ADVANCING TEACHING EVALUATION: INVESTIGATING THE ROLE OF READINESS AND LEADERSHIP IN DEPARTMENTAL CHANGE

by

HANNAH ERICSON

(Under the Direction of Tessa Andrews)

ABSTRACT

Teaching evaluation at many institutions is insufficient to effectively support, recognize, and reward high-quality teaching, necessitating systemic change. This dissertation examines the factors affecting teaching evaluation reform in STEM departments through three interrelated studies. The first study investigates the effects of an intervention designed to support department heads in advancing teaching evaluation practices. We examine the amount of change different departments achieved. We also investigate department head readiness for change and how it related to the reforms they made. The second study explores the ideas and actions department heads used to lead these changes. We applied Kotter's 8-step model for leading change as an analytical framework, and suggested modifications that can make it more applicable to the context of change in academic departments. Finally, the third study focuses on the development and validation of the Teaching Evaluation Readiness Assessment (TERA), a survey designed to measure faculty readiness for changing teaching evaluation practices. The TERA is capable of detecting differences in readiness across time and between departments, making it useful for offering insights for change agents and researchers. Collectively, these studies contribute to a deeper understanding of how department head readiness and leadership and faculty readiness impact efforts to reform teaching evaluation practices.

INDEX WORDS: Departments, department heads, readiness for change, teaching evaluation, leadership, faculty, organizational change

ADVANCING TEACHING EVALUATION: INVESTIGATING THE ROLE OF READINESS AND LEADERSHIP IN DEPARTMENTAL CHANGE

by

HANNAH ERICSON

B.S., University of Iowa, 2020

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2025

ADVANCING TEACHING EVALUATION: INVESTIGATING THE ROLE OF READINESS AND LEADERSHIP IN DEPARTMENTAL CHANGE

by

HANNAH ERICSON

Major Professor: Tessa Andrews

Committee: Erin Dolan
Julie Stanton
Gaelen Burke

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2025

ACKNOWLEDGEMENTS

This would not have been possible without the support, guidance, and encouragement of many individuals. I want to express my deepest gratitude to my advisor, Tessa Andrews, for their unwavering support, insightful feedback, and patience throughout this process. I would like to thank the members of my research group for all of their support and feedback over the years. Additionally, I would like to thank the members of the DeLTA project, as well as the members of BERG for taking the time to give feedback and suggestions on my work many times, which has been immensely helpful. I would like to thank my family and friends for their constant support, understanding, and encouragement. Finally, I want to thank Maya for being my writing buddy and keeping me sane throughout this whole process. Completing this dissertation has been a challenging but rewarding journey, and I am deeply grateful to everyone who has been a part of it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In the United States, there has long been concern about the number of students persisting in earning STEM degrees, with roughly 60% of students enrolled in STEM majors switching to non-STEM majors or dropping out of college before graduation in 2012 (Presidential Council of Advisors on Science and Technology, 2012). This number has risen recently, with over half of continuing-generation, White and Asian STEM students persisting through to STEM degrees. However, this is in comparison with just over 30% of first-generation, or Latinx students and 28% of Black students persisting through to obtain their degrees (Nolan et al., 2025). As a result, demands for workers in STEM fields will not be met, which has the potential to have major economic and societal impacts (Augustine, 2005). Students cite many reasons for leaving, including feeling discouraged due to low grades, loss of interest or motivation, competitive STEM culture, weed-out classes, poor course design, and notably, the poor quality of teaching in STEM courses (Hunter, 2019). In 2012, the President's Council of Advisors on Science and Technology put out a report that presented five recommendations to transform undergraduate STEM education, the first of which being to "catalyze widespread adoption of empirically validated teaching practices" (Presidential Council of Advisors on Science and Technology, 2012).

One such empirically validated practice is active learning. Active learning is an approach that engages students in the learning process, through activities like discussion, problem-solving, or reflection, rather than passive receipt of information, like listening to a lecture. The adoption of active learning in the classroom has been shown to be beneficial to students in multiple ways. The use of active learning increases student performance, with instructors seeing

increases in grades, and reductions in failure rates (Freeman et al., 2014). It also results in better conceptual understanding of course material (Knight & Wood, 2005; Freeman et al., 2007). Students develop more favorable attitudes towards learning (Springer et al., 1999), and engage more in the course (Armbruster et al., 2009). Students who participate in active learning in the classroom also show a greater persistence in STEM (Springer et al., 1999). While this is beneficial for all students, the effects of active learning are particularly influential for students from underrepresented minority groups (Haak et al., 2011, Theobald et al., 2020).

Even with these known benefits, active learning is not the most prevalent style of teaching in STEM classrooms (Stains et al., 2018). STEM instructors have cited a variety of barriers to implementing it, including lack of time, training, and incentives to do so, instructional challenges such as not being able to cover all necessary course content, and class sizes (Brownell & Tanner, 2012; Andrews & Lemons, 2015; Shadle et al., 2017). Most instructors are not pedagogy experts, and lack the expertise necessary to effectively implement active learning in their classrooms (Andrews et al., 2011). Additionally, instructors often face institutional barriers. One such barrier is the lack of weight placed on teaching effectiveness for processes like annual evaluation and promotion and tenure (Walczyk et al., 2007). Another is the conflict between teaching and research that many faculty face. Research is more highly valued than teaching, and faculty are more highly rewarded for their research activity than for their teaching, further disincentivizing them from devoting time to improving their teaching (Brownell & Tanner, 2012; Lester & Kezar, 2012). If improvements are to be made in how STEM courses are being taught, there needs to be a shift in how teaching is viewed.

One way to put more emphasis on teaching is through how it is evaluated. In order to reward good teaching, teaching evaluation systems must be able to recognize it. Unfortunately, teaching evaluation systems at most institutions are unable to do so (e.g. Brickman et al., 2016; Dennin et al., 2017). Most institutions rely mainly on student end-of-course surveys as their only source of evidence about an instructor's teaching (Keig, 2000; Brickman et al., 2016). This is

problematic as these surveys are known to be biased against instructors based on their social identities (e.g., race, gender, native language), which is confounded by issues related to class type (Cashin, 1990; Ramsden, 1991; Greenwald & Gillmore, 1997; Bedard & Kuhn, 2008; Boring, 2017; Fan et al., 2019; Esarey & Valdes, 2020; Aragón et al., 2023). These surveys do not correlate with student learning outcomes, and do not provide constructive feedback for instructors (Bouwma-Gearhart & Hora, 2016; Brickman et al., 2016). More holistic evaluation practices are needed to improve the quality of data that is being collected about an instructor's teaching. This can include relying on more sources of evidence, such as information from peer observations and an instructor's own systematic self-reflections (Finkelstein et al., 2020; Weaver et al., 2020). Chapter 2 of this manuscript goes into more detail on what robust and equitable teaching evaluation practices look like.

To create better experiences for students, teaching evaluation systems need to shift. This dissertation seeks to explore some of the factors affecting teaching evaluation change. Chapters 2 and 3 focus on department heads, and their role in the change process. Department heads are important figures within their departments. They are heavily involved in teaching evaluation and processes such as annual review and promotion and tenure. Heads are also the ones that can prioritize certain changes over others. For example, they are responsible for setting agendas within the department and creating and charging departmental committees. For a change to be successful, support from department heads is important. Understanding how heads interact with teaching evaluation change is therefore important for implementing that change.

Chapter 2 examines the effects of an institutional change project that was aimed at advancing departmental teaching evaluation practices. It investigates the intervention's impact on teaching evaluation practices in participating departments. It also examines department head readiness for change, and how it related to the reforms they achieved while a part of the intervention. This work allows us to examine what factors support heads in implementing more

robust and equitable evaluation practices. Chapter 3 examines the ideas and actions department heads used to enact changes to teaching evaluation. Little scholarly work has been done to understand what leadership strategies are effective for department heads leading change in their units. This work uses a popular model for leading change as an organizing framework to understand how department heads enacted change.

Faculty are also important in leading change in their units. They often are heavily involved in the decision-making process, due to the culture of shared governance in academia. Therefore, it is important to understand faculty readiness for changing teaching evaluation practices. Chapter 4 describes the development of an instrument to measure faculty readiness for changing teaching evaluation. The instrument aims to provide both change agents and researchers with useful information that can be used to aid change initiatives.

CHAPTER 2

ARE DEPARTMENT HEADS READY FOR CHANGE? LEVERAGING A LEADERSHIP ACTION TEAM TO ADVANCE TEACHING EVALUATION PRACTICES[1]

[1] Ericson, H. C., Lemons, P. P., Dolan, E. L., Brickman, P., Krishnan, S., & Andrews, T. C. 2025. *CBE-Life Sciences Education.* 24 (1).

ABSTRACT

Teaching evaluation at many institutions is insufficient to support, recognize, and reward effective teaching. We developed a long-term intervention to support STEM department heads in advancing teaching evaluation practices. We describe the intervention and systematically investigate its impact on departmental practices within a research-intensive university. The outcomes varied considerably by department, with four departments achieving extensive teaching evaluation reform and seven departments achieving more limited reform. We used qualitative content analysis of interviews and meetings to investigate department head readiness for change and how it related to the reforms they achieved. All department heads perceived inadequacies in their current evaluation practices, but this dissatisfaction did not reliably predict the changes they pursued. Heads only pursued changes that they perceived to have clear benefits. All heads worried that faculty might resist new practices, but heads who were most successful in facilitating change saw ways to work around resistance. Heads who led the most change questioned their own expertise for reforming teaching evaluation and delegated the work of developing new evaluation practices to knowledgeable colleagues. We discuss emergent hypotheses about factors that support heads in challenging the status quo with more robust and equitable evaluation practices.

AUTHOR CONTRIBUTIONS

As the first author of this manuscript, I was the primary contributor to the completion of this work. Sandhya Krishnan and Tessa Andrews designed the protocol for the first round of interviews, and Sandhya conducted them. Tessa and I designed the protocol for the second round of interviews, and I conducted them. I led the development of the qualitative codebook, and conducted all analyses presented in this chapter in collaboration with Tessa. I wrote the manuscript (except for the sections entitled “Social cognition perspective guided LAT design,” “Cultural perspective guided LAT design,” and “LAT Design”, which were written by Tessa). I

created all the graphs and figures present in this chapter. Tessa provided iterative feedback on the manuscript. Paula Lemons, Erin Dolan, and Peggy Brickman ran the LAT, as described in the chapter. All co-authors provided feedback on the manuscript prior to submission and on revisions following peer review. Every co-author agreed that this work may be presented in this dissertation.

INTRODUCTION

Reward systems at many higher education institutions inadequately support, recognize, and incentivize effective teaching and teaching improvement (e.g., Brickman et al., 2016; Dennin et al., 2017). Research-intensive institutions often make reward decisions (e.g., promotion, tenure, raises, titled positions) based primarily on an individual's research contributions, missing an opportunity to incentivize faculty investment in improving teaching (Bradforth et al., 2015). Even if an institution or department wants to give meaningful weight to teaching in reward decisions, the evidence they collect to evaluate teaching is often limited and unreliable (Bradforth et al., 2015; Dennin et al., 2017).

Teaching evaluation in most institutions of higher education fails to support teaching improvement or provide robust evidence of teaching quality. Student evaluations are the most common form of teaching evaluation (e.g., Brickman et al., 2016), and student ratings show bias based on instructor's social identities (e.g., race, gender, native language) and may not correlate with student learning outcomes. (e.g., Bedard & Kuhn, 2008; Boring, 2017; Fan et al., 2019; Esarey & Valdes, 2020; Aragón et al., 2023). Additionally, instructors feel dissatisfied with student evaluations because they only measure satisfaction, often suffer from low response rates, and fall short of providing constructive and insightful feedback that instructors can use to improve their teaching (e.g., Bouwma-Gearhart & Hora, 2016; Brickman et al., 2016). Peer evaluation is another form of teaching evaluation available to some faculty, but these processes typically lack structure or trained observers and, as a result, tend to provide feedback on superficial aspects of teaching rather than the substantive feedback instructors desire (e.g., Bouwma-Gearhart & Hora, 2016; Brickman et al., 2016). Not surprisingly then, instructors often do not perceive information from teaching evaluation as meaningful and may not rely on these data to inform their teaching (e.g., Blaich & Wise, 2010; Bouwma-Gearhart & Hora, 2016). If institutions aim to support, recognize, and incentivize high-quality teaching, teaching evaluation practices must be reconsidered.

Multiple projects have responded to the need to advance teaching evaluation practices, often in the context of science, technology, engineering, and mathematics (STEM) departments. At a national scale, the Association of American Universities and Cottrell Scholar Collaborative have gathered institutional leaders and change agents to propose ways that teaching could be better recognized and rewarded in research-intensive universities (e.g., Bradford et al., 2015; Dennin et al., 2017). At the institutional level, the TEval project has developed new approaches to teaching evaluation and supported their implementation at four institutions (e.g., Finkelstein et al., 2020; Weaver et al., 2020). Key outcomes of TEval include a common framework and rubric that define dimensions of effective teaching and articulate criteria for each dimension in a rubric, as well as tools for collecting evidence of teaching effectiveness from multiple perspectives. Researchers at Boise State University also created a framework to define effective teaching and evaluate teaching formatively and summatively (Simonson et al., 2022). Outside of the peer-reviewed literature, numerous departments and institutions have developed resources for teaching evaluation (see a list in supplemental materials from Krishnan et al., 2022). These efforts have contributed to national conversations about the importance of shifting teaching evaluation practices, produced resources, and demonstrated that change to teaching evaluation practices is feasible in research-intensive institutions and STEM departments.

An important next step for the work of transforming teaching evaluation is developing, testing, and reporting the effectiveness of different models of facilitating change. Currently, evidence of the impact of such efforts remains scarce. Here we begin to address this gap by presenting the design, implementation, and evaluation of a novel Leadership Action Team (LAT) intervention at one research-intensive university. The LAT is a facilitated working group of STEM department heads developing and piloting new teaching evaluation practices in their departments over 2-3 years.

In this paper, we start by describing the changes to teaching evaluation practices that we aimed to achieve with the LAT intervention, its theoretical underpinnings, and how the LAT

worked. Then we present evidence about changes to teaching evaluation practices in participating departments and how the change achieved related to readiness for change among department heads.

Desired Change: Robust and Equitable Teaching Evaluation

Instead of relying solely on student end-of-course evaluations, a more robust and equitable approach to teaching evaluation involves collecting evidence from multiple sources (Andrews et al., 2020; Weaver et al., 2020; Krishnan et al., 2022). Considerable work has pointed toward more holistic ways of evaluating teaching (e.g., Glassic et al., 1997; Smith, 2008; Lyde et al., 2016; Dennin et al., 2018; Weaver et al., 2020). The three-voice framework for teaching evaluation stems from this work, and uses evidence from students, trained peers, and the instructor to create a holistic picture of teaching effectiveness and improvement (Reinholz et al., 2018). The *student voice* can include end-of-course evaluations, as well as mid-semester evaluations, assessments of student outcomes, and student interviews (Weaver et al., 2020). Data about student learning, such as from comparisons of pre- and post-assessments, can be an especially powerful source of evidence. Notably, student evaluations do not necessarily correlate positively with student learning and therefore cannot be assumed to serve as evidence of learning (e.g., Bedard & Kuhn, 2008; Boring, 2017; Fan et al., 2019; Esarey & Valdes, 2020; Aragón et al., 2023) Nonetheless, students, as the intended beneficiaries of teaching, can offer important perceptions of their experiences in the course (Krishnan et al., 2022). The *peer voice* collects evidence about an instructor's teaching from other instructors, usually through teaching observations or review of teaching materials (Weaver et al., 2020). Peers have relevant expertise in the discipline and teaching experience, situating them to provide more constructive feedback than students (Thomas et al., 2014). The *self voice* involves an instructor reflecting on their teaching and teaching improvement (Weaver et al., 2020). The instructor knows best their goals for the course and changes they have made and can synthesize and contextualize

evidence from the student and peer voices to characterize teaching accomplishments, strengths, and opportunities for improvement (Krishnan et al., 2022). Using three voices provides more holistic evidence because it provides evidence from important and different perspectives and mitigates the potential bias present in any single perspective.

Research-based teaching evaluation practices for each voice are structured, reliable, and longitudinal (Krishnan et al., 2022). *Structured* evaluation involves formalizing processes and expectations, such as using standard forms and consistently enacting processes across faculty (Table 2.1). Adding structure standardizes the experiences of faculty and makes processes and expectations transparent, both of which can result in more equitable experiences for faculty. *Reliable* evaluation draws from multiple sources of evidence and appropriately analyzes and interprets this evidence, making the findings more trustworthy (Table 2.1). Considering reliability is critical to mitigating potential biases and engendering trust in the conclusions drawn from evidence of teaching effectiveness. *Longitudinal* evaluation is able to document change over time (Table 2.1). These practices make it possible to value improvement, not just excellence in teaching.

Table 2.1: Research-based target practices for peer voice. These align with three characteristics of robust and equitable evaluation (Krishnan et al., 2022). See the target practices for student and self voice in Krishnan et al. (2022) or Appendix C.

Structured	Department uses a formal observation form to guide what is observed and which other data are being collected (e.g., class materials, assessments, pre-observation meeting).
	Department has a formal template for writing a report based on peer review, potentially distinguishing between formative and summative review.
	Department uses formal processes or criteria to select peer observer(s) for all instructors.
	Department enacts policy about the number of peer observations & observers during a review period and/or across review periods.
	Department designates a coordinator, leader, or committee to carry out and refine peer observation practices.
	Department has a process for allocating and recognizing workload related to coordinating and conducting observations.

	Department periodically discusses and improves peer evaluation practices to maximize utility to instructors and the department.
	Department provides or arranges formal training about the departmental peer review process for peer observers.
Reliable	Department relies on multiple observations for all instructors, such as using multiple observers, observing multiple lessons, and/or observing multiple courses.
	Department specifies which class materials (e.g., syllabi, exams, homework, slides, handouts) are collected and evaluated as part of peer observation.
	Department expects observers to talk with instructors to properly contextualize observations and review materials. This might include discussing course goals, lesson goals, class structure, and students.
Longitudinal	Department conducts peer observation over multiple time points in a review period for all instructors to document teaching improvements.
	Department ensures that the peer observation process provides feedback to instructors via follow-up discussion that covers strengths and areas for improvement.

Guiding Theory Related to Change

Second-order change is needed to achieve the long-term goal of STEM departments enacting robust and equitable teaching evaluation. Second-order change occurs when organizations question and then alter their operating systems, underlying values, and culture (Argyris & Schon, 1996; Kezar, 2018). In contrast, first-order change modifies existing practices without shifting the status quo and is therefore easier to achieve but less likely to have lasting effects. For example, altering the questions asked in student end-of-course evaluations is likely to be a first-order change to teaching evaluation practices. This change could produce less biased ratings from students (e.g., Peterson et al., 2019), but without changing the approach to evaluation as a whole, this change may not impact how these less-biased ratings can inform instructor evaluations. In contrast, adopting a three-voice approach with research-based practices for each voice would be a second-order change in many STEM departments because it would change the operating systems.

In order to move beyond implicit assumptions about how change occurs, we used theory to ground the design of the LAT intervention and research (Connolly & Seymour, 2015; Reinholz et al., 2021). Achieving second-order change in STEM higher education requires flexibly drawing on multiple change perspectives (Corbo et al., 2016; Kezar, 2018). Two theoretical perspectives on change are best suited for achieving second-order change: social cognition and cultural perspectives. These change perspectives served as anchors in our design and implementation.

Social cognition perspective guided LAT design

A social cognition perspective defines change as the development of new ways of thinking among individuals (i.e., learning; Kezar, 2018). Therefore, facilitating change requires surfacing underlying ideas, providing feedback and new information, challenging prior beliefs, and pointing out leaps in logic (Kezar, 2018). Developing new ways of thinking may involve the use of new language and concepts, such as the contrast between formative and summative evaluation (Eckel & Kezar, 2003). It may also involve attaching new meaning to familiar concepts, such as viewing teaching evaluation as involving three distinct voices, rather than equating teaching evaluation with student evaluations (Eckel & Kezar, 2003). Guided by this perspective, the LAT was designed to engage department heads in repeated opportunities for reflection on current thinking and learning about robust and equitable teaching evaluation.

Cultural perspective guided LAT design

A cultural perspective defines change as a gradual shifting of values, beliefs, and underlying assumptions within an organization (e.g., department). Underlying assumptions and values in an organization tend to be implicit, rarely challenged, and hard for members to articulate (Schein, 1999; Kezar, 2018), yet they influence an organization's functioning both implicitly and explicitly. Cultural change tends to be slow, involves many members of an

organization, and requires long-term intervention (Kezar, 2018). Change agents create opportunities to make underlying assumptions visible to members of the organization, so they can be reconsidered. Change agents can also promote scholarly engagement, shared decision-making, and rationality around the change initiative (Bergquist & Pawlak, 2007; Kezar, 2018). Guided by this perspective, the LAT was designed to engage department heads in reflecting on underlying values, beliefs, and assumptions that they held and that undergirded departmental teaching evaluation practices.

Readiness for change guided research

We drew on a specific change theory, the readiness for change framework, to ground our research into what distinguished department heads who achieved more change. Readiness for change is the extent to which an individual has positive beliefs and attitudes about (a) the need for a given change in their organization and (b) the ability of the organization to achieve the change (Armenakis et al., 1993). The readiness for change framework was originally developed by organizational management researchers to explain why so many organizational change efforts failed (e.g., Armenakis et al., 1993). The framework has also been used as a tool to help leaders more successfully engage employees in organizational change (e.g., Armenakis & Harris, 2002; Holt et al., 2007). An individual's or group's readiness reflects their thoughts and is considered a precursor to engaging productively in a change process (Armenakis et al., 1993). Therefore, leaders hoping to lessen resistance to change can work to increase readiness for change (Holt et al., 2007). This framework relates to both the social cognition and cultural perspectives on change in that it centers individuals' thoughts and learning about change (i.e., social cognition) and is influenced by underlying values, beliefs, and assumptions about the organization (i.e., cultural).

Five components contribute to an individual's readiness for change (Rafferty et al., 2013). In this study, we examined readiness for change among department heads, and we

describe each component in that context. *Discrepancy* is the head's belief that changing teaching evaluation in the department is necessary (Armenakis & Harris, 2002) Discrepancy can result from comparing the current status of departmental practices with some goal or value. *Appropriateness* is the head's sense that the solution under consideration (e.g., 3-voice framework for teaching evaluation) will address the identified need for change. Next, *efficacy* is a department head's assessment of their individual and the department's capacity to effectively advance teaching evaluation practices. *Valence* is the head's perception of the costs and benefits of reformed teaching evaluation practices, and could focus on themselves, the department, or faculty (Holt et al., 2007). Finally, *principal support* is the head's belief that the university will provide necessary support for advancing teaching evaluation (Rafferty et al., 2013). We used the readiness for change framework to guide our investigation of how heads who achieved more change in their department differed from those who achieved less change.

Department Heads as the Focus

We designed the LAT to engage department heads for several reasons. We chose to target change in departments because these are the organizational units that most impact hiring, mentoring, evaluating, and rewarding faculty in research-intensive institutions. Departments are also often the organizational unit that enacts practices for peer and self voice and that rely on data from student evaluations to make decisions. Furthermore, departments have unique histories, cultures, and practices, even within the same institution, and may opt to pursue different changes and take different paths toward a change (Bouwma-Gearhart et al., 2016).

We choose to prioritize department heads because of their roles and power. At our institution, department heads serve an indefinite number of three year appointments and are appointed by the dean. However, they often serve no more than two or three terms. Department heads are central to teaching evaluation and reward decisions in our institution. They conduct

annual evaluations of faculty and lead evaluations for promotion and tenure. Heads also have the power to set agendas within the department, populate departmental committees, and orient new faculty toward departmental priorities. As a result, they can prioritize particular changes to departmental practices and steer the department away from other changes. Therefore, changing thinking among department heads (i.e, social cognition perspective) has the potential to impact faculty thinking and departmental direction. Cultural change requires more than changing individual thinking, and department heads can create conditions favorable to cultural change. Department heads have the positioning and power to elevate and concentrate on particular values and priorities, to advocate for and direct resources toward changing existing structures, and to navigate the complexity of individual differences within the department, each of which are important components of culture and culture change (Reinholz & Apkarian, 2018).

LAT Design

We convened STEM department heads in a LAT with the goal of providing support to learn about and enact more robust and equitable teaching evaluation in their departments. We assumed such support would be needed because many individuals who become department heads have not served in similar positions previously and feel unprepared to step into the role (Wolverton et al., 2007). In addition, department heads often report concerns about a lack of adequate resources within the department (Cipriano & Riccardi, 2017), and difficulty balancing their various responsibilities (Gmelch et al., 2017).

In alignment with our guiding theory, we designed the LAT to create space for learning and reflection about robust and equitable teaching evaluation, to prompt critical examination of assumptions that underlie existing departmental practices, and to scaffold and provide differentiated support for learning and action to develop new departmental practices (Table 2). We engaged department heads in considering the three-voice framework and curated teaching evaluation tools and discussing with each other the changes they wanted to pursue and the

challenges they faced. We used directive facilitation to keep conversation on track and to challenge ideas that acted as barriers. We also used meeting time to accomplish work, knowing that department heads had very limited time outside of meetings to make progress.

The LAT met five times per year for three years for 1-hour facilitated sessions. A lead facilitator (P.P.L) started each meeting with a brief presentation (<10 minutes) in order to frame the work of the meeting and then engaged department heads in small-group and whole-group discussion (Table 2.2). Additional team members (E.L.D, P.B, and two other faculty) facilitated small-group discussion and contributed to the whole-group discussion. In addition to LAT meetings, we provided individualized support based on the interests and desires of each head. This included 1-on-1 meetings, small working sessions, and joining faculty meetings to present and/or facilitate discussions.

Department heads willingly participated in the LAT. We recruited by meeting with heads to explain the project goals and inquiring about their interest. Every head that we asked agreed to participate and attended most meetings. We did not provide personal incentive, nor did heads receive institutional resources or goodwill as a result of their participation. The external credibility of National Science Foundation funding and institution-level celebration of the grant may have encouraged participation, but upper administrators, to our knowledge, did not directly encourage departmental participation. One head declined to participate in one element of the research, but otherwise heads consented to all LAT activities and research.

Table 2.2: Typical LAT meeting structure, including activities and their intended purposes.

Length (min.)	Activity	Purposes
5-10	Information sharing from lead facilitator	● Frame the work of the LAT ● Frame the focus of the meeting ● Introduce new information and resources ● Provide instructions for small group discussion
30-45	Small group discussion with 2-4 department heads and facilitators.	● Prompt reflection on current thinking and current department practices

	Individual thinking time about prompting questions or example materials, followed by discussion. Facilitators kept discussions on topic, asked probing questions, and moderated participation to quiet dominators and invite listeners to share.	● Provide a chance to critically consider teaching evaluation materials and practices and culture in other departments ● Create conditions to recognize underlying assumptions ● Troubleshoot barriers and challenges with peers
5-15	Share out from small group discussion in larger group. Share out led by facilitator and/or department head from each group. Brief period of questions and answers.	● Provide a chance to hear ideas from other department heads ● Amplify ideas from small groups to highlight progress, point toward next steps, or problematize what is needed to make progress
5	Next steps offered by the lead facilitator and/or participating department heads asked to commit to one small next step to accomplish before the next meeting.	● Provide accountability for making progress ● Provide concrete ideas for next steps ● Frame the upcoming work of the LAT

Relevant Context

Institutions function as important contexts for departmental reform because institutions set expectations and priorities that departments must respond to in order to secure resources. The LAT was designed, enacted, and investigated within the context of a large, public land-grant and research-intensive university. Though research is a key priority at UGA, the institution also prides itself on providing quality undergraduate education. The project both benefited from and contributed to interest in teaching evaluation reform among upper administrators. Prior to our intervention, the President charged a Task Force on Student Learning and Success that led to 12 recommendations to further enhance the education experiences for undergraduates inside and outside the classroom. One of the 12 recommendations made by the group was to "Strengthen systems to document and promote effective teaching" (Task Force on Student Learning and Success, 2017). Based on this recommendation, the Office of Instruction convened a committee to consider changes to teaching evaluation. A member of our team served on the task force and then on the more focused committee. When the committee work stalled due to a few members' concerns about faculty workload, our team stepped in to

contribute to a new teaching evaluation policy and help shepherd it through faculty governance structures. The policy formally went into effect just before the end of the 3-year intervention reported here. This series of events demonstrates interest in teaching evaluation reform among some upper administrators. Andrews et al. (2021) describes in more detail the university-level work that our team undertook alongside this department-level intervention.

The LAT also occurred within a larger, National Science Foundation (NSF)-funded project. Four life sciences faculty (P.P.L, P.B., E.L.D, T.C.A) led the project, each of whom is a celebrated teacher and discipline-based education researcher. We used a shared leadership approach, in which each leader spearheaded one part of the project, in alignment with their strengths and spheres of influence, and the team worked together closely to plan and enact all project components. P.P.L led the LAT and the larger project, E.L.D and P.B co-facilitated LAT meetings and led other aspects of the larger project, and T.C.A led project research. We secured support for the project from department heads and upper administrators as part of our proposal to NSF. When we received the award, we used formal communication channels within the institution to promote the project. As the project proceeded, we annually reported on successes to a wide range of upper administrators. The positive perception of the project among administrators may have contributed to some department heads' willingness to stay involved, though not, we suspect, to the change they actually achieved. The larger project also included efforts to support the uptake of evidence-based teaching that engaged over 65 faculty. This raised awareness of the project within and beyond STEM departments. Additionally, some heads may have perceived that their involvement in the LAT supported the involvement of their faculty, which may have encouraged their continued engagement.

METHODS

This section and the subsequent results address two questions by examining departmental practices and thinking among department heads:

RQ1: To what extent did participating departments change their teaching evaluation practices to align with research-based practices?

RQ2: How did department head readiness for change relate to change achieved in departmental practices?

Participants

We studied 11 STEM departments, including a range of disciplines (e.g., life sciences, physical sciences, engineering) and the heads that served during the 3-year duration of the study (n = 16). This work was determined to be exempt by the University of Georgia Institutional Review Board (PROJECT00009085).

Collecting Data About Teaching Evaluation Practices (RQ1)

We gathered data about departmental teaching evaluation practices using interviews with department heads and other faculty. We conducted initial interviews at the start of the project and a second round of interviews three years later. Each interview used a semi-structured protocol that asked department heads to describe how their department evaluated teaching (see Appendix A). The initial interviews asked more general questions because department heads were not yet familiar with the details of robust and equitable teaching evaluation. Questions included “Can you please talk me through how teaching effectiveness is evaluated for promotion and tenure?” and “How is the evaluation of teaching effectiveness different for annual review than for promotion & tenure?” The interviewer specifically prompted participants to address whether and how each voice was used. After several years in the LAT, department heads were much more familiar with the range of research-based teaching evaluation practices, and we were able to ask more specific questions about the nuances of their departmental practices. The second interview included both general questions, such as

“Can you walk me through how peer voice works in your department?”, and more specific questions, such as “How do observers provide feedback to instructors who are observed?”

In addition to interviews with department heads, we relied on two additional data sources. First, recordings of LAT meetings provided an additional source of information about departmental practices because several meetings invited each department head to share their progress on changing practices. Occasionally, what a department head said in an LAT meeting and what they reported in an interview differed. In those cases, we sought an additional perspective on the current departmental practices. Additionally, two department heads were not available for interviews three years into the project, so we relied on interviews of other faculty likely to be knowledgeable about teaching evaluation practices, such as associate department heads or someone involved in departmental teaching evaluation reforms. These additional sources ensured we had comprehensive and accurate information about practices for each department.

Analyzing Teaching Evaluation Practices (RQ1)

We post-hoc assigned a score to each department’s use of robust and equitable teaching evaluation practices using the Guides to Advance Teaching Evaluation (GATEs) (Krishnan et al., 2022). The GATEs includes lists of research-based target practices for each of the three voices (see examples in Table 2.1). The target practices in the GATEs are comprehensive and aspirational, so we did not expect most departments to achieve all target practices for each voice. We made minor modifications to the GATEs lists so we could reliably judge the presence and absence of each target practice for each voice (see Appendix B). For example, we divided one target practice into two because these practices are distinct and needed to be evaluated separately. The divided practices were “Department recognizes known biases, such as bias against women, marginalized groups, and large class size,” and “Department limits comparisons of mandatory student evaluations between instructors.”

We characterized teaching evaluation practices in two steps, analyzing each voice separately. First, we characterized whether a department used a voice to evaluate teaching or did not use that voice. A department could use a voice (e.g., peer observation) as part of teaching evaluation and not meet any of the target practices. Therefore, the second step involved determining which research-based target practices a department had in place, for those departments using a given voice. Two researchers made these characterizations independently, and we discussed any disagreements to consensus. We achieved high interrater reliability using a weighted Cohen's kappa for each step of analysis (weighted Cohen's kappa = 0.82 and 0.75 respectively).

Collecting Data About Readiness for Change (RQ2)

We used two sources of data to identify and characterize readiness for change among department heads: recorded LAT meetings and semi-structured interviews. Each data source had affordances. First, we audio-recorded LAT meetings and created verbatim transcripts. These data captured authentic and impromptu interactions among department heads and facilitators. In total, we analyzed transcripts spanning the first seven LAT meetings, which occurred in the first 1.5 years of the intervention. We selected the earlier meetings to analyze for evidence of readiness for change because the latter six meetings focused more narrowly on discussing specific resources for teaching evaluation and provided scant information about readiness for change.

Second, we interviewed LAT members at the end of the 3-year span reported in this paper. Interviews allowed us to collect targeted, comprehensive information from each participating LAT member. Some heads talked in LAT meetings more than others, resulting in a range of the quantity of data from LAT meetings alone. In addition to asking about department teaching evaluation practices (see section called Collecting data about teaching evaluation practices), these interviews aimed to elicit department heads' readiness for advancing teaching

evaluation. We designed the protocol based on the readiness for change framework, asking questions targeted at each component (see Appendix A). For example, a question focused on the dimension of *valence* asked, “How would changes to teaching evaluation benefit your department, or faculty in your department?” A question about efficacy asked, “How comfortable do you feel leading changes to teaching evaluation in your unit?” Interviews were transcribed verbatim for analysis.

Qualitative Analyses of Readiness for Change (RQ2)

We conducted qualitative content analysis to address our second research question. We began by conducting provisional coding of LAT meeting transcripts, which is a deductive process using codes based on an established framework (i.e., readiness for change) (Saldaña, 2013). We developed and refined codes to capture the variation in our data for each readiness for change component. During this process, we determined that the data we had elicited about *appropriateness* did not capture the full range of ideas that heads likely held. Within meetings and interviews, heads sometimes commented positively about the fit between a voice or practice and their department, but only rarely commented about a lack of fit with their department. We were concerned that they had refrained from sharing more negative views of fit, meaning that our data could not accurately represent appropriateness. Therefore, we excluded data about appropriateness from further analyses. After we refined the codebook for the remaining components of readiness for change, we re-coded all previously coded data, first coding independently, then discussing all disagreements to consensus. H.C.E, T.C.A, and an undergraduate researcher completed all coding.

We next analyzed transcripts from the interviews with LAT members. We started with the codebook created for the LAT meeting recordings and refined it to better align with data in interviews. The interviews provided more comprehensive information, allowing us to add new codes and refine existing codes. We again worked iteratively and collaboratively, with two

coders working independently and then discussing to consensus. In order to draw on both the meeting and interview data for further analysis, we re-coded all meeting recordings using the final codebook created with the interview data (see Appendix C).

Readiness for change profile development and analysis (RQ2)

We constructed profiles for each department head to organize and synthesize qualitative data and enable comparisons. This approach has a similar purpose as creating a case record in case study research, as it aims to record the data comprehensively and is a manner that is manageable (e.g., Patton, 1990; Yin, 2014). Each profile summarized the data related to the included components of readiness for change from LAT meeting recordings and one-on-one interviews. For each LAT member, one researcher read and briefly summarized every coded data segment across both types of data, and then summarized all data related to a single readiness component in a brief paragraph. This process resulted in a 4-10 page profile for each department head. We then considered each component of readiness for change separately. Two researchers read the entirety of evidence for each department head and noted trends in the data. We specifically looked for similarities and differences between the department heads who achieved more and less change, and also for relationships between readiness dimensions and the changes that a department head achieved. As we noticed an emergent pattern, we revisited all relevant data to scrutinize whether the evidence supported or refuted the pattern. The process was iterative and highly collaborative. After agreeing on emergent patterns, the researchers returned to consider each profile as a whole, as an additional opportunity to identify patterns and test conclusions against the data. The outcome of these analyses are patterns in how readiness for change related to the teaching evaluation changes achieved, by component.

Trustworthiness

Trustworthiness in a qualitative analysis encompasses four components: credibility, transferability, dependability, and confirmability (Anfara et al., 2002; Shenton, 2004). In addressing both research questions, we used well-established research methods, and used several different methods of collecting data from our participants to obtain a more cohesive picture of their thoughts and ideas (triangulation) to help ensure the credibility of our work. In order to support the potential for transferability, we present details about the context of our study and the LAT and later discuss potential implications of our context. This allows an understanding of our unique situation to contextualize our results. We took several steps to improve the dependability and confirmability of our work. First, we provided a detailed description of our methods to enable potential replication. Second, we engaged in constant comparison to help mitigate the biases of any one researcher. We completed analyses in two phases, with researchers first working independently and then discussing to reach consensus. We also protected against inconsistency over time in our coding by repeatedly comparing quotes within a code to each other. Third, we intentionally masked the names and departments of each department head while coding, relying instead on randomized pseudonyms in each transcript, in order to lessen the impact of any impressions we had formed through our work with the LAT.

RESULTS

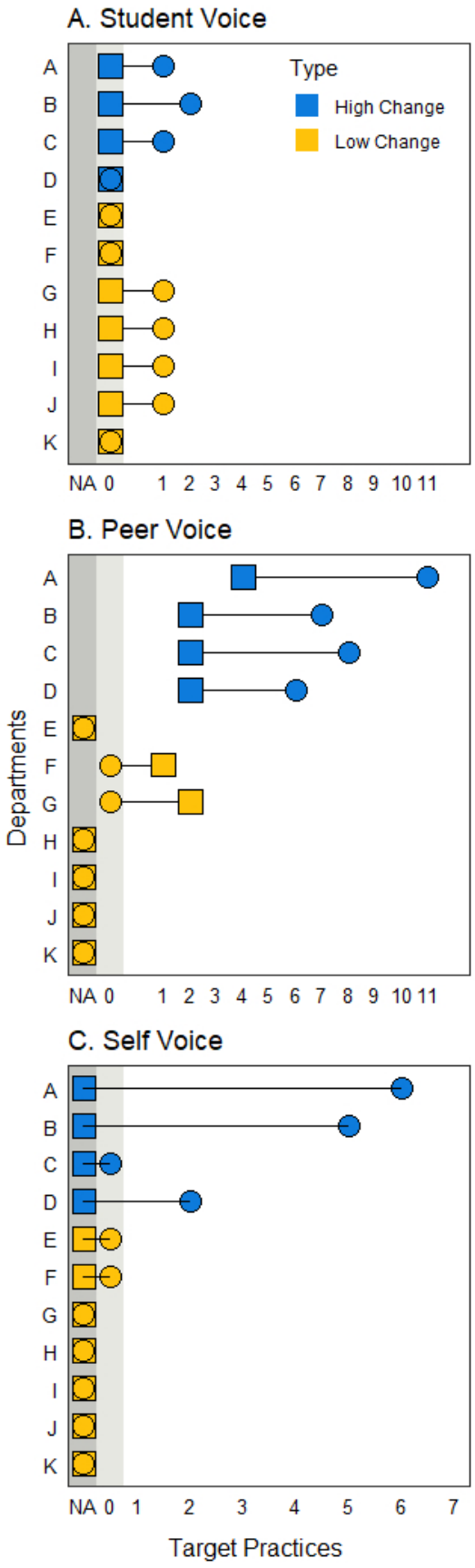
Our systematic investigation of the impact of the LAT on departmental teaching evaluation practices revealed considerable variation. Here we present these findings and the results of our investigation of how department head readiness for change related to the outcomes achieved in a department.

RQ1: To what extent did participating departments change their teaching evaluation practices to align with research-based practices?

Four departments (A, B, C, and D) achieved extensive changes to their teaching evaluation practices (Figure 2.1). These departments made progress on multiple voices and enacted a range of research-based target practices (Table 2.1). They are hereafter referred to as "High Change" departments. The remaining seven departments (E, F, G, H, I, J, and K) achieved more limited change to teaching evaluation practices, often without the adoption of target practices. In the first few months of the LAT, we asked three of the participating departments (A, C, & F) to serve as "pilot" departments for adopting new teaching evaluation practices. We hoped that work in pilot departments could chart a path forward for other departments and that the heads of pilot departments would feel accountable for making changes. Two of these are categorized as High Change and one is not (Figure 2.1). The willingness to serve as a pilot department could be indicative of the readiness the department heads had for changing teaching evaluation practices in their units, but it was not a perfect predictor. Departments advanced the three voices differently. Departments that made progress with peer observation had used this voice prior to their involvement in the LAT (Figure 2.1). Each had a few target practices in place at the start and then tripled or quadrupled the number of target practices. In contrast, no departments used self voice at the start of the project. Of those that added self voice, half enacted target practices (Figure 2.1). Most commonly, departments began providing support for instructors engaging in self-reflection, such as sharing examples of self-reflections or providing a guide for engaging in the reflection process. All departments used student voice at the beginning of the project, because the institution required end-of-course student evaluations, but none had target practices in place. Seven departments advanced this voice, typically by adding a single target practice (Figure 2.1). Departments most commonly began expecting faculty to take steps to achieve higher response rates on their end-

of-course evaluations. Next, we describe the changes that took place in one High Change and one Low Change department in more detail.

Figure 2.1: Use of research-based teaching evaluation practices in STEM departments. 11 STEM departments (A-K) are shown at the start of the LAT (squares) and three years later (circles), by voice. X-axis denotes number of target practices used. Dark gray band labeled “NA” denotes absence of a voice in departmental practices and light gray labeled “0” denotes the use of voice with no target practices (Krishnan et al., 2022). Blue symbols highlight departments that advanced their practices in multiple voices and with multiple research-based target practices, referred to as “High Change” departments. Yellow symbols highlight departments that achieved more limited change, referred to as “Low Change” departments.


A. Student Voice
Type
High Change
Low Change
A
B
C
D
E
F
G
H
I
J
K
NA 0 1 2 3 4 5 6 7 8 9 10 11
B. Peer Voice
Departments
A
B
C
D
E
F
G
H
I
J
K
NA 0 1 2 3 4 5 6 7 8 9 10 11
C. Self Voice
A
B
C
D
E
F
G
H
I
J
K
NA 0 1 2 3 4 5 6 7
Target Practices

High Change example: Department B

At the start of their involvement, Department B used two voices to evaluate teaching, student and peer, with two research-based target practices. Three years later, the department used all three voices, each with multiple target practices (Figure 2.1). The department integrated all three voices via written self-reflection, which was new to the department, and also added new target practices to their use of peer voice.

The change process in this department started when the head joined the LAT and asked a faculty member to lead changes to teaching evaluation. This faculty member, who was seen as a teaching leader in the department, developed an initial plan for both peer evaluation and written self-reflection in collaboration with a few departmental colleagues. At the same time, the head raised the topic of teaching evaluation at multiple faculty meetings. At two meetings, the head invited members of the project team to present about robust and equitable teaching evaluation and to answer faculty questions. At a later meeting, the faculty leader presented the pilot plan and solicited feedback from faculty. After revisions based on feedback, the faculty voted in favor of adopting the proposed teaching evaluation practices.

Department B's new practices included a teaching self-reflection form that all faculty submit as part of their annual evaluation. The 2-page form leveraged student and peer voices, and included two or three sections: (1) Faculty reflection on what went well in their teaching and what was challenging for them in the prior year; (2) Data from faculty's student end-of-course evaluations from the last three years (including response rates and reflection on student comments); and (3) Information from a peer observation if one was performed during the year.

The department also developed and piloted a new system of peer observation. Their peer observation process relied on a standard form, which specified that peer evaluation should include a pre-observation meeting (with some guiding instructions) and a post-observation meeting. The form included guidance for the meetings and details about the focus of observations. It contained a table for the observer to fill out, prompting them to look at specific

aspects of the lesson, such as content, clarity, student engagement, etc. The faculty also agreed on designated timepoints in a faculty member's career when their teaching would be evaluated by peers.

Low Change example: Department F

Department F changed one voice (self) and did not implement any target practices (Figure 2.1). The department head decided to add self reflection to the annual evaluation process. In the first year, the head allowed faculty to decide whether to include a written self-reflection. In subsequent years, the head conveyed the expectation that all faculty submit a written self-reflection. In neither year did the head provide guidance about what questions faculty should answer or what process they should use to reflect, which falls short of using any research-based target practices (Krishnan et al., 2022). The department did not advance student nor peer voice. Their peer evaluation practices remained ad hoc, with no structures in place to guide the process. During the 3-year period of this study, a new head took over and they described a peer evaluation practice with even fewer target practices than the prior head. The department's use of student evaluations remained stagnant without target practices.

The data presented in this section demonstrate that the LAT had mixed results. About one third of the participating departments made impressive progress in transforming their teaching evaluation practices to be more robust and equitable. And they did so during a global pandemic that affected every aspect of how academic departments function and the work of teaching. During the same time period, two thirds of participating departments achieved limited progress. Since we designed the LAT with the department heads as the learners and leaders of change, we wanted to better understand how differences among heads might have impacted the change they achieved. We investigated this question using the readiness for change framework.

RQ2: How did department heads' readiness for change relate to changes in departmental practices?

Some components of readiness for change predicted the changes department heads achieved and others did not. Each head described inadequacies they saw in their current teaching evaluation practices (i.e., *discrepancy*), but this did not reliably lead them to pursue change. However, one particular discrepancy distinguished the two heads who led the most change. Considering *valence*, heads pursued changes to teaching evaluation practices only when they saw clear potential benefits. Each head expressed concerns about potential costs of changing teaching evaluation practices, but heads of High Change departments did not see these as insurmountable barriers. The heads of High Change departments most distinguished themselves when it came to *efficacy*. Their personal efficacy for changing teaching evaluation was low. They reported that they personally lacked relevant expertise, yet they believed the department could rely on knowledgeable colleagues and they delegated accordingly. Lastly, department heads were not convinced or dissuaded from pursuing changes in their departments based on their perceptions of *principal support* from upper administration. We do not report findings about *appropriateness* because the evidence we collected did not provide sufficient detail (see Qualitative analysis of readiness for change).

The remainder of this section elaborates each of these findings and provides supporting evidence in the form of quotes. Quotes have been edited lightly for grammar and we have used ellipses to indicate portions cut from quotes for the sake of brevity and clarity. Brackets indicate text added to convey the meaning implied by the quote or previous utterances that are not shown here. Quotes from meeting recordings often require these additions for clarity because the speaker may be referencing something said by others in the room.

Discrepancy: Seeing problems with teaching evaluation practices did not reliably translate into action

Department heads all saw a need for changes to teaching evaluation in their departments (i.e., discrepancy), but this did not reliably translate into action. Department heads frequently expressed a desire to add a voice that their department did not currently use to evaluate teaching. For example, the head of Department E explained that they wanted to move beyond sole reliance on student evaluations:

> "We had a faculty meeting a couple of weeks ago and discussed this... We're starting from the point that none of us are happy with student evaluations. It's more or less useful, but as responsible faculty, we don't want to rely upon that. We want to have feedback from our peers, but we're not used to that."

Despite this stated need for change, Department E did not add peer nor self voice to their teaching evaluation practices (Figure 2.1).

Heads also often talked about problems they saw with the specific way in which the department used a voice. For example, heads lamented low response rates on course evaluations, as described here by the head of Department B:

> "I think the biggest challenge is always getting students to participate [in end-of-course evaluations] because what usually happens is you get 10-15% of your class who have an ax to grind one way or the other, and then you don't really get information from the bulk of the class."

Though this head expressed dissatisfaction with response rates to student evaluations, they did not set an expectation that faculty take steps to achieve a higher response rate, which is a target practice for increasing the reliability of student voice. As another example, the head of Department D shared that their department needed to more regularly conduct peer evaluation, a dissatisfaction shared by other participants:

> "I think everyone needs feedback [from peer evaluation]. And more periodically instead of at random times. Typically, if there's only two stages of promotion, some people have not been evaluated for many years."

Yet, Department D did not work to ensure that peer evaluations occurred over multiple time points throughout a faculty member's career, a target practice. In addition to wanting peer evaluation to occur more regularly, department heads commonly reported that observers needed training and that the workload for peer observation had not been equitably distributed, but seeing these needs did not necessarily result in pursuing change.

One particular discrepancy was unique to the heads of the two departments that achieved the most change. The heads of Departments A and B felt that their departments undervalued teaching and that teaching evaluation needed to change to address this problem. As the head of Department A explained, historically their department had not known how to evaluate teaching, and this had consequences for how teaching was valued:

> "This is... an ongoing issue in the department... We don't have a good sense of how to evaluate people's teaching... If the majority of your appointment is classroom teaching, how do you evaluate that?... [This] creates some logical resentments within the department because there's not a good appreciation [for teaching]. We've spent a fair amount of time trying to appreciate each other's science, and giving each other the benefit of the doubt for science. But we don't have a similar way of doing that for teaching because we just don't think of teaching as an activity that has known skills that make you good at it."

This head felt that faculty tended to view teaching and research differently in the department. Whereas faculty viewed research as requiring particular expertise, they did not necessarily see teaching as requiring the same level of expertise. The head felt this resulted in the department underappreciating the work invested in teaching. For this head, advancing teaching evaluation

practices would enable better recognition of teaching, which had the potential to improve teaching effectiveness. Similarly, the head of Department B aimed to set an expectation in their department that teaching would be valued and rewarded,

> "We have an amazing research enterprise here, but sometimes teaching is given a short shrift. It's viewed as a necessary evil, or it's a bitter pill to swallow. You get done with it in two or three weeks and you survive. I personally enjoy and value highly the part of our job that is related to instruction. And I think that developing a system for evaluation is a way to positively highlight the importance of that job and the value that we place on it in the department and in the university."

These two heads believed that teaching was a critical component of faculty work, that effective teaching follows from expertise and investment in improving, and that existing systems did not allow for tangible recognition and reward for effective teaching. Though these ideas align with national calls to improve undergraduate education (i.e., AAAS, 2011), most participating department heads did not espouse these views.

Valence: Heads only pursued changes they perceived to offer clear benefits, and heads of high change departments saw ways to mitigate costs

Department heads pursued changes in teaching evaluation practices only when they perceived concrete benefits for faculty in their department. The heads of High Change departments perceived benefits to changing multiple voices, and then pursued these changes. As an example, the head of Department C discussed the benefits that they saw in implementing new peer and self-evaluation practices. In this quote, they described how peer evaluation could be beneficial for the observer and the observed:

> "I actually learned a lot [participating in peer evaluations]. Especially because I teach a non-majors class with 250 students, and I evaluated somebody [else] teaching it in the

> other semester. I learned a lot about the classroom dynamic in the corners of the classroom that you can't see and things like that. It actually gave me a lot of valuable perspective. I think there's value in the evaluating that goes on. I think that it's beneficial for people to watch others teach and get ideas, see what works and what doesn't and think about things in new ways."

The head of Department C also saw concrete benefits to self-reflection, including helping faculty focus on how they could improve, even if they believed they did not need to improve:

> "Just because things [in the classroom] are going well doesn't mean that they can't be improved. For some people, I think they think things are going perfectly well, but maybe they're not. I see student evaluations, I see [the students' responses], and sometimes those don't agree [with a faculty's evaluation of their own teaching]. I think asking people [in a written self-reflection] to identify an area that they would like to work on, or a new practice that they would like to implement would have some value."

In alignment with the benefits that the head of Department C anticipated, Department C made progress in implementing new target practices in the peer voice, and initiated a system of faculty self-reflection on teaching (Figure 2.1).

In contrast, the heads of Low Change departments anticipated benefits from more limited changes to teaching evaluation. For example, the head of Department G described student end-of-course evaluations as a valuable source of data for them:

> "I read all of the evaluations from the students. If I see, for instance, a good number of students saying the same thing, then I really pay attention to that one thing. Sometimes, if there's a problem, I talk to that faculty member. I think it is extremely helpful for the faculty. And we can actually help prevent issues by going back to the students."

This head added one target practice to their use of the student voice (Figure 2.1). They started expecting their instructors to take all necessary steps to improve their response rates on end-of-

course evaluation. This head did not discuss benefits of any of the other voices and did not pursue changes to them.

All department heads had concerns about potential costs to changing their department's teaching evaluation practices. However, these concerns did not dissuade the heads of High Change Departments from leading change. Heads raised a variety of concerns, but the most common dealt with faculty resistance and, relatedly, the potential to increase faculty workload. Some heads had already experienced resistance, whereas others anticipated future resistance. For example, when considering implementing self-reflection as a part of annual evaluations, the head of department K said:

> "I don't think [my department] will ever be in a place where I can tell all the faculty they have to do [self-reflections]. There will be a significant fraction that will just simply refuse this, not do it... This actually requires some effort on their part. I think for some people, some of the senior faculty that have already been promoted to full professor, for example, I can imagine some of them being like 'Well, okay, this is just some other thing that the department is asking me to do...' I'm not entirely optimistic that I could get buy-in from everyone at all levels."

The quote above illustrates that the department head anticipated that faculty would resist new teaching evaluation practices because these practices would increase their workload. More than half of the participating heads shared this concern. For example, when asked if they thought their faculty would be supportive of making changes to their peer evaluation practices, the head of Department E said:

> "I want [peer evaluation] to be an efficient process that doesn't add significant burden to the lives of faculty... It's a burden to do [peer evaluations]... There'll be a perception that this is one more thing that needs to be done on top of everything else."

Even though all heads had similar concerns about potential costs, heads of High Change departments saw ways to move forward productively. Heads of High Change departments tended to view resistance as a normal part of making change. For example, the head of Department B explained that most efforts to change how things worked in a department would engender resistance from some faculty. They felt they could balance costs and benefits for faculty, and as a result, most faculty would be supportive. They explained,

> "I think the thing that people would be most worried about is the amount of time required [for them] to spend in other people's lectures. That could add up. We're trying to strike a balance when doing this. It has to be often enough so that it's part of our lives... I'm worried even doing it once a year for everyone in the department, it's going to be tough... I mean, there's nothing that doesn't engender complaints. In the entire world, that's the way it is. And in our little microcosm of [Department B], that's the way too. So yeah, I think most people will be mostly supportive, and some people will complain. It's life."

The head of Department A similarly anticipated resistance from some faculty and felt equipped to successfully navigate the resistance by having open conversations with faculty. When asked how they might handle resistance from faculty as they pursued new teaching evaluation practices, the head of Department A explained,

> "The [faculty] who come to you directly… are the easiest ones [to deal with] because they've already started the dialogue... You can say, 'Look, I'm sorry that you find this to not be very valuable. If you want to be involved in the process, we could try and make it a more valuable experience for you... I could put you on the committee.' Or just tell them 'Look, I'm sorry that you don't find this valuable, but we have to have a way to fairly evaluate instruction. This is a way that is relatively well supported with data and that provides us with actual feedback, as opposed to the haphazard way this has been done in the past. If you don't want to be engaged in your teaching, that's up to you, but it will be reflected in your annual evaluations going forwards.' ... The key thing is not to just tell people 'Shut up and do it,' because that never works with academics... It's to try and have the dialogue and engage with people."

This department head is able to describe what responses they anticipate from faculty and different ways they could respond. Thus, they were not dissuaded from pursuing change they felt was important for the department. Other heads similarly anticipated resistance from some faculty but did not articulate clear ideas about how they could foster buy-in among faculty or work with or around a few resistant individuals.

Efficacy: Heads of High Change departments questioned their own knowledge and delegated to others

Heads of High Change departments reported that they lacked expertise in teaching and changing teaching evaluation and sought help. For example, when asked if they had felt prepared to lead changes to teaching evaluation, the head of Department A replied,

> "Not even a little bit. This is so far outside my area of expertise that when I started on this project, I was myself resistant to it. I had no idea what this was going to look like. I didn't have the vocabulary to talk about it. I had no idea."

Each head of a High Change department expressed similar doubts about their knowledge of teaching and teaching evaluation. In contrast, not a single head of a Low Change department questioned their own expertise in teaching or preparation for leading change to teaching evaluation. For example, when asked whether they felt comfortable leading these changes, the head of Department H responded,

> "I feel very comfortable... We've implemented a number of changes over the past five years for our academic programs... I think change management is a fairly large topic, there is a skill to change management."

Heads of Low Change departments generally felt that they were equipped to lead change, but our data suggest that they did not choose to use these leadership skills to engage in the work of changing teaching evaluation practices.

Following their assessment that they lacked relevant expertise, heads of High Change departments delegated the work of developing and piloting new teaching evaluation practices to knowledgeable faculty. For example, the head of Department C explained,

> "In terms of leading change in this specific area, I'm not, by training, a pedagogical expert. And so I would like to have partners in it. In our department, we have a lot of really good instructional faculty, tenure track and non-tenure track... I wouldn't be comfortable implementing everything or deciding what needs to be implemented, but I would be perfectly comfortable supporting the implementation and prompting change."

The heads of High Change departments each identified one or two faculty who were interested and willing to spearhead the development of new departmental practices, and who the head felt had expertise that they personally lacked. Of the four High Change departments, one created a new committee to undertake teaching evaluation reform, one charged an existing committee with this work, and in the other two, the work largely fell to individual faculty.

Principal Support: Department heads' pursuit of change was unaffected by their perceptions of administrative support

Department heads did not rely on their perceptions of administrative support in their decisions to pursue changes to teaching evaluation. Each head, including those of both High and Low Change departments, felt that the upper administration of the university supported advancements to teaching evaluation. For example, the head of Department D stated,

> "I think they're supportive at the college level and even at the Provost level. I can't say

> that I've seen a push per se. There are new directives now that we have to adjust our annual evaluation and promotion documents to include [the three voice framework]. So in that way they're supportive."

This head and others pointed toward policy changes as evidence that the upper administration would be supportive of departmental work advance teaching evaluation. Two heads perceived that there would not be resources allocated for any potential changes, but still felt that the administration was generally in favor of changes to teaching evaluation.

DISCUSSION

One clear finding of this work is that achieving second-order change in departmental teaching evaluation practices is challenging. The LAT intervention facilitated extensive teaching evaluation reform for some departments but a greater number of departments accomplished only first-order change. Pursuing extensive reforms relied on seeing a need for change, perceiving concrete benefits of the change, seeing ways to address potential costs, humility about one's expertise, and tapping human resources in the department. These findings have informed our ongoing approaches to advance teaching evaluation within our own institution. We elaborate in this section about hypotheses that emerge from this work that warrant consideration in other change efforts and testing in future research.

We hypothesize that department heads who leverage the human resources in their department will more successfully advance teaching evaluation. This requires a department head to (a) acknowledge the limitations of their own expertise related to teaching and teaching evaluation, and (b) delegate the work to departmental colleagues who are more knowledgeable and/or eager to learn. Though most department heads in this study had a lot to learn about robust and equitable teaching evaluation, only a few reflected openly on what they knew or had room to learn in this area. This humility played a role in their ability to lead change because it prompted them to rely on other departmental faculty who were more knowledgeable or eager to

learn. A team with diverse knowledge and skills can more successfully lead change within an organization (e.g., Kotter & Cohen, 2012).

Delegating the work of developing new teaching evaluation practices to other faculty had two additional benefits as well. First, it meant that someone besides the department head was working to make progress. This is likely critical because department heads often struggle to balance the workload of their positions (Gmelch & Miskin, 1993) and therefore may also struggle to find time to work on teaching evaluation. We observed this first hand in the LAT. Second, including other faculty can ultimately foster faculty buy-in because the process and products are guided by what faculty view as both useful and feasible (e.g., Cheldelin, 2000; Lucas, 2000). Relatedly, second-order change requires cultural shifts and culture change involves many members of a department shifting their thinking and practices, not just the leader. Department heads who felt they needed to do all of the work of developing new teaching evaluation practices themselves, rather than delegating, made limited progress in enacting new practices and likely had no impact on the culture surrounding teaching evaluation.

We also hypothesize that department heads need to be adequately prepared to effectively lead change. Many department heads take the job out of a sense of duty to serve their department (Gmelch & Miskin, 1993), which is commendable. In turn, they deserve adequate preparation to maximize their effectiveness and efficiency in this role. Department heads occupy a difficult position, caught between serving their faculty and their administration (Wolverton et al., 2005; Kruse, 2022). This challenge can be exacerbated by a lack of preparation for the department head role (Gmelch et al., 2017). Department heads have both management and leadership responsibilities, which require different skills. Management involves maintaining day-to-day operations of the department (e.g., budgeting, supervising staff, facilities management, interfacing with upper administrators, addressing problems), whereas leadership involves defining a vision for the future and inspiring faculty to make it happen (Gmelch & Miskin, 1993; Kotter, 2012). Change leadership is a specific type of leadership, and

requires a distinct set of skills (e.g., advocating for a change, spreading an inspiring vision, encouraging faculty involvement) (Yukl, 2012). Most department head training provides instruction on campus policies and procedures (Normore & Brooks, 2014), rather than building leadership capacity. In the absence of leadership training, department heads resourcefully rely on their own experiences and ideas about leadership to guide their actions, which vary considerably from person to person.

The heads of High Change departments successfully led teaching evaluation reform and their cases may be instructive regarding the skills and strategies that leadership training should foster. In particular, High Change department heads functioned in two ways that align with scholarship about organizational change and departmental leadership. First, High Change heads built a team to do the change work; team building is an essential step in organizational change and change teams are most effective when they include individuals with political power, expertise, credibility, and leadership skills (Kotter, 2012). Furthermore, academia values shared decision-making and governance, requiring heads to be equipped to manage these processes (Lucas, 2000). Building a team ensures that these values are built into the change process. Second, high change heads dealt effectively with resistance. Resistance is a normal part of any organizational change. Change involves letting go of how things have worked previously and uncertainty about how things will work in the future, both of which can cause stress (Cheldelin, 2000). High change heads recognized different forms of resistance, heard concerns without taking them personally, offered reassurances about shared values and a clear vision for the future, and moved forward (Cheldelin, 2000). These ideas about effective leadership and others, if put into practice, could help department heads lead change. Therefore, we see a need for leadership training to be encouraged and incentivized for department heads.

Finally, we hypothesize that achieving second order change in teaching evaluation is more likely when a department head perceives that teaching is undervalued and interprets this as a problem. This hypothesis is based on two departments who achieved the most change and

national conversations that have called on institutions and departments to reconsider how teaching is being valued and rewarded (Dennin et al., 2017). Inadequate teaching evaluation is not the only structure in the academy that systematically places lower value on teaching than research. Reward systems often also devalue teaching. Within the institution where this work occurred, faculty who have teaching as most or all of their responsibility are paid less than tenure-track faculty. There are also fewer awards and titles available for teaching-focused faculty, which are symbols of prestige in the academy. Lastly, teaching-focused faculty are excluded from consideration for particular leadership roles, limiting their power to address systemic structures that undervalue teaching. All of our participants had the opportunity to notice that the systems of higher education in which they work undervalue teaching relative to research, yet only two expressed this idea. It may be incumbent upon change agents to create learning opportunities for leaders to critically consider the systemic devaluing of teaching. The department heads in this study expressed value for teaching, but we struggled to help everyone see how evaluation practices were misaligned with these espoused values, a step that may be necessary for cultural change.

The findings of this project have informed our ongoing work to advance teaching evaluation. We expanded our approach to include faculty leaders in addition to department heads. We continue to convene department heads in an LAT and focus their time on learning about teaching evaluation and leading change. We now support their work by making delegation part of our intervention, since this was the strongest predictor of being a high change department. As a condition of their participation, we ask heads to identify one or more faculty to lead the development and piloting of new evaluation practices. We gather these faculty bi-weekly for an academic year to support their learning and development of new practices, and compensate them for their time. The department head helps facilitate piloting new practices and fostering buy-in among faculty.

Readiness for Change as a Guiding Theory

Though it was developed to describe change in the context of private businesses, the readiness for change framework provided a valuable lens for looking at change in the context of STEM higher education. We gained new insights about the department heads' motivations to advance teaching evaluation and how we might more directly foster motivation in future interventions. Though it shares similarities with theories of personal motivation, such as situated expectancy-value theory (Eccles & Wigfield, 2020), the readiness for change framework encompasses organizational level components, and therefore likely has more explanatory power for work on changing departments or institutions. For example, within the readiness for change framework, valence considers an individual's perceptions of costs or benefits of a change for themselves and the organization. Department heads prioritized the department (and faculty therein) over personal costs and benefits when they considered teaching evaluation change, suggesting the need for theories that reach beyond the individual. Though no participants in this study were familiar with readiness for change, they talked extensively about securing "buy-in" from faculty about teaching evaluation change. Some of the ways they talked about buy-in align with components of readiness for change. Accordingly, one addition we have made to the LAT are meetings focused on fostering buy-in, in which we help department heads craft messaging about change that addresses components of readiness for change (e.g., Armenakis & Harris, 2002).

The component of readiness for change that most distinguished heads of High Change departments was efficacy, but the pattern was actually the opposite of what the framework predicts. The readiness for change framework stipulates that employees are more likely to productively engage in with change if they feel capable of implementing the change (e.g., Armenakis et al. 1993; Rafferty et al. 2013). Yet in this study, leaders achieved more change if they lacked personal efficacy and sought others they viewed as more capable than themselves in the domain of the change. This deserves attention in future research as it may be specific to

the organizational structure of higher education, where departments use more shared governance and responsibility than may be common in businesses.

A Multitude of Factors May Influence Departmental Change

Though useful, readiness for change alone is insufficient to explain the differences in teaching evaluation reform achieved by participating departments. We must also consider the larger context surrounding departments and department heads. We discuss some potential contextual influences on departmental change that future interventions and research may wish to consider.

Departmental history and culture can influence how change unfolds. For example, a department whose culture prioritizes full consensus for major decisions may take longer to gain approval for new practices and policies than a department that relies on majority votes to finalize decisions. In contrast, a department that has hired a cohort of new faculty who they are eager to support may see teaching evaluation change as a promising mechanism for supporting these new faculty with constructive feedback and valuable evidence for their promotion dossiers.

Factors at the college and institutional level could also impact changes in teaching evaluation. For example, in a department experiencing pressure from their college to increase research productivity, the head and faculty may feel that they are already overburdened with carrying out other changes and cannot also pursue teaching evaluation reform. Indeed, experiencing too many changes in a short amount of time can create stress for employees, and may negatively impact the outcome of those changes (Bernerth et al., 2011). In contrast, a department experiencing pressure from the college to reduce DFW rates in introductory courses may see research-based peer evaluation practices as a way to provide constructive feedback to faculty, spark more conversations across sections of a course, and demonstrate to upper administration that they are taking action.

Institutional and/or college policies and structures may also influence departmental autonomy in teaching evaluation. For example, college or institutional promotion and tenure guidelines could require that faculty include their quantitative student evaluation scores compared to a departmental average. This approach is highly problematic due to the many factors that influence course evaluation scores that are out of the instructor's control, including instructor race, gender, country of origin and course time and size (e.g., Bedard & Kuhn, 2008; Boring, 2017; Fan et al., 2019; Esarey & Valdes, 2020; Aragón et al., 2023). Yet a department may have to comply until policies at a higher level change. Faculty unions represent another institutional factor that can influence change, for example by mandating or limiting the use of peer evaluations for faculty promotion and tenure consideration.

Disciplinary context also matters. For example, some disciplines undergo extensive accreditation of their undergraduate programs, which can be a barrier to change processes (e.g., Laursen et al., 2019). The Accreditation Board for Engineering and Technology, Inc. (ABET) is an accrediting agency for engineering and computing disciplines that expects departments to define their program's education objectives, as well as outcomes for students, and focus on continuous improvement over time. This requires considerable planning and ongoing work for departments, who then may feel both overburdened by assessment and convinced that they already evaluate teaching sufficiently. Yet assessments for ABET do not focus on individual instructors. These data may not provide feedback that helps faculty improve or evidence on which to base reward decisions for individual faculty. In these ways and others, accreditation may act as a barrier to teaching evaluation reform (Laursen et al., 2019).

Limitations

Given that this work examines 11 STEM departments at a single research-intensive institution, the findings are best considered exploratory and most appropriately used to generate and test hypotheses in other contexts. We caution readers about generalizing these findings to

other disciplines, institutions, or institution types. Additionally, all of our department heads agreed to participate voluntarily. We note that multiple participants made limited progress, so a willingness to attend meetings appears unrelated to a willingness to change.

The timing of this work creates another set of limitations. We examined change over three years, but change in academic departments is slow and this project is ongoing. Change to teaching evaluation is continuing in some of these departments, so our findings may under-represent the change accomplished by prioritizing those able to achieve change more quickly. We also did not follow departments long enough to study how faculty use the information provided by more robust teaching evaluation practices. It may be the case that faculty need additional support to build their confidence and skills for relying on these data to inform their teaching (Hora et al., 2017; Lenhart & Bouwma-Gearheart, 2021). Additionally, the years of this work included the two academic years most negatively impacted by the COVID-19 pandemic. This undoubtedly slowed progress. In later iterations of the LAT, we have been able to facilitate learning much more quickly, which may have resulted from our own development as change agents, the greatly reduced impact of the pandemic on academic functioning, and growing awareness of the successes of high change departments.

Our research is limited by the fact that we did not have the exact same data from each department. Several department heads were unavailable during the second round of interviews due to turnover, meaning we relied on other sources of information to characterize practices and had only data from LAT meetings to assess readiness for change. We see that two of these departments F and G, appeared to regress their peer evaluation practices (Figure 2.1). It is unclear whether the initial or second report of practices was inaccurate, or if practices actually moved away from research-based practices.

Lastly, we were not able to fully apply the readiness for change framework in this context. Our data lacked sufficient resolution to compare appropriateness, which is a sense that

the solution under consideration will address the identified need for change. Future work can consider how to elicit data about appropriateness and relate this to change achieved.

Conclusions

Without robust and equitable teaching evaluation practices, departments are unable to determine who is teaching effectively and who is working to improve their teaching over time. As a result, they cannot reward investments in teaching and teaching improvement. Evidence-based teaching will only become widespread when evaluation and reward systems illuminate and incentivize effective teaching. Changing the status quo in these systems requires long-term interventions. As a community, we will be best equipped to succeed if we take a scholarly approach to these changes and share what we learn. We hope that others can learn from our efforts, including what went well and where our efforts feel short.

CHAPTER 3

NAVIGATING RESISTANCE AND BUILDING BUY-IN: HOW DEPARTMENT HEADS LEAD TEACHING EVALUATION REFORM[2]

[2]Ericson, H. C., Lemons, P. P., Dolan, E. L., Brickman, P., Smith, K., & Andrews, T. C. To be submitted to a peer-reviewed journal

ABSTRACT

Department heads are important in the process of implementing changes within their departments. However, not much is known about the strategies they use when leading change. We examined the ideas and actions department heads used to lead change to teaching evaluation in their departments. We used Kotter's 8-step model for leading change as an organizing framework to do so. Department heads found it important to involve their faculty in the planning process, seeking their input on how to implement new teaching evaluation practices. They also worked with specific colleagues to both develop new practices and to build buy-in amongst the rest of the department. Heads were strategic in how they presented the changes to faculty, often emphasizing the benefits faculty would see. They worked to help faculty feel confident they could be successful in implementing new practices. They were also strategic in starting the implementation process with smaller, more manageable changes, rather than doing everything all at once. We discuss Kotter's 8-step model for leading change, and how the ideas and actions department heads used to lead change overlap with and deviate from the model.

AUTHOR CONTRIBUTIONS

I designed the interview protocol for the interviews used in this chapter in collaboration with Tessa Andrews, and I conducted all interviews. I led the development of the qualitative codebook, and conducted all analyses presented in this chapter, in collaboration with Tessa Andrews and Kylie Smith. I wrote the manuscript, and created all the graphs and figures present in this chapter. Tessa provided iterative feedback on the chapter.

INTRODUCTION

Department heads are important figures within the lives of faculty and for the functioning of their departments. Department heads[3] have many different significant responsibilities including, but not limited to: dealing with the governance of the department including scheduling and presiding over meetings, scheduling classes, presiding over merit and annual review processes, mentoring and coaching faculty, advising students and handling issues and complains, serving as a representative of the department at events or formal proceedings, and preparing and maintaining departmental budgets (Berdrow, 2010; Kruse, 2022, Cervato et al., 2024). They also face the challenges of bridging between the upper administration of the college and university and faculty in the department and balancing managerial and academic demands on their time (Bolden et al., 2008; Gmelch et al., 2017). Because of their role in the department, heads are uniquely positioned to be agents of change in their departments (Cervato et al., 2024). However, most have had few formal opportunities to develop leadership skills, and what little training they receive is focused on the managerial side of running a department (Normore & Brooks, 2014; Gmelch et al., 2017).

Despite their critical role in departments and their potential as change agents, very little scholarly work has focused on academic department heads' role in leading change. There is not a clear understanding of the specific leadership strategies department heads can use to navigate change efforts. Without this knowledge, department heads lack the support to become more effective change agents, potentially slowing progress on important change initiatives. By gaining insight into how they approach leading change, we can better prepare department heads for this important role. This study examines how STEM department heads thought about and approached leading departmental change.

[3] Institutions use different names for department heads based on their location, the culture of their universities, and the details behind appointments (Bryman, 2007). In this institution, department heads serve at the pleasure of the Dean, are selected with input from faculty, and may serve one or more three-year terms.

Guiding Framework

We used an 8-step model for leading organizational change as a framework for analyzing department heads' expertise for leading change (Kotter, 2012). The model, popularized in the discipline of organizational management, emerged from Kotter's work guiding businesses as they tried to change aspects of how their organizations functioned, and especially from studying cases of successful and unsuccessful organizational change (Kotter & Cohen, 2002; Kotter, 2012). The 8-step model is prescriptive and practical, and professionals in a range of fields, including health care, public services, government agencies, and higher education have used it to guide change efforts (Wentworth et al., 2020).

In the following section, we describe the eight steps of Kotter's model and relate the model to other relevant literature about leadership, organizational change, higher education, and department heads. This model focuses on organizational, rather than individual level change, which might include changes in departmental practices, policies, leadership structure, curriculum, etc. Though Kotter's model focuses on change in businesses, we describe the 8 steps of the model using academic departments as the focal organizational level. The model may also be useful for college- or university-level change, or changes that depend on coordination across these levels.

Step 1: Increase urgency

The first step of the 8-step model for leading change is to *increase urgency* surrounding a change in the department (Kotter, 2012). A sense of urgency emerges when change agents convince faculty that a given change is necessary to address a problem, and must happen quickly (Kotter, 2012). Studies of organizational change posit that seeing a need for a change is a key component of an individual's readiness to engage positively in a change process (Armenakis et al., 1993; Rafferty et al., 2013). Without a sense of urgency, faculty may be unmotivated or resistant to investing in the change, leading to stalled or failed change efforts

(Kotter, 2012). As an example, certain changes, such as curricular reform, may languish for years without a strong sense of urgency behind the process.

Two broad strategies can help create a sense of urgency: leveraging external forces and helping people see a problem (Kotter & Cohen, 2002; Kotter, 2012). Effective leaders monitor events and changes internal and external to the organization (Yukl, 2013), which may necessitate an urgent change, or be framed as a driver for urgent change. In business, dropping profits or shifts in the market may be crises that create a sense of urgency (Kotter, 2012). In higher education, dropping enrollments, a shift in policy, or new faculty recruits could be framed as time-sensitive opportunities for change in a department. A second strategy involves helping organization members see and appreciate a problem. Information, especially if it can be visually represented and is personally relevant, can help foster a sense of urgency. A visualization or other communication that sparks feelings tends to be especially powerful. Though faculty rely heavily on empirical evidence in their research, personal experiences and stories may be more persuasive for some decisions (e.g., Andrews & Lemons, 2017). Testimonials from trusted colleagues or other stakeholders in the change (e.g., students) can make a problem relatable and salient, convincing faculty that change is needed and pressing.

Different problems are likely to be more motivating to different faculty, meaning that leaders may need to describe several problems that need to be addressed. For example, reforming teaching evaluation practices can help mitigate bias in evaluations, provide teaching support for early-career faculty, add to a sense of community around teaching in the department, and improve students' experiences. Some faculty will view disappointing student outcomes as an urgent problem, whereas others will be more concerned that the department is inadequately supporting the success of new hires.

Step 2: Build the guiding team

The second step of the 8-step model is *building a guiding team*. Change requires a team because individuals rarely have all of the necessary information, credibility, connections, and time to make optimal decisions and foster necessary buy-in among colleagues (Kotter, 2012). An effective group can work more quickly than a single individual and has more bandwidth to tackle the difficult situations likely to arise when leading change (Kotter, 2012). Additionally, engaging a team generates a wider sense of ownership of the project, which fosters buy-in (Kanter, 1991; Cheldelin, 2000). Department heads may already be in the habit of using committees to accomplish important organizational work because they tend to function more as team leaders than as autocrats (Lucas, 2000).

Effective guiding teams for change include members who, collectively, are well-connected and influential within the organization, and have the necessary skills and experience developing and leading the implementation of new practices (Kotter & Cohen, 2002). Leading change requires managing operations of the change, setting direction, and inspiring people to bring them on board, so prior leadership experience may be especially important (Kotter, 2012; Yukl, 2012). For example, a guiding team for changing teaching evaluation practices might benefit from faculty who are respected for their teaching and research leaders who will be seen as credible by research-centric faculty who worry about new evaluation taking time away from research. It might also benefit from junior faculty who can articulate how new evaluation practices would serve their development and senior faculty who have led policy development previously. Teams function more effectively when group members trust each other and when members feel their time is being used effectively, so a convener who can effectively facilitate inclusive and productive meetings can help facilitate change (Kotter & Cohen, 2002; Kotter, 2012).

Step 3: Get the vision right

The third step in the model is *getting the vision right*, a process led by the guiding team. A vision is a picture of a desirable future that could be accomplished through successfully achieving the change. It lays out the end state that change plans and strategies will achieve (Kotter, 2002). A well-constructed and effective vision can inspire faculty to action through a clear sense of direction, motivate them to work together to achieve it, and help to soothe resistance (Angelo, 2000; Lucas, 2000; Kotter, 2012; Yukl, 2013). Making changes to the status quo can be difficult because it forces faculty out of their comfort zone, and it may feel as though they are being asked to act against their own self-interests (Kotter, 2012). Additionally, faculty have many responsibilities competing for their attention and being asked to do something that they don't explicitly view as beneficial may lead to resistance. A sufficiently well-designed and inspiring vision can help spur people to action, by showing them what the future could look like.

Effective visions for a change have several key characteristics. First, the change vision is imaginable; faculty should be able to picture what the future of the department could look like after the change is implemented. Second, the vision is desirable, meaning that the outcome of the change is appealing to faculty. This requires emphasizing a range of benefits of the change, to appeal to members with different priorities, and addressing concerns about costs. Faculty will reasonably worry about how a change may affect them. Specifically and directly addressing these concerns can mitigate resistance. Addressing concerns often requires offering enough details to convince organization members that the proposed changes have been carefully considered and emphasizing how benefits outweigh cost (Armenakis & Harris, 2002). Third, the change vision is feasible. Organization members who worry that the organization or members are not prepared or able to successfully enact the change will sensibly push back on proposed changes (Rafferty et al., 2013). Fourth, the change vision is flexible, so that it allows for creativity and tailoring in how it is executed, while also being specific enough to engender excitement. Finally, the change vision is easily communicable; it is able to be easily explained

and understood by faculty within the department. (Kotter & Cohen, 2002; Kotter, 2012). Visions with these characteristics can assuage faculty concerns about how a change might negatively affect them or the organization and build excitement about the future that the change could create (Armenakis & Harris, 2002; Kotter, 2012).

Step 4: Communicating the vision

The fourth step of the model is to *communicate the vision*. An excellently crafted vision will not be effective if the individuals within the organization do not know about it or do not understand it. This is not an easy step of the model, but it is essential for change agents to be effective in this work (Bryman, 2007). Insufficient or inconsistent communication about the vision is common in organizational change efforts and can lead to a change initiative failing to gain the support it needs (Kotter, 2012).

Effective communication about organizational change can involve a few strategies. First, communication about a change vision is more effective when it is focused, succinct, and avoids the use of jargon (Kotter, 2012). Second, repeated communication, using a range of formats, helps foster familiarity and reiterates the benefits of the change. A department head might use presentations and discussions in meetings, emails, and informal one-on-one or small group conversations to reiterate the change vision (Kotter, 2012). As part of repeated discussions, leaders will want to be prepared to mediate any conflict that may occur and answer any questions that may be asked. This can help faculty process their ideas surrounding the change, and ultimately build buy-in (Lucas, 2000; Bryman, 2007). Third, matching actions to words by participating in new practices can show the feasibility and benefits of the change. Individuals will question the credibility of the leader or the organization's commitment to the change if leaders communicate one thing and then are seen to be doing another, quickly undermining buy-in (Angelo, 2000, Kotter & Cohen, 2002; Kotter, 2012). For example, a department head leading

revision to teaching evaluation practices can volunteer to be among the first faculty to be observed by peers, conveying their personal investment in the process.

Step 5: Remove barriers

The fifth step of the model is to *remove barriers* that would prevent individuals from acting in alignment with the vision (Kotter & Cohen, 2002). Even if a team fully engages with the first four steps of the model, barriers will still stop individuals from engaging fully in the change. The goal of this step is to empower as many faculty as possible to take action by removing barriers (Kotter, 2012). Examples of barriers that may impact change efforts include supervisors who have a negative or apathetic attitude about the change (in academia, this could be a department head or committee chair), inadequate training and/or information for faculty to successfully engage in new practices, and evaluation and reward structures that are not aligned with change efforts. Kotter and Cohen (2002) make clear that not all of these barriers can be solved at once, but that eventually addressing them is necessary for successful and sustained change.

Different barriers require different approaches. Influential people in the department who are not on board, including department heads, can undermine change efforts. Change leaders may benefit from working strategically to build buy-in among formal and informal leaders by fostering a sense of urgency, reiterating the intended benefits of the change, and conveying its feasibility (Armenakis & Harris, 2002; Kotter & Cohen, 2002). To ensure faculty feel confident and are capable of participating successfully in the change, leadership may need to provide or organize appropriate and timely training and support (Bryman, 2008; Kotter, 2012). Additionally, asking individuals who have experienced the change to share about the processes and outcomes can build confidence that the change is feasible and beneficial. Existing practices, especially those related to evaluation and rewards, can hamper change efforts if the work people are asked to do as part of the change will not contribute positively to their advancement

(Kotter & Cohen, 2002; Kotter, 2012). Changing these structures to be in line with the change can motivate faculty to participate in the new practices (e.g., Reinholz & Apkarian, 2018). For example, a department that is adding systematic peer evaluation might change their promotion and tenure guidelines to expect the inclusion of these evaluations in dossiers. Identifying and then rewarding individuals who are working towards enacting the change vision can also be an effective way to empower faculty members (Lucas, 2000; Kotter & Cohen, 2002).

Step 6: Create short-term wins

The sixth step of the model is *creating short-term wins*. Change moves slowly and takes a long time, especially in academia (Kezar, 2014). Celebrating progress that has been made reinforces the change effort and shows people their efforts are paying off (Kotter, 2012). Without this proof, change efforts stagnate, resistance grows, and people lose enthusiasm for the project. Prioritizing smaller, shorter-term goals throughout the process can provide evidence that change efforts are worthwhile, address concerns about the change, and help keep faculty excited about and engaged with the change (Kotter & Cohen, 2002; Kotter, 2012).

Leaders can plan strategically to accomplish short-term wins. Celebrating these wins early and often helps maintain momentum and recruits new supporters for the change. Leaders can ensure early wins by targeting highly visible aspects of the change first, especially ones that are particularly meaningful to faculty in their department. Another good starting place comes from tackling changes that will be easier to achieve first (Kotter, 2012). Focusing on a discrete part of the change, rather than spreading the leadership team's focus out over many parts of the change, may also more quickly achieve short-term wins (Kotter & Cohen, 2002). An example of a win could be developing and piloting new peer evaluation practices with early-career faculty in a department. These faculty could then be invited to share their experience with the rest of the department at a faculty meeting, ensuring that everyone in the department knew about what had happened and how it was beneficial to their colleagues. Rather than trying to start the process

by involving faculty at all levels, this approach focuses on a specific group that is eager to participate because they value feedback and need evidence for their promotion dossiers. This approach would also produce highly visible results when faculty share their experiences with others.

Step 7: Don't let up

The seventh step of the model is *don't let up*. This step recognizes that change takes a long time to accomplish, that resistance may never fully dissipate, and that the harder parts of a change tend to be tackled later in the process (Kotter & Cohen, 2002). After working hard to secure early wins, the change team may be ready to turn their attention elsewhere, but achieving the vision often requires maintaining momentum and addressing the more entrenched parts of the system that stand in its way (Kotter & Cohen, 2002; Kotter, 2012).

Not letting up involves capitalizing on positive perceptions of shorter-term wins to initiate more difficult aspects of the change and otherwise working opportunistically to tackle more difficult parts of the change. This also involves maintaining and stoking a sense of urgency and recruiting additional faculty to contribute to the change so that the initial leaders don't burn out. Extending the example from the last section, after a department has successfully piloted a new system for peer observation, they can capitalize on the positive attitudes about the benefits for early-career faculty to expand the system to all faculty in the department. To avoid the resurgence of resistance, leaders can continue using strategies from previous steps while pursuing the next visible, meaningful, and feasible change in service of accomplishing their larger goal. Faculty members who do not need to be reviewed soon may not be excited about participating in a new peer observation system. In these cases, change agents could talk about how they could serve as important mentors for less experienced faculty. This next step in expanding the process is then seen as important and necessary, which can cut down on any faculty members resisting participating in the process.

Step 8: Anchor the Change in the Culture

The eighth, and final, step of the model is to *anchor the change in the culture*. This model primarily defines culture as norms of behavior and shared values in the organization (e.g., Kotter & Cohen, 2002). These norms and values influence how people interact with change and can shift as people become convinced that the new ways of doing things are superior to older practices (Kotter, 2012). Tradition is a powerful force, and it is easy for people to fall back into old norms of behavior (Kotter & Cohen, 2002). Changes may be held in place by only one or two key people who were instrumental in enacting them, and if these individuals leave their positions, it can be easy to backslide into how the department operated previously (Kotter & Cohen, 2002; Kotter, 2012). However, if the changes lead to shifts in the norms and shared values within the department, they are more likely to persist.

Shifting the culture of an organization is hard and can take a long time. It involves continuous messaging, with evidence, about the benefits of new practices. This serves to remind faculty that the new practices work as intended, providing benefits to faculty and the department. Leaders can shape norms and values by repeatedly conveying who the department is, what it does, and why it succeeds in relation to the changes (Kotter & Cohen, 2002). Relatedly, introducing new faculty to the change practices and their benefits and engaging them in the new processes from the beginning starts to set new norms of behavior as new people see the change as the way the department has always operated (Kotter & Cohen, 2002). Inviting additional faculty to lead parts of the change broadens the base of support and helps ensure that the change does not depend on just a few individuals. Giving credit to those engaging in new practices, such as recognizing service work and valuing their contributions in promotions, lends credibility and sets new expectations. Finally, leadership matters and new leaders who fully embrace the change will contribute to shifts in the culture, whereas leaders who do not buy-in to the change can undermine the shift toward new norms of behavior and shared values.

Many actions contribute to shaping new norms and shared values, and these actions can be relatively small. For example, a department head might set aside a 90-minute session at a faculty retreat for one or more faculty to share a teaching approach that's relevant across classes. They might introduce the session by saying, "Our department has a long history of valuing teaching. Most recently, we launched peer observation as a way to learn from each other and continue to improve our instruction. One of the many benefits of this is that we get to learn more about our colleagues' innovations in the classroom. Based on the recommendation of our peer observation committee, today I've invited some of your colleagues to share a teaching approach that might also be useful to you." This message reminds the department of the new practices, emphasizes a benefit, elevates faculty who engage in peer observation, and implicitly and explicitly conveys that teaching is valued in the department.

Study Goals

This study took place in the context of one institutional change project, called DeLTA (Departmental and Leadership Teams for Action), which aimed to advance departmental teaching evaluation practices (Andrews et al., 2021). Department heads participated in the project for one year. Their role was to learn about robust and equitable teaching evaluation practices, empower faculty advocates in their department to select and pilot new evaluation practices, and facilitate the adoption of the practices in the department. This paper examines heads' ideas and actions related to leading changes to teaching evaluation. Readers can learn more about robust and equitable teaching evaluation and how it was promoted in the DeLTA project in Andrews et al. (2021); Krishnan et al. (2022); and Ericson et al. (2025). In this paper, we seek to address the following research questions:

1. What ideas and actions about leadership did department heads use to lead change to teaching evaluation in their departments?

2. To what extent are these ideas and actions aligned with the 8-step model for leading change?

METHODS

This work was determined to be exempt by the University of Georgia Institutional Review Board (PROJECT00009085).

Data Collection

We interviewed 19 heads from 17 STEM departments at one research-intensive university to collect data about their leadership ideas and actions. We conducted two rounds of semi-structured interviews, with all 19 heads participating in the first round. Some heads had just joined the DeLTA project, and some had been involved for close to a year. We conducted a second round one year later with eight of the heads, to ensure that we had data from all 19 heads toward the end of their 1-year participation. The dataset therefore includes two interviews, conducted one year apart, for eight of the 19 heads. We made only minor changes to the interview protocol between the first and second rounds (see Appendix B). Questions were asked about what department heads viewed as their role in leading change, strategies they would use to lead change, and advice they would give for leading change to other heads. For example, one question was “What strategies would you use to lead the change? Why are [these strategies] useful?”. Another example is “Imagine you are giving advice to a department head who has been in place for about a year, has a feel for the job, and has identified a few changes they want to make to departmental practices. What advice would you give this person about how to lead the change they want to achieve? Why do you think this is important in achieving change?”

Qualitative Content Analysis

We aimed to identify the leadership ideas and actions that department heads shared as they discussed their plans for leading change or advice they would give to other change agents. We conducted structural coding (Saldaña, 2013), where quotes were organized into categories (codes) based on the broad ideas and actions department heads were sharing related to leading change in their departments. These codes were developed in vivo, although special attention was paid to quotes that were potentially related to the 8-step model. This was a highly iterative process: after roughly half of the interview transcripts had been coded, we examined all of the quotes within each code. We made sure that each quote belonged within that code, as well as checked that each code definition was clearly distinguished from the others. This led to the creation of new codes, as well as more precise definitions for each. We then re-coded all of the transcripts using the revised codebook, as well as coded the rest of the data. Finally, we examined each of the quotes in each code again, to ensure that they belonged in that code category, as well as the code being well-defined enough to capture all variation. Authors H.C.E., T.C.A., and K.S. performed all coding independently, discussing all disagreements to consensus. We finished the process by once again examining all quotes within each code. We then grouped these codes into broader themes reflecting the ideas department heads had on leading change to teaching evaluation and examined how these themes aligned with, or diverged from, the 8-step model.

Trustworthiness

Throughout the analysis process, we worked to ensure the trustworthiness of our data. The trustworthiness of an analysis is impacted by four different domains: Credibility, transferability, dependability, and confirmability (Anfara et al., 2002; Shenton, 2004).

Credibility relates to whether the research measures what it was intended to (Shenton, 2004). To strengthen the credibility of our findings, we used well-established qualitative research methods, conducted multiple interviews with several department heads to ensure we had similar data from all participants, and designed interview questions that approached the topic of leading change from multiple angles.

Transferability is whether the findings from one study can be applied to other projects (Shenton, 2004). To facilitate this, we provide a detailed description of our study's context and intervention, along with a discussion of how these factors may have influenced our findings. This allows for an understanding of our unique situation, to better contextualize our results.

Dependability deals with if the research was repeated in the same context with the same methods and participants, if similar results would be obtained, while confirmability deals with the objectivity of the work (Shenton, 2004). To improve both of these domains, we present our methods in detail, to enable replication. We used constant comparison to mitigate any biases coming from any one researcher. We also masked the identity of each department head while coding, to ensure that any prior knowledge or perceptions of the participants did not affect our interpretation of the data.

RESULTS

Department heads planned to lead change to teaching evaluation practices by building buy-in and addressing resistance within their units. Heads discussed a collection of strategies they would use to achieve these goals. Department heads anticipated resistance, and most considered resistance a normal part of any change process. For example, one head explained that "faculty don't like change; that is human nature." Heads described six ideas and actions they would use to lead change to departmental teaching evaluation practices. We describe these below, with quotes as supporting evidence, and then examine how these overlap with the 8-step model. Quotes are verbatim except minor edits for grammar and clarity.

Department Heads Sought Faculty Input

Department heads actively sought faculty input on new teaching evaluation practices to build buy-in and shape implementation plans. They sought faculty feedback about the teaching evaluation practices they would like to see happen, and how to implement them. They heard faculty perspectives in faculty meetings, executive committee meetings, and one-on-one interactions, and often sought input repeatedly.

Some heads invited faculty input early in the change process in order to avoid the perception that top-down changes were being imposed on faculty. One department head explained their rationale this way:

> “Faculty don’t like being told what to do...I’m obviously generalizing here, and this doesn’t apply to all faculty, but as a whole, faculty are very skeptical and cynical of [the] upper administration and leadership. Whenever something is imposed on them, there’s always resistance. And so if the faculty can develop things themselves, and they feel like it’s something they’ve developed, then they’re more likely to be receptive. Because it feels like it’s something that’s been developed by them or their peers rather than something they’re being told to do.”

This department head anticipated that involving faculty in the planning process would help foster a sense of ownership for changes, thereby mitigating resistance. Similarly, another department head described why they wanted to involve faculty this way:

> "It's important to give people a voice in changes... Let people know that they're being listened to, and have a say in shaping the process, and that their concerns aren't being ignored... I think just having that conversation helps with buy-in. It makes them feel part of the process... I think that makes people more willing or more open to change."

This department head expected more openness to the idea of changing teaching evaluation practices if faculty felt heard and had agency in decisions about new departmental practices and processes.

Department heads also asked for faculty input in order to learn what concerns faculty had regarding new teaching evaluation practices. They more commonly pursued this feedback by talking to faculty one-on-one or in small groups. One department head wanted to talk to as many faculty as possible and described why:

> "What's really important is that you, so to speak, walk the corridors. You talk to people. You give everyone the space and time to think about this issue and you explain [it] to them... So you almost want to purposely engage in conversation with as many faculty as possible about this issue, have a conversation, hear their concerns, listen to them, really listen to them. And ultimately maybe try to modify what you're doing based on what you're hearing."

This department head planned to seek out these conversations with faculty to hear their ideas and concerns and to have the chance to explain it to faculty. They expected to then modify their plans for implementing new evaluation practices based on those concerns. This would reduce resistance, as it would ensure that any major concerns were addressed quickly.

Additionally, department heads expected faculty to have useful ideas about how the department should advance teaching evaluation practices. One department head explained their process this way:

> "And so it's listening to what the faculty have to say, because [they] have great ideas and maybe [we'll] make use of them. As department head, I like to delegate in that regard, to get everybody's ideas on change. And then we implement it as a whole."

This head planned to use faculty's ideas to develop a better plan for how to change teaching evaluation in the department. Some heads planned for this to be an iterative process, where they would use faculty feedback to revise the plan, and then bring a new proposal back to faculty for more input.

Department Heads Strategically Proposed Changes to Faculty

Department heads were careful in how they presented ideas about changing teaching evaluation, with the goal of building buy-in. Most commonly, heads emphasized the benefits of the changes. Some heads planned to slowly introduce the idea of changes to evaluation practices, to allow time for faculty to warm to these ideas. Others planned to leverage the influence of forces outside the department to motivate faculty to pursue new evaluation practices. Finally, some department heads talked about how they would consider mandating the change as a last resort.

Department heads planned to or had emphasized the benefits they expected the department and faculty to experience as a result of new teaching evaluation practices (i.e., desirability of the change). Heads explained that they would "talk about the positive impacts" and "try to sell [the change] as something that could be beneficial for [faculty]." For example, one head shared their strategy:

> "I think [the change process should] be framed in a way for continuous improvement [of teaching], and a way of collecting information to help guide people in their professional development of their teaching... Talking about [the change process] in a faculty meeting, there's going to be some resistance. I fully expect that, but I think that faculty want to be evaluated fairly in their annual evaluation. I think [the change] would have to be presented in a way that [the faculty] could see [it] as being beneficial to them, as opposed to a way of criticizing them."

Heads, like this one, expected to mitigate resistance from faculty by highlighting how it would help, not hurt faculty. Another head saw their role as framing the change as something the department could do for their own benefit:

> "Part of that is just me being positive about [the change being] a good thing. This will be good for us. This will help us, instead of saying 'this is something we have to do because the upper administration wants us to do so, even though we do not want to do it. We have to suck it up.' Those are two very different approaches. And I think it's important for the department head to present the positive. 'We should do this for our own sake. And because it also satisfies something the upper administration wants.' That's just sort of an extra bonus."

Though new university policies asked departments to develop teaching evaluation practices that were new to many departments, this head did not expect faculty to respond positively to the argument that they make changes to respond to policy. Instead, the head thought faculty would be more motivated to make a change if they thought it would benefit them. One other head relied on their experience within the department to predict faculty resistance, and framed their arguments accordingly:

> "I've introduced a lot of changes to this department, and I have a playbook now, you know? I can anticipate where certain kinds of reactions will come, and what quarters it will come from. If I feel so moved, I can game that out ahead of time a little bit. Like faculty X is probably going to make this kind of point or raise this kind of an issue, and it is pretty predictable, so I might as well figure out how I am going to respond upfront. I have done it a bunch of times and I do not think it is going to be hard to appeal to their self-interest... My approach to this has been to continually evoke the collective interest. The fact that we are not individuals working independently with one another. We are part of a whole. Everyone needs to be rowing the boat."

This department head relied on pointing out how a change would benefit the department to mitigate the habitual resistance they anticipate from certain faculty. By anticipating where resistance was likely to come from, they were able to prepare arguments ahead of time about how the change would benefit everyone.

In the same vein, department heads felt it was important to proactively address faculty's concerns about the change. They felt that "hearing out people's concerns" was an important way to mitigate resistance. If they could "find out why they're resistant and see if [they] can mitigate those concerns," then more faculty would support the new practices. One head described why they felt this was an important process:

> "Faculty meetings can go disastrously. It just takes one person at the beginning to get up and start being aggressively against something. And then once people sense there's blood in the water, it's all over. So it is very important to get out and talk to people, particularly if the change is essential... It's important to not be surprised at that meeting by arguments against [the change] because if you are not prepared to address the arguments in the meeting, it is unlikely to be successful."

This head felt it was important to have heard faculty's concerns about a change, to both have a chance to talk through them with faculty, as well as prepare counterarguments. Otherwise, they were unlikely to be able to break through strong resistance.

Some department heads expected faculty to need time to form positive impressions about making changes to teaching evaluation. Heads discussed how faculty may not have had opportunities to think about robust and equitable evaluation practices previously, and also recognized that the topic of evaluation could feel threatening since it relates to job security. One head explained it this way:

> “I think the important thing in achieving change is that the majority of people are on board with [it]. I think people take their own time to realize that certain changes are a good thing... Some people are quite new to this and feel slightly threatened by a lot of these changes... It takes a little bit longer for some people to process and deal with, ‘Okay, I feel threatened, but that is sort of a little bit ridiculous’... I think that’s very understandable and I think it’s important to not back people into a corner. If you surprise them, they might say something publicly that they may regret later.”

This head anticipated that, given time, faculty would move beyond feeling threatened and come to appreciate the benefits of changes to teaching evaluation. They felt that faculty deserved time to consider the issue before a public discussion. Many heads planned to have low stakes discussions in faculty meetings early on, and then eventually “have a motion to approve [the changes].” This approach allowed faculty time to consider proposed changes individually and to consider the perspectives of peers before being asked to make any decisions.

Some department heads found it useful to leverage external forces to help convince faculty that a change was needed. For example, when asked what advice they would give to a new department head on how to lead change, one head responded with:

> “Use your upper administration where you can, and often the upper administration are great at being the bad guy. ‘We have to do something about this because the Dean is concerned about [this]. So we have to address this concern that the Dean [has] and we want to address this issue [ourselves] rather than the Dean getting so frustrated that he imposes a solution on us.”

Based on prior experiences, this head knew that their faculty preferred to initiate a change themselves, rather than waiting for a change to be imposed on them by their upper administration. The head felt that this approach could be used to convince faculty to act more quickly than they otherwise might. Another head discussed how they could use a new emphasis

on post-tenure review within the state university system to argue that the department should conduct peer observations of teaching for post-tenure review:

> "In terms of peer evaluations of teaching, we only do that for new faculty, [for] third year review and then when they're coming up for promotion. So [we don't do any] periodic peer review [of] teaching at the moment. [When I first became involved in this project] I had proposed to do [peer evaluations] at the time of the post-tenure review. But the view of many faculty was [that] a post-tenure review was just a box to check, and doesn't provide anything. But now with the new policies from the Board of Regents, I think it's time to reintroduce that idea to have at least a teaching evaluation done at every post-tenure review. That hasn't been approved by the department, but that's a plan that I would try to get going."

This head planned to use outside pressure (i.e., increased focus on post-tenure review in the university system) to re-open a discussion that had not gained traction with faculty previously. Some heads had not yet settled on a strategy, but had considered "the university changing the annual evaluation process" and "accreditation processes" as possible external forces they could leverage to motivate faculty to support changes to teaching evaluation.

While department heads mainly focused on building buy-in within their departments, some discussed how it might not be realistic to convince every faculty member to productively engage in new evaluation practices. In these cases, heads imagined that they might "mandate" the change as a last resort to force faculty participation. One department head talked about their opinion on this strategy, and when they felt it may be necessary to use:

> "Remember how I said it's not effective to pound your fist on the table? Sometimes it is. It should never be your first option, but you have to hold it in reserve because people can be very stubborn and rationalize the criticisms of their colleagues with all kinds of convoluted justifications that invalidate them. So there's a point where you do have to be able to say, 'This is what we're doing'... There are times when you have to call in

> everybody's cards and say 'This is the end. We are doing this.' I have only done it when I am supported by a strong consensus. We take a departmental vote, count the number of hands. So it's not like I am making it up."

This head felt comfortable asserting that a new practice would be implemented once most faculty had agreed on the change, even if not everyone agreed.

Department Heads Relied on Colleagues to Contribute to Change

Many department heads relied on departmental colleagues to lead the development and implementation of new teaching evaluation practices, and to help foster buy-in among other faculty. Most often, heads tapped multiple faculty for this work, though some relied on a single person. Some departments convened a new or *ad hoc* committee and others asked an existing committee related to teaching to lead the change. One head described their plan for engaging faculty in the process in this way:

> "I will form an ad hoc committee [to revise our teaching evaluation guidelines]. And there are people that I would focus on that I know are going to provide good input. I'm always a member of every committee in the department. And so I will charge a committee [to do this work], and we'll brainstorm on how we make changes. And then we present it to the whole faculty for further discussion, and then it might come back to the committee to implement whatever the faculty recommended for changes."

Heads planned to work with faculty to advance teaching evaluation in order to benefit from their expertise and influence within the department. Some department heads explained that they personally lacked important expertise about teaching and evaluation, and that others in the department were more knowledgeable about the subject. For example, this department head describes how delegating the work to other faculty made it possible to achieve relatively swift change:

> “I would say that we cheated because we have [two people] in our department. [One] is very conscientious and right in the middle of [campus teaching evaluation reform efforts]. And then [the other] has an enormous amount of teaching experience and experience with instructional issues... The two of them [were] the co-chairs of my instructional evaluation committee... It was so much easier for us to adopt [teaching evaluation reforms] because it wasn’t driven by me... I have to do nothing. [They] do all the work... I take extraordinarily small amounts of credit for anything that [my department] has been able to do in the last five years [because it] is really driven by them. They’ve done 90% of the work. I’ve been extremely fortunate as department head to have both the general faculty engagement that I have, and the assistance from really skilled and engaged faculty leadership on this issue.”

Department heads frequently lacked time that could be dedicated to developing and piloting new evaluation practices and could make more progress by empowering other faculty to lead the change.

Department heads also saw faculty as the most effective messengers about new teaching evaluation practices. They anticipated that the perspective of certain faculty would be especially convincing to their colleagues and strategically created opportunities for these “well-respected” faculty and “thought leaders” to share their perspectives with others. One department head spoke about this in detail when describing advice they would give a department head looking to lead change:

> “You have got to have at least a couple of ringers in the room who you have already talked to, who already are going to be supportive. Give them the talking points ahead of time so that when you start talking, they can chime in and be supportive... If you pick them well, if these are people who already have a strong and positive voice in the department, then that means anybody who would speak against them is sort of automatically putting themselves... on the other side from a colleague who in general is well respected in the department... Change is hard. The first thing people most usually think about when you present them with change is why it is going to be bad and difficult and maybe impossible. That’s just the way people are. So if you have multiple voices in the room speaking up and

> saying ‘No, I think this would be a good thing. I think we should try it. I think it would be positive,’ it sort of changes the tone in the room. And it prevents the discussion from devolving into all of them against you, which is unfortunately a dynamic that can happen really quickly if you don’t have those pre-identified supporters in the room.”

This department head ensured they had support from certain faculty ahead of time to help convince others. Some department heads also talked about how supportive faculty could help others see the feasibility of a change. They identified faculty who had already participated in the new practices, and asked them to share their experiences with the department. When asked about how they might navigate resistance during their change process, one department head said:

> “I’m a big fan of building consensus, and when [a faculty member’s] peers are participating in a process that maybe they haven’t fully bought into and [the participating faculty] can share their feelings that it’s beneficial... You know, students bring other students along, faculty bring other faculty along, so we basically help convince each other. That’s the way to go.”

This department head thought that by hearing their peers talk about their positive experiences with the new practices, resistant faculty members could be brought on board.

Department Heads Fostered Credibility and Confidence Regarding New Practices

Department heads used several strategies to help faculty feel more confident that the department would be successful in implementing new practices. Most commonly, they invited experts from outside of the department to present about what new practices might entail. Department heads also participated in the practices themselves in order to build credibility for their arguments in favor of new practices. Finally, a small number of heads planned to celebrate

the implementation of new practices, so faculty could see progress, and feel confident that it would continue to be successful.

Department heads relied on experts to help their faculty see the feasibility of changing teaching evaluation practices. They were able to leverage visits from DeLTA members to show faculty how feasible new practices could be. DeLTA offered to attend faculty meetings and host sessions about teaching evaluation change. These sessions typically started with a presentation covering the benefits of reforming teaching evaluation practices, and how that had already occurred in other departments at their institution, followed by ample time for questions. The department heads that invited these presentations felt that they were useful in getting faculty on board with making changes to teaching evaluation in their own departments. One head described why they felt this was helpful:

> "I feel that what helped [our change initiative] a lot were the two visits we had from [members of the DeLTA project] about [using multiple sources of evidence when evaluating teaching], and about peer evaluation specifically. We basically set aside two faculty meetings... These presentations were focused on [things] like how you actually do [peer evaluations], not the broad objectives [of how to make changes more generally.] And I think once people saw, 'Wow, you know, we're just going to do steps 1, 2, 3, and 4,' then it made a lot of sense. It's like [when] you go to a bakery, and you see a beautiful cake... and you think 'Oh wow! I could never do that,' [but] then you look at the recipe, and it's [just] 5 ingredients and 7 steps, and [you're like, 'Yeah,] I can do that.'"

The head felt these visits showed faculty how a change might look in their department, and how it had been successfully done in other departments. These visits made the change seem more feasible, which helped faculty feel more confident and empowered to participate in the new practices.

Some department heads planned to participate in new teaching evaluation practices themselves in order to lead by example. This most commonly looked like inviting a peer observer into their classroom to give feedback. Heads felt that "when someone can speak to their own personal experience, it brings credibility" to their arguments, and that leading by example "gives you some sort of authority on the subject" of new teaching evaluation practices. Heads aimed to convey the value they saw in new practices by participating themselves. One head described why they felt this was important:

> "I'm a firm believer in leading by example, and I fully expect to have a little group of my colleagues come in and evaluate my lectures... I think leading by example is effective in everything on the planet earth. I think leading by mandate, by pounding your fist on the table, generates cynicism and pushback. But if you say, 'Look, I'm a faculty member and I'm part of this, and I believe in it, and I'm going to do it,' then people respect that and it makes a more persuasive case."

By participating in the change, the department head is able to model how new practices will function, as well as "demonstrate some sort of buy-in on [their] own part," which may further convince other faculty to buy-in to the process.

Two heads planned to celebrate small successes in the process of changing teaching evaluation practices in order to build confidence that the department could achieve the changes they desired. One head explained their rationale:

> "Celebrating successes is important. If it's a long process, you're going to need some benchmarks. You need some early wins. I've tried to do that with some bigger changes, just chunk out early wins so that people are like 'Oh, we can do this!'"

This head planned to strategically divide larger changes into smaller chunks, so they would have successes to celebrate along the way. They saw early wins as a way to help convince faculty that the larger planned changes were feasible.

Department Heads Started with Smaller Changes

Department heads planned to start implementing new teaching evaluation practices with smaller changes to avoid resistance from faculty. This idea manifested differently between heads. Some planned to start by focusing on implementing only a handful of new practices, others planned to start enacting new practices with faculty who volunteered to be involved, whereas others invited a small group of faculty to participate in a pilot run of new practices. One head explained why they felt this would be the best approach:

> "I think that when there are big changes, it's good to set up a pilot year. Particularly if you think people really might resist them. Say 'it's not that we're changing this irrevocably, but we're going to do this for a year, and then we're going to take stock.'"

This head expected faculty to more readily accept the idea of trying something new on a trial basis rather than a permanent change to practices. Other heads similarly favored "rolling out something new in phases."

Heads also saw other benefits to piloting practices first, including the chance to try and refine new practices. One head elaborated on this idea:

> "You're not sure what will work and what won't. So starting to implement something, but being willing to change and modify it as you go along, rather than trying to presuppose the exact right way things will work and drop that all at once."

This head appreciated this approach because it allowed them to be flexible based on feedback from their faculty. It allowed them to address any concerns that arose during the process rather than "dropping it in everyone's lap and mandating it from the word go."

Department Heads Considered Culture Related to Teaching Evaluation

Department heads considered how teaching evaluation fit into the culture of their departments. They recognized that "change is very difficult to implement," and that just changing teaching evaluation practices would not be sufficient to ensure that they continued in the long term. Heads talked about wanting to see a "culture shift" so that the new teaching evaluation practices "become part of the normal order of business." One department head talked about why they felt this would be necessary:

> "You can't just put [the changes] in place overnight. I could come to my faculty and by edict say 'We are going to start doing this,' and it might stick so long as I stay in my position as department head. And then as soon as I'm gone, if the faculty didn't really buy into it, if it was always just a hoop jumping thing, then as soon as I'm no longer holding up the hoop, they will no longer jump through it... I watched faculty time and time again be forced to do something. They would kick and scream and complain and do everything they could to prevent doing it until they were ultimately forced by the administration to do it. And then they would only do it under protest. And as soon as the dean changed, or the department head changed, it all went away."

This head felt that without some sort of larger shift, any new practices were likely to return to what they were, after there was no longer any pressure to continue them. They went on to talk about how they aimed to incorporate their new practices into the culture of the department by helping faculty see the value of them, and by incorporating them into their "culture of practice." While department heads talked about how it would be important to shift culture, they did not discuss many strategies they may use to do so. A few heads had ideas about how they might change annual evaluation procedures to incorporate the new teaching evaluation practices. However, many heads did not talk about any concrete steps they intended to take to incorporate new teaching evaluation practices into the culture of the department.

In What Ways Did Heads' Ideas and Actions Overlap With and Diverge From The 8-Step Model?

The ideas and actions that department heads used to lead change to teaching evaluation overlapped with the steps of the 8-step model in some ways, but not in others. In this section, we describe how the ideas and actions department heads did and did not overlap with each of the steps of the 8-step model. Later, in the discussion, we discuss modifications that could be made to the model to tailor it to the context of departmental change in higher education and lessons that change agents and leaders could take from the model.

Department heads typically did not report that they tried to increase urgency (Step 1). Some discussed how they could leverage external forces to help motivate faculty, but none explained that they would intentionally try to convince faculty that current teaching evaluation practices were a problem that needed attention, immediate or otherwise. Heads may have had several reasons for not taking this approach. First, they may have assumed that the problems with teaching evaluation would be obvious to faculty. This may be a reasonable assumption because most departments use only student end-of-course evaluations to evaluate teaching (Ericson et al., 2025), which faculty tend to agree are inadequate (Brickman et al., 2016). Or, heads may have worried that introducing the idea of teaching evaluation reform by focusing only on problems would spark too much immediate resistance or apathy. This seemed to be the case for at least one department head, who discussed why they wanted to start the process by focusing on the "positive change" rather than the "negative problem:"

> "I think the key to getting faculty buy-in is starting with the motivation for the change that you want to make. So focus on the positive change first, and then on the negative problem that you are trying to solve second. If you ever start a conversation with faculty saying, 'this thing sucks, right? This is terrible. We do this badly,' then a certain subset of your faculty are going to get defensive. 'I've been in this department for 20 years. It's never been a problem before.' And then others are just going to say, 'Well, this is hopeless. You

> know, I don't know anything about teaching evaluation. How are we going to do this? It's just going to be more work. So why do we even have to?' So if you focus instead on what the positive benefits of doing it are: 'If we did this, it would be really positive.' Then you could say, 'and it solves these problems we have,' then that makes it seem like, wow, this is a positive thing. And it has a secondary good thing that is also positive because it solves this problem. I think that is always better. If you just walk in the room and say something is bad, you start out antagonizing a certain subset of people and demoralizing everybody else."

A department head with this sense of their department would sensibly opt to downplay the problem in favor of emphasizing the benefits of change. Yet, the 8-step model stipulates that organization members need to feel a sense of urgency to solve a problem in order to avoid complacency at the start and throughout the change process (Kotter, 2012). Given that the professional identities of faculty prioritize rationality and evidence, it is tempting to assume that efforts to help faculty visualize a problem and spark an emotional reaction would be unnecessary, condescending, or ineffective. Efforts to increase urgency may also require time, creativity, and possibly people who are willing to share their personal experiences. Thus, it is not surprising that department heads did not take this step, and it is also worth exploring what strategies could effectively foster a sense of urgency among faculty.

Department heads almost all recruited and relied on a team to guide changes to teaching evaluation practices (Step 2). Collectively, heads considered the skills and expertise that team members would bring to advancing teaching evaluation practices and how the credibility of the team could influence attitudes about the change. Heads did not describe the prior leadership experiences of the faculty to whom they delegated this work. Areas of consideration for future change leaders include convening a team with relevant expertise, influence, and leadership experience, who can be supported to build a trusting and functional team.

In the 8-step model, leaders work with a guiding team to craft a vision of the change (Step 3), and strategically and repeatedly communicate that vision to organization members (Step 4) to foster buy-in and productive engagement with the change (Kotter, 2012). Department heads took a different approach. Instead of working with a smaller team to craft a vision of a change they had already decided to make, most heads talked to their whole department to gather input and decide what changes to make and how. This difference in approach aligns with the culture of higher education, in which faculty are, to varying degrees, granted academic freedom in their teaching and research and expected to engage in shared decision-making within the institution (American Association of University Professors [AAUP]). Heads expected significant pushback on any new practice that the faculty had not had the chance to consider and critique and therefore began the change process by discussing potential teaching evaluation reforms with faculty. Often, they expected to have repeated discussions with faculty, one-on-one or in meetings, which aligns with the focus in the 8-step model on deliberate and repeated communication.

Though heads did not preemptively craft a vision of the desirable future that changing teaching evaluation practices could achieve, their ideas and actions about communicating with faculty aligned with key pieces of Steps 3 and 4 of the 8-step model. They emphasized the benefits of the change, aiming to make the change desirable and they tried to hear faculty concerns so that they could be addressed. They also took steps, such as inviting the DeLTA team to present in the department and participating in new practices themselves, to build confidence in the feasibility of changing evaluation practices. Lastly, they came to faculty with flexible ideas, ready to alter plans or head in a different direction based on faculty input.

Other aspects of an effective vision and communication of that vision may similarly have aided the change process for department heads. In particular, faculty may need help imagining what new teaching evaluations practices would look like, and the range of benefits they could provide. A little bit of time invested in clearly illustrating new potential practices and developing

specific, succinct talking points about the benefits might have served department heads in meetings and conversations and ensured their arguments in favor of new practices stayed consistent over time. This deliberate messaging could easily be paired with a process of gathering and using faculty input to determine the specific evaluation practices that the department would adopt.

Department heads attended to removing some barriers (Step 5) to faculty participation in new teaching evaluation practices, but not others. They used various strategies to address faculty concerns about feasibility, as well as helping them feel confident they could successfully implement the change. However, they did not attend to other barriers, such as significantly modifying departmental structures. This presents an opportunity for change agents, in that change efforts may be more successful if steps are taken to align departmental policies with new practices.

The majority of department heads did not strategically create short-term wins (Step 6) or discuss the importance of not letting up (Step 7). Only one head talked about how they would create opportunities for short-term wins in their plans. Other heads may rely more on these ideas as they progress in the change process, as many were just getting started. Creating short-term wins might be easy to add because department heads often planned to start with a small step, such as piloting new practices with eager faculty. If the head then created opportunities for faculty to notice and appreciate the positive outcomes of the pilot, this could be a powerful addition that fed enthusiasm for the new practices and set the stage for launching a practice for the full faculty.

Finally, some department heads discussed the importance of anchoring the change in the culture (Step 8), lest it disappear after their departure from the role, but may have lacked concrete ideas about how to do so. Without concrete strategies, heads may struggle to shift culture in their department towards valuing teaching and robust teaching evaluation. Strategies aligned with the 8-step model include continuous messaging about the benefits of new

practices, intentionally orienting new faculty to these practices and their importance, strategically inviting new faculty to take leadership roles in enacting the practices and prioritizing departmental leaders who embrace the change. Departments in this study had not reached a stage where most of those strategies made sense, but they could be useful later in their process.

DISCUSSION

This study aimed to transfer the 8-step model from organizational leadership into the context of higher education. Our analysis suggests that while the model aligns with some of the ideas and actions department heads used to lead changes to teaching evaluation, there are opportunities for leaders to learn from the 8-step model and to tailor the model to the context of higher education. Modifying the model can make it more specific to the context of departmental change in higher education and therefore more useful (Kezar, 2018).

Implications for Change Agents and Leaders

After systematically analyzing the ideas and actions of department heads and comparing them to the 8-step model, we see potential for the 8-step model to provide an overarching structure for leading change. For department heads who have not had many, if any, formal opportunities to learn about change leadership, the 8-step model provides a starting place for planning the process. It can provide guidance on what they will need to do first, as well as things they should keep in mind throughout the process. However, modifications must be made to help it better fit within the context of academia and departmental change.

For example, the step of increasing urgency (Step 1) can be reframed to emphasize widespread recognition of the problem that needs to be solved. Change agents and leaders could focus on eliciting faculty's critiques about the current system of evaluating teaching. Department heads could prompt these conversations with questions about what faculty felt

dissatisfied with current evaluation systems, or what faculty's experiences had been getting feedback from student end-of-course evaluations. These conversations would foster discussion that would bring out faculty's critiques of the current practices. By the end of the process, department heads could be confident that faculty understood the problems surrounding teaching evaluation, felt their concerns had been heard, and were ready to consider implementing changes. This achieves the same effect as increasing urgency, in that it builds buy-in for the changes. It more directly involves faculty, which department heads felt provided them with a greater sense of ownership over the process, leading to more buy-in.

More modifications could be made to the steps of getting the vision right and communicating the vision (Steps 3 and 4). Change agents could first develop a version of a vision that focuses more on why new teaching evaluation practices would be beneficial and feasible for the department. They could then bring it to the faculty and ask for their feedback on what changes to prioritize and what concerns they have. The vision could then be fleshed out to incorporate these ideas. This modified process puts more focus on developing a vision of the change before starting conversations with faculty, while still allowing for the important step of involving them in the planning process. A clear, cohesive vision is important for building buy-in and change agents may benefit from coming into conversations with faculty with one already developed.

Future Research

There is significant room for future research about leading change in higher education. Understanding the effectiveness of strategies used by department heads for leading change could help facilitate more successful change initiatives in the future. This study focuses on changes to teaching evaluation, but it is possible that heads would have other ideas and actions for leading other types of change in their departments. Future work could examine how department heads lead change in other areas, such as curricular reform, new approaches to

how workload is shared among faculty, and departmental restructuring or merging. Faculty are an important part of the change process, as they can serve as informal leaders that influence the opinions of their colleagues. Future work could focus on their role in the enacting change, and how informal leaders can be more effectively utilized. Department heads described the importance of factors external to the department in leading teaching evaluation change, and other scholarship emphasizes the important role of external pressures (Kezar, 2018). Studying how external factors influence change leadership could inform future change agents and efforts. Additionally, departmental culture around governance influences the change process. For example, some departments seek to reach complete consensus before making a change, whereas others delegate work to committees, and then vote to enact changes. Future work could examine the benefits or drawbacks of these approaches. It is possible that the first approach may foster more faculty buy-in, while the second may allow change agents to move more quickly. Finally, department heads have limited opportunities to engage in training about how to lead change (Gmelch et al., 2017). Some heads had more developed leadership ideas than others. Future work could investigate how heads developed their ideas about leading change, and how they translate them into action.

Limitations

The sample in this study provided more insight into heads' ideas and actions prior to, and in the early stages of leading change, than in the later stages of implementation. As a result, we have extensive findings regarding the latter stages of the 8-step model. Future work should prioritize samples with heads at more varied stages of the change process, or with heads that had already completed the implementation of their change. This would allow for more comprehensive characterization of the ideas and actions heads use throughout the change process.

What department heads shared in interviews may not fully represent the expertise they used and the actions they took while leading change. Given that department heads have few formal opportunities to develop leadership expertise, they may rely on tacit ideas about leadership. Meaning, they may have expertise for leadership that they may not be fully able to articulate. For example, department heads may take certain actions when leading a faculty meeting, demonstrating their leadership expertise, but then not share that expertise in an interview setting. Future research might benefit from incorporating other sources of data in addition to interviews, such as meeting recordings.

This study occurred within a single institutional context, and heads in different contexts may lead change differently. The institutional context for this study may have been especially conducive to teaching evaluation change due to support from upper administration and efforts by the DeLTA project to shift policies related to teaching evaluation (e.g., Andrews et al. 2021). A Presidential Task Force on Student Learning and Success in 2017 recommended that the systems for documenting and rewarding teaching be strengthened (Task Force on Student Learning and Success, 2017). This recommendation ultimately led to a university-wide teaching evaluation policy, enacted in 2022, that set the expectation that departments would develop processes for peer observation of teaching and systematic self-reflection on teaching. Relatedly, and around the same time, guidelines for the information faculty could provide to demonstrate teaching excellence for promotion and tenure were updated to expect evidence of effectiveness, evidence collected in ways that mitigate biases, and increased emphasis on continuous teaching improvement. Collectively, these policies and the administrative and faculty support that made them possible, created an institutional context that encouraged department heads to advance teaching evaluation practices. Nonetheless, departments retained considerable autonomy and agency regarding their evaluation practices and faculty could determine, from a long list, which evidence to include in their dossier regarding teaching excellence. It is

impossible to separate these heads' ideas and actions from this context and that should be considered in generalizing these findings to other contexts.

CHAPTER 4

TEACHING EVALUATION READINESS ASSESSMENT (TERA): DEVELOPMENT OF A TOOL TO MEASURE FACULTY READINESS FOR ADVANCING DEPARTMENTAL TEACHING EVALUATION PRACTICES[4]

[4]Ericson, H. C., Lemons, P. P., Dolan, E. L., Brickman, P., & Andrews, T. C.
To be submitted to a peer reviewed journal

ABSTRACT

How teaching is evaluated at many institutions is insufficient, and shifts are necessary to support, recognize, and reward good teaching. Faculty members must be on board with making changes to teaching evaluation in order for it to be successful. The aim of this study was to develop the Teaching Evaluation Readiness Assessment (TERA), which is intended to measure faculty's readiness for changing teaching evaluation practices in their departments. We utilized expert review, think-aloud interviews, and exploratory and confirmatory factor analysis to gather validity evidence to support this new instrument. The TERA is capable of detecting differences in change readiness over time, as well as between departments. The survey can be used as a tool for both change agents and researchers to assess faculty's change readiness in conjunction with teaching evaluation change initiatives.

AUTHOR CONTRIBUTIONS

I designed the instrument presented in this chapter in collaboration with Tessa Andrews. We collaboratively designed the interview protocol for the think-aloud interviews, and each conducted the interviews. I coordinated survey distribution, and conducted all analyses presented in this chapter, in collaboration with Tessa Andrews. Paula Lemons, Erin Dolan, and Peggy Brickman provided feedback on several versions of the instrument we designed, as described in the manuscript. I wrote the manuscript, and created all the graphs and figures present in this chapter. Tessa provided iterative feedback on the chapter.

INTRODUCTION

Many higher education institutions inadequately evaluate teaching (Brickman et al., 2016; Dennin et al., 2017). Teaching is often evaluated using only student end of course evaluations (Keig, 2000; Brickman et al., 2016). These evaluations do not provide reliable information because they can be biased against instructors based on their social identities (e.g., race, gender, native language), confounded by issues related to class type (Cashin, 1990; Ramsden, 1991; Greenwald & Gillmore, 1997; Bedard & Kuhn, 2008), and may not correlate with student learning outcomes (Bedard & Kuhn, 2008; Boring, 2017; Fan et al., 2019; Esarey & Valdes, 2020; Aragón et al., 2023). Furthermore, they often do not provide constructive feedback for instructors (Bouwma-Gearhart & Hora, 2016; Brickman et al., 2016). Evaluation that is more holistic has the potential to better support and reward effective teaching (e.g. Glassic et al., 1997; Smith, 2008; Bradforth et al., 2015; Lyde et al., 2016; Dennin et al., 2018). More holistic evaluation draws on multiple perspectives to provide evidence regarding teaching effectiveness, including students, trained peers, and instructors (Finkelstein et al., 2020; Weaver et al., 2020). Together these voices draw on a range of relevant experiences and expertise, and collectively they can mitigate the biases present in any one perspective. Teaching evaluation has the potential to provide feedback that instructors can use to improve their teaching and evidence that can be used to reward teaching effectiveness, but to realize this potential, higher education institutions and departments need to advance their teaching evaluation practices.

Reforming teaching evaluation practices can be a large undertaking and faculty and leaders often worry that their colleagues will resist such changes (e.g., Ericson et al. 2025). Yet numerous institutions and departments have taken steps toward better evaluation practices, creating resources and paths forward that other departments can then tailor to their unique context and needs (e.g., Finkelstein et al., 2020; Weaver et al., 2020; Krishnan et al., 2022). As departments consider changes to their evaluation practices, those supporting their work (e.g., educational developers, change agents) and those investigating change processes (e.g.,

change researchers) would benefit from a way to take the pulse of a department to determine their openness to new evaluation practices. This paper describes the development and utility of a survey instrument to measure faculty readiness to advance departmental teaching evaluation practices.

This survey is grounded in the readiness for change theoretical framework. Readiness for change is the extent to which an individual has positive beliefs and attitudes about the need for a change in their organization, and the ability of the organization to successfully achieve that change (Armenakis et al., 1993). In essence, someone who has a high level of readiness will be more likely to engage productively in new behaviors that align with the organizational change. Organizational change efforts often fail, and unproductive engagement is a common culprit (Kotter, 2012). In response, scholars in organizational management developed this theoretical framework to explain data about failed change efforts and to help leaders more successfully engage their employees in organizational change processes (Armenakis et al., 1993; Armenakis & Harris, 2002; Holt, Armenakis, Feild & Harris, 2007).

There are five components that encompass an individual's readiness for change (Rafferty et al., 2013). *Discrepancy* is whether the individual believes that change is necessary. Individuals must see that there is a problem that needs to be addressed within their organization (Armenakis & Harris, 2002). *Appropriateness* is the extent to which an individual feels that the change being proposed adequately addresses the problem they perceive. *Valence* is the individual's perception of the cost and benefits of implementing the change, for their organization (organizational valence) and for them as an individual (personal valence) (Holt, Armenakis, Feild & Harris, 2007). *Efficacy* is the individual's assessment of whether their organization is capable of effectively implementing the proposed change. Finally, *principal support* is whether the individual feels that the change would have sufficient support from organization leaders to be successful.

We developed the Teaching Evaluation Readiness Assessment (TERA) described in this paper to assess faculty readiness for changes to departmental teaching evaluation practices. We aimed for the TERA to be useful to departments and to researchers. For departments and those supporting them, we intended for results to provide an assessment of the overall readiness among faculty and how much faculty vary in their readiness. For researchers, we intended for the survey to distinguish between departments with more and less readiness for changes to teaching evaluation practices and to be capable of detecting shifts in readiness over time. Given that readiness is theorized to influence the success of change efforts, change agents and researchers may aim to foster readiness and measure changes in readiness. Though scholars have developed instruments to measure readiness for change in other contexts (see Holt, Armenakis, Harris & Feild, 2007 and Weiner, 2008), none exist for the context of higher education or for changes to evaluation practices. The goal of this paper is to describe the development of the TERA, to provide evidence of the validity and reliability of the results among STEM faculty, and to demonstrate the utility of the findings.

METHODS

This work was determined to be exempt by the University of Georgia Institutional Review Board (PROJECT00009085).

Context for Development

We developed this survey in the context of the DeLTA project, an institutional change project aimed at fostering departmental and institutional change in teaching evaluation practices and expectations (Andrews et al., 2021). DeLTA advocated for the three-voice framework, which includes the use of three voices (i.e., perspectives) to provide evidence of teaching effectiveness, including students, trained peers, and instructors (Finkelstein et al., 2020; Weaver et al., 2020). The student voice can provide important information about students' experiences

and outcomes (Reinholz et al., 2019). The peer voice involves trained peers providing feedback to instructors based on evaluations of their teaching and teaching materials, which leverages their disciplinary and teaching expertise (Thomas et al., 2014). The self voice involves the instructor reflecting on their teaching, analyzing evidence and their own observations, and providing critical context about their students, course, goals, and efforts to improve (Reinholz et al., 2019).

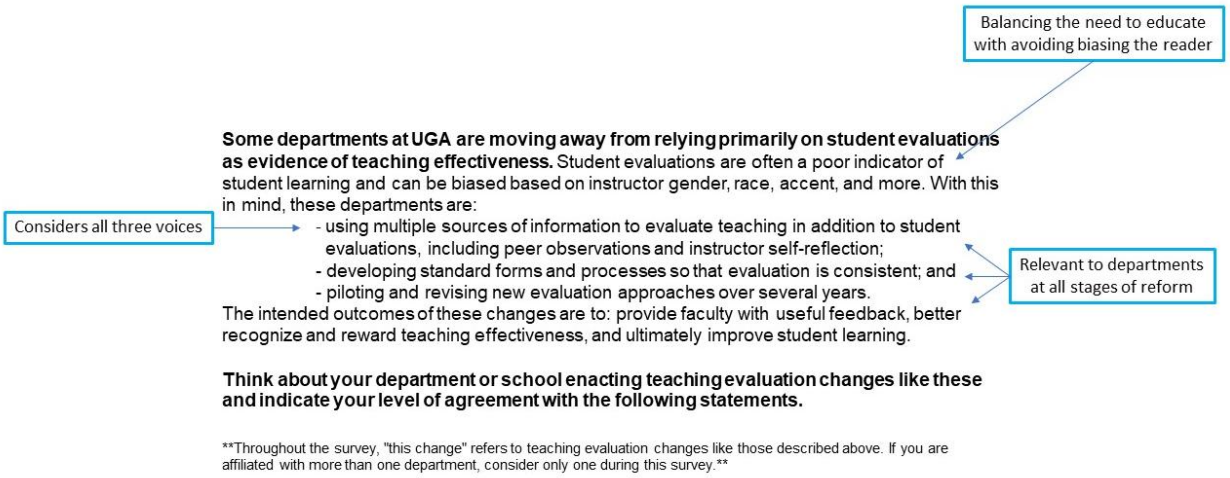
DeLTA worked with departments by engaging department heads and faculty in learning about robust and equitable teaching evaluation practices, developing and implementing new practices, and building readiness for new practices among departmental colleagues (Krishnan et al., 2022; Ericson et al., 2025). Departments participated in the project for one year, and piloted new evaluation practices either late in that year, or during the following academic year. DeLTA also worked at the university level by contributing to the development, approval, and implementation of new policies and practices related to teaching evaluation (e.g. reforming teaching evaluation policy to consider the use of the three voices). DeLTA contributed to revisions to the expectations for demonstrating excellence in teaching in promotion and tenure dossiers, which moved toward expected evidence of effectiveness, using strategies to mitigate bias, and valuing continuous improvement (Andrews et al., 2021). DeLTA also led the development of the first university-wide teaching evaluation policy, which set the expectation for departments to develop processes for each of three voices (Task Force on Student Learning and Success, 2017). Though departments do not have to engage in self-reflection or peer evaluation, according to university policy, departments must create robust and equitable structures to make these options available to faculty, and faculty can use the products of self-reflection and peer observation as evidence of effectiveness in promotion and tenure dossiers. These policies, and the administrative and faculty support that made them possible, make this context favorable for advancing teaching evaluation practices in departments.

We developed the survey to measure readiness in the departments participating in the DeLTA project, both so that we could report back to heads and faculty working to advance evaluation in their units and so that we could study whether participation in DeLTA increased readiness and how readiness impacted the change that departments achieved.

Survey Design

The TERA has two parts: (a) a description of changes to departmental teaching evaluation practices (Figure 4.1), and (b) items to assess respondents' degree of readiness for teaching evaluation change. Respondents read the description of change, which each item then refers to as "this change." The description encompasses changes to all three voices (e.g., student, peer, self) to minimize the chance that a respondent focuses on one voice about which they may hold strong positive or negative perceptions. One intended use of the TERA is as a pre- and post-intervention assessment of readiness, so it needed to be understandable and relevant to faculty who had not yet had the opportunity to consider teaching evaluation reform and to those who had engaged extensively in teaching evaluation reform. Thus, the change description summarizes shortcomings of the status quo (i.e., student evaluations), new practices that departments have adopted, and why they chose to do so.

Figure 4.1: Description of the change survey respondents read at the start of the Teaching Evaluation Readiness Assessment. The boxes and arrows highlight rationales underlying aspects of the description.



We developed the survey items based on the readiness for change framework and previously published work, including surveys of organizational readiness for change (Armenakis et al., 2007; Holt, Armenakis, Feild & Harris, 2007). These surveys were used because they were both written to assess readiness for a specific change, rather than readiness for change in general. They were also both written to assess changes that were organizationally relevant. Finally, they both underwent thorough psychometric analysis, unlike many other similar instruments (Holt, Armenakis, Harris & Feild, 2007; Weiner et al., 2008). We used these surveys as models when developing the TERA, tailoring previous items to the context of teaching evaluation reform in higher education and adding new items. Authors H.E. and T.C.A. generated the first draft of the description and items (final items in Table 4.1), aiming to encompass appropriateness, valence, efficacy, and principal support. We did not write items aligned with discrepancy because we wanted to focus the survey on positive aspects of teaching evaluation rather than problems. Appropriateness items dealt with the change making sense for the department. Valence items addressed both personal and organizational (departmental) benefits that could result from new teaching evaluation practices. These items referred to benefits in general and specific benefits. Efficacy items dealt with the departments' capabilities for making the change. Finally, principal support items addressed department

heads' support for the change. Items focused on heads because these leaders are key figures in making successful, lasting change in academic departments (Cervato et al., 20024; Ericson et al., 2025).

Table 4.1: Final Teaching Evaluation Readiness Assessment (TERA) item list. The items included here are the final items from the TERA. Items are organized by factor and include the component of the readiness for change framework they were originally intended to address. For the full, original list of items, see Appendix E.

Item	Intended component of readiness for change
Factor 1: Valence	
1. I think this change would be a worthwhile use of the department's time.	Appropriateness
2. It would make sense for my department to initiate this change.	Appropriateness
3. I think this change would have a favorable effect on teaching in my department.	Valence (Organizational)
4. This change would improve student learning in my department.	Valence (Organizational)
5. I think that my department would benefit from this change.	Valence (Organizational)
6. This change could improve my department's overall teaching effectiveness.	Valence (Organizational)
7. This change would prove helpful for student outcomes in my department.	Valence (Organizational)
8. I feel it would be worthwhile for me if the department made this change.	Valence (Personal)
9. I believe this change would provide me with better feedback on my teaching.	Valence (Personal)
10. I feel this change would lead to fairer evaluations of my teaching.	Valence (Personal)
Factor 2: Efficacy	
11. The expertise in my department is sufficient to accomplish this change.	Efficacy
12. My department would be able to make this change.	Efficacy
13. I am confident that my department could implement this change successfully.	Efficacy
14. I think my department is well-equipped to implement this change.	Efficacy
Factor 3: Principal Support	
15. I think the head of my department would take steps to support this change.	Principal Support
16. I think the head of my department would be in favor of this change.	Principal Support
17. I anticipate that the head of my department would encourage us to support this change.	Principal Support

The TERA asks respondents to use a six-point, ascending likert scale for each item, from left to right: Strongly disagree, Disagree, Slightly disagree, Slightly agree, Agree, and Strongly agree. This approach counters known response biases (Chyung et al., 2018). Respondents have a tendency to select options on the left side of the scale, while also tending to select options that are presented at the beginning of the scale (Hartley & Betts, 2008; Nicholls et al., 2006). In contrast, respondents tend to agree with the statements that are provided to them, and to select options that they perceive to be more socially desirable (Callegaro, 2008; Liu & Keusch, 2017). The use of an ascending scale places these biases in opposition to each other, with the negative options on the left side of the scale, and the positive items on the right. We did not include a neutral point, in an effort to encourage respondents to consider each item, instead of opting out by answering neutrally.

We improved the test content validity of the TERA by asking experts to evaluate the survey and revising based on their feedback. We iteratively refined the instrument multiple times throughout the design process. We gathered initial feedback from members of the DeLTA leadership team (P.P.L, E.D., P.B), who serve as change agents for departmental teaching evaluation reform, as well as from a larger group of biology education researchers at different career stages. The feedback we received allowed us to refine the instrument so that items better aligned with the readiness for change framework and the context of teaching evaluation reform in higher education (Reeves & Marback-Ad, 2016). For example, based on this feedback, we added the language to the change description about how departments were moving away from relying solely on student evaluations, to help respondents less familiar with teaching evaluation recognize why others felt motivated to pursue these changes.

We conducted think-aloud interviews to examine and improve the response process validity of the TERA with faculty (see Appendix F for full protocol). We conducted six interviews with STEM faculty strategically selected to represent a range of positions and departments. Their departments varied by discipline and in their progress on teaching evaluation reform.

Interview participants held positions as Lecturers, Assistant, Associate, and full Professors across five departments. We used these interviews to access the thought processes of faculty as they completed the TERA within Qualtrics. We asked them to read the change description and items aloud, sharing everything they were thinking. In addition to sharing their reasoning, interview participants sometimes shared feedback about alternative wording for items and the change description. We ended the interviews by asking participants concluding questions to understand any additional thoughts or responses to the survey they may not have yet articulated.

We iteratively revised the instrument based on participants' thoughts in these interviews. Following each interview, we revised based on feedback and then presented the revised version to the next participant. These revisions included fine-tuning phrasing that was confusing or unclear, clarifying the change description, and revising or deleting items that were not interpreted as intended. For example, several participants either expressed confusion or misinterpreted the item "There are legitimate reasons for us to make this change," resulting in its removal. Another change that we made as a result of these interviews was adding an "I don't know" option for each item. Respondents indicated that for several items they did not feel they knew enough to answer definitively and felt uncomfortable trying to make a judgement to answer the item. We continued this iterative revision process until we no longer received new feedback from participants.

Following the interviews, we collected final feedback from our project team about the instrument as a whole and made minor revisions. This feedback considered the content validity after revisions based on the think-aloud interviews. For example, we revised to reduce the appearance of repetition in the items aimed at a single factor, which multiple interview participants had noted. We also aimed to refine items so that they varied in their degree of agreeableness.

Survey Deployment

We deployed the TERA to faculty in departments involved in the DeLTA project before, during, or after their departments' involvement. This paper reports data collected in two academic years, about 20 months apart, in spring 2023 (n = 294 faculty in 15 departments) and fall 2024 (n = 498 faculty in 22 departments), with 10 departments surveyed at both time points. We invited all faculty with teaching appointments to complete the survey, which comprised the vast majority of faculty in each department. We asked respondents to report their home unit (i.e. department or school) and their current position (e.g., Assistant Professor, Lecturer, Academic Professional, etc.). Respondents represented a range of disciplines, though the majority were STEM, and a wide range of positions (see Appendix G).

We hosted the survey on Qualtrics and invited faculty to participate anonymously via personalized emails sent through the platform. We sent 2-3 reminder emails to faculty who had not yet participated, spaced roughly a week apart. The final recruitment email included a specific response rate goal for their department. For example, the final email might read "If just three more faculty in the Department of "X" complete this survey about teaching evaluation, your department will achieve a 75% response rate. Please be one of those three!". We also asked the heads of each department if they were willing and able to set aside 10 minutes in an upcoming faculty meeting to allow time for faculty to complete the survey, and/or otherwise encourage faculty to participate. Some heads were willing to do so, and in those cases, we sent text that they could use to introduce the survey to their faculty, or someone from our project team attended their meeting and introduced the survey. Response rates varied across departments, with a mean of 63% (SD = 16%), and a range of 30-92%. In departments where the head set aside time in faculty meetings, the average response rate was 60% (SD = 15%), whereas other departments had an average response rate of 53% (SD = 20%).

We intended to use this instrument to collect data at multiple time points and therefore wanted to be able to match responses from the same person. However, we also expected more

candid responses from an anonymous survey. To achieve both goals, we used two questions to generate an anonymous identifier for each respondent. We asked respondents “What day of the month were you born?” with a drop-down menu from 1 to 31, and “What are the first three letters of your mother’s first name?” We then used the answers to these questions to match responses from the same individual across time points.

Survey Analysis

We investigated the internal structure validity of the TERA by conducting exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). We started by checking the properties of our data, to determine if our data met the assumptions for conducting factor analysis. We looked for univariate or multivariate outliers, examined the factorability of the dataset, and checked for normality and multicollinearity, following the guidance of Knekta et al. (2019).

We started with an EFA, rather than a CFA, because our instrument included many new items and addressed readiness for change in a different context than prior instruments (Armenakis et al., 2007; Holt, Armenakis, Feild & Harris, 2007). We used survey responses collected during a department’s participation in DeLTA so that all respondents were experiencing a similar phase of teaching evaluation reform. We also limited these analyses to complete cases (i.e., no missing data). We divided the dataset into equal halves and used one half to conduct the EFA (n = 155) and one half for a later CFA (n = 155). We performed all analyses using R Statistical Software (v4.3.1, R Core Team, 2023). We conducted EFA via the psych R package (v2.3.5, Revelle, 2024), and CFA via the lavaan R package (v0.6-18, Rosseel, 2012).

Starting with exploratory factor analysis allowed us to determine the number of factors present in our data. We used parallel analysis and a visual inspection of the scree plot in conjunction with theoretical considerations to determine the number of factors to test (psych

package, v2.3.6, Revelle, 2023). We used a weighted least square (WLS) estimator to extract the variances from our data. We also used an oblique rotation method (oblimin), as theory and prior readiness instruments indicated that the different components were correlated. We then determined the best fitting models by examining the total variance explained by the factors, the pattern coefficients and communalities for each item and factor correlations.

We performed confirmatory factor analysis to confirm the results from the EFA, as well as to identify any items with sub-optimal fit. We used the robust maximum-likelihood estimation (MLR) to extract the variances from our data, as the data were both ordinal and nonnormal. We considered multiple fit indices to evaluate model fit (chi-square value from robust maximum likelihood estimation (MLR X^2), standardized root mean square residual (SRMR), root mean square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis index (TLI)). These fit indices represent absolute (X^2, SRMR), parsimony-adjusted (RMSEA), and incremental (CFI and TLI) fit indices (Bandalos & Finney, 2010). We also examined the Akaike's information criterion (AIC) for two models, to determine the best model (Akaike, 1987). Finally, we also calculated McDonald's omega (ω) (McDonald, 1999) as a measure of internal consistency reliability, as it is considered more robust than coefficient alpha (Hayes & Coutts, 2020; Kalkbrenner, 2023). We calculated McDonald's omega using the semTools (v.0.5-6, Jorgenson et al., 2025) package in R.

Demonstrating the Utility of the TERA

We intended for the TERA to be capable of discriminating between higher and lower readiness for change to teaching evaluation across departments and across time. We examined our data in a few ways to determine if the TERA could serve these purposes. First, we created graphs to visualize differences between departments. All graphs were created in R using the packages ggplot2 (v.3.5.1, Wickham et al., 2016) and patchwork (v.1.3.0, Pedersen, 2024). For each respondent, we averaged their responses to each item together, creating an average

factor score. We plotted the average scores for each respondent at both timepoints, along with a boxplot, for each factor (Figure 4.3). We also plotted the average scores for just the first timepoint for each department individually, along with a boxplot, for all three factors (Figure 4.4).

Second, we examined whether our data indicated any statistically detectable differences in readiness for change across time and departments. We used linear mixed-effects modeling to determine if there were differences in readiness for change, for any component, before and after DeLTA participation and among departments. We fit a separate model for each component of readiness for change, using the R packages lme4 (v.1.1-36, Bates et al., 2015) and lmerTest (v.3.1-3, Kuznetsova et al., 2017). For each factor, the model used was Factor Score ~ Timepoint + Department + (1|Respondent ID). The random effect of respondent ID was used to account for repeated measures.

We narrowed the datasets used in these analyses in a few ways. First, we maximized the data used for each model by including data from respondents who had answered at least three items for the focal factor. This criteria meant that we excluded responses that answered less than 100% of the items for principal support (n = 229), 75% of the items for efficacy (n = 245), and 30% of the items for valence (n = 254). Only 19 responses (7.5%) answered fewer than 70% of the items for valence.

Second, we limited these analyses to departments (n=6) that were involved in the DeLTA project during both survey timepoints. DeLTA engaged both department heads and faculty advocates and encouraged and supported work to advance teaching evaluation practices. We expected DeLTA to increase readiness in these departments because they were having more discussions about teaching evaluation, and the head and faculty advocates were learning about the benefits of new evaluation practices, how to enact those practices, and how to foster readiness among colleagues. Ultimately, some of these departments made extensive changes to teaching evaluation practices, while others did not.

RESULTS

Descriptive Statistics

Before performing our initial EFA, we had to determine if our data would meet the assumptions of factor analysis. We started this process by determining if there were any outliers in our dataset. We examined minimum and maximum values for each item (Table 4.2), as well as frequency histograms of the responses. This let us ensure there were no univariate outliers. Univariate outliers are responses that have an extreme value on one item (Tabachnick & Fidell, 2013). Mean values for the items ranged from 4.55 to 5.17. To test for multivariate outliers, we calculated the Mahalanobis distance for each respondent. The majority of responses (n = 298) had high Mahalanobis distance ($p < 0.001$), indicating they may be potential multivariate outliers. Multivariate outliers are responses where there is an unusual combination of scores on two or more items (Tabachnik & Fidell, 2013). We looked at each case to determine if respondents had answered the items in a concerning way that may indicate they were not considering each item. We found no reason to remove any of these responses.

Table 4.2: Descriptive statistics of the Teaching Evaluation Readiness Assessment items. This includes the mean response value, standard deviation, minimum, maximum, skewness, and kurtosis values.

Item	Mean	SD	Min	Max	Skew	Kurtosis
1	4.87	1.14	1	6	-1.39	2.10
2	4.95	0.98	1	6	-1.33	2.69
3	4.98	1.09	1	6	-1.48	2.77
4	4.75	1.05	1	6	-1.14	1.98
5	4.97	1.03	1	6	-1.55	3.21
6	4.83	1.07	1	6	-1.27	2.20
7	4.78	1.05	1	6	-1.00	1.21
8	4.99	1.10	1	6	-1.38	2.17
9	5.07	1.03	1	6	-1.51	2.93
10	4.90	1.10	1	6	-1.35	2.17
11	4.55	1.32	1	6	-0.90	0.17
12	4.74	1.01	1	6	-1.11	1.86
13	4.73	1.04	1	6	-0.98	0.98
14	4.57	1.12	1	6	-0.82	0.38
15	5.14	0.91	1	6	-1.61	4.44
16	5.17	0.79	1	6	-1.06	2.38
17	5.09	0.83	1	6	-1.08	2.65

We next examined the factorability of our dataset by calculating Kaiser's measure of sampling adequacy, as well as examining the inter-item correlation matrix. Factorability is whether there are strong enough correlations between the items in an instrument that the data would be able to be used for factor analysis (Tabachnick & Fidell, 2013). The Kaiser-Meyer-Olkin (KMO) measure tests for sampling adequacy for factor analysis, and values > 0.6 suggest good factorability (Tabachnick & Fidell, 2013). We used the psych package (v.2.3.6, Revelle, 2023) to conduct this test. The overall KMO value for our data was 0.935, suggesting the data would be suitable for factor analysis. We also examined the inter-item correlation matrix (see Appendix H), generated using the corrplot package (v.0.92, Wei & Simko, 2024). The matrix showed that all items that were intended to factor together had correlations above 0.66. All of the appropriateness and valence items also had correlations above 0.66 with each other, indicating that these items may factor together when running our factor analysis.

We then examined our dataset for normality and linearity. We started by assessing univariate normality, which can be measured by skewness and kurtosis for each item, using the psych package (v.2.3.6, Revelle, 2023). Skewness is related to the symmetry of the distribution of the data (Tabachnick & Fidell, 2013). Items with skewness less than |2.0| are considered sufficiently normally distributed (Bandalos & Finney, 2010). All items in the TERA had a skewness below |1.61| (Table 4.2). Kurtosis is related to the peakedness of the distribution of a set of data. A dataset can either be too peaked (having long, thin tails), or too flat (having short, heavy tails), neither of which is ideal for factor analysis (Tabachnick & Fidell, 2013). Values for kurtosis should also be < |2.0|, although some researchers suggest a more liberal cut-off of < |7.0| (Bandalos & Finney, 2010). About a third (35%) of items in the TERA had kurtosis below |2.0|, and all items had values below |7.0| (Table 4.2). The results of these tests indicated the items were sufficiently univariately normal. We also assessed the multivariate normality of the data, using Mardia's multivariate normality test, using the psych package (v.2.3.6, Revelle, 2023). The test showed significant multivariate skewness and kurtosis ($p < 0$ for both values),

indicating multivariate non-normality. We therefore used robust estimation methods to handle this non-normality in subsequent factor analyses.

Finally, we examined whether the dataset was multicollinear. Multicollinearity is when variables are too highly correlated (Tabachnick & Fidell, 2013). We again examined the inter-item correlation matrix, in addition to variance inflation factors (VIFs) and tolerance values from multiple regressions (implemented using the olsrr package, v.0.5.2, Hebbali, 2024). Problematic multicollinearity is indicated when the VIF is above 10, or tolerance values are less than 0.1 (Knekta et al., 2019). The highest inter-item correlation was 0.85, the highest VIF was 7.78, and the lowest tolerance was 0.129.

Exploratory Factor Analysis

We started our exploratory factor analysis by using several different methods to determine how many factors to retain. Visual examination of the scree plot indicated a leveling out at two factors, parallel analysis indicated three factors, and theory suggested that the items encompassed four factors. For completeness, we tested all three potential solutions.

The pattern matrices for these three solutions revealed important differences (Table 4.3). A pattern matrix represents the relationships between each item and the underlying factors. Pattern coefficients greater than 0.3 indicate that the item is strongly associated with that factor. In the four factor model, the valence items (OV and PV) did not behave as a single factor. A few items (OV1, OV2, and OV5) had pattern coefficients above 0.3 on the fourth factor, whereas other valence items had high pattern coefficients on the first factor, along with the appropriateness items (Table 4.3). This was not unexpected, as one of the instruments we drew on (Holt, Armenakis, Feild & Harris, 2007) found that some valence and appropriateness items factored together in their instrument. The three-factor solution aligned best with theory, with appropriateness and valence items factoring together, and efficacy and principal support items aligning with their own unique factors (Table 4.3). The two-factor solution retained the factor

with appropriateness and valence items and combined principal support and efficacy items in a single factor (Table 4.3). Each of the solutions explained similar proportions of the total variance (0.75 for four factors, 0.72 for three factors, and 0.65 for two factors).

Table 4.3: Pattern matrix from the four-factor, three-factor, and two-factor models. WLS 1/2/3/4 pattern coefficients for the first/second/third/fourth factor retained from the EFA using the WLS estimator are listed for each of the models tested. Pattern coefficients < 0.2 are not shown for clarity.

	4-Factor Solution				3-Factor Solution			2-Factor Solution	
Item	WLS1	WLS2	WLS3	WLS4	WLS1	WLS2	WLS3	WLS1	WLS2
1	0.82				0.78			0.85	
2	0.82				0.82			0.83	
3	0.23			0.52	0.68			0.74	
4				0.89	0.80			0.81	
5	0.78				0.82			0.89	
6	0.63			0.28	0.88			0.91	
7	0.36			0.54	0.83			0.87	
8	0.86				0.83			0.82	
9	0.76				0.89			0.84	
10	0.89				0.85			0.83	
11		0.76				0.75			0.73
12		0.71				0.74			0.78
13		0.89				0.90			0.90
14		0.87				0.87			0.83
15			0.83				0.83	0.39	0.37
16			0.97				0.95	0.45	0.47
17			0.86				0.85	0.43	0.35

We proceeded with the three factor solution because it aligned with theory and avoided item fragmentation (i.e., items intended to measure the same component split across multiple factors). We termed the factor that encompassed the appropriateness, organizational valence, and personal valence items as “Valence,” as the majority of items were originally intended to measure valence. All pattern coefficients were greater than 0.40, and no items had coefficients over 0.20 in any secondary factors. We then examined the communality and complexity values for each item to determine if any were not behaving as expected, and should be considered for removal (Table 4.4). Communality (h2) is the proportion of variance within the item explained by the factors (Tabachnick & Fidell, 2013). If this value is close to one, then more of the variance in the item is being explained by the solution (Knekta et al., 2019). Low h2 values indicate an item may be a candidate for removal. All h2 values were above 0.50, with the lowest being 0.52,

meaning none were considered for removal. Hoffman's index of complexity (com) describes the average number of factors necessary to explain that item. This value should ideally be 1.0, which means that exactly one factor would be necessary to explain the item (Knekta et al., 2019). All com values were either 1.0 or 1.1, which did not indicate that any items might be candidates for removal.

Table 4.4: Communality and complexity for the three-factor solution. h2 is a measure of communality. com is Hoffman's index of complexity.

Item	h2	com
1	0.74	1.1
2	0.73	1.0
3	0.58	1.1
4	0.62	1.0
5	0.81	1.0
6	0.82	1.0
7	0.72	1.0
8	0.68	1.0
9	0.71	1.0
10	0.67	1.0
11	0.52	1.0
12	0.65	1.1
13	0.84	1.0
14	0.74	1.0
15	0.72	1.0
16	0.91	1.0
17	0.77	1.0

Confirmatory Factor Analysis

We next conducted a CFA using the 3-factor model to determine if the model fit the second half of the dataset. Fit statistics indicated suboptimal fit (Table 4.5), and exploration of

factor loadings, correlation residuals, item correlation matrix, and modification indices pointed to two items with sub-optimal performance. One pair of personal valence items had a particularly large correlation residual (0.149), meaning that the observed correlation between these items was notably higher than what the model predicted, suggesting potential redundancy or shared variance not accounted for by the factor. Additionally, the modification index, which is a list of modifications that can be made to the model which would improve its overall fit, listed adding a term that correlated the errors between these two items as the highest impact modification. Given these data and the fact that the valence factor included 10 other items, we removed these two items from the instrument.

Table 4.5: Fit statistics for the CFA models.

Model	df	X^2	RMSEA (90% confidence interval)	SRMR	CFI	TLI
Original 3-factor model	149	497.497	0.125 (0.111 – 0.140)	0.056	0.898	0.883
Modified 3-factor model	116	369.136	0.121 (0.104 - 0.138)	0.058	0.916	0.901

This modified 3-factor model demonstrated high factor loadings and substantial variance explained by each factor, indicating that the items effectively measured their intended factors. The model has universally high factor loadings, with 0.789 as the lowest (Figure 4.2). For each factor, the average amount of variance explained by the items surpassed 75% (Bandalos & Finney, 2010), with valence explaining 79.1%, efficacy 75.7%, and principal support 79.7%. The efficacy and principal support factors had the highest correlation, and each had moderate correlations with valence (Figure 4.2). We also calculated McDonald's omega for each factor. Omega values, which indicate the internal consistency of a factor, a measure of reliability, were 0.97, 0.92, and 0.92 for valence, efficacy, and principal support, respectively.

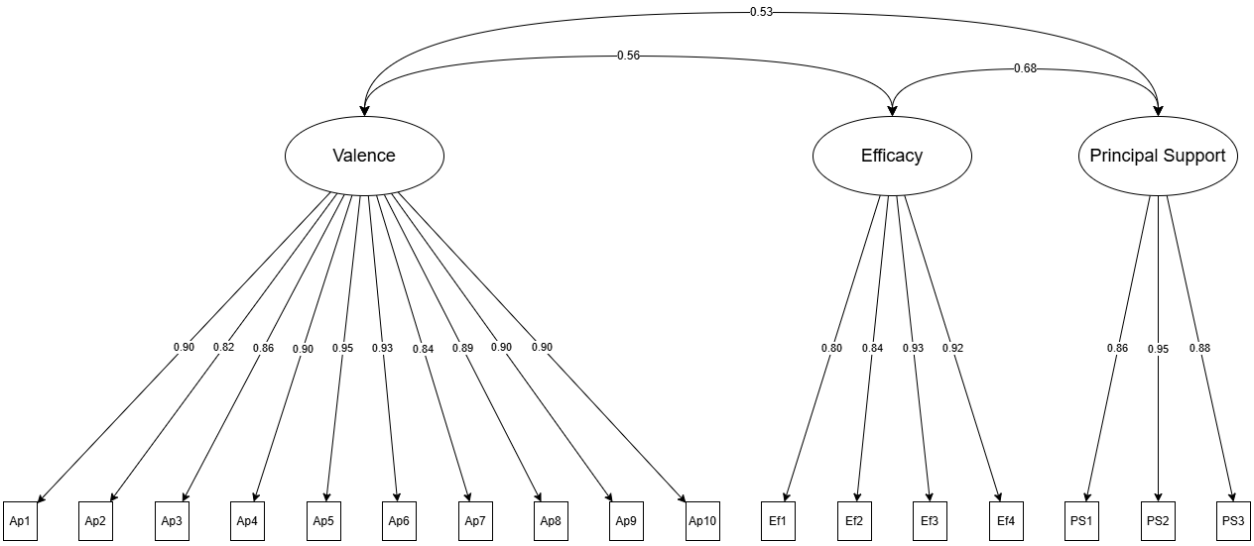


Figure 4.2: Results from the final three-factor CFA model. Survey items are represented by squares and factors are represented by ovals. The numbers on the double headed arrows represent the correlations between factors. The numbers on the single headed arrows represent the standardized factor loadings.

The modified model had marginally better fit than the original 3-factor model (Table 3). The modified model had a lower AIC value than the original model (5366.6 vs 6036.4). AIC values are used to evaluate model fit between several models, with lower values indicating better fitting models (Akaike, 1987). The modified 3-factor model exceeded the cutoff values for SRMR, but not those for RMSEA, CFI, nor TLI (Table 5). For a model to be considered a good fit, researchers have recommended that the RMSEA index or its 90% confidence interval should be < 0.05 for a well-fitting model, or < 0.08 for an acceptable one, the SRMR index should be <.08, and the CFI and TLI indices should be > 0.95 (Hu & Bentler, 1999; Bandalos & Finney, 2010). Many researchers, including those who recommended the original cutoff values (e.g., Hu & Bentler, 1999), recommend against adhering strictly to the values, as it might lead to incorrectly rejecting acceptable models (Marsh et al., 2004; Perry et al., 2015; McNeish & Wolf,

2021). Given the strong construct validity of the instrument and the benefits for users of a simpler model, we moved forward with the modified 3-factor model.

Demonstrating the Utility of the TERA

Results from the TERA survey in our context could detect changes in readiness over time, and distinguish among departments with higher and lower readiness (Figure 4.3, Figure 4.4, Table 4.6). We used three different datasets, one for each factor, to conduct these analyses, in order to maximize the amount of responses in each. On average and across departments, faculty reported higher efficacy at the second time point compared to the first, whereas their perceptions of valence and principal support did not differ between these time points (Figure 4.3, Table 4.6). This suggests that involvement in the DeLTA project coincided with increases in confidence among faculty that their department could successfully change their teaching evaluation practices. However it did not impact the value faculty saw in changing teaching evaluation, nor their perception of their department heads' support of these changes.

Departments differed in their readiness for changing teaching evaluation practices (Figure 4.4, Table 4.6). The results of the mixed model indicate that three departments (3, 5, and 6) reported higher efficacy than Department 1 (Table 4.6). Departments similarly differed in their perceptions of principal support, with Department 3 and 6 reporting higher values than Department 1 (Table 4.6). Thus, even when we only compare departments to one reference department (Department 1), the results discriminate between higher and lower levels of efficacy and principal support.

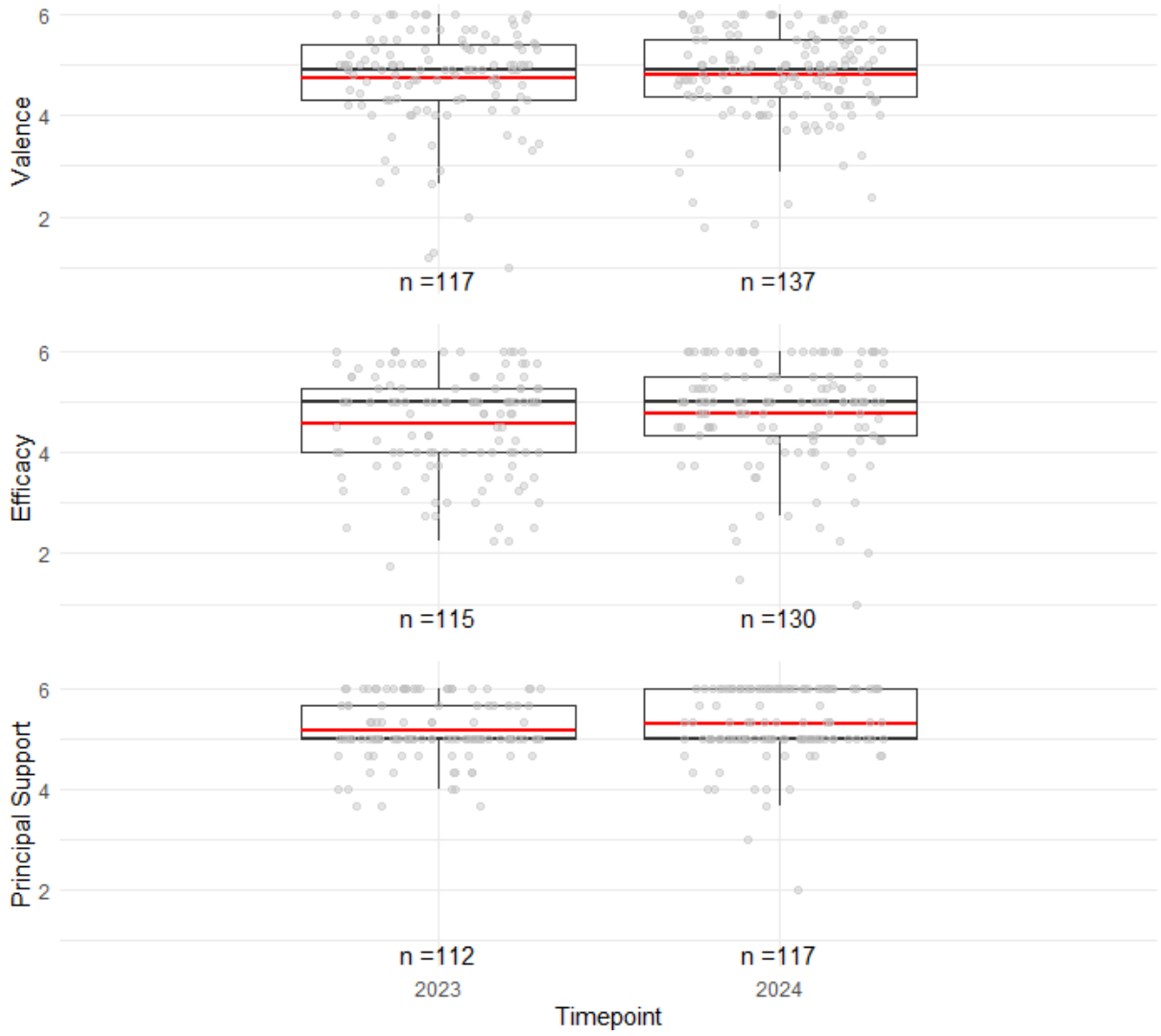


Figure 4.3: Readiness for change in six departments at the first and second time point, by factor (i.e., Valence, Efficacy, Principal support). The first timepoint was spring 2023, with the second being roughly 20 months later, in fall 2024. Black boxplots indicate the median (thick line), the interquartile range (IQR) from the first quartile to the third quartile (the box itself), and whiskers extending 1.5 times the IQR. Red lines indicate the mean. Grey dots represent the average factor score for each respondent, with n below the plot. Sample size varies among factors as we limited analyses to respondents who responded to 3+ items per factor. The average departmental response rate was 71.5% (SD = 14.0) and ranged from 53% to 84%.

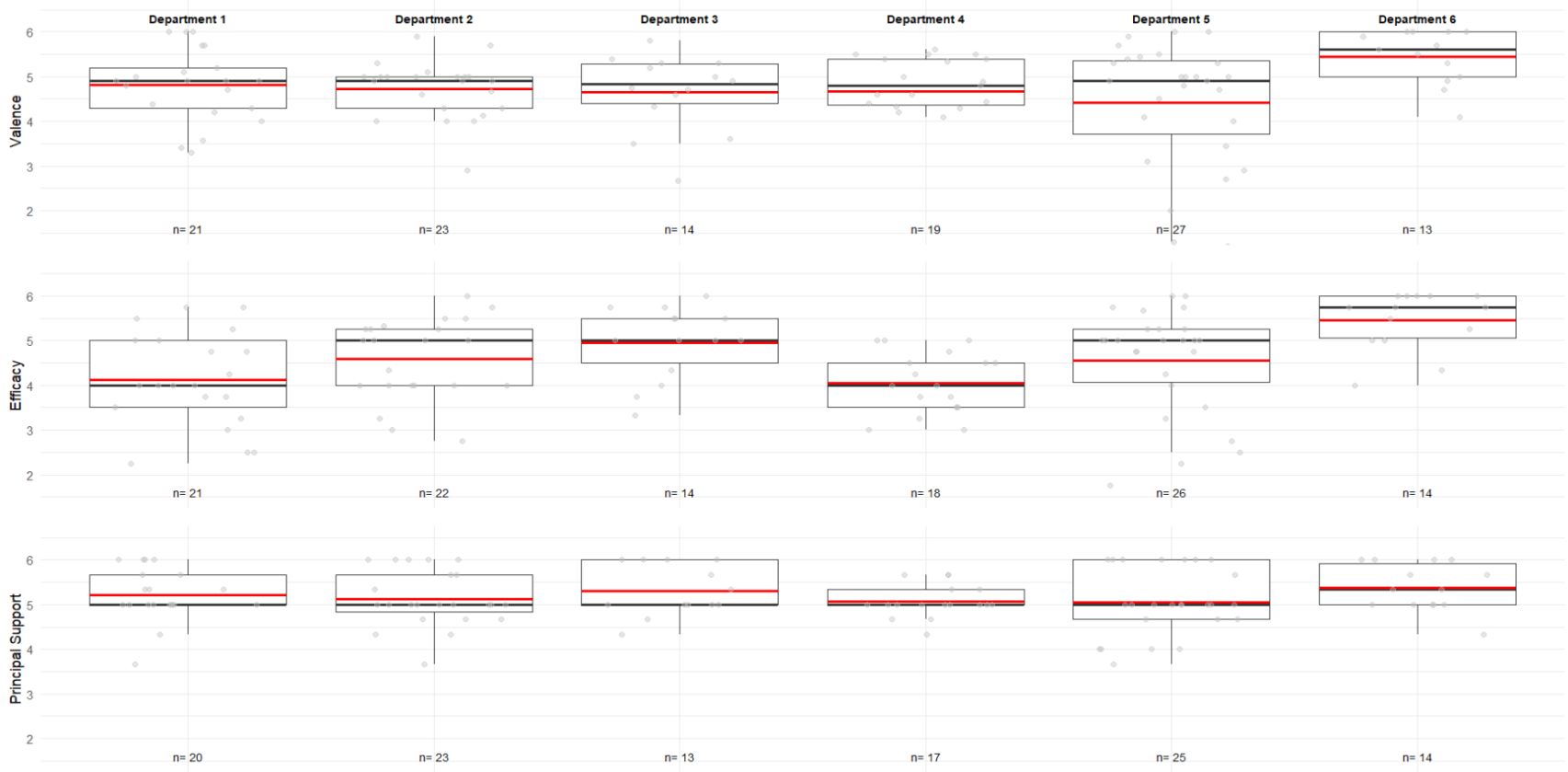


Figure 4.4: Readiness for change in six departments at the first time point, by factor (i.e., Valence, Efficacy, Principal support). Black boxplots indicate the median (thick line), the interquartile range from the first quartile to the third quartile (the box itself), and whiskers extending 1.5 times the interquartile range. Red lines indicate the mean. Grey dots indicate the average factor score for each respondent, with n below the plot. Sample size varies among factors because we limited analyses to respondents who responded to 3+ items per factor. The average departmental response rate for these data was 71.5% (SD = 14.0) and ranged from 53% to 84%.

Table 4.6: Results from the linear mixed-effects models examining the association between factors of readiness and time (2023 as reference) and department (Department 1 as reference). Estimates represent the fixed effects and significant effects denoted in bold.

Predictor	Estimate (β)	SE	df	t-value	p-value
Valence					
(Intercept)	4.84	0.19	188.08	25.09	**<0.001**
Time (2024)	0.07	0.08	84.22	0.86	0.39
Department 2	-0.24	0.24	156.36	-1.01	0.32
Department 3	-0.10	0.27	153.99	-0.37	0.72
Department 4	-0.16	0.25	156.56	-0.65	0.52
Department 5	-0.38	0.23	160.97	-1.68	0.09
Department 6	0.44	0.27	162.50	1.64	0.10
Efficacy					
(Intercept)	4.14	0.20	187.46	20.71	**<0.001**
Time (2024)	0.21	0.10	93.82	2.21	**0.03**
Department 2	0.12	0.26	158.64	0.47	0.64
Department 3	0.50	0.25	156.73	2.03	**0.04**
Department 4	0.27	0.24	165.41	1.14	0.26
Department 5	0.60	0.27	153.19	2.20	**0.03**
Department 6	1.24	0.27	160.56	4.52	**<0.001**
Principal Support					
(Intercept)	4.80	0.15	178.63	32.96	**<0.001**
Time (2024)	0.08	0.07	82.92	1.13	0.26
Department 2	0.34	0.19	149.02	1.79	0.08
Department 3	0.36	0.18	147.73	2.02	**0.04**
Department 4	0.10	0.18	153.49	0.58	0.56
Department 5	0.32	0.20	145.69	1.63	0.11
Department 6	0.56	0.20	150.43	2.86	**0.01**

DISCUSSION

The Teaching Evaluation Readiness Assessment (TERA) provides evidence of the degree to which faculty: (1) see value in moving towards teaching evaluation that relies on multiples voices and best practices, for themselves and for the department; (2) feel confident that the department can successfully change their practices; and (3) think that the department head supports these changes. It is the first tool for measuring readiness for change within higher education, and the only measurement tool within higher education focused specifically on teaching evaluation change. This theoretically- and empirically-grounded tool has the potential to provide useful information for those undertaking and leading changes to evaluation practices, and change researchers. We elaborate on potential future uses of the TERA, relate our findings back to theory, and discuss limitations.

Recommendations for use of the TERA in change work

Faculty and other change agents working within departments, or in collaboration with departments, can use the TERA to glean insights about readiness and to assuage concerns about widespread resistance. By examining average scores, variation across factors, and responses to specific items, one can identify areas where faculty are in agreement, as well as areas that might need more attention. Responses to the TERA could help change agents strategize about what is needed in the departments they are working with to be successful. For example, if someone found they were championing change in a department with lower valence, they could strategically build valence by asking faculty who had experience with the types of changes to teaching evaluation they were looking to implement to share their experiences. By hearing about the benefits that other people experienced, faculty could see how the change would benefit them and their department (Ericson et al., in prep). For departments with lower efficacy, change agents could recruit experts from outside the department to give a presentation about the changes to faculty. Hearing about what exactly new teaching evaluation practices would require can make faculty feel that changes would not be overly burdensome, leading to them feeling more confident the department would be able to implement the changes successfully (Ericson et al., in prep). Finally, in departments with lower principal support, change agents may consider participating in the new practices themselves, showing faculty their commitment to the new practices (Kotter, 2012, Ericson et al., in prep).

The results of the TERA can also assuage concerns about widespread resistance. Faculty in our context had high valence (Figure 4.3), meaning they agreed that changing teaching evaluation practices would have benefits for them and their departments. It was also true that every department head and faculty leader worried about resistance from departmental colleagues (Ericson et al., 2025). This same concern comes up every time we discuss teaching evaluation change outside of our institution. This suggests that resistant voices may be louder and more memorable, resulting in an outsized sense that the department is resistant, when in

fact most faculty feel ready for changes to teaching evaluation practices. The results of the TERA can provide evidence to counter concerns about resistance.

In the DeLTA project, we shared the TERA results with department heads and faculty leaders, and discussed the findings in a meeting. We created simple reports that provided stacked bar plots of the average responses to each factor, summarized the factors, and provided an overview of the findings (see example in supplemental material). Typically, faculty expected more resistance than the data indicated, and seeing evidence of readiness helped empower them to pursue substantial changes to current practices.

Using the TERA for Research

Researchers can use results from the TERA to identify differences among departments and across time, and to advance our knowledge of the role of readiness in change efforts. The TERA could be a useful tool in determining the effectiveness of an intervention focused on changing teaching evaluation practices, as it could be used to measure shifts in readiness for changing teaching evaluation. This includes investigating whether readiness within a department before the start of the intervention is predictive of the amount of change a department achieves, as is predicted by theory (Armenakis et al., 1993). The instrument could also be used to investigate how readiness for change shifts over time when a department is involved in shifting teaching evaluation practices.

Relating back to Readiness for Change Theory

The readiness for change framework emerged from scholarship in organizational management, a field that studies businesses rather than higher education. Thus, we must consider which aspects of the theory transfer usefully to the context of higher education, which lose relevance in this new context, and which might be adapted based on insights within higher education. The TERA is unique from other instruments assessing readiness for change in that it

was designed to be change specific. Most instruments focus on readiness for change in general, which is not always ideal, as readiness in an organization can vary depending on the change (Weiner, 2009). A department may be ready to change teaching evaluation practices, but they may not be ready to engage in curriculum reform, necessitating an instrument specific to the focal type of change.

In developing the TERA, the readiness for change framework helped us operationalize two concepts that department heads and faculty consistently discussed: resistance and buy-in. Every head who participated in DeLTA anticipated some resistance or push-back from faculty, and could imagine the specific concerns faculty might raise (Ericson et al., 2025). They saw buy-in as a counterpart to resistance, and aimed to address resistance and build buy-in. And though they, and the DeLTA team, had a range of ideas about how to address resistance and build buy-in (Ericson et al., in prep), the readiness for change framework formally operationalizes buy-in as "readiness," and defines its components (i.e., discrepancy, appropriateness, valence, efficacy, principal support). Mitigating resistance, then, involves fostering readiness. As change agents, we found this operationalization useful for our own planning to support heads and faculty in leading change, and we also introduced the framework to these leaders, so that they could think more specifically about what it might take to foster buy-in.

The TERA addresses just three factors of the readiness for change framework. We found that the items we wrote for appropriateness and valence factored together. This was not entirely unexpected as one of the instruments we drew on (Holt, Armenakis, Feild & Harris, 2007) found that their discrepancy, organizational valence, and some personal valence items factored together. As faculty did not distinguish between personal and organizational valence when responding to the TERA, it is possible that they differentiate between what would be beneficial for themselves personally, and what would be beneficial for their department. They may have also not distinguished between what would be beneficial for their department

(organizational valence), and what was important or worthwhile for their department to do (appropriateness), resulting in both components factoring together. The TERA excludes discrepancy because we opted to focus the survey on the positive aspects of teaching evaluation rather than the problems. Including items that addressed each potential issue with teaching evaluation would have been a challenge, as each department has its own idiosyncratic teaching evaluation practices, and faculty all perceive different problems with teaching evaluation. Additionally, when talking about teaching evaluation change, department heads tend not to focus on the problems inherent in the current practices, instead focusing on potential new practices, to avoid faculty resistance (Ericson et al., in prep). To avoid prompting this kind of resistance, we did not include items aimed at discrepancy. Future work is needed to explore the role of discrepancy in teaching evaluation change. One way to approach this would be to write a set of discrepancy items to add to the TERA and then re-validate the survey.

Limitations

We invite readers to consider the limitations of the TERA and the development and validation processes when considering how this tool could be useful in other contexts. One potential limitation of deploying the TERA is that faculty with less interest in teaching or teaching evaluation may be less likely to complete the TERA when invited to do so. This could result in overestimating readiness in a department because those most willing to participate might be most willing to complete the survey. We included individual data points in Figures 4.3 and 4.4 to demonstrate that individual answers varied considerably in our sample, suggesting that respondents were not limited to only the most ready faculty. Nonetheless, we observed that faculty who only responded to the TERA once were more likely to report lower readiness than faculty who responded at both timepoints, though those reporting lower readiness were about equally represented at each time point. Trends in which faculty are mostly likely to respond may affect the accuracy of estimates of readiness, and we encourage users to take measures to

increase response rates as much as possible. We tried to do this by asking department heads if they would be willing to allocate time in a faculty meeting to the survey, or to send an email to their faculty asking them to participate. This raised the average departmental response rate by 7%, but it remained challenging to recruit the vast majority of faculty as respondents.

The data collected to date with the TERA are from a single research-intensive institution. Institutional context may influence how faculty perceive teaching evaluation reform and change more broadly. In acknowledgement of this limitation, we described the context in which this work took place. We also intentionally developed the instrument to be accessible to those more and less familiar with teaching evaluation reform. We expect that the TERA can produce valid results at other research-intensive institutions, though this should be tested in future studies. Faculty in institutions with different organizational structures, such as faculty unions or college-level evaluation policies, may interpret the change description and/or some items differently. We encourage future users to consider this before deploying the TERA on a broad scale.

Finally, including the option for respondents to select “I don’t know” for each item increases the frequency of missing data, while also more accurately representing faculty thinking regarding some items. While the majority of respondents only selected this option a couple of times, some faculty chose this option for many items, and we ultimately had to exclude some or all of their data. When possible, we maximized the data included by making exclusions by factor, while only using cases with no missing data for the EFA/CFA.

CHAPTER 5

CONCLUSION

Collectively, this dissertation adds to the literature surrounding the process of leading change in academic departments. Chapter 2 investigated department head readiness for changing teaching evaluation. It characterized how the amount of change that was achieved by department heads was correlated with their level of readiness. Chapter 3 characterized the ideas and actions department heads used to lead change to teaching evaluation in their departments. We transferred the 8-step model for leading change (Kotter, 2012; Kotter & Cohen, 2012) into the context of academic departments. We then investigated how the strategies department heads used did or did not overlap with the model. Finally, Chapter 4 described the development of an instrument to assess faculty readiness for changing teaching evaluation practices. We described the validity evidence we collected for the instrument, and how it could be utilized by change agents and researchers. Our hope is that this work provides useful insight and guidance for anyone looking to undertake their own change processes.

These studies highlight the importance of understanding the factors at play when leading change in an academic department. However, there is still work to be done to more fully understand everything that impacts change processes. All studies presented here were done at just one research-intensive institution. Future work should investigate what change processes look like at other kinds of academic institutions, or ones with different governance structures, such as faculty unions, or college-level evaluation structures. These studies also centered only around leading change to teaching evaluation practices. This raises the question of what change processes look like for other types of changes (such as curriculum reform or departmental restructuring), and whether department heads and faculty approach these

differently. Finally, future work could investigate the impact of factors outside of the department. Things such as university level policies or accreditation practices have the potential to be hugely impactful to change efforts.

REFERENCES

American Association for the Advancement of Science (2011). Vision and change in undergraduate biology education: A call to action, Washington, DC.

American Association of University Professors. *Statement on Government of Colleges and Universities*. Retrieved March 16, 2025, from https://www.aaup.org/report/statement-government-colleges-and-universities

Andrews, T. C., Brickman, P., Dolan, E. L., & Lemons, P. P. (2021). Every Tool in the Toolbox: Pursuing Multilevel Institutional Change in the DeLTA Project. *Change: The Magazine of Higher Learning*, *53*(2), 25-32. https://doi.org/10.1080/00091383.2021.1883974

Andrews, T. C., & Lemons, P. P. (2015). It's Personal: Biology Instructors Prioritize Personal Evidence over Empirical Evidence in Teaching Decisions. *CBE—Life Sciences Education*, *14*(1), ar7. https://doi.org/10.1187/cbe.14-05-0084

Andrews, T. M., Leonard, M. J., Colgrove, C. A., & Kalinowski, S. T. (2011). Active Learning Not Associated with Student Learning in a Random Sample of College Biology Courses. *CBE—Life Sciences Education, 10*(4), 394-405. https://doi.org/10.1187/cbe.11-07-0061

Anfara Jr, V. A., Brown, K. M., & Mangione, T. L. (2002). Qualitative analysis on stage: Making the research process more public. *Educational researcher*, *31*(7), 28-38.

Angelo, T. (2000). Transforming departments into productive learning communities. *Leading academic change: Essential roles for department chairs*, 74-89.

Aragón, O. R., Pietri, E. S., & Powell, B. A. (2023). Gender bias in teaching evaluations: the causal role of department gender composition. *Proceedings of the National Academy of Sciences*, *120*(4), e2118466120. https://doi.org/doi:10.1073/pnas.2118466120

Armbruster, P., Patel, M., Johnson, E., & Weiss, M. (2009). Active Learning and Student-centered Pedagogy Improve Student Attitudes and Performance in Introductory Biology. *CBE—Life Sciences Education, 8*(3), 203-213. https://doi.org/10.1187/cbe.09-03-0025

Armenakis, A. A., Bernerth, J. B., Pitts, J. P., & Walker, H. J. (2007). Organizational Change Recipients' Beliefs Scale: Development of an Assessment Instrument. *The Journal of Applied Behavioral Science*, *43*(4), 481-505. https://doi.org/10.1177/0021886307303654

Armenakis, A. A., & Harris, S. G. (2002). Crafting a change message to create transformational readiness. *Journal of organizational change management*, *15*(2), 169-183.

Armenakis, A. A., Harris, S. G., & Mossholder, K. W. (1993). Creating readiness for organizational change. *Human relations*, *46*(6), 681-703.

Augustine N. R. (Chair). (2005). Rising above the gathering storm: Energizing and employing America for a brighter economic future. Committee on Prospering in the Global Economy of the 21st Century. Washington, DC: National Academies Press.

Bandalos, D. L., & Finney, S. J. (2010). Factor Analysis. Exploratory and Confirmatory. In Hancock, G. R., Mueller, R. O. (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (pp. 93-114). New York: Routledge.

Barab, S. (2014). Design-based research: A methodological toolkit for engineering change. *The Cambridge handbook of the learning sciences*, *2*, 151-170.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software,* 67(1), 1-48. doi:10.18637/jss.v067.i01.

Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review*, *27*(3), 253-265. https://doi.org/https://doi.org/10.1016/j.econedurev.2006.08.007

Berdrow, I. (2010). King among kings: Understanding the role and responsibilities of the department chair in higher education. *Educational Management Administration & Leadership*, *38*(4), 499-514.

Bergquist, W. H., & Pawlak, K. (2007). *Engaging the six cultures of the academy: Revised and expanded edition of the four cultures of the academy*. John Wiley & Sons.

Bernerth, J. B., Walker, H. J., & Harris, S. G. (2011). Change fatigue: Development and initial validation of a new measure. *Work & Stress*, *25*(4), 321-337.

Blaich, C. F., & Wise, K. S. (2010). Moving from assessment to institutional improvement. *New Directions for Institutional Research*, *2010*(S2), 67-78. https://doi.org/https://doi.org/10.1002/ir.373

Bolden, R., Petrov, G., & Gosling, J. (2008). *Developing collective leadership in higher education*. Leadership Foundation for Higher Education.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, *145*, 27-41. https://doi.org/https://doi.org/10.1016/j.jpubeco.2016.11.006

Bouwma-Gearhart, J., Sitomer, A., Fisher, K. Q., Smith, C., & Koretsky, M. (2016). Studying organizational change: Rigorous attention to complex systems via a multi-theoretical research model. 2016 ASEE Annual Conference & Exposition.

Bouwma-Gearhart, J. L., & Hora, M. T. (2016). Supporting faculty in the era of accountability: How postsecondary leaders can facilitate the meaningful use of instructional data for continuous improvement. *Journal of Higher Education Management*, *31*(1), 44-56.

Bradforth, S. E., Miller, E. R., Dichtel, W. R., Leibovich, A. K., Feig, A. L., Martin, J. D., Bjorkman, K. S., Schultz, Z. D., & Smith, T. L. (2015). University learning: Improve undergraduate science education. *Nature*, *523*(7560), 282-284. https://doi.org/10.1038/523282a

Brickman, P., Gormally, C., & Martella, A. M. (2016). Making the Grade: Using Instructional Feedback and Evaluation to Inspire Evidence-Based Teaching. *CBE—Life Sciences Education*, *15*(4), ar75. https://doi.org/10.1187/cbe.15-12-0249

Brownell, S. E., & Tanner, K. D. (2012). Barriers to Faculty Pedagogical Change: Lack of Training, Time, Incentives, and…Tensions with Professional Identity? *CBE—Life Sciences Education, 11(*4), 339-346. https://doi.org/10.1187/cbe.12-09-0163

Bryman, A. (2007). Effective leadership in higher education: A literature review. *Studies in Higher Education*, *32*(6), 693-710.

Callegaro, M. (2008). Social Desirability. In Lavrakas, P. J. (Ed.), *Encyclopedia of Survey Research Methods* (Vol. 2, pp. 825-826). SAGE Publications. https://link.gale.com/apps/doc/CX3073300551/GVRL?u=bois91825&sid=bookmark-GVRL&xid=0373927a

Cashin, W. E. (1990). Students do rate different academic fields differently. *New directions for teaching and learning*, *1990*(43), 113-121.

Cervato, C., Peterson, S., Johnson, C. A., Bilen-Green, C., Koretsky, C., Minerick, A., & Kremer, G. O. (2025). Department Chairs as Change Agents: A Virtual Cross-Institutional Professional Development Model for Chairs. *Innovative Higher Education*, *50*(1), 59-84. https://doi.org/10.1007/s10755-024-09714-8

Cheldelin, S. I. (2000). Handling resistance to change. In Lucas, A. F. (Ed.), *Leading Academic Change: Essential Roles for Department Chairs*. Sand Francisco, CA: Jossey-Bass Publishers.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, *49*(4), 219-243.

Choi, M. (2011). Employees' attitudes toward organizational change: A literature review. *Human resource management*, *50*(4), 479-500.

Chyung, S. Y., Kennedy, M., & Campbell, I. (2018). Evidence-Based Survey Design: The Use of Ascending or Descending Order of Likert-Type Response Options. *Performance Improvement*, *57*(9), 9-16. https://doi.org/https://doi.org/10.1002/pfi.21800

Cipriano, R. E., & Riccardi, R. L. (2017). The Department Chair: A Decade-Long Analysis. *The Department Chair*, *28*(1), 10-13. https://doi.org/https://doi.org/10.1002/dch.30144

Connolly, M. R., & Seymour, E. (2015). *Why Theories of Change Matter* Working Paper No. 2015-2. Retrieved http://wcer-web.ad.education.wisc.edu/docs/working-papers/Working_Paper_No_2015_02.pdf

Corbo, J. C., Reinholz, D. L., Dancy, M. H., Deetz, S., & Finkelstein, N. (2016). Framework for transforming departmental culture to support educational innovation. *Physical Review Physics Education Research*, *12*(1), 010113.

Dawson, S. M., & Hocker, A. D. (2020). An evidence-based framework for peer review of teaching. *Advances in Physiology Education*, *44*(1), 26-31. https://doi.org/10.1152/advan.00088.2019

Dennin, M., Schultz, Z. D., Feig, A., Finkelstein, N., Greenhoot, A. F., Hildreth, M., Leibovich, A. K., Martin, J. D., Moldwin, M. B., O'Dowd, D. K., Posey, L. A., Smith, T. L., & Miller, E. R. (2017). Aligning Practice to Policies: Changing the Culture to Recognize and Reward Teaching at Research Universities. *CBE—Life Sciences Education*, *16*(4), es5. https://doi.org/10.1187/cbe.17-02-0032

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary educational psychology*, *61*, 101859. https://doi.org/https://doi.org/10.1016/j.cedpsych.2020.101859

Eckel, P. D., & Kezar, A. (2003). Key strategies for making new institutional sense: Ingredients to higher education transformation. *Higher Education Policy*, *16*, 39-53.

Edelson, D. C. (2002). Design research: What we learn when we engage in design. *The Journal of the Learning sciences*, *11*(1), 105-121.

Ericson, H. C., Lemons, P. P., Dolan, E. L., Brickman, P., Krishnan, S., & Andrews, T. C. (2025). Are Department Heads Ready for Change? Leveraging a Leadership Action Team to Advance Teaching Evaluation Practices. *CBE—Life Sciences Education*, *24*(1), ar8.

Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education*, *45*(8), 1106-1120. https://doi.org/10.1080/02602938.2020.1724875

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLOS ONE*, *14*(2), e0209749. https://doi.org/10.1371/journal.pone.0209749

Finkelstein, N., Greenhoot, A., Weaver, G., & Austin, A. (2020). A department-level cultural change project: Transforming evaluation of teaching. In White, K., Beach, A., Finkelstein, N., Henderson, C., Simkins, S., Slakey, L., Stains, M., Weaver, G., & Whitehead, L. (Eds.), *Transforming institutions: Accelerating systemic change in higher education*. Pressbooks. Retrieved May 27, 2022, from http://openbooks.library.umass.edu/ascnti2020/

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410-8415. https://doi.org/doi:10.1073/pnas.1319030111

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., Dirks, C., & Wenderoth, M. P. (2007). Prescribed Active Learning Increases Performance in Introductory Biology. *CBE—Life Sciences Education, 6(*2), 132-139. https://doi.org/10.1187/cbe.06-09-0194

Glassick, C. E., Huber, M. T., & Maeroff, G. I. (1997). *Scholarship assessed: Evaluation of the professoriate*. John Wiley & Sons.

Gmelch, W. H., & Miskin, V. D. (1993). *Leadership skills for department chairs*. Anker Publishing Company, Inc.

Gmelch, W. H., Roberts, D., Ward, K., & Hirsch, S. (2017). A retrospective view of department chairs: Lessons learned. *The Department Chair*, *28*(1), 1-4.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American psychologist*, *52*(11), 1209.

Haak, D. C., Hillerislambers, J., Pitre, E., & Freeman, S. (2011). Increased Structure and Active Learning Reduce the Achievement Gap in Introductory Biology. *Science, 332*(6034), 1213-1216. https://doi.org/doi:10.1126/science.1204820

Hartley, J., & Betts, L. R. (2010). Four layouts and a finding: the effects of changes in the order of the verbal labels and numerical values on Likert-type scales. *International Journal of Social Research Methodology*, *13*(1), 17-27. https://doi.org/10.1080/13645570802648077

Hayes, A. F., & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But…. *Communication Methods and Measures*, *14*(1), 1-24. https://doi.org/10.1080/19312458.2020.1718629

Hebbali, A. (2024). olsrr: Tools for Building OLS Regression Models. Retrieved March 13, 2025 from https://olsrr.rsquaredacademy.com/.

Holt, D. T., Armenakis, A. A., Feild, H. S., & Harris, S. G. (2007). Readiness for organizational change: The systematic development of a scale. *The Journal of Applied Behavioral Science*, *43*(2), 232-255.

Holt, D. T., Armenakis, A. A., Harris, S. G., & Feild, H. S. (2007). Toward a comprehensive definition of readiness for change: A review of research and instrumentation. *Research in organizational change and development*, 289-336.

Hora, M. T., Bouwma-Gearhart, J., & Park, H. J. (2017). Data driven decision-making in the era of accountability: Fostering faculty data cultures for learning. *The Review of Higher Education*, *40*(3), 391-426.

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Hunter, A. B. (2019). Why Undergraduates Leave STEM Majors: Changes Over the Last Two Decades. In E. Seymour & A.-B. Hunter (Eds.), Talking about Leaving Revisited: Persistence, Relocation, and Loss in Undergraduate STEM Education (pp. 87-114). Springer International Publishing. https://doi.org/10.1007/978-3-030-25304-2_3

Jorgenson, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2025). semTools: Useful tools for structural equation modeling. Retrieved March 13, 2025, from https://CRAN.R-project.org/package=semTools

Kalkbrenner, M. T. (2023). Alpha, omega, and H internal consistency reliability estimates: Reviewing these options and when to use them. *Counseling Outcome Research and Evaluation*, *14*(1), 77-88.

Kanter, R. M. (1991). Change-master skills: What it takes to be creative. *Henry and Walker, Managing Innovation, Sage*, 54-61.

Keig, L. (2000). Formative Peer Review of Teaching: Attitudes of Faculty at Liberal Arts Colleges Toward Colleague Assessment. *Journal of Personnel Evaluation in Education*, *14*(1), 67-87. https://doi.org/10.1023/A:1008194230542

Kezar, A. (2014). Higher education change and social networks: A review of research. *The journal of higher education*, *85*(1), 91-125.

Kezar, A. (2018). *How colleges change: Understanding, leading, and enacting change*. Routledge.

Knekta, E., Runyon, C., & Eddy, S. (2019). One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE—Life Sciences Education*, *18*(1), rm1.

Knight, J. K., & Wood, W. B. (2005). Teaching More by Lecturing Less. *CBE—Life Sciences Education*, *4*(4), 298-310. https://doi.org/10.1187/05-06-0082

Kotter, J. P. (2012). *Leading change*. Harvard business press.

Kotter, J. P., & Cohen, D. S. (2012). *The heart of change: Real-life stories of how people change their organizations*. Harvard Business Press.

Krishnan, S., Gehrtz, J., Lemons, P. P., Dolan, E. L., Brickman, P., & Andrews, T. C. (2022). Guides to Advance Teaching Evaluation (GATEs): A Resource for STEM Departments Planning Robust and Equitable Evaluation Practices. *CBE—Life Sciences Education*, *21*(3), ar42. https://doi.org/10.1187/cbe.21-08-0198

Kruse, S. D. (2022). Department chair leadership: Exploring the role's demands and tensions. *Educational Management Administration & Leadership*, *50*(5), 739-757.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1-26. doi:10.18637/jss.v082.i13.

Laursen, S., Andrews, T., Stains, M., Finelli, C., Borrego, M., McConnell, D., & Foote, K. (2019). *Levers for change: An assessment of progress on changing STEM instruction.* Washington, DC: American Association for the Advancement of Science. Retrieved May 27, 2022, from www.aaas.org/sites/default/files/2019-07/levers-for-change-WEB100_2019.pdf

Lenhart, C., & Bouwma-Gearhart, J. (2021). STEM Faculty Instructional Data-Use Practices: Informing Teaching Practice and Students' Reflection on Students' Learning. *Education Sciences*, *11*(6), 291. https://www.mdpi.com/2227-7102/11/6/291

Lester, J., & Kezar, A. (2012). Faculty grassroots leadership: Making the invisible visible. *Journal of the Professoriate, 6*(2).

Liu, M., & Keusch, F. (2017). Effects of Scale Direction on Response Style of Ordinal Rating Scales. *Journal of Official Statistics*, *33*(1), 137-154. https://doi.org/doi:10.1515/jos-2017-0008

Lucas, A. F. (2000). A teamwork approach to change in the academic department. In Lucas, A. F. (Eds.), *Leading academic change: Essential roles for department chairs*. (pp. 7-32.) Jossey-Bass Publishers.

Lyde, A. R., Grieshaber, D. C., & Byrns, G. (2016). Faculty teaching performance: Perceptions of a multi-source method for evaluation. *Journal of the Scholarship of Teaching and Learning*, *16*(3), 82-94.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural equation modeling*, *11*(3), 320-341.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press.

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, *28*(1), 61.

Mohr, D. C., Cuijpers, P., & Lehman, K. (2011). Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. *Journal of medical Internet research*, *13*(1), e30.

Nicholls, M. E. R., Orr, C. A., Okubo, M., & Loftus, A. (2006). Satisfaction Guaranteed: The Effect of Spatial Biases on Responses to Likert Scales. *Psychological Science*, *17*, 1027-1028. https://doi.org/10.1111/j.1467-9280.2006.01822.x

Nolan, S., Kim, B. H., Armstrong, E., Freeman, M., Hughes, R., Nicola, T., & Kajikawa, T. (2025). Tracking persistence in STEMM: From application aspirations to college degrees. Retrieved April 11, 2025 from https://www.commonapp.org/files/DAR/Common-App-Persistence-in-STEMM.pdf

Normore, A. H., & Brooks, J. S. (2014). The department chair: A conundrum of educational leadership versus educational management. In Lahera, A. I., Hamdan, K., & Normoe, A. H. (Eds.), *Pathways to excellence: Developing and cultivating leaders for the classroom and beyond* (pp. 3-19). Emerald Group Publishing Limited.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. SAGE Publications, inc.

Pedersen, T. L., (2024). patchwork: The Composer of Plots. Retrieved March 13, 2025 from https://patchwork.data-imaginist.com

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in physical education and exercise science*, *19*(1), 12-21.

Peterson, D. A. M., Biederman, L. A., Andersen, D., Ditonto, T. M., & Roe, K. (2019). Mitigating gender bias in student evaluations of teaching. *PLOS ONE*, *14*(5), e0216241. https://doi.org/10.1371/journal.pone.0216241

President's Council of Advisors on Science and Technology (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering and Mathematics, Washington, DC: U.S. Government Office of Science and Technology.

R Core Team. (2023). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved March 13, 2025, from https://www.R-project.org/

Rafferty, A. E., Jimmieson, N. L., & Armenakis, A. A. (2013). Change readiness: A multilevel review. *Journal of management*, *39*(1), 110-135.

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education*, *16*(2), 129-150.

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE—Life Sciences Education*, *15*(1), rm1.

Reinholz, D. L., & Andrews, T. C. (2020). Change theory and theory of change: what's the difference anyway? *International Journal of STEM Education*, *7*, 1-12.

Reinholz, D. L., & Apkarian, N. (2018). Four frames for systemic change in STEM departments. *International Journal of STEM Education*, *5*, 1-10.

Reinholz, D. L., Corbo, J. C., Bernstein, D. J., & Finkelstein, N. D. (2018). Evaluating scholarly teaching: A model and call for an evidence-based approach. In Lester, J., Klein, C., Johri, A., & Rangwala, H. (Eds.), *Learning Analytics in Higher Education* (pp. 69-92). Routledge.

Reinholz, D. L., White, I., & Andrews, T. (2021). Change theory in STEM higher education: a systematic review. *International Journal of STEM Education*, *8*(1), 1-22.

Revelle, W. (2024). psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois: Northwestern University. Retrieved March 13, 2025, from https://CRAN.R-project.org/package=psych.

Rodríguez-Dorans, E., & Jacobs, P. (2020). 'Making narrative portraits: a methodological approach to analysing qualitative data'. *International Journal of Social Research Methodology*, *23*(6), 611-623. https://doi.org/10.1080/13645579.2020.1719609

Rossel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36. doi:10.18637/jss.v048.i02

Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed.) SAGE Publications Ltd.

Schein, E. H. (1999). Kurt Lewin's change theory in the field and in the classroom: Notes toward a model of managed learning. *Systems practice, 9*, 27-47. https://doi.org/10.1007/BF02173417

Scott, E. E., Wenderoth, M. P., & Doherty, J. H. (2020). Design-based research: A methodology to extend and enrich biology education research. *CBE—Life Sciences Education*, *19*(2), es11.

Schön, D., & Argyris, C. (1996). Organizational learning II: Theory, method and practice. *305*(2), 107-120. Reading: Addison Wesley.

Shadle, S. E., Marker, A., & Earl, B. (2017). Faculty drivers and barriers: laying the groundwork for undergraduate STEM education reform in academic departments. *International Journal of STEM Education, 4*(1), 8. https://doi.org/10.1186/s40594-017-0062-7

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for information*, *22*(2), 63-75.

Simonson, S. R., Earl, B., & Frary, M. (2022). Establishing a Framework for Assessing Teaching Effectiveness. *College Teaching*, *70*(2), 164-180. https://doi.org/10.1080/87567555.2021.1909528

Smith, C. (2008). Building effectiveness in teaching through targeted evaluation and response: connecting evaluation to teaching improvement in higher education. *Assessment & Evaluation in Higher Education*, *33*(5), 517-533. https://doi.org/10.1080/02602930701698942

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of Small-Group Learning on Undergraduates in Science, Mathematics, Engineering, and Technology: A Meta-Analysis. *Review of Educational Research, 69*(1), 21-51. https://doi.org/10.3102/00346543069001021

Stensaker, B., & Vabø, A. (2013). Re-inventing Shared Governance: Implications for Organisational Culture and Institutional Leadership. *Higher Education Quarterly*, *67*(3), 256-274. https://doi.org/https://doi.org/10.1111/hequ.12019

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Tanner, K. D. (2013). Structure matters: twenty-one teaching strategies to promote student engagement and cultivate classroom equity. *CBE—Life Sciences Education*, *12*(3), 322-331.

Task Force on Student Learning and Success. (2017). *Report of Progress and Recommendations*. Retrieved on December 17, 2024, from https://president.uga.edu/wp-content/uploads/final_task_force_report.pdf

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., . . . Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences, 117*(12), 6476-6483. https://doi.org/doi:10.1073/pnas.1916903117

Thomas, S., Chie, Q. T., Abraham, M., Jalarajan Raj, S., & Beh, L.-S. (2014). A Qualitative Review of Literature on Peer Review of Teaching in Higher Education: An Application of the SWOT Framework. *Review of Educational Research*, *84*(1), 112-159. https://doi.org/10.3102/0034654313499617

Walczyk, J. J., Ramsey, L. L., & Zha, P. (2007). Obstacles to instructional innovation according to college science and mathematics faculty. *Journal of Research in Science Teaching, 44*(1), 85-106. https://doi.org/https://doi.org/10.1002/tea.20119

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag.

Weaver, G. C., Austin, A. E., Greenhoot, A. F., & Finkelstein, N. D. (2020). Establishing a Better Approach for Evaluating Teaching: The TEval Project. *Change: The Magazine of Higher Learning*, *52*(3), 25-31. https://doi.org/10.1080/00091383.2020.1745575

Wei, T., & Simko, V. (2024). R package ‘corrplot’: Visualization of a Correlation Matrix. Retrieved March 13, 2025, from https://github.com/taiyun/corrplot.

Weiner, B. J. (2020). A theory of organizational readiness for change. In *Handbook on implementation science* (pp. 215-232). Edward Elgar Publishing.

Wentworth, D. K., Behson, S. J., & Kelley, C. L. (2020). Implementing a new student evaluation of teaching system using the Kotter change model. *Studies in Higher Education*, *45*(3), 511-523. https://doi.org/10.1080/03075079.2018.1544234

Wolverton, M., Ackerman, R., & Holt, S. (2005). Preparing for leadership: What academic department chairs need to know. *Journal of Higher Education Policy and Management*, *27*(2), 227-238.

Yin, R. K. (2014). *Case Study Research*. SAGE Publications. https://books.google.com/books?id=Cdk5DQAAQBAJ

Yukl, G. (2012). Effective leadership behavior: What we know and what questions need more attention. *Academy of Management perspectives*, *26*(4), 66-85.

APPENDIX A: EXAMPLE OF THE IMPLEMENTATION OF DESIGN PRINCIPLES DURING AN LAT MEETING

This is a portion of the meeting where a facilitator was talking to a small group of heads. Highlighted text indicates places where the design principles are being demonstrated, accompanied by an explanation for each.

Meeting transcript excerpts	Design principle and explanation
Small-group facilitator: Take a minute. And while you're reading, start thinking about how this could work in your unit. How might you adapt this or use this in your unit? [Long pause].	Create space for thought and learning: Department chairs considered an example self-reflection form and were asked to reflect on how it would work in their department. Facilitators allowed time for department chairs to read and think.
Small-group facilitator: Okay... If you think of this as a resource for helping your faculty, think about how to engage in self-reflection, do you think this would be useful to them, not useful to them? How might you imagine using it with your faculty?	
Head of Department K: I think it would be useful because it pretty much covers the two main examples of where this is really useful. Like for the early career faculty that are just getting started with teaching, and then also people more mid-career or late career that are trying new methodologies, new pedagogies. So in my department, I've asked them to provide more of a narrative. Like what did you publish?... What talks did you give? What classes did you teach? It's really student evaluations centric. It's always been that way. This could become part of [the process] I imagine, and we could even do that this year. I think it would be really useful to have [it] as an attachment to an email... as an example. I can see that going over pretty nicely.	
Small-group facilitator: Okay. So, you imagine maybe including a field or a couple of fields on your annual evaluation reporting, perhaps	Scaffold learning and action & Expect proactive

along with, what did you teach, what teaching problem did you tackle? One of the things that [another department] did is they actually asked some faculty to be the pilot testers the first go-round and then that even allowed them to collect [discipline-] specific examples and also to get some feedback on how that worked as a process before they expected all their faculty to do it.

work: Facilitator clarifies the concrete action that the department head plans to take and offers an example of how another department approached this.

Head of Department K: Yeah. So that's something I'd like to talk about a little bit. I don't think we'll ever be in a place where I can tell all the faculty “You have to do this”. I mean, there will be a significant fraction that will just simply refuse, just not do it.

Small-group facilitator: Do you have that happen on your annual review reports now? Where they just say “No, I'm not going to fill out that section?”

Compel reflection: Facilitator questions the assumption that faculty will not engage in teaching reflection in annual review.

Head of Department K: No. It's kind of culturally accepted within the department that you are expected to provide the scores, the quantitative outcomes from the student evaluation process. So, I guess people are accepting of that. They're used to doing that because it doesn't require any effort on their part. They just download the results and paste them in. But this actually requires some effort on their part. I think for some people, some of the senior faculty that have already been promoted to full professor, for example, I can imagine some of them being like, “Well, okay, this is just some other thing that the department is asking me to do.”...I guess I'm not entirely optimistic that I could get buy-in from everyone in all levels of [my department], definitely the younger faculty would be more amenable to this for sure.

Small-group facilitator: So that sort of brings two questions to my mind. One is, is that okay? To just put [the policy in place] and that way folks who are ready to engage with it can, and folks who are resistant... there's a spot there that they might feel some obligation to fill in in some way, but that maybe the quality of what they put in that spot is not what you'd hope for, but it's at least, you know, a spot.

Compel reflection: Facilitator questions the assumption that a change should only be taken if everyone will engage productively.

Head of Department K: Yeah, I think it’s important to have the spot, and also these examples, because I do believe that on the flip side, there are faculty that want this. They really want to have an opportunity to put their own words. We talked about this before, it

empowers them to create their own narrative around their teaching and allows them to have some say. But we're going to have a bi-modal distribution in the beginning. And then with time, I think that the group that resists will shrink. Then at some point it can be part of a formal process by which merit based raises and things like this are brought to bear on this.

Small-group facilitator: I have one other question and then [another department chair], I'd love to hear what you think about this. Is there maybe a senior faculty member who you think would be amenable [to this] and would garner the respect of other faculty in the department? Maybe someone who is more teaching interested, but also research accomplished and who'd be willing to speak up about why they might see value in this and maybe offer their own example?

Provide differentiated support: Facilitator suggests a specific strategy that a department chair could use to help foster buy-in for a new teaching evaluation practice.

Head of Department K: Yeah. Probably so. Yeah, that's something I can work on, is trying to identify a model citizen.

Small-group facilitator: Yeah. I know that happened in my department. We have a couple of really accomplished researchers who are really student oriented. If they say, “Hey, this thing we're doing that's innovative for students, I'm on board because I think it's helpful for students.” Then it's a little bit harder for some of our colleagues who say, “Well, I'm just a researcher” because they can't say that. Right? They can't when they have someone who they view as also just a researcher.

Head of Department K: That's a really good idea.

Small-group facilitator: So [other department chair], what do you think, how would this, something like this be useful in your department?

APPENDIX B: DEPARTMENT HEAD INTERVIEW PROTOCOLS

These interview protocols were used to gather data from department heads on readiness for change, leadership ideas and actions, and departmental teaching evaluation practices. Interview protocols 1 and 2 were used for data collection for Chapter 2, while protocols 2 and 3 were used for data collection for Chapter 3.

Interview protocol 1:

The goal of this interview is to learn about departmental practices related to undergraduate teaching, particularly the practices that may not be part of formal policies and might not be obvious to an outsider. We will start with some questions about teaching evaluation since that is the initial focus of the work of the Leadership Action Team.

Alternate text for non-LAT: The goal of this interview is to learn about departmental practices related to undergraduate teaching, particularly the practices that may not be part of formal policies and might not be obvious to an outsider like me. I'd like to start by asking you about how teaching effectiveness is evaluated in your department.

I'll be asking about a wide array of practices and I don't expect any department to be engaging in all of these practices. So please don't hesitate to say "no" when I ask if something is happening.

1. Can you please talk me through how teaching effectiveness is evaluated for promotion & tenure?

- Student evaluations?

- How are peer evaluations used?
    - Is there a particular form that guides this process? OR
    - Are there particular things that you ask observers to pay attention to?
    - Is that written down somewhere?
- Some departments indicate that colleagues might be very hesitant to provide critical feedback in a peer review for promotion. Do you that happen in your department?
- Summary of what they said. Is there more I should know about how teaching effectiveness is evaluated from promoted & tenure?

2. To what extent is teaching effectiveness discussed when the department votes on tenure and promotion?

- What evidence of effectiveness do you think is most strongly valued by <u>your</u> faculty? What makes you think that?
- How does this vary by rank or position type of the candidate?

3. How is the evaluation of teaching effectiveness different for annual review than for promotion & tenure?

- How are student evaluations used?
- IF peer evaluation is mentioned, ask these:
    - Is there a particular form that guides this process? OR
    - Are there particular things that you ask observers to pay attention to?
- Summary of what they have said. Is there more I should know about how teaching effectiveness is evaluated annually?

4. I understand that the university asks you to rate each faculty member with teaching responsibilities as exceeding, meeting, or not meeting expectations for teaching. How do you know when someone exceeds expectations versus meeting expectations?

- If there is a written document, ask if they would be willing to share it.

LAT members only:

5. As you know, the initial focus of the LAT is considering how teaching effectiveness is evaluated in our departments. What would you like to see as an outcome for your department in this area?

6. Do you think your departmental colleagues would also value that outcome?

- Alternative text IF they do not have ideas about an outcome they would like to see: Do you think your departmental colleagues see a need for any changes related to how teaching effectiveness is evaluated?

Interview protocol 2:

The goal of this interview is to check in about your department and hear some of your ideas about leadership. I'm going to ask you about your current teaching evaluation practices and any changes you've made a bit later, but I want to start our conversation by asking you about any plans you have for the future.

To start out today….

Are there changes to teaching evaluation you plan to prioritize in the next year or two, or is that not a priority at this time? WHAT changes to teaching evaluation are priority for you?

- *WHY do you think these changes are a good next step for your department?*

- *OR What convinces you that teaching evaluation is serving its purpose in your department?* Skip to leadership section.

I heard you say that you want to change [x, y, and z]. Do I have that right? [they answer] And of these, would some be higher priority than others? [if needed, follow up with “why?”]

For outgoing heads: What changes to teaching evaluation do you think the department should prioritize in the next 2-3 years?

- *WHY do you think these changes are a good next step for your department?*

DeLTA has presented the three voice framework in the LAT. Three voices refers to relying on students, peers, and the instructors themselves as sources of information for teaching evaluation. We have had the chance to talk about _____. Are you satisfied with how your department engages with [other voices]?

Returning to your plans to [describe teaching eval change they intend to make], how would [name change] benefit your department, or faculty in your department?

To what extent do you expect faculty in your department to support [describe the changes they want to make]?

- *WHY is that what you expect?*

What might you do to navigate the resistance that you might encounter?
To what extent do you think the upper administration (college, university) supports changes to teaching evaluation in your department?

How comfortable do you feel leading [changes to teach eval they want to make]?

- *Tell me more about that.*
- *We've observed that department chairs have different levels of comfort with leading different changes in their department. How prepared are you feeling to lead these changes?*
- *I want to hear about how ready you feel to move forward with [change they want], do you have particular worries about how to go about leading [change they want to pursue]?*
- *What makes you feel [however they described feeling less than prepared]?*

Now I would like to ask you some questions about the teaching evaluation practices that are currently in place in your department. I appreciate your patience as I ask these questions so that I have all the information that I need for my dissertation research. **Keep in mind that the answer to many of these questions may be "no."** Most departments are just starting to make changes to teaching evaluation

Peer Voice

Do you currently use peer voice as a way to evaluate teaching?

Can you walk me through how that works in your department?

I want to follow up about a few details. I'm particularly interested in what is standard across observations, more so than what may have occurred once or twice but not consistently:

Do you have a formal observation form to guide peer observations? *Is it okay if I follow up by email to get a copy of that form?*

- Is collecting class materials part of the process? And does that happen consistently for all faculty?

    - If yes, what else is considered?
- Is it a standard part of the process that observers meet with the instructor prior to observations, to better understand the course?

How do observers provide feedback to instructors who are observed? And is the way feedback is provided the same regardless of the observer?

- Is there a template for a report?
- Do observers meet with the instructor to discuss feedback after the observation?

How do you select peer observers? Does this process vary across faculty?

- What is the process for selecting peer observers?
- *Is this a formal process, or just how it's done?*
- Is there training for peer observers?

Is there a standard number of lessons that are observed?

- Is there a standard number of observers?
- Is there a particular timeline for reviews across someone's career?
- And would someone be observed in more than one semester for a given review period, such as prior to consideration for tenure?

Who is in charge of carrying out and refining peer observations?

How is your department considering the workload that is required for coordinating peer observations?

Have you had departmental discussions about these peer evaluation processes recently?

- Do you anticipate having those in the future?

Student Voice

Can you walk me through how data from students is used in your department? Most commonly these are data from mandatory student evaluations, but it may also include data about student learning outcomes.

I want to follow up about a few details. I'm particularly interested in what is standard across all faculty, more so than what individual faculty do:

Is it expected that instructors collect any data beyond the mandatory end of course evaluations?

Do you have standard expectations for how faculty analyze student evaluation data? To give an example, one expectation could be that faculty thematically analyze comments from student evaluations rather than cherry picking positive comments. Another example is analyzing quantitative data using distributions rather than means. *Is it okay if I follow up by email to get a copy of those expectations?*

-How are the data from quantitative questions handled? Are they analyzed as distributions or as averages?

-How are student comments from end of course surveys analyzed? Are there any guidelines for systematically examining them?

Do you have different expectations for the student evaluation data faculty present for different review periods (annual review, 3rd year review, promotions)?

When faculty present data from student evaluations for review periods, are they expected to show data over time, so that patterns of change are evident?

Are there different expectations depending on years of teaching experience?

Are there expectations for the response rate that instructors get on student evaluations?

Does your department make comparisons in student evaluation scores among instructors? This might happen formally, maybe as part of annual evaluation and merit raise decisions, or it may happen informally, such as in P & T discussions?

Have you had departmental discussions about these student evaluation processes recently?

Does the department provide any training or resources for faculty about how to achieve high response rates and appropriately analyze student data?

Self Voice

Does your department currently ask faculty to provide a written self-reflection about teaching for annual review or another purpose?

- If yes: Can you walk me through how self-reflections are used in your department?

I want to follow up about a few details. I'm particularly interested in what is standard across faculty, more so than what some faculty may choose to do:

Do you have a formal template or guide for written teaching self-reflections? *Is it okay if I follow up by email to get a copy of that form?*

When an instructor is writing a self-reflection, what are they asked to focus on?

-Are they asked to focus on teaching challenges?

-Are they asked to systematically consider evidence as part of their self-reflection process? One example of this would be to systematically analyze student evaluation data and reflect on what they learned from this.

-Are they asked to build on prior self-reflections?

-Are they asked to discuss or describe their efforts to grow and learn as educators?

Does the department provide any training or resources about how to engage in systematic teaching self-reflection?

When self-reflections are evaluated as evidence of teaching effectiveness, how do you account for different levels of experience among instructors?

Have you had departmental discussions about teaching self-reflection expectations and processes recently?

Interview protocol 3:

The goal of this interview is to check in about your department and hear some of your ideas about leadership. To start out today….

READINESS FOR CHANGE

Discrepancy/Appropriateness: Do they recognize shortcomings in teaching evaluation practices?

Starting big picture, why do you want to pursue changes to teaching evaluation in your unit?

- Prompt to make sure you understand the problem they want to solve.
    - And do you want to [positive thing] because you think that is currently lacking in your department?
- Possible prompting to get THEM to elaborate on the problem:
    - Why do you think that's an important problem to solve?
    - How will changes to teaching evaluation help address this problem?
    - What are the consequences of lacking_____?
    - (*Appropriateness):* And why would changing teaching evaluation be a way to address _____?

Valence: What departmental benefits do they anticipate from changing teach. eval. Practices?

How do you think changing teaching evaluation would benefit your department, or faculty in your department?

At this point in time, what concerns do you have about changing teaching evaluation in your department?

Support: Do they perceive sufficient support from departmental faculty & upper admin? (+ leadership)

To what extent do you think the upper administration (college, university) supports changes to teaching evaluation?

What kinds of responses do you anticipate from colleagues in your department as changes to teaching evaluation are made?

How might you move forward given the variety of responses you may receive from your faculty?

- Are there any particular strategies you anticipate using?
- If they aren't understanding what we mean by strategies: how do you think you might work with colleagues who express concerns about changing teaching evaluation?
- Why do you think that would be useful?

<u>Efficacy: how capable of facilitating change do they perceive themselves to be?</u>

(Departmental efficacy): Do you think your department has what it needs to make changes to teaching evaluation, that might include expertise, resources, etc.?

(Personal efficacy): And how prepared do you feel personally to help lead these changes?

- Tell me more about that.
- Why do you feel that way?

<u>Leadership:</u>

You've told me that you're interested in seeing [insert change here] occur in your department. Can you walk me through the steps you anticipate taking to achieve that change in the department?

APPENDIX C: GUIDES TO ADVANCE TEACHING EVALUATION (GATES) MODIFICATIONS

The following shows the target practices used to analyze teaching evaluation practices in this study. Black text denotes the original research-based target practices from Krishnan et al., 2022. Blue text denotes modifications for this study that allowed us to more reliably judge the presence and absence of target practices. Crossed out text denotes text from the original target practices that was removed for brevity or clarity. The target practices are separated by each voice, and then by the features of robust and equitable teaching evaluation.

Peer Voice

Structured:

1. Department has formal observation form to guide what is observed~~ and which other data are collected (e.g., class materials, assessments, pre-observation meeting). Forms may be adopted or adapted from other departments~~.
2. Department has a formal template for writing a report based on peer review, potentially distinguishing between formative and summative reviews.
3. ~~Department uses formal processes or criteria to select peer observer(s) for all instructors.~~
4. Department ~~enacts policy about~~ has a consistent practice concerning the number of peer observations and observers during a review period and/or across review periods.
5. Department designates and empowers a coordinator, leader, or committee to carry out and refine peer observation.
6. ~~Department has a process for allocating and recognizing workload related to coordinating and conducting observations.~~

7. ~~Department periodically discusses and improves peer evaluation practices to maximize utility to instructors and the department.~~
8. Department provides or arranges formal training about the departmental peer review process for peer observers. Formal training here means a session where faculty interact about the peer review process in their department. This could include a norming session, a practice session, etc. It's more than just being given materials to do the observations.
9. Department conducts peer observations for ALL instructors in their department, not just a subset of faculty.

Reliable:

1. Department relies on multiple observations for instructors, such as multiple observers, observing multiple lessons, and/or observing multiple courses.
2. Department specifies which class materials ~~(e.g., syllabi, exams, homework, slides, handouts)~~ are collected and evaluated as part of peer observation.
3. Department expects observers to ~~talk~~ meet with instructors ~~to properly contextualize observations and review of materials. This might include discussing course goals, lesson goals, class structure, and students.~~ before the observation(s) to properly contextualize observations and review materials.

Longitudinal:

1. Department conducts peer observations ~~over multiple time points in a review period for all instructors~~ for instructors over multiple time points in their career to document teaching improvements.
2. Department ensures that the peer observation process provides feedback to instructors ~~via follow-up discussion that covers strengths and areas for improvement~~. This could come in the form of a follow-up discussion, or written feedback.

Student Voice

Structured:

1. Department has formal standards for how and when instructors collect, analyze, and report student data (e.g., standard quantitative and qualitative analysis).
2. Department intentionally selects or designs student evaluation questions.
3. Department makes appropriate distinctions in their expectations about student data ~~for different review periods (e.g., annual review, 3rd year review, promotion) and different levels of teaching experience within a given course.~~ for instructors with different levels of teaching experience.
4. ~~Department periodically discusses and improves expectations for collecting and analyzing data from students to maximize utility to instructors and the department.~~
5. Department provides or arranges formal training, or other support, for instructors about collecting and analyzing student data, ~~including achieving high response rates, analyzing quantitative and qualitative data systematically and appropriately, gathering data beyond mandatory evaluations, and making comparisons across time.~~ Other support here could mean providing materials for instructors. Examples include an excel file to aid in making graphs or results, directions on how to conduct a thematic analysis, or resources on how to get a higher response rate.

Reliable:

1. Department expects instructors to take steps to achieve higher response rates on mandatory student evaluations ~~(e.g., course credit offered, class time set aside)~~.
2. Department recognizes and attends to known biases~~, such as bias against women, minoritized groups, and large class size, and limits comparisons of mandatory student evaluations between instructors~~ when considering data from mandatory student evaluations.

3. Department limits comparisons of mandatory student evaluations between instructors. This includes not comparing instructors to a departmental average.
4. Department specifies that quantitative questions on mandatory student evaluations be analyzed as distributions of scores, rather than averages. ~~Because quantitative questions often use an ordinal rating scale (excellent, very good, good, poor), average scores and standard deviations are inappropriate. We cannot assume the points on ordinal scales are equidistant.~~
5. ~~Department specifies which set of quantitative student evaluation questions are used for each review period (e.g, annual, promotion).~~
6. Department specifies that feedback in student comments on mandatory evaluations be systematically examined. This is NOT just putting comments into a word cloud and looking at the results, actual systematic examination must be conducted.
7. Department expects instructors to collect, analyze, ~~and~~ interpret, and report some data beyond mandatory student evaluations.

Longitudinal:

1. Department expects instructors to document change (or consistently exemplary results) by comparing data from students across multiple timepoints

Self Voice

Structured:

1. Department uses a formal self-reflection form to guide the scope and content of written self-reflection narratives, including standards for what constitutes evidence-based self-reflection. ~~Forms may be adopted or adapted from other departments.~~

2. ~~Department periodically discusses and improves standards for written teaching reflections to maximize utility to instructors and the department.~~
3. Department provides or arranges formal training, or other support, for instructors about the self-reflection process ~~and to help instructors meet departmental expectations for documenting self-reflection~~. Other support here can mean hosting discussions where faculty can engage in talks about self-reflection, sharing examples of self-reflections, providing models of what self-reflection should look like, providing training on how to engage with self-reflection, making a guide for faculty on how to engage in self-reflection, etc.

Reliable:

1. Department expects instructors to engage in a self-reflection process~~, and written documentation thereof,~~ that is focused on tackling teaching challenges ~~(e.g., concerns raised in student evaluations or peer observation, student learning difficulties, lack of engagement)~~.
2. Department expects the self-reflection process~~, and written documentation thereof,~~ to rely on the systematic analysis of evidence about student learning and experiences.
3. Department expectations for self-reflection consider the experience level of instructors. ~~For example, instructors new to a course or teaching may primarily rely on informal sources of data (e.g., notes, brief written feedback from students), whereas more experienced instructors rely on formal sources of data (e.g., assessment data) and systematic observation (e.g., feedback from trained peers).~~

Longitudinal:

1. Department expects that written reflections discuss how instructors have built on prior self-reflection, including the outcomes of planned improvements and innovations.

2. Department expects that written reflections discuss efforts to grow and learn as educators. ~~This can include learning from both successes and failures.~~

APPENDIX D: READINESS FOR CHANGE CODEBOOK

We developed and refined this codebook to systematically examine evidence of department head readiness for change in interview transcripts and LAT meeting recordings. The codes are organized by the component of readiness for change they were designed to capture (in bold).

Code Name	Description
Discrepancy	
Departmental	This modifier is used when the discrepancy being talked about is a departmental discrepancy, not a personal one.
Lack of discrepancy	This code is used when a department head is specifically asked what they want to change, and they clearly indicate that they do not see a need and/or intend to change something related to teaching evaluation.
Practice discrepancy	This code is used when a department head recognizes a discrepancy with the details of how teaching evaluation is performed in their department. An example would be them stating that they only have one peer observer perform an observation for each review period, and they think that that number needs to be larger. This should not be used when a head is recognizing a broad, overarching problem with teaching evaluation, or when they are recognizing a problem with how a voice is implemented or used overall in their department.
Voice discrepancy: Voice is not useful as is/used	This code is used when a department head indicates that they see major shortcomings with all or part of a voice for teaching evaluation. They often do not indicate that they want to add or majorly modify the voice; instead, they are just discussing problems. This often focuses on how student voice is meaningless or unhelpful, or maybe just that the comments or numbers are not useful. Other times they discuss how peer voice is not useful because faculty only share positive feedback.
Voice discrepancy: I want to add this voice	This code is used when a department head indicates that they would like to add a voice or significantly modify

	a voice for teaching evaluation. They might offer no details about how they would do this (no clear practices) or they might indicate some problems they see (needs to be more routine and robust) without identifying practices that would achieve that. Indicating that there are problems with a voice is NOT enough to get this code. They need to clearly state a desire to add or substantially change a voice.
Overall discrepancy	This code is used when a department head recognizes a very broad or overarching discrepancy. An example would be them stating that they feel that the entire teaching evaluation system needs to be overhauled. This should not be used when a head is recognizing a problem with how a specific voice is implemented in their department, or how there are problems with the details of how teaching evaluation is performed in their department.
Appropriateness	
General appropriateness	This code is used if the head expresses a preference for making changes in evaluation to one or more voices over others, or if they believe that specific practices are more important than others that need to be changed in the teaching evaluation process. This is often talked about as priority, or the next thing they're going to do.
Efficacy	
DeLTA resources	This code is used when a department head is talking about the DeLTA project, or resources from the DeLTA project being a source of their efficacy.
Experience	The department head already has experience leading change and talks about how that will impact how they pursue change in their department.
I have help	This code captures evidence that a department chair's readiness for change or accomplishment of change is at least partly based on leveraging the expertise/human resources in their department. This shows up most obviously when we ask about efficacy. Department heads present the fact that they have help as part of what makes them feel comfortable pursuing changes to teaching evaluation. It also shows up when heads talk about changes they have already achieved or planned. They explain the important role that faculty in the department have played in leading/planning/imagining the change.
It's my job	This code is used when department heads express that making change is a part of their job responsibilities, or a part of their job, which then influences their overall level of efficacy. They could also be talking about how they took the job specifically to make the changes that they see as necessary.

Believes in change	This code is used when department heads are expressing that they believe this change will have a positive impact for their department. They believe that their department will benefit from making the change. This is NOT just any evidence that they see benefits or believe in the change. This should only be used when they're using their belief as evidence of their efficacy.
Efficacy for teaching evaluation	This modifier is used when the efficacy being referred to is efficacy in relation to changing teaching evaluation practices.
Efficacy for change in general	This modifier is used when the efficacy being referred to is efficacy in relation to change in general.
Level of efficacy: Low self-efficacy	This code is used when a department head describes themselves as having a low level of efficacy for leading change to teaching evaluation.
Level of efficacy: High self-efficacy	This code is used when a department head describes themselves as having a high level of efficacy for leading change to teaching evaluation.
Valence	
Cost	This code is used when a department head is describing what they view as a cost of making a change to teaching evaluation, or a cost of a specific practice. This could be something they view as personally costly, or something that they view will be costly for their faculty or their department. This is different from a departmental cost, as this is the head's own opinion, and not their perception of what their faculty's opinions are. Each quote represents the discussion of only ONE cost. Examples of costs include: using too much faculty time, causing distress or low morale if people feel over-evaluated, causing inequities in faculty workload Keep in mind, this isn't the cost of current practices, that would likely be discrepancy.
Benefit	This code is used when a department head is describing what they view as a benefit of making a change to teaching evaluation, or a benefit of a specific practice. This could be something they view as personally beneficial, or something that they view will be beneficial for their faculty or their department. This is different from a departmental benefit, as this is the head's own opinion, and not their perception of what their faculty's opinions are. Each quote represents the discussion of only ONE benefit. Examples of benefits include: student comments being informative, allowing the department to recognize good teaching, building departmental conversations around

	teaching, mid-semester evaluations helping current students, forcing faculty to reflect on their teaching, observers learning from watching someone else, providing actionable data for improvement, helping junior faculty with P&T, and allowing faculty to look longitudinally at their teaching.
Principal Support	
General Principal Support	This code is used when a department head is describing their perception of administrative support for changing teaching evaluation in their department.
Unsure about upper admin support	This code captures instances when the participant does not have a sense of whether the upper admin would be supportive (in spirit or otherwise) of changes to teaching evaluation. It's common for the heads to qualify their answer by saying they haven't talked to anyone, or they don't have evidence, but if the overwhelming idea is that they perceive supportiveness, code accordingly.
Department Behavior	
Realized	This modifier is used when the departmental behavior being talked about is behavior that the department head has actually seen occurring in their department.
Anticipated	This modifier is used when the departmental behavior being talked about is only anticipated by a department head, not behavior that has actually happened.
Unproductive engagement	Department heads are talking about how they perceive their faculty have interacted or may interact with teaching evaluation or proposed changes to teaching evaluation. They can comment on how they perceive their faculty to be against the new changes, or how they anticipate resistance to look in the future. They can also comment on how they perceive their faculty to have engaged with teaching/teaching evaluation in an undesirable way in the past. BEWARE: don’t use this code if they are not talking about teaching evaluation specifically. This will often discuss resistance, but it may also discuss hesitancy or concerns.
Lack of engagement	This code is used to capture when a department head is talking about apathy of their faculty towards a new change/teaching evaluation practice.
Productive engagement	Department heads are talking about how they perceive their faculty have interacted or may interact with teaching evaluation or proposed changes to teaching evaluation. They can comment on how they perceive their faculty to be on board with the new changes, or how they don't anticipate them to resist. They can also comment on how they perceive their faculty have engaged with teaching/teaching evaluation in a positive light in the past. BEWARE: don’t use this code if they are not talking about teaching evaluation specifically.

Peer Voice	This modifier is used to tag discrepancies and appropriateness referring to peer voice, or practices related to peer voice.
Student Voice	This modifier is used to tag discrepancies and appropriateness surrounding student voice, or practices related to student voice.
Self Voice	This modifier is used to tag discrepancies and appropriateness about self voice, or about practices related to self voice.

APPENDIX E: LIST OF ORIGINAL TERA ITEMS

This table lists all items that were originally developed for the Teaching Evaluation Readiness Assessment (TERA) before modifications and deletions were made based on expert feedback, think-aloud interviews, exploratory factor analysis, and confirmatory factor analysis. The source column indicates where each item originated from, prior to being revised to fit our context. Arm indicates the item originally came from (Armenakis et al., 2007), Holt indicates the item originally came from (Holt, Armenakis, Feild & Harris, 2007), and Orig indicates the item was newly written for this survey.

Item	Source	Component of Readiness for Change
I think this change would be a worthwhile use of the department's time	Holt	Appropriateness
This change would match the priorities of my department	Holt	Appropriateness
It would make sense for my department to initiate this change	Holt	Appropriateness
I think this change would have a favorable effect on teaching in my department	Arm	Valence (Organizational)
This change would improve student learning in my department	Orig	Valence (Organizational)
I think that my department would benefit from this change	Holt	Valence (Organizational)
This change could improve my department's overall teaching effectiveness	Holt	Valence (Organizational)
This change would prove helpful for student outcomes in my department	Orig	Valence (Organizational)
I feel it would be worthwhile for me if the department made this change	Holt	Valence (Personal)
I think this change would provide me with better feedback on my teaching	Orig	Valence (Personal)
I feel this change would lead to fairer evaluations of my teaching	Orig	Valence (Personal)
I believe this change could benefit me	Arm	Valence (Personal)
I think this change would provide me with new ideas for my teaching	Orig	Valence (Personal)
The expertise in my department is sufficient to accomplish this change	Holt	Efficacy
My department would be able to make this change	Arm	Efficacy
I am confident that my department could implement this change successfully	Holt	Efficacy
I think my department is well-equipped to implement this change	Arm	Efficacy
I think the head of my department would take steps to support this change	Orig	Principal Support
I think the head of my department would be in favor of this change	Arm	Principal Support

I anticipate that the head of my department would encourage us to support this change	Arm	Principal Support

APPENDIX F: THINK-ALOUD INTERVIEW PROTOCOL

This interview protocol was used to conduct think-aloud interviews about the Teaching Evaluation Readiness Assessment (TERA) with faculty members:

The goal of this interview is to fine-tune a survey that we will soon be sending out to faculty across STEM departments. This survey is part of the research of DELTA, which is an NSF-funded project to advance undergraduate education by working with faculty, departments, and the university. DeLTA works with departments to reconsider how we evaluate teaching. The survey will gather information about faculty perceptions about whether changing teaching evaluation makes sense for their department.

What I'm going to ask you to do is called a think-aloud interview. It's just how it sounds; you will try to share everything you are thinking by saying it out loud. I'm going to ask you to read through a survey, select your own responses, and talk me through everything you are thinking as you do so.

This helps us determine if the questions are being interpreted as we intend, where confusions are, and generally how we can make the survey more functional. The survey answers you enter today won't be used as data. We use what you share as data to help us create a better survey. Feel free to ask questions at any time.

Start survey

I'm going to ask you to open the survey and then share your screen. That will allow me to see the same thing you are seeing.

I've pasted a link in the chat. Let's just start with the instructions. Please read those and then tell me everything you think in response to the instructions. If possible, wait to read the matrix below.

Question

Now we will proceed through the questions. As you read a statement and determine your own level of agreement, please share anything you are thinking.

Interviewer instructions:

If they struggle to understand a question... rather than explaining it, ask:

Which part of that sentence is unclear? OR What is most confusing about it?

Once you understand the confusion, offer clarification. If you can think of a simple fix, suggest it in the chat and ask if the fix is better. Make sure to write the fix you suggested in your notes, and their assessment of it.

Wrap-up questions

Are there any negative feelings coming up for you as you consider this survey?

Do you anticipate any negative feelings coming up for your departmental colleagues? If so, is there a way the survey could be changed to prevent that?

What other thoughts or impressions do you want to share?

If someone is taking a super long time, cut them off with 5 minutes remaining and ask the wrap-up questions.

APPENDIX G: RESPONDENT DEMOGRAPHIC INFORMATION

Respondents to the Teaching Evaluation Readiness Assessment (TERA) were asked to answer two questions about their demographics: what department they were in, and what position they held. We report here to position information shared by respondents. Each column reports data collected either in spring 2023 or fall 2024. Values are reported as the percentage of respondents that selected that option.

Position	**2023**	**2024**
Assistant Professor	14.59	14.14
Associate Professor	18.51	20.37
Professor	42.35	37.84
Lecturer	10.32	9.36
Senior Lecturer	4.27	4.99
Principal Lecturer	0.00	1.46
Academic Professional Associate	0.00	1.04
Academic Professional	2.85	0.62
Senior Academic Professional	0.36	0.62
Clinical Assistant Professor	0.36	0.42
Clinical Associate Professor	0.71	0.00
Clinical Professor	0.71	0.21
Professor of Practice	0.36	0.21
Other	1.07	1.46
No response/ Prefer not to respond	3.56	7.27

APPENDIX H: INTER-ITEM CORRELATION MATRIX FOR THE TEACHING EVALUATION READINESS ASSESSMENT

This matrix displays Pearson correlation coefficients between the final survey items for the TERA. These coefficients illustrate the relationships between items within and across factors. Higher correlations indicate stronger associations between items. The stronger the correlation between two items, the darker the blue used in that box.