f*ormative* assessments strive to gather evidence of student learning in order to modify instruction (Cauley & McMillan, 2010). In other words, the primary role of formative assessment is diagnostic – to inform the instructor what each student knows or does not know over some area of content. While there are numerous ways to assess students' knowledge, multiple-choice tests are the most widely used assessments in K-16 as they can be the most efficient to administer while simultaneously being quick and objective to grade (Rodriguez, 2011; Haladyna & Rodriguez, 2013). We use a typical College Algebra item to contextualize multiple-choice item terminology in mathematics.

[Stem]

Solve the linear equation below.

[Problem]

$$\frac{-3x-6}{3}-\frac{-8x-8}{5}=\frac{7x+6}{2}$$

[Options]

A. $x=\frac{-40}{29}$ **[Distractor]**

B. $x=\frac{-66}{29}$ **[Distractor]**

C. $x=\frac{-34}{29}$ **[Solution]**

D. $x=\frac{-17}{10}$ **[Distractor]**

Figure 1: Example of a typical multiple-choice item.

A *multiple-choice item* consists of a stem and options. The *stem* includes the context, content, and problem for the student to answer. In Figure 1's example, this includes the instructions (context) and the problem. By *problem*, we refer to the content issue that must be solved. In the example in Figure 1, this would be solving the linear equation. Solving this problem leads to the *solution*. Plausible, but incorrect, answers to the problem are referred to as *distractors*. The solution and distractors are used to create the *options*, or choices presented that the student must choose from.

Numerous guides for constructing quality multiple-choice questions exist and they largely agree on the best practices for developing assessments (Moreno, Martinez, & Muniz, 2015; Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005). These guides are routinely used by content specialists to create multiple-choice items, which are then disseminated for general use. Guidelines commonly focus on writing the content and choices of an item.

For example, Haladyna et al. (2002) proposed 31 suggestions when writing multiple-choice items: 8 related to content and 14 related to choices. These suggestions can be vague (e.g., “avoid trick or ambiguous items”) and do not provide a way to systematically develop multiple-choice items. In fact, the authors state ``The science of MC item writing is advancing, but item writing is still largely a creative act” (p. 329). The development of a systematic guide to create distractors based on common errors and misconceptions would provide an avenue to advance multiple-choice item writing in a “non-creative” way.

Literature on Distractor Generation

Creating the stem, problem, and solution for a multiple-choice item in K-14 mathematics is a relatively straightforward task. Item content development follows the objectives laid out in the associated textbook, developed by the textbook author(s) to focus on specific content. No such blueprint exists for developing the distractors though. For example, consider a question that asks students to expand the expression $(x-y)^2 = x^2 - 2xy + y^2$. A student with incomplete knowledge of polynomial expansion may choose $x^2 + y^2$ as the expansion and ignore the second term. Another student with partial knowledge of polynomial expansion may choose $x^2 - y^2$ and consider ‘distributing the exponent’ as a valid mathematical operation (Filloy & Rojano, 1989). These two examples illustrate common student misconceptions with polynomial expansions – misconceptions instructors want to capture during formative assessment so that these conceptions can be challenged and subsequently modified. This illustrates one of the biggest hurdles for creating quality distractors: the misconceptions a student may hold can be item-specific, requiring an item-by-item analysis. Without a systematic method to develop these distractors efficiently, creating a single assessment can be a timely endeavor.

It is well-known that distractors play a fundamental role in multiple-choice tests for any topic (Haladyna & Rodriguez, 2013). Gierl et al. (2017) consider distractors to (i) require a significant amount of time and resources to create, (ii) affect item quality and learning outcomes, and (iii) provide diagnostic inferences about students' knowledge (e.g., inferences about what students know or do not know). The authors go on to say that “Distractor development, in fact, is often considered by content specialists to be the most daunting and challenging component of writing a multiple-choice item” (p.1086). Yet, research on empirically-supported development of quality distractors for multiple-choice items is relatively sparse, even in the context of mathematics specifically (Gierl, Lai, Hogan, & Matovinovic, 2015). The following paragraphs will review the recent advances in generating quality distractors and how this paper will expand on these advances.

There are currently three general strategies to generate distractors (Chamberlain, Jr. & Jeter, 2019). The first focuses on common misconceptions in student thinking while they reason about the problem. We illustrated this with polynomial expansion as students hold two pronounced misconceptions about polynomial expansion. These misconceptions can be recalled and utilized by experienced content specialists reflecting on the common errors they have seen in the past (Collins, 2006) or identified through evidence-based research on students' work during open-ended items (Briggs, Alonzo, Schwab, & Wilson, 2006).

As such, this approach creates high-quality distractors that mirror mistakes (based on misconceptions) students may make during an assessment. This quality comes at a steep price – a great deal of time and resources must be used to develop these distractors, especially for items developed through evidence-based research (Gierl, Bulut, Guo, & Zhang, 2017).

The second strategy focuses on similarities between the solution and distractors. For example, a numeric solution such as $\frac{3}{4}$ could be manipulated in some form (e.g., being negated, divided by a factor, or divided by 1) to provide a host of distractors like $\frac{-3}{4}, \frac{4}{3}, -\frac{4}{3}$. In contrast to the first strategy, manipulating the solution in some way to make similar responses does not require a great deal of time and resources, and thus is commonly utilized (Gierl, Bulut, Guo, & Zhang, 2017). The disadvantage to this method is that distractors may not reflect actual mistakes a student would make on the assessment. Students with incomplete knowledge may be able to eliminate these types of distractors and thus arrive at the solution (or, at least, more easily guess at the solution). Alternatively, students who completed the problem correctly may accidentally choose visually similar distractors and thus their multiple-choice answer would not accurately reflect their knowledge. Due to these limitations, some authors have suggested multiple-choice assessments cannot provide diagnostic information (Lissitz, Hou, & Slater, 2012) rather than the more nuanced position that multiple-choice assessments are not commonly written to provide diagnostic information.

The third strategy relies on utilizing research on how students develop an understanding of concepts to model student responses at different levels of conception. For example, the Precalculus Assessment by Carlson, Oehrtman, and Engelke (2010) utilized a theoretical model for how students develop an understanding of covariational reasoning, the Covariation Framework (Carlson, Jacobs, Coe, Larsen, & Hsu, 2002), along with interview-based research to create common student reasoning based on each level of understanding. These responses were used as the option choices for multiple-choice assessments that targeted students' level of understanding. Similar to the first strategy, developing interview-based items was a resource-heavy endeavor (Carlson, Oehrtman, & Engelke, 2010).

In short, creating distractors based on conceptions and/or misconceptions is preferred but not always feasible, and thus distractors are commonly developed based on small variations of the solution. One avenue for creating quality distractors based on conceptions and common misconceptions is *Automatic Item Generation* (AIG). AIG utilizes computer technologies and content specialists to automatically generate problems, solutions, and quality distractors. By *automatically*, we mean that an item structure can be developed ahead of time that some technology would use to create many items without the need for future human intervention. Few examples of AIG currently exist, even in the context of mathematics (Gierl, Lai, Hogan, & Matovinovic, 2015; Gierl, Bulut, Guo, & Zhang, 2017). We now review one of the most recent, relevant works in AIG to set the stage for our method. This method was written to be general and used medical science as the context for their examples.

Gierl and Lai (2013) described a three-step process for generating multiple-choice items. First, an *item model*, or the general scaffolding of the stem and problem, is developed. Then, the content knowledge required to solve the problem to be used in the item is determined. Finally, computer-based algorithms are used to place content from step 2 into the item model from step 1. The authors suggest that "Using this three-step process, hundreds or even thousands of items can be generated using a single item model" (p. 37). We provide a short overview of each of these three steps as described by Gierl and Lai (2013), along with our own examples in a mathematical context.

Step 1: Item Model

There are currently two types of item models: 1-layer and n-layer item models. A 1-layer item model manipulates some small number of elements in the model, all at the same level. We can think of this as choosing 1 element from some set. For example, to generate a linear equation of the form $y = mx + b$, we could choose m and b to be rational numbers. Choosing a single rational pair (m, b) provides the single set needed to change the mathematical problem at hand. This would generate a 1-layer item asking students to solve the equation $y = mx + b$ 1-layer item models are ubiquitous in current multiple-choice tests (Gierl & Lai, 2013).

An n-layer item model manipulates many elements at multiple levels in a model. We can think of this as choosing 1 element from numerous sets. For example, an item model may ask students to solve a linear equation of *any* form. The item model could first choose the form the linear equation would be displayed in (e.g., standard, point-slope, slope-intercept). The item model could also choose the types of numbers that would be used in the linear equation (e.g., Naturals, Integers, Rationals). This would create a 2-layer item that chooses one element from the sets of equation type and number type. After making these two choices, the problem equation can be generated. In summary, the *n*-layer structure has multiple layers of elements, where each element can be varied *simultaneously* to produce varying items. The n-layer model can thus quickly develop test items that address many content objectives based on how the elements in the structure are chosen, though care needs to be taken to ensure consistent item difficulty. This will be addressed in step 2.

Step 2: Content

After determining the item model, content specialists are used to identify the content. Two general approaches to identifying content exists: weak and strong theory (Gierl & Lai, 2013). *Weak theory* uses design guidelines to create new item models that remain similar (in terms of difficulty and structure) to the original item model and is commonly employed in 1-layer item models. For example, to create similar linear equations to solve, the content specialist would choose a single type of linear structure and a single type of elements for this structure, as introducing additional changes may fluctuate the difficulty of the item.

That is, the linear equation $4 = 2x + 5$ may be easier for a student to solve than the equation $\frac{-3x-6}{3} - \frac{-8x-8}{5} = \frac{7x+6}{2}$ as they utilize different structures of a linear equation. Similarly, the linear equation $\frac{4}{7} = \frac{2}{3}x + 5$ would likely be more difficult for a student than $4 = 2x + 5$ as it introduces rational numbers to the same structure. This illustrates why 1-layer items are ubiquitous in assessment generation.

Strong theory utilizes a cognitive model to identify and manipulate items that may change the difficulty level of the item. While relatively few cognitive theories exist to guide general item development practices (Gierl & Lai, 2013), many have been proposed in the last 30 years in undergraduate mathematics education (Leatham, 2014). These can be utilized to model the knowledge and skills a theoretical student may need to solve the mathematical problem, which in turn can provide guidance to develop item models and manipulate the elements of the item model. This potential was illustrated in the Precalculus assessment by Carlson, Oehrtman, and Engelke (2012).

Step 3: Computer-Based Algorithms

Once the item model is created and the content for the model determined, a computer program is needed to assemble the two to create specific items. While software has been developed specifically for generating test items, Gierl and Lai (2013) state "… it is also important to note that any linear programming method can be used to solve the type of combinatorial problem found within AIG" (p. 43-44).

The three-step method above focuses on item generation holistically. Gierl and Lai (2013) showcased an n-layer structure with a possible solution list that remained static while the stem was changed, resulting in different solutions from the static solution list. The resulting distractors were the rest of the possible solutions, which may or may not have mirrored student misconceptions based on the randomly generated problem. This illustrates how the n-layer structure does not inherently describe how distractors could be automatically generated based on student misconceptions for the particular problem generated. As generating distractors is the most difficult aspect of multiple-choice item generation (Gierl, Bulut, Guo, & Zhang, 2017), we will introduce a novel method to automatically generate distractors by manipulating the *problem* within the stem in a way that reflects students' misconceptions and mistakes. The following section details this distractor-generation process.

Automated Assessment Generation Method

We present a method for generating assessments that is grounded in the idea of creating *nearby problems* based on common errors made while solving the original problem as well as on common misconceptions students have with the content being evaluated. From these nearby problems, one can create a set of *distractor solutions* that can be used as answer choices in a multiple-choice item.

Question and Solution Generation

Before we can discuss the process by which we create *plausible* distractors, the reader must have a clear understanding of how questions can be randomly generated, and by extension, the solutions (correct answers) to those questions. Figure 2 introduces the sample question that we use to walk the reader through the methodology for the automated assessment algorithm conceptually, before presenting the algorithm more generally.

Question 1. Solve the linear equation below.

$$\frac{-3x-6}{3}-\frac{-8x-8}{5}=\frac{7x+6}{2}$$

A. $x=\frac{-40}{29}$

B. $x=\frac{-34}{29}$

C. $x=\frac{-66}{29}$

D. $x=\frac{-17}{10}$

Figure 2: College Algebra example item.

A question of this type can be randomly generated from a template for questions that involve solving rational equations. To create this template, all coefficients in the numerators and the denominators are replaced with unknown integers that are randomly chosen at the time the problem is generated. The general form of this type of problem is:

$$\frac{a_1x+b_1}{c_1}-\frac{a_2x+b_2}{c_2}=\frac{a_3x+b_3}{c_3},$$

where $a_i, b_i,$and c_i are integers. Typically, these numbers are chosen within a range that will not make the problem too computationally unwieldy, though with the ubiquity of calculators, this can be relaxed. The following limitations are placed on these unknown integers to ensure exactly one solution: (i) $c_1, c_2, c_3 \neq 0$ and (ii) $c_2c_3a_1 - c_1c_3a_2 - c_1c_2a_3 \neq 0$. After guaranteeing that a unique solution exists, we methodically generate the general solution to this problem template.

Generating Plausible Distractors

In problem solving, a *plausible* distractor would be one that corresponds to a specific, common error that a student can make when solving a problem or an observed student misconception. Plausible distractor solutions provide a way to evaluate specific content issues a student is having by consistently providing answer choices that correspond to common misconceptions. Moreover, they provide a more reliable assessment by avoiding the confounding of artificially similar answer choices. The process for creating plausible distractor solutions is nearly identical to the process for creating the correct solution, in that an exact, unique solution to a problem is found. The difference is that for distractor solutions, we construct *nearby problems* that are based on common errors students make when solving the original problem. Based on these errors, we can reverse engineer a problem, and then solve that problem algorithmically to obtain a *nearby solution* to the original problem. To make this more concrete, we present the creation of a distractor for the original problem.

A potential error that students may make when solving rational equations is that they do not divide each term in the numerator by the denominator. Essentially, students who do not have a complete understanding of rational expressions are solving the problem

$$\frac{a_1x}{c_1}+b_1-\frac{a_2x}{c_2}-b_2=\frac{a_3x}{c_3}+b_3.$$

In a similar way, distractors can be created for not dividing the first term in the numerator by the denominator or failing to distribute the minus sign in the numerator of the second term of the rational equation. These distractors are summarized in Figure 3 below.

Question 1. Solve the linear equation below.

$$\frac{-3x-6}{3}-\frac{-8x-8}{5}=\frac{7x+6}{2}$$

A. $x=\frac{-40}{29}$ *This corresponds to not distributing division throughout.*

B. $x=\frac{-34}{29}$ *This is the correct solution.*

C. $x=\frac{-66}{29}$ *This corresponds to not distributing division in the first term.*

D. $x=\frac{-17}{10}$ *This corresponds to failing to distribute the minus sign in the second term.*

Figure 3: The problem introduced in Figure 2 with the distractor solutions revealed and explained.

This method to automatically generate quality distractors can easily be extended to other observed issues that instructors see in students' work. Specifically, to generate a distractor from a known misunderstanding, solve the general template of the problem while committing the error(s) associated with the misunderstanding. Then, reverse engineer a nearby problem in the form of the original problem template, so that the nearby solution can be obtained in the same way as the original solution. This creates a plausible nearby solution that can be used as a distractor answer choice for the problem.

We have created plausible distractors that mirror common student errors made while completing an open-response version of this question. However, a student can find the correct solution to the previous example by taking each option and plugging it into the question, thereby rendering these distractors moot. The next section addresses this critical loophole in multiple-choice assessments by masking these distractors (and the solution) in intervals.

Disguising the Plausible Distractors and Solution

Solving algebraic problems presents a unique challenge for creating quality multiple-choice questions. When presented with a collection of options for the solution to a problem, students can test each of the potential solutions in the original equation and determine whether a given solution is valid. Considering this, additional measures must be taken to mask the answer choices to preserve the integrity of the distractors and ultimately generate quality assessments.

Conceptually, the additional layer for masking the answer choices is straightforward: replace the single-number answers with intervals that contain not only the corresponding single-number answer, but also infinitely many nearby numbers. This detaches students from the idea that they can test all the answer choices, because each answer choice contains an interval of infinitely many values that can be tested in the original problem. In Figure 4 below, we show an example of how the assessment question looks with the disguised answer choices.

Question 1. Solve the linear equation below. Then choose the interval that contains the solution.

$$\frac{-3x-6}{3}-\frac{-8x-8}{5}=\frac{7x+6}{2}$$

A. $x \in [-1.47, -1.21]$

B. $x \in [-1.21, -0.94]$

C. $x \in [-2.30, -2.10]$

D. $x \in [-1.78, -1.61]$

Figure 4: Multiple-choice example with masked solution and distractors.

Random, algorithmic interval generation itself is simple, compared to the method for generating distractor solutions described in the previous section. However, the problem-specific requirements for masking the answer choices can be a little more nuanced than the general algorithm for creating intervals. To create a quality disguise, it is necessary that the interval does not give clues as to the specific value that it is disguising. We do so by creating intervals that must satisfy two criteria: (i) there is minimal overlap between intervals (as any overlap will not contain a solution) and (ii) the intervals do not reveal much information about the solutions they are disguising. We achieve this generation utilizing a normal standard distribution and interval checking using Python, but the interval generation need not be done in this way.

Method for Generating Multiple-Choice Items

We walked through how to generate a multiple-choice item based on a “Solve the equation” type question utilizing the 3-step model described by Gierl and Lai (2013). Here we explicitly describe how to include distractor generation into the 3-step model.

Step 1: Item Generation

In this step, the stem-type should be determined. This is equivalent to writing a free-response question and must include the stem and problem. In order to procedurally-generate versions of the question, elements of the stem and problem that can be modified must be identified at this point. A 1-layer model would be developed if only some small number of elements in the model can be modified. An n-layer model would be developed if many elements at multiple levels in a model can be modified.

Step 2: Content

In this step, the content knowledge required to solve the problem is determined. To accommodate the development of plausible distractors, any common errors or misconceptions associated to the problem should also be determined here. This can be collected by content specialists recalling common errors or misconceptions they are familiar with, recording any common errors identified in educational research experiments, or theoretically predicted errors or misconceptions according to published mathematics education theoretical perspectives.

Step 3: Computer-based Algorithms

In this step, the content knowledge collected in step 2 is utilized to procedurally solve the problem. In addition, distractor solutions should also be generated by:

a) Isolating common conceptual misunderstandings or common errors related to the topic assessed by the problem.
b) Using these misunderstandings and/or errors to construct ``nearby problems”.
c) Algorithmically solving these nearby problems to create a list of distractor solutions.

If the solution and distractor solutions are numeric in nature, the options can be disguised by algorithmically generating intervals that must satisfy two criteria:

a) There is minimal overlap between intervals (as any overlap will not contain a solution).
b) The intervals do not reveal much information about the solutions they are disguising.

To create distinct nearby problems based on common misconceptions or errors, the original stem/problem may need to be modified or a check may need to be created to regenerate the question until common misconceptions or errors do not produce the same solution as the correct solution.

Discussion of the Merits

Efficient assessment generation - Distractor generation is simultaneously the most costly and critical component of writing multiple-choice assessments (Gierl, Bulut, Guo, & Zhang, 2017). In the literature, there were effectively three options when generating multiple-choice exams: (i) generate distractors based on similarity to the solutions (*weak theory*), (ii) generate every distractor manually by relying on previous experiences with students or through experimental data (*strong theory*), or (iii) relying on education research that describes how students could develop their conception (Chamberlain, Jr.& Jeter, 2019). While methods (ii) and (iii) are preferred to develop strong assessments, method (i) is commonly used due to the high costs of generating every distractor (Gierl & Lai, 2013). Our method generalizes and automates these distractors so that numerous items may be generated. In fact, some student errors (such as not distributing a negative) are so ubiquitous that they can be considered for a wide range of questions. This further reduces the time and effort a content specialist would need to generate distractors based on common student errors and misconceptions. Thus, our method for automatic item generation would allow for the cost-efficient development of numerous multiple-choice tests.

Multiple-Choice Assessment Integrity - One of the limitations of multiple-choice tests is the ability to assess students' procedural knowledge with integrity. This limitation is especially prevalent in K-14 mathematics, where questions will commonly require students to solve an equation (or system of equations) and provide possible solutions. A student needs only check these options in order until one satisfies the equation to arrive at the correct solution. To counter this limitation, we introduced a method to automatically generate intervals for each solution that effectively mask these options to prevent students from gaming the assessments. In unison with our distractor generation, we can automatically generate and mask multiple-choice options to assess students' procedural knowledge with integrity.

Formative assessment - Traditional multiple-choice assessments are used to determine whether students know or do not know some content. This is akin to knowing whether there is an issue with students' knowledge but does not effectively allow instructors to diagnose *why* there may be an issue. By considering the distractors a student chooses over the course of one or more assessments, instructors can more accurately pinpoint *why* a student is not answering a question correctly. For example, during a multiple-choice assessment, a student may answer $5/20$ questions incorrectly. This student may have some minor issue with multiple content ideas, but it is also possible they are making the same common student errors (such as not distributing a negative correctly) over multiple questions. By tracking which solutions *and* distractors a student chooses throughout an entire assessment, we can more accurately assess if their issues are with the content or common mistakes. Moreover, this allows instructors to continuously evaluate foundational knowledge while simultaneously evaluating new content knowledge. These benefits illustrate that multiple-choice assessments *can potentially* provide diagnostic information, contrary to prevalent beliefs about multiple-choice assessments (Lissitz, Hou, & Slater, 2012).

Consequential merits from those described above include:

- **Potential for widespread use –** Unlike assessments developed by hand, these assessments can be used widespread once they are developed as they are efficient to generate and maintain their integrity even when the generation methods are shared.
- **Practical and Research Usefulness** – Assignments can be created for formative assessment in the classroom as well as for large-scale research use to test theoretical conception development.
- **Standardization of assessment** – Makes standardization of easy-to-generate assessments (e.g. aligned to State/National standards) possible.
- **Potential to use calculators –** By providing a method to disguise numeric-type options, the method allows for students to utilize calculators without

dampening the integrity of the assessment.

Limitations

The method is not without limits. We discuss the most pressing issues with the method below, while also describing how these limitations can be mitigated.

High Start-Up Cost - Generating high-quality multiple-choice items normally requires a content specialist for distractor design. Our method would require either both a content specialist and someone with programming experience working side-by-side, or a content specialist with programming knowledge. For questions attempting to utilizing a theoretical perspective for how students with a misunderstanding or under-developed conception may answer, this would also require an education specialist. This further increases the start-up costs of developing multiple-choice assessments, making the method impractical for instructors with limited resources. However, once a series of items are created, they can be easily disseminated to other instructors. This task can be performed by those with the resources to do so and mass disseminated to other instructors.

Complication of Multiple-Choice Options - Masking the multiple-choice options, while effective in protecting the integrity of the assessment, does complicate students' choice of the solution. Rather than searching for the exact match of their answer, students would need to parse the interval notation language. Moreover, this may become confusing when the solution itself is an interval. For example, consider the inequality item in Figure 5.

Question 2. Solve the linear inequality below. Then, choose the constant and interval combination that describes the solution set.

$8x - 6 > 10x$ or $5x - 5 < 8x$

A. $(-\infty, a) \cup (b, \infty)$, where $a \in [1.5, 4.1]$ and $b \in [2, 4]$.

B. $(-\infty, a) \cup$, where $a \in [-9, -2]$ and $b \in [-8, 2]$.

C. $(-\infty, a) \cup$, where $a \in [-3, 5]$ and $b \in [-1, 5]$.

D. $(-\infty, a) \cup (b, \infty)$, where $a \in [-4.9, -1.6]$ and $b \in [-3, 0]$.

E. $(-\infty, \infty)$.

Figure 5: Automatically generated problem-solving systems of inequalities with interval answer choices.

While it may be second nature to instructors, students may struggle to interpret a phrase such as (a, ∞), where $a \in [a_1, a_2]$ for some a_1, a_2. This could lead to students solving the inequality correctly but choosing the wrong option. Addressing this limitation is a topic of future research.

Preliminary Results

Overall, our method is promising. It has been used to generate multiple exams for a large (800-1000 students annually), hybrid course of College Algebra. By leveraging Python, SageMath, and shell scripts written over the course of a year, complete exams and keys are generated *without any human input* in approximately 2.5 minutes. Two points to emphasize:

(1) *No technological skill is needed to create the exams at this point* (though the instructor may need assistance downloading the open-access software and files utilized by the authors) and
(2) Exam generation would cost nothing to the instructors nor to the students.

While data analysis for these assessments is ongoing, a summary of statistics for the Final Exam in Fall 2017 and Fall 2019 is provided below.

The Final Exam in Fall 2017 consisted of 25 multiple-choice questions with 4 options each. The majority of questions, 20/25, were taken from Pearson's College Algebra test bank while the other 5 were previous free-response questions (written by the instructor) and modified to be multiple-choice. The Final Exam in Fall 2019 consisted of 22 multiple-choice questions with 5 options each. These questions were generated using the procedure described in this paper. We analyzed three parts of each exam: (1) distractors, (2) how well individual items predicted student success, and (3) how consistent the exam was as a whole.

The procedure described in this paper for generating distractors provided ways to create *plausible* distractors – mistakes and misconceptions students could theoretically have. Literature suggests an *effective* distractor is one that is chosen at least 5% of the time (Hingorjo & Jaleel, 2012). We could then consider *quality* distractors as those that students both theoretically could make (based on misconceptions or common errors) and do make during exams. We analyzed the distractors in both Final Exams in two ways: (1) categorizing the frequency of each individual distractor being selected (DS) some percentage of time and (2) calculating the number of items with *x* many distractors chosen at least 5% of the time. Tables 1-4 present a summary of these results. These percentages were done by version of the exam and then averaged for ease of discussion.

Table 1 illustrates the percent of distractors that were selected by frequency. For example, Fall 2017 AVG 16% means that 16% of the distractors in Fall 2017 were not chosen by students for their respective question. Both exams had similar percentages of their distractors chosen through the exam. This suggests the novel procedure introduced in this paper was at least as effective as the non-computer-generated exam. This could also be considered a success for the computer-generated exam as it provided an additional distractor for each question and had the potential to provide an overabundance of theoretical distractors that students did not actually choose.

Table 1:

Percentage of distractors selected by students out of total number of distractors in Fall 2017 and Fall 2019.

	Fall 2017				Fall 2019			
Distractor Selected (DS)	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG
0%	9%	17%	23%	16%	19%	27%	15%	20%
0% < DS < 5%	33%	29%	40%	34%	38%	27%	33%	33%
5% < DS < 10%	29%	29%	17%	25%	20%	28%	30%	26%
10% < DS < 15%	16%	11%	11%	12%	15%	7%	10%	11%
15% < DS < 20%	3%	4%	1%	3%	3%	5%	7%	5%
DS > 20%	8%	9%	8%	8%	5%	6%	5%	5%

Table 2 illustrates a by-question analysis of the distractors by considering the number of items with *x* many distractors chosen at least 5% of the time. Again, we note that the computer-generated exams provided 4 distractors, while the non-computer-generated exam had only 3 distractors. Here we see clear advantages to the computer-generated distractors. It averaged generating 5% of the exam items with *all 4 distractors* chosen by students and an additional 15% average of exam items with 3 effective distractors. The largest difference was in the number of questions with no effective distractors: 9% average for the computer-generated exam versus 21% average for the non-computer-generated exam. This is a clear success of the distractor generation method – it provided at least one quality distractor for a large majority of the exam (91%).

Table 2:

Percent of questions with x distractors chosen by more than 5% of students.

	Fall 2017				Fall 2019			
Items with x distractors chosen >5%	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG
4	NA	NA	NA	NA	0%	9%	5%	5%
3	28%	24%	8%	20%	18%	14%	14%	15%
2	28%	32%	28%	29%	45%	32%	27%	35%
1	32%	24%	32%	29%	27%	41%	41%	36%
0	12%	20%	32%	21%	9%	5%	14%	9%

In Item Response Theory, statistics are used to measure the relationship between performance on individual assessment items and the overall assessment (Varma, 2006). The Point-Biserial Correlation (PBC) is a common correlation measure for assessments, where a positive PBC corresponds to a high-achieving student marking the question correctly while low-achieving students marking the question incorrectly. A PBC of 0.1 or higher is considered desirable while simultaneously avoiding negative PBCs, which are indicative of low-achieving students marking correctly what high-achieving students mark incorrectly (Varma, 2006).

Table 3 categorizes the percentage of questions that fall within the identified ranges. First, it should be noted that both exams have a large percentage of predictive questions: 88% and 91% respectively. They both also have similar numbers of problematic questions (1% and 2%) and suspect questions (5% and 8%). Like the Distractor Selection analysis, this suggests the computer-generated exam is at least as effective at generating quality distractors as the non-computer-generated exam.

Table 3:

Percentage of assessment items in point biserial coefficient ranges.

	Fall 2017				Fall 2019			
Point Biserial Correlation	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG
PBC < 0	0%	0%	4%	1%	0%	0%	5%	2%
0 < PBC < 0.15	8%	0%	8%	5%	5%	14%	5%	8%
0.15 < PBC < 0.25	4%	12%	16%	11%	0%	14%	18%	11%
PBC > 0.25	84%	80%	68%	77%	95%	73%	73%	80%

Finally, the KR-20 reliability coefficient is used to estimate the internal consistency reliability of an assessment (Salvucci, Walter, Conley, Fink, & Saba, 1997). In other words, the reliability coefficient attempts to measure whether another group of similar students achieve in a similar way. Salvucci, Walter, Conley, Fink, & Saba (1997) proposed the following interpretations of KR-20 coefficients:

- Less than 0.5, the reliability is low;
- Between 0.5 and 0.8, the reliability is moderate;
- Greater than 0.8, the reliability is high (p. 115).

Table 4 illustrates that both exams are in the upper-moderate range. As with much of the other data, this suggests the computer-generated exams are at least as good as the non-computer-generated exams. However, this is another clear win for the computer-generated exams as this result illustrates that the code effectively controlled for changing the individual items *without changing their difficulty,* a potential issue described when detailing the procedure.

Table 4:

KR-20 correlation coefficients

	Fall 2017				Fall 2019			
	Ver A	Ver B	Ver C	AVG	Ver A	Ver B	Ver C	AVG
KR20	0.80	0.79	0.60	0.73	0.73	0.65	0.69	0.69

At each level of the data analysis (distractor, item, and overall exam), the computer-generated exam was shown to be at least as effective as the exam created by Pearson and the instructor. The one major difference was in the number of items with at least one quality distractor: 91% versus 79%. Combining this with the clear advantages in amount of time to create an exam (2.5 minutes for all 3 versions of Fall 2019) and dynamic nature of the computer-generated exams, the procedure appears to be effective at generating quality distractors and quality multiple-choice exams in general.

Conclusions

Automated item generation is not a novel concept in the assessment literature and has been discussed as early at 1969 (Bormuth, 1969). Since then, copious guidelines for developing multiple-choice items have been developed and agree that distractors play a fundamental role in multiple-choice tests. For example, Gierl et al. (2017) consider distractors to (i) require a significant amount of time and resources to create, (ii) affect item quality and learning outcomes, and (iii) provide diagnostic inferences about students' test performance. The authors go on to say that “distractor development, in fact, is often considered by content specialists to be the most daunting and challenging component of writing a multiple-choice item” (p. 1086). Yet, automated distractor generation has received relatively little attention, even in the context of mathematics (Gierl, Lai, Hogan, & Matovinovic, 2015). When automatic distractor generation has been explored, it has largely been relegated to manipulating the solution of an item in some minor way or by mapping all possible solutions to the (relatively simple) structure of an item (Gierl, Bulut, Guo, & Zhang, 2017).

Note that distractor generation is distinct from the approach Intelligent Tutoring Systems (such as ALEKS) take that utilize student *Knowledge Spaces* – a pair (Q, K) consisting of some set of Questions (Q) and a subset of questions (K) that represent the questions a student could answer correctly (Cosyn & Thiery, 2000). At a fundamental level, Knowledge Spaces operate by identifying the questions a student can and cannot complete in some progression to determine the next question their knowledge would allow them to start on. For a simplistic example, consider a concept to have 6 linear questions that build up to a robust understanding. The system would start by asking the student to complete Q1 – if the student is correct, it could move on to Q2 or beyond. If the student was correct with Q1, it moved to Q3, and the student was then incorrect, the student would have the knowledge space K = {Q1} and thus be taught the knowledge needed to answer Q2.

While an oversimplification of Knowledge Spaces, this example illustrates that Intelligent Tutoring Systems work through correct/incorrect and not *theoretically why* a student is incorrect. Moreover, these systems purposely avoid multiple-choice items to further correlate a correct answer to sufficient knowledge to answer the question. Thus, our work on distractor generation is fundamentally different than the computer-generated questions Intelligent Tutoring Systems employ.

We presented a novel method for dynamically generating distractors by creating *nearby problems* that can be algorithmically solved via computers. For a given concept, the instructor decides what content will be evaluated, and chooses the corresponding stem template, from which the problem is algorithmically generated. This formal statement of the problem can be procedurally solved to find the correct solution. Then, using common student misconceptions, *nearby problems* can be constructed and then solved using the same procedure that solved the original problem. These *nearby solutions* are plausible distractor solutions corresponding to specific content areas with which students struggle.

Moreover, we introduced a method of masking these solutions and distractors to prevent students from working backwards from answer choices the correct solution, thus preserving the integrity of these automatically generated multiple-choice items. For problems with single-number answers, we propose hiding the solutions (distractor or otherwise) within non-overlapping intervals that contain not only the corresponding single-number answer, but also infinitely many nearby numbers. This detaches students from the idea that they can test all possible choices, because each answer choice contains an interval of infinitely many values that can be tested in the original problem. However, students who obtained a solution (distractor or otherwise) will be able to easily identify the appropriate answer choice. Numerous methods for generating these intervals can be effective, as long as there is minimal overlap in the intervals and the intervals do not reveal information about the option it is disguising.

Dynamically generating distractors associated with student misconceptions and errors holds a variety of theoretical merits. First and foremost, it allows for the cost-efficient development of numerous multiple-choice assessments. Constructing a single multiple-choice, n-layer item can result in hundreds (or even thousands) of questions with relevant distractors. In unison with our method to mask options, multiple-choice assessments can be efficiently used to assess students' procedural knowledge with integrity. Dissemination of these automatically generated assessments can help solve a practicality issue with educational research (Van Velzen, 2013) by bridging the gap between the research and practice.

Theoretically speaking, as generated distractors are associated to student misconceptions and errors (rather than small perturbations of the correct solution), these assessments can be used to help diagnose *why* a student did not answer a question correctly and could counter the misconception that multiple-choice assessments cannot provide diagnostic information (Lissitz, Hou, & Slater, 2012). This method also allows instructors to track misconceptions and small errors through multiple assignments, allowing for the continuous evaluation of foundational knowledge while simultaneously evaluating new content knowledge. Tracking misconceptions and small errors could potentially lead to partial credit on multiple-choice items and create free-response-like grading. It can also allow the development of semester-long feedback systems that track student development of critical concepts.

Our method is not without limitations. It further increases the start-up costs of developing multiple-choice assessments, making the method impractical for instructors with limited resources. This, however, can be mitigated by the generality of the method. In addition, masking the distractors and solutions complicates the option decision process, which may lead to students solving the problem correctly but choosing the wrong option.

Conflicts of Interest

No potential conflict of interest is reported by the authors.

References

Bormuth, J. (1969). *On a theory of achievement test items.* Chicago, Illinois: University of Chicago Press. doi:10.1086/443056

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33-63. doi:10.1207/s15326977ea1101_2

Carlson, M., Jacobs, S., Coe, E., Larsen, S., & Hsu, E. (2002). Applying covariational reasoning while modeling dynamic events: A framework and a study. *Journal for Research in Mathematics Education, 33*(5), 352-378.

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction, 28*(2), 113-145. doi:10.1080/07370001003676587

Cauley, K., & McMillan, J. (2010). Formative assessment techniques to support student motivation and achievement. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 81*(1), 1-6. doi:10.1080/00098650903267784

Collins, J. (2006). Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics, 26*, 543-551. doi:10.1148/rg.262055145

Cosyn, E., & Thiery, N. (2000). A practical procedure to build a knowledge structure. *Journal of Mathematical Psychology, 44*(3), 383-407. doi:10.1006/jmps.1998.1252

Filloy, E., & Rojano, T. (1989). Solving equations: The transition from arithmetic to algebra. *For the Learning of Mathematics, 9*(2), 19-25. Retrieved August 16, 2018, from https://www.jstor.org/stable/40247950

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education, 21*, 357-364. doi:10.1016/j.tate.2005.01.008

Gierl, M. J., & Lai, H. (2013). Using automated processes to generate test items. *Educational Measurement: Issues and Practice, 32*, 36-50. doi:10.1111/emip.12018

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082-1116. doi:10.3102/0034654317726529

Gierl, M. J., Lai, H., Hogan, J., & Matovinovic, D. (2015). A method for generating test items that are aligned to the common core state standards. *Journal of Applied Testing Technology, 16*, 1-18. Retrieved August 16, 2018, from http://www.jattjournal.com/index.php/atp/article/view/80234

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items.* New York, NY: Routledge.

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *JPMA-Journal of the Pakistan Medical Association, 62*(2), 142-147.

Leatham, K. R. (2014). *Vital directions for mathematics education research.* Springer Science & Business Media.

Lissitz, R., Hou, X., & Slater, S. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology, 13*(3), 1-50.

Moreno, R., Martinez, R. J., & Muniz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema, 27*, 388-394. doi:10.7334/psicothema2015.110

Rodriguez, M. C. (2011). Item-writing practice and evidence. In S. N. Elliott, R. J. Kettler, P. A. Beddrow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all student: Bridging the gaps between research, practice, and policy* (pp. 201-206). New York, NY: Springer.

Salvucci, S., Walter, E., Conley, V., Fink, S., & Saba, M. (1997). *Measurement error studies at the National Center for Education Statistics.*

Van Velzen, J. (2013, August). Educational researchers and practicality. *American Educational Research Journal, 50*(4), 789-811. doi:10.3102/0002831212468787 Varma, S. (2006). *Preliminary item statistics using point-biseral correlation and p-values.* Morgan Hill CA: Educational Data Systems Inc.

JAHE | The Journal of Assessment in Higher Education

Volume 1 | Number 1 April 2020

https://journals.flvc.org/assessment/ ISSN: 2642-0694

The Impact of an Assessment Certificate on Faculty Perceptions and Knowledge

Katherine Perez, Eilyn Sanabria, Suzanne Lebin & Jennifer Doherty-Restrepo*

Florida International University

Abstract

Administrators are struggling to understand how to best promote and implement a culture of evidence-based decision making to stakeholders. The research study presented explored best practices on creating meaningful professional development experiences using both direct and indirect evidence of learning -- this article will describe the effectiveness of a hybrid certificate program designed to educate faculty about assessment and its impact on faculty learning gains, perceptions, and self-efficacy. The study used a pre-/post-test design to measure participant knowledge using quizzes for each of the four modules of the certificate and participant perceptions using a survey. The modules covered writing student learning and program outcomes, curriculum mapping, developing assessment methods, creating assessment instruments, collecting data, analyzing and reporting results, and using results for improvement. Certificate completers demonstrated increased knowledge of assessment terminology, procedures, and best practices, as well as improved assessment-related self-efficacy. However, their perception regarding assessment did not change. Data gathered through this study can help inform decisions on needed assessment-related faculty professional development activities.

Keywords: Assessment, student learning outcomes, program outcomes, professional development, evidence-based decision making

* Corresponding author: Jennifer Doherty-Restrepo, Office of Academic Planning & Accountability, Florida International University, 11200 SW 8th Street, PC 112. Miami, FL, 33199. jennifer.doherty@fiu.edu

Recommended citation: Perez, K., Sanabria, E., Lebin, S., & Doherty-Restrepo, J. (2020). The impact of an assessment certificate on faculty perceptions and knowledge. *Journal of Assessment in Higher Education, 1*(1), 50-79. doi: 10.32473/jahe.v1i1.116915

©Journal of Assessment in Higher Education. Published under Creative Commons License CC BY- 4.0 International.

Introduction

Over time, the role of faculty has expanded from teaching to include research and service (Boyer, 1990). Institutions of higher education are placing greater demands on faculty with higher teaching loads and increasing expectations of scholarly productivity (e.g., research publications, funding). Engaging in the scholarship of teaching and learning, which includes the assessment of student learning, is a central focus for institutes of higher education because it is required to maintain regional accreditation. In addition, the assessment of student learning is critical to ensure students are prepared to enter the workforce with the necessary knowledge, skills, and abilities to engage in continued learning beyond graduation (Boud & Falchikov, 2007). However, as Boud and Falchikov (2007) argue, the assessment discourse is "…commonly dominated by the needs of certification" (p. 4) and fails to truly capture its impact on student learning.

Faculty are subject-matter experts rather than pedagogical experts; therefore, they often lack the knowledge and skills necessary to transmit concepts essential to the discipline in ways that optimize student learning (Boyer, 1990; Saroyan & Amundsen, 2004). This knowledge gap between content and pedagogy in the classroom hinders faculty efficacy in the assessment of student learning and in the development and implementation of strategies to improve student learning. For this reason, higher education institutions need to support faculty not only in their research and service appointments, but also in their pedagogical development (Hott & Tietjen-Smith, 2018), which encompasses assessment of student learning for the purpose of continuous academic improvement. In this study, we examine the effectiveness of a faculty professional development certificate program created to (1) instill and enhance knowledge of assessment best practices in student learning, (2) improve attitudes and beliefs about assessment, and (3) improve participants' self-efficacy regarding assessment.

Review of the Literature

Content and pedagogical knowledge are essential to developing and implementing meaningful assessment practices that lead to academic improvement. However, this can only occur when faculty are knowledgeable of effective teaching, learning, and assessment strategies; therefore, faculty development is needed in these areas. The following sections will explore relevant literature related to content and pedagogical knowledge, the current state of assessment practices in higher education, and the need for faculty development in pedagogy.

Content Knowledge and Pedagogical Content Knowledge

Content knowledge is the deep understanding of discipline-specific concepts and principles necessary to make pedagogical and curricular judgments (Shulman, 1986). To ensure learning, faculty must manage numerous components in the classroom (e.g., learner differences, context, background knowledge) while effectively transmitting content knowledge to the students in a meaningful context. This ability to transmit content knowledge to students was coined by Shulman (1986) as *pedagogical content knowledge*. Shulman (1986) differentiates it from content knowledge by arguing that pedagogical content knowledge is the "…amalgam of content and pedagogy (Shulman, 1987, p. 5)" or "…the dimension for subject matter knowledge *for teaching* (Shulman, 1986, p. 9)," which involves understanding, organizing, and adapting content for all learners. Pedagogical content knowledge is necessary for faculty to employ strategies that optimize student learning, which is determined through the assessment of important competencies to be attained by the students. Then, faculty are informed by these assessment data to make data-driven improvements in their classroom, which further expands their pedagogical content knowledge.

Assessment in Higher Education

There has been little change in assessment practices within higher education. According to Ronald Barnett, assessment is not utilized to its maximum potential (Boud and Falchikov, Chapter 3, 2007) and continues to be mostly a bureaucratic process for faculty. David Boud calls for a transformation of the current discourse in higher education into one that reframes the purpose of assessment, how we talk about it, and how we describe it (Boud and Falchikov, Chapter 2, 2007). This transformation can only take place if we empower stakeholders (e.g., faculty, administration) with the knowledge and understanding of assessment best practices. In order to accomplish this, active collaboration (Banta, Jones & Black, 2009) is needed to "…find ways of thinking about assessment that have positive consequential influence on learning (Boud and Falchikov, 2007, p. 19)." Aligning scholarship and teaching for the improvement of assessment practices, beginning with faculty development, is necessary to transform the assessment discourse.

Faculty Development

Higher education institutions hire faculty as subject-matter experts and expect them to effectively convey knowledge to their students; yet, many faculty lack pedagogical training. De Golia et al. (2019) surveyed psychiatry faculty to assess faculty development needs and *teaching skills workshops* including teaching methods, assessment skills, and pedagogy was identified as an unmet need. Faculty teaching in an online education doctorate program identified the need for more professional development in the area of pedagogy in a qualitative study conducted by Berry (2019). A study conducted by Behar-Horenstein, Garvan, Catalanotto, Su, and Feng (2016) assessed faculty development needs amongst dental faculty; their findings suggest the need for faculty development that enhances teaching. More importantly, Rutz, Condon, Iverson, Manduca, and Willett's (2012) study looked at the relationship between faculty development, pedagogy, and student achievement and identified a direct relationship between faculty development and improved pedagogy. In summary, the findings of these studies demonstrate the need, and faculty desire, for professional development opportunities in the areas of pedagogy and assessment of student learning.

Problem Statement

Current literature (Allan & Driscoll, 2014; Behar-Horenstein et al., 2016; Berry, 2019; De Golia et al., 2019; Hott & Smith, 2018; Pawlyshyn & Hitch, 2016; Rutz et al., 2012) notes the importance of faculty development and its positive impact on student learning. With increasing faculty expectations and growing accountability measures from regional accrediting agencies, higher education needs to expand faculty development in the area of assessment. As Boud and Falchikov (2007) state, "…focus on assessment *practices* is needed, not simply on labelled *methods* considered independently of their consequences (p. 12)." In response to this identified need, our intervention – an assessment certificate program – was provided to faculty to impart knowledge related to best practices in the assessment of student learning and program efficiency for the purpose of continuous academic improvement.

Research Questions

Grounded on literature that stresses the importance of faculty development (Allan & Driscoll, 2014; Boyer, 1990; Hott & Smith, 2018; Pawlyshyn & Hitch, 2016; Saroyan & Amundsen, 2004) and the need for meaningful assessment practices (Allan & Driscoll, 2014; Boud & Falchikov, 2007), this study examined the effectiveness of an institution-wide assessment certificate created for faculty to learn assessment best practices in student learning and program efficiency. The following research questions served as a guide for the design of this intervention:

1. Did knowledge about assessment terminology, procedures, and best practices improve?
2. Does participation in a certificate program improve participants' perception of the effectiveness of support systems?
3. Did participation in the certificate program improve participants' self-efficacy?

The current study took place in the fall 2018 term at Florida International University (FIU). FIU is a large urban public research university with over 58,000 students and 2,300 faculty members. A pre-test post-test experimental design was used to analyze learning gains and participant perception data gathered in this study.

Methodology

The Institutional Effectiveness team within the Office of Academic Planning and Accountability at Florida International University developed a hybrid (delivered partially on-line and face-to-face) certificate program designed to educate faculty about assessment and its impact on faculty learning gains, perceptions, and self-efficacy. The certificate program consisted of four modules delivered over the 12 weeks. The modules covered (1) writing student learning and program outcomes, (2) curriculum mapping, (3) developing assessment methods, creating assessment instruments, collecting data, analyzing and reporting results, and (4) using results for improvement.

Each on-line module consisted of interactive activities to facilitate learning of assessment best practices. The information included in each module was guided by the work of Banta and Blaich (2010), Kuh et. al. (2014, 2015), and Suskie (2009). Participants completed assigned readings, participated in discussion boards, and developed a comprehensive assessment plan. Participants also attended two in-person workshops during which the institutional effectiveness team provided detailed feedback on the assessment plans developed by the faculty. Furthermore, the in-person sessions were designed to reinforce course content by reviewing learning outcomes for each module and providing active learning opportunities for participants (e.g., discussions to where faculty were asked to apply competencies to their programs and courses, activities to create outcomes). This study used a pre-/post-test design to measure participant knowledge using quizzes for each of the four modules and participant perceptions using a survey.

Participants

The participants in this study consisted of faculty, instructors, and staff who enrolled in a semester-long, hybrid assessment certificate program. Recruitment for participants was initiated as an e-mail to a convenience sample of faculty who work on assessment reports for academic programs, certificate programs, and general education courses. Recruitment was focused on faculty from the College of Arts, Sciences and Education (CASE) and the Steven J. Green School of International & Public Affairs (SIPA). The researchers targeted these two colleges as participants in a pilot study since they represented the bulk of the general education courses and programs at the institution.
Out of the 81 people who demonstrated interest, 50 were selected based on their availability and willingness to complete all requirements of the certificate program. By the end of the semester, only 45 participants completed the fourth and final post-test of the program. The analysis of the data reflects the drop-out of two participants from Module 1 through Module 3 and the drop-out of three remaining participants in Module 4. Demographic data for the 48 participants that completed the first three modules are presented in Table 1.

Instruments

Two instruments were used to answer the three research questions. The first instrument given to the participants was a perception survey. The survey consisted of 59 items and was designed by the researchers to measure faculty perceptions of assessment practices, knowledge, and utility.

A Likert scale was used to quantify the rating in the perception survey. For items related to beliefs about assessment and willingness to do assessment tasks, a 4-point Likert scale was used where 1 represented “strongly disagree” and 4 represented “strongly agree.” For items related to self-efficacy or utilization of assessment best practices, a 4-point Likert scale was used where a score of 1 indicated “not at all” and 4 indicated “to a very great extent.” Finally, items related to assessment support and culture in their department, college, and institutional effectiveness office were rated using a 3-point scale where 1 indicated “no”, 2 indicated “sometimes”, and 3 indicated “yes.” The survey questions were broken down into the following categories:

1. Effectiveness & utility of assessment (10 items)
2. Willingness to participate in assessment activities (6 items)
3. Self-efficacy (10 items)
4. Perception of assessment – Department Level (9 items)
5. Perception of assessment – College Level (9 items)
6. Perception of assessment – Institutional Effectiveness Office Level (9 items)
7. Extent to which assessment results are perceived to be useful (6 items)

The same survey was distributed twice, the first iteration before the certificate program began and second iteration after participants completed all modules and assignments of the certificate program (including the in-person session and mid-term and final projects). Both validity and reliability tests of the instrument were conducted. The survey validation method selected was construct validity, which is "the instrument's ability to relate to other variables (Burton, & Mazerolle, 2011)." Thus, Exploratory Factor Analysis was the appropriate statistical technique to identify instrument constructs (Turocy, 2002). To conduct this analysis, the seven categories were grouped according to their possible answer choices (refer to Table 2).

For the categories *Effectiveness and Utility* and *Willingness*, the Kaiser-Meyer-Olkin (KMO) and Bartlett's value (chi-square=761.591, p<.05, sig=.000) exceeds the heuristic of .60, indicating adequate correlations to continue with factor analysis (Burton, & Mazerolle, 2011); refer to Table 3. Communalities range from .476 to .894 (refer to Table 4). Since they all exceed the 0.4 minimum, factor analysis was conducted using all items.

And as shown on Table 5, using the Kaiser rule with eigenvalues greater than 1, two factors emerged (eigenvalues 9.119 and 2.773). In total, 16 components were extracted, accounting for 100% of the variance. The first factor accounts for 56.99% of the variance and the second factor accounts for 17.33% of the variance. The total amount of variance accounted for by the first two principal components solution is 74.32%.

For the categories of *Self Efficacy* and *Results Utility*, KMO and Bartlett's value (chi-square=728.430, p<.05, sig=.000) exceeds the heuristic of .60, indicating adequate correlations to continue with factor analysis (Burton, & Mazerolle, 2011); refer to Table 7. Communalities range from .524 to .898 (refer to Table 8). Since they all exceed the 0.4 minimum, factor analysis was conducted using all items.

As shown on Table 9, using the Kaiser rule with eigenvalues greater than 1, three factors emerged (eigenvalues 7.770, 3.402, and 1.164). In total, 16 components were extracted, accounting for 100% of the variance. The first factor accounts for 48.56% of the variance, the second factor accounts for 21.26% of the variance, and the third factor accounts for 7.28% of the variance. The total amount of variance accounted for by the first three principal components solution is 77%.

The Rotated Component Matrix indicates (boxed in black) the items belonging to each of the three components (refer to Table 10). Factor cross loading at 0.5 or above occurred for two items (boxed in red). As Burton and Mazerolle (2010) suggest, these items should be removed in future administrations.

For the categories of *Perception of Department*, *College*, and *Institutional Effectiveness*, KMO and Bartlett's value (chi-square=1190.405, p<.05, sig=.000) exceeds the heuristic of .60, indicating adequate correlations to continue with factor analysis (Burton, & Mazerolle, 2011); refer to Table 11. The question asking participants to rate whether assessment is valued by the Institutional Effectiveness team was removed from this analysis, as this variable had zero variance (i.e., all responses were "Yes"). Communalities range from .627 to .921 (refer to Table 12). Since they all exceed the 0.4 minimum, factor analysis was conducted using all items.

As shown on Table 13, using the Kaiser rule with eigenvalues greater than 1, five factors emerged (eigenvalues 11.050, 3.127, 2.357, 2.058, and 1.289). In total, 26 components were extracted, accounting for 100% of the variance. The first factor accounts for 42.5% of the variance, the second factor accounts for 12.03%, the third factor accounts for 9.07%, the fourth accounts for 7.92%, and the fifth accounts for 4.99%. The total amount of variance accounted for by the first two principal components solution is 76.51%. These results indicate further refinement of these categories is also necessary, as participants' responses were grouped and analyzed based on three constructs.

The Rotated Component Matrix indicates (boxed in black) the items belonging to each of the five components (refer to Table 14). Factor cross loading at 0.5 or above occurred for three items (boxed in red). As Burton and Mazerolle (2011) suggest, these items should be removed in future administrations. In addition to Exploratory Factor Analysis, a Cronbach's Alpha was done to test the reliability of the items. Reliability indicated high internal consistency (α > 0.70) per category, as shown on Table 15.

The second instrument consisted of pre-test and post-test quizzes for each of the four modules of the certificates to measure participant knowledge about assessment terminology, procedures, and best practices. The pre-tests and post-tests used the same items. Table 16 describes the total number of questions per module and a breakdown of the learning areas the questions focused on.

Participants completed the pre-test before each module and did not see the answers to the questions. The same questions were then presented at the end of the module as a post-test and they had one opportunity to respond correctly. Items were scored dichotomously (1=correct, 0=incorrect). A higher score indicated greater knowledge.

Data Analysis

Descriptive statistics and paired sample t-tests were used to answer the first research question: Did learning about assessment terminology, procedures, and best practices improve? Descriptive statistics were used to calculate averages of the pre-test and post-test and were used as a direct measure for assessing learning gains for each of the four modules. A secondary analysis using paired sample t-tests was conducted to calculate significance of the differences in means. The last two research questions related to changes in attitudes/beliefs, self-efficacy, and effectiveness of support services were answered using independent sample t-tests since the survey was anonymous and we were not able to pair pre-test scores with post-test scores. Alpha levels for all tests were set at the .05 level.

Results

Results for the first research question indicate a significant increase in mean assessment knowledge from pre- to post-test for each of the four modules (Refer to Table 17).

Results for the second research question, change in perceptions regarding assessment after completion of the certificate program, did not indicate significant differences between pre- and post-test scores. However, results for the third research question indicate a post-test significant increase in participants' perceived self-efficacy on most of the areas surveyed (Refer to Table 18). It is important to note pre- and post-survey data were not paired. Since the pre-survey was completed anonymously, multiple submissions resulted in a larger sample size (n=60); this was rectified prior to completion of the post-survey, which yielded an accurate sample size (n=47). This is reflected by the 105 degrees of freedom of the two-tailed *t*-test (refer to Table 3).

Discussion and Future Directions

Results of this study suggest that structured professional development activities are effective in teaching faculty assessment best practices. Though average increases and significant differences could be attributed to test-retest validity since the same questions were used for pre- and post-test quizzes and survey, the need remains for higher education institutions to invest in assessment-related professional development activities for faculty, as they (for the most part) are only subject-matter experts (Allan & Driscoll, 2014; Boyer, 1990; Hott & Smith, 2018; Pawlyshyn & Hitch, 2016; Saroyan & Amundsen, 2004). However, as Banta (2009) states, this must be done with the support of senior-level administrators and with faculty who are committed to the process. Hence, collaboration amongst senior-level administrators, institutional effectiveness teams, and faculty is pivotal in creating and sustaining a culture of meaningful assessment practices within higher education institutions.

Beyond the formal instruction provided through the certificate program, this intervention served as a springboard to initiate a culture of faculty-driven assessment practices throughout our institution. A longitudinal follow-up study to this research is forthcoming to examine whether or not the learning achieved by the participants affects the quality of assessment reports as measured by a standardized rubric. Improving assessment practices among faculty could facilitate more impactful improvement strategies that lead to enhanced student learning; thereby perpetuating a positive cycle of continuous improvement in both teaching and learning.

It is also important to note that the impact of learning experiences such as this certificate on self-efficacy should be further explored. The results indicating that self-efficacy was significantly increased can be meaningful since research shows that self-efficacy is correlated with motivation as it relates to learning and applying/transferring learned concepts from the training program (Chiaburu & Lindsay, 2008). Sorrenti, Filippello, Buzzai, Butto & Costa (2017) found that self-efficacy was positively correlated with traits of conscientiousness, extraversion, openness to experience, and agreeableness and negatively correlated with learned helplessness. These studies suggest that self-efficacy needs to be taken into consideration when creating learning environments for learners such as the faculty and staff in this study. Self-efficacy may be an important factor in influencing perceptions of assessment and motivation to improve and apply assessment best practices. Follow-up studies should include an investigation of whether self-efficacy is correlated with competency mastery and application of learned skills in program/course assessment practices.

Future research should further explore the effectiveness of assessment-related professional development activities, as well as perhaps identify additional assessment-related competencies faculty should master to ensure student success. A phenomenological study would also be beneficial to better understand the challenges institutions and faculty face when developing and implementing assessment practices; survey responses cannot capture the depth, intricacies, and differences amongst institutions. Another untapped area is the long-term effect of assessment-related faculty development activities and its impact on student learning. Beyond faculty preparedness, the preparation of administrative staff (e.g., registrar, student affairs) to assess the quality and effectiveness of their processes, initiatives, and areas of oversight should also be explored, as they too are tasked with supporting the institution’s mission. Finally, further refinement of the survey instrument used in this study should be explored, as it will provide researchers with a valid and reliable instrument to assess categories discussed.

References

Allan, E. G., & Driscoll, D. L. (2014). The three-fold benefit of reflective writing: Improving program assessment, student learning, and faculty professional development. *Assessing Writing*, 21, 37-55.

Banta, T. W., & Blaich, C. (2010). Closing the assessment loop. *Change: The Magazine of Higher Learning*, 43(1), 22-27.

Banta, T. W., Jones, E. A., & Black, K. E. (2009). *Designing effective assessment: Principles and profiles of good practice*. John Wiley & Sons.

Behar-Horenstein, L. S., Garvan, C. W., Catalanotto, F. A., Su, Y., & Feng, X. (2016). Assessing faculty development needs among Florida's allied dental faculty. *American Dental Hygienists' Association*, 90(1), 52-59.

Berry, S. (2019). Professional development for online faculty: instructors' perspectives on cultivating technical, pedagogical, and content knowledge in a distance program. *Journal of Computing in Higher Education,* 31(1), 121-136.

Boud, D. (2007). Reframing assessment as if learning was important. In Boud, D. & Falchikov, N. (Eds.) Rethinking Assessment for Higher Education: Learning for the Longer Term. London: Routledge, 14-25

Boud, D., & Falchikov, N. (Eds.). (2007). Rethinking assessment in higher education: Learning for the longer term. New York, NY, US: Routledge/Taylor & Francis Group.

Boyer, E. L. (1990). *Scholarship reconsidered: Priorities of the professoriate*. Princeton University Press, 3175 Princeton Pike, Lawrenceville, NJ 08648.

Burton, L. J., & Mazerolle, S. M. (2011). Survey instrument validity part I: Principles of survey instrument development and validation in athletic training education research. *Athletic Training Education Journal*, 6(1), 27-35.

Chiaburu, D. S., & Lindsay, D. R. (2008). Can do or will do? The importance of self-efficacy and instrumentality for training transfer. *Human Resource Development International*, *11*(2), 199-206.

De Golia, S. G., Cagande, C. C., Ahn, M. S., Cullins, L. M., Walaszek, A., & Cowley, D. S. (2019). Faculty development for teaching faculty in psychiatry: where we are and what we need. *Academic Psychiatry,* 43(2), 184-190.

Hott, B. L., & Tietjen-Smith, T. (2018). The Professional Development Needs of Tenure Track Faculty at a Regional University. Research in Higher Education Journal, 35.

Kuh, G. D., Ikenberry, S. O., Jankowski, N. A., Cain, T. R., Ewell, P. T., Hutchings, P., & Kinzie, J. (2015). Beyond compliance: Making assessment matter. *Change: The Magazine of Higher Learning*, 47(5), 8-17.

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. L. (2014). Knowing what students know and can do: *The current state of student learning outcomes assessment in US colleges and universities*. Champaign, IL: National Institute for Learning Outcomes Assessment.

Pawlyshyn, N., & Hitch, L. (2016). New Models for Faculty Work: A Critical Need for Community. *The International Journal of Adult, Community and Professional Learning*, 23(1), 39-53. doi:10.18848/2328-6318/cgp/v23i01/39-53

Rawlusyk, P. E. (2018). Assessment in Higher Education and Student Learning. *Journal of Instructional Pedagogies*, 21.

Rutz, C., Condon, W., Iverson, E. R., Manduca, C. A., & Willett, G. (2012). Faculty professional development and student learning: What is the relationship? *Change: The Magazine of Higher Learning*, *44*(3), 40-47.

Saroyan, A., & Amundsen, C. (2004). Rethinking teaching in higher education: From a course design workshop to a faculty development framework. Stylus Publishing, LLC.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1-23.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4-14.

Sorrenti, L., Filippello, P., Buzzai, C., Buttò, C., & Costa, S. (2018). Learned helplessness and mastery orientation: The contribution of personality traits and academic beliefs. *Nordic Psychology*, *70*(1), 71-84.

Turocy, P. S. (2002). Survey research in athletic training: the scientific method of development and implementation. *Journal of athletic training*, 37(4 suppl), S-174.

Table 1.

Demographics of Study Participants (n=48)

Factor	n (%)
Gender	
Male	16 (33%)
Female	28 (58%)
Prefer not to say	4 (8%)
Age	
25 years or younger	0 (0%)
25-34 years	2 (4%)
35-44 years	17 (35%)
45-54 years	14 (29%)
55-64 years	8 (17%)
65-74 years	3 (6%)
75 years and older	0 (0%)
Prefer not to say	4 (8%)
Position	
Instructor	10 (21%)
Assistant Professor	5 (11%)
Associate Professor	14 (30%)
Professor	6 (13%)
Administrator	7 (15%)
Other	5 (11%)
Faculty Rank	
Non-tenure earning	22 (47%)
Tenure earning	3 (6%)
Tenured	20 (43%)

Table 2.

Category Groupings for Factor Analysis of the Survey

Strongly Agree Somewhat Agree Somewhat Disagree Strongly Disagree	To a Very Great Extent To a Considerable Extent To Some Extent Not at All	Yes Somewhat No
Effectiveness and Utility	Self-Efficacy	Perception – Department
Willingness	Results Utility	Perception – College
		Perception – Institutional Effectiveness

Table 3.

Kaiser-Meyer-Olkin and Bartlett's Test for Effectiveness and Utility and Willingness

Statistical Analysis		Results
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.864
Bartlett's Test of Sphericity:	Approximate Chi-Square	761.591
	df	120
	Significance	.000

Note. df=Degrees of Freedom.

Table 4.

Communalities for Effectiveness and Utility and Willingness

Communalities	Initial	Extraction
I believe assessment practices:		
Improve curriculum	1.000	.734
Improve student learning	1.000	.749
Improve student success	1.000	.701
Improve faculty teaching practices	1.000	.656
Provide more meaningful information than course grades	1.000	.476
Lead to shared program goals	1.000	.742
Lead to shared student expectations	1.000	.709
Lead to program or course improvements	1.000	.794
Lead to a better understanding of the curriculum	1.000	.793
Lead to faculty engagement in data-driven improvement actions	1.000	.585
I am willing to:		
Learn about assessment	1.000	.800
Undertake assessment responsibilities	1.000	.867
Teach colleagues about assessment	1.000	.779
Support other faculty to conduct assessment	1.000	.769
Review my course/program curriculum to incorporate assessment best practices	1.000	.841
Analyze assessment results to develop improvement plans	1.000	.894

Note. Extraction Method: Principal Component Analysis.

Table 5.

Total Variance Explained for Effectiveness and Utility and Willingness

Component	Initial Eigenvalues			Extraction SS Loadings			Rotation SS Loadings		
	Total	% of Variance	Cum %	Total	% of Variance	Cum %	Total	% of Variance	Cum %
1	9.119	56.993	56.993	9.119	56.993	56.993	6.807	42.543	42.543
2	2.773	17.329	74.322	2.773	17.329	74.322	5.085	31.779	74.322
3	.985	6.156	80.478						
4	.651	4.066	84.544						
5	.445	2.780	87.324						
6	.405	2.532	89.856						
7	.334	2.085	91.941						
8	.284	1.772	93.713						
9	.250	1.561	95.274						
10	.192	1.198	96.472						
11	.174	1.085	97.558						
12	.119	.742	98.300						
13	.088	.551	98.852						
14	.085	.529	99.381						
15	.053	.332	99.713						
16	.046	.287	100.00						

Note. Extraction Method: Principal Component Analysis. SS=Sums of Squared. Cum=Cumulative

Table 6.

Rotated Component Matrix[a] for Effectiveness and Utility and Willingness

	Component	
	1	2
I believe assessment practices:		
Lead to a better understanding of the curriculum	.876	.161
Lead to program or course improvements	.857	.244
Lead to shared program goals	.845	.169
Lead to shared student expectations	.834	.116
Improve student learning	.808	.309
Improve curriculum	.804	.295
Improve student success	.799	.249
Lead to faculty engagement in data-driven improvement actions	.763	.049
Improve faculty teaching practices	.707	.395
Provide more meaningful information than course grades	.621	.300
I am willing to:		
Review my course/program curriculum to incorporate assessment best practices	.178	.900
Learn about assessment	.145	.883
Support other faculty to conduct assessment	.038	.876
Analyze assessment results to develop improvement plans	.388	.862
Undertake assessment responsibilities	.359	.859
Teach colleagues about assessment	.393	.791

Note. Extraction Method: Principal Component Analysis; Rotation Method: Varimax with Kaiser Normalization. a. Rotation converged in 3 iterations.

Table 7.

Kaiser-Meyer-Olkin and Bartlett's Test for Self-Efficacy and Results Utility

Statistical Analysis		Results
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.772
Bartlett's Test of Sphericity:	Approximate Chi-Square	728.430
	df	120
	Significance	.000

Note. df=Degrees of Freedom.

Table 8.

Communalities for Self-Efficacy and Results Utility

Communalities	Initial	Extraction
Regarding assessment, I am able to:		
Create a mission statement for my department or program	1.000	.524
Create measurable outcomes	1.000	.770
Create a curriculum map	1.000	.784
Differentiate between direct and indirect measures	1.000	.805
Create a rubric	1.000	.824
Assess student work using a rubric	1.000	.753
Collect data related to outcomes and methods	1.000	.875
Analyze assessment results	1.000	.760
Use assessment results to generate improvement actions	1.000	.898
Document implementation and effectiveness of improvement actions	1.000	.738
To what extent are assessment results used within your courses or program:		
To make changes to the curriculum	1.000	.773
To develop best teaching practices	1.000	.825
To create faculty development opportunities	1.000	.621
To engage faculty in discussions about the curriculum	1.000	.838
To evaluate the effectiveness of improvement strategies	1.000	.802
To evaluate whether outcomes are met at the expected level of achievement	1.000	.745

Note. Extraction Method: Principal Component Analysis.

Table 9.

Total Variance Explained for Self-Efficacy and Results Utility

Component	Initial Eigenvalues			Extraction SS Loadings			Rotation SS Loadings		
	Total	% of Variance	Cum %	Total	% of Variance	Cum %	Total	% of Variance	Cum %
1	7.770	48.562	48.562	7.770	48.562	48.562	4.583	28.644	28.644
2	3.402	21.262	69.824	3.402	21.262	69.824	4.435	27.719	56.362
3	1.164	7.275	77.099	1.164	7.275	77.099	3.318	20.737	77.099
4	.842	5.262	82.361						
5	.672	4.203	86.564						
6	.464	2.903	89.467						
7	.405	2.532	92.000						
8	.280	1.752	93.752						
9	.250	1.563	95.315						
10	.203	1.266	96.581						
11	.154	.962	97.543						
12	.145	.907	98.450						
13	.090	.565	99.015						
14	.075	.470	99.485						
15	.050	.314	99.799						
16	.032	.201	100.000						

Note. Extraction Method: Principal Component Analysis. SS=Sums of Squared. Cum=Cumulative

Table 10.

Rotated Component Matrix[a] for Self-Efficacy and Results Utility

	Component		
	1	2	3
Regarding assessment, I am able to:			
Create a rubric	.877	.068	.222
Differentiate between direct and indirect measures	.854	.090	.261
Create a curriculum map	.809	.335	.131
Assess student work using a rubric	.801	.172	.288
Document implementation and effectiveness of improvement actions	.671	.036	.535
Create a mission statement for my department or program	.629	.080	.349
Analyze assessment results	.622	.009	.611
To what extent are assessment results used within your courses or program:			
To engage faculty in discussions about the curriculum	.082	.912	-.013
To evaluate the effectiveness of improvement strategies	.022	.884	.144
To develop best teaching practices	.053	.868	.261
To evaluate whether outcomes are met at the expected level of achievement	.121	.855	.002
To create faculty development opportunities	.276	.738	-.010
To make changes to the curriculum	.136	.736	.461
Regarding assessment, I am able to:			
Collect data related to outcomes and methods	.289	.119	.882
Use assessment results to generate improvement actions	.427	.217	.818
Create measurable outcomes	.413	.168	.756

Note. Extraction Method: Principal Component Analysis; Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Table 11.

KMO and Bartlett’s Test for Perception of Department, College, and Institutional Effectiveness

Statistical Analysis		Results
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.698
Bartlett's Test of Sphericity:	Approximate Chi-Square	1190.405
	df	325
	Significance	.000

Note. df=Degrees of Freedom.

Table 12.

Communalities for Perception of Department, College, and Institutional Effectiveness

Communalities Please rate how each of the following statements is represented at the following levels:	Initial	Extraction
Institutional Effectiveness Office –		
Faculty are encouraged to participate in assessment activities	1.000	.734
Faculty are encouraged to align their courses with their outcomes	1.000	.687
Faculty are encouraged to conduct meaningful program improvement	1.000	.692
Institutional Effectiveness Office - Faculty are encouraged to participate in making long-term plans for their program	1.000	.859
It is easy for faculty to meet regularly to discuss assessment issues	1.000	.741
The assessment process is transparent	1.000	.662
Assessment expertise is readily available	1.000	.779
Adequate resources are provided for assessment training	1.000	.627
Department Level –		
Assessment is valued	1.000	.688
Faculty are encouraged to participate in assessment activities	1.000	.786
Faculty are encouraged to align their courses with their outcomes	1.000	.778
Faculty are encouraged to conduct meaningful program improvement	1.000	.851
Faculty are encouraged to participate in making long-term plans for their program	1.000	.713
It is easy for faculty to meet regularly to discuss assessment issues	1.000	.732
The assessment process is transparent	1.000	.729
Assessment expertise is readily available	1.000	.727
Adequate resources are provided for assessment training	1.000	.720
College Level –		
Assessment is valued	1.000	.668
Faculty are encouraged to participate in assessment activities	1.000	.853
Faculty are encouraged to align their courses with their outcomes	1.000	.908
Faculty are encouraged to conduct meaningful program improvement	1.000	.848
Faculty are encouraged to participate in making long-term plans for their program	1.000	.734
It is easy for faculty to meet regularly to discuss assessment issues	1.000	.829
The assessment process is transparent	1.000	.778
Assessment expertise is readily available	1.000	.846
Adequate resources are provided for assessment training	1.000	.921

Note. Extraction Method: Principal Component Analysis.

Table 13.

Total Variance Explained for Perception of Department, College, and Institutional Effectiveness

Component	Initial Eigenvalues			Extraction SS Loadings			Rotation SS Loadings		
	Total	% of Variance	Cum %	Total	% of Variance	Cum %	Total	% of Variance	Cum %
1	11.050	42.502	42.502	11.050	42.502	42.502	5.409	20.803	20.803
2	3.127	12.029	54.530	3.127	12.029	54.530	4.906	18.868	39.671
3	2.357	9.065	63.595	2.357	9.065	63.595	3.447	13.257	52.928
4	2.058	7.917	71.512	2.058	7.917	71.512	3.093	11.895	64.823
5	1.298	4.993	76.505	1.298	4.993	76.505	3.037	11.682	76.505
6	.935	3.597	80.102						
7	.909	3.496	83.598						
8	.754	2.899	86.497						
9	.539	2.072	88.569						
10	.492	1.891	90.460						
11	.439	1.688	92.148						
12	.326	1.253	93.401						
13	.281	1.079	94.481						
14	.236	.906	95.387						
15	.199	.764	96.151						
16	.188	.722	96.873						
17	.178	.685	97.558						
18	.149	.574	98.132						
19	.134	.514	98.646						
20	.113	.433	99.080						
21	.085	.328	99.407						
22	.053	.204	99.611						
23	.042	.161	99.772						
24	.027	.105	99.877						
25	.017	.065	99.942						
26	.015	.058	100.000						

Note. Extraction Method: Principal Component Analysis. SS=Sums of Squared. Cum=Cumulative

Table 14.

Rotated Component Matrix[a] for Perception of Department, College, and Institutional Effectiveness

Please rate how each of the following statements is represented at the following levels:	Component				
	1	2	3	4	5
College Level -					
Faculty are encouraged to conduct meaningful program improvement	.860	.200	.076	.043	.246
Faculty are encouraged to align their courses with their outcomes	.855	.308	.059	.036	.278
Faculty are encouraged to participate in assessment activities	.839	.314	.144	.000	.173
Adequate resources are provided for assessment training	.698	.211	.339	.520	-.064
Assessment expertise is readily available	.697	.117	.319	.475	-.137
Assessment is valued	.683	-.029	-.071	.328	.296
The assessment process is transparent	.668	.364	.344	.274	-.079
Faculty are encouraged to participate in making long-term plans for their program	.648	.099	.366	.121	.395
Department Level –					
Faculty are encouraged to conduct meaningful program improvement	.107	.905	.064	.025	.126
Faculty are encouraged to participate in assessment activities	.318	.819	.042	.097	.051
Faculty are encouraged to participate in making long-term plans for their program	.138	.796	.229	.091	.009
Faculty are encouraged to align their courses with their outcomes	.103	.773	-.078	.237	.328
The assessment process is transparent	.197	.712	.184	.386	.000
Assessment expertise is readily available	.334	.646	-.021	.446	-.005
Institutional Effectiveness Office -					
Assessment expertise is readily available	.160	-.057	.861	-.056	.084
Adequate resources are provided for assessment training	.078	.075	.784	.032	-.021
It is easy for faculty to meet regularly to discuss assessment issues	.128	.017	.728	.328	.296
The assessment process is transparent	.216	.317	.679	.134	.192
It is easy for faculty to meet regularly to discuss assessment issues	.130	.294	.127	.779	-.077
College Level -					
It is easy for faculty to meet regularly to discuss assessment issues	.552	.204	.307	.621	.058

Table 14 Continued

Department Level -					
Adequate resources are provided for assessment training	.308	.482	.100	.604	.130
Assessment is valued	.090	.536	-.099	.553	.278
Institutional Effectiveness Office -					
Faculty are encouraged to participate in assessment activities	.133	.039	.030	.062	.842
Faculty are encouraged to participate in making long-term plans for their program	.216	.197	.398	.066	.782
Faculty are encouraged to align their courses with their outcomes	.188	.160	.035	-.168	.773
Faculty are encouraged to conduct meaningful program improvement	.138	.054	.457	.380	.563

Note. Extraction Method: Principal Component Analysis.; Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 8 iterations.

Table 15.

Survey Reliability

Category	Reliability Cronbach's Alpha
Effectiveness and Utility	.943
Willingness	.948
Self-efficacy	.943
Perception - Department	.922
Perception - College	.945
Perception - Institutional Effectiveness	.830
Results Utility	.925

Table 16.

Pre-test and Post-test Items

	Total # of Items	Area of Focus (# of Items)
Module 1	5	Outcomes (4) Curriculum Mapping (1)
Module 2	8	Methods (6) Rubrics (2)
Module 3	7	Data Analysis (7)
Module 4	14	Improvement Actions (14)

Table 17.

Assessment Knowledge Quiz Pre- to Post-test Mean Change Scores

	n	M (SD)	*t*-test
Module 1	48	1.1 (0.85)	-9.1*
Module 2	48	1.7 (1.48)	-8.2*
Module 3	48	1.8 (1.23)	-10.4*
Module 4	45	0.7 (1.03)	-4.7*

**p*<0.00

Note. M = Mean. SD = Standard Deviation.

Table 18.

Participants' Perceived Assessment Self-Efficacy at Post-test

	df	t-test	p value
Statistical difference (increase)			
Creating measurable outcomes	105	-3.60	0.000
Creating a curriculum map	105	-3.17	0.002
Differentiating between direct and indirect measures	104	-6.05	0.000
Collecting data related to outcomes and methods	105	-2.49	0.014
Analyzing assessment results	105	-2.61	0.010
Using assessment results to generate improvement actions	105	-2.18	0.032
Documenting implementation and effectiveness of improvement actions	105	3.71	0.000
No statistical difference			
Create a mission statement for my department or program	105	-1.48	0.141
Create a rubric	105	-1.49	0.138
Assess student work using a rubric	105	-1.68	0.097

Note. *df* = Degrees of freedom.