Towards a Comprehensive Teaching Evaluation Framework

A Summary of Literature, Recommendations, and Examples from Other Institutions

Bridget Arend, PhD, Office of Teaching and Learning
December 2018

Table of Contents

All resources, documents related to initiatives at DU and materials related to the institutional examples cited in this report have been uploaded to the Teaching Excellence Initiative (TEI) Portfolio Site: http://portfolio.du.edu/teachingexcellence/

Overview

Higher education institutions have high expectations for faculty members to create and facilitate meaningful, creative and transformational learning experiences. Teaching is a complex endeavor that, much like research, takes time, effort and ongoing development. However, the reward structure in higher education typically encourages faculty to focus their time and energy away from teaching (Wieman, 2015). Faculty members tend to be assessed and promoted primarily on the basis of research success (Bradforth, et al., 2015). If teaching is truly to be taken seriously, it must be factored into the overall faculty reward structure in a significant and meaningful way.

Student end-of-course rating forms, commonly referred to in the literature as Student Ratings of Instruction (SRIs) remain the long-standing default method for collecting feedback about teaching performance in higher education. Despite widespread discussion and debate, nearly all scholars who study SRIs agree on one important aspect: they should never be used as the sole or primary basis of determining a faculty member's teaching performance (Arreola, 2000; Benton & Ryalls, 2016; Berk, 2005). Despite this articulation, overreliance on SRIs for determining a faculty member's teaching performance continues to exist (Seldin, 1999; Pallett, 2006).

In recent years, calls have been made for college teaching to be evaluated in a more meaningful, balanced, and comprehensive way. Formal policies at DU as well as recommendations from national organizations state that multiple measures and multiple sources should be used in any evaluation of teaching. Many attempts at comprehensive frameworks have been proposed throughout the years, along with decades of research about alternative methods for showing evidence of teaching development. The tools and frameworks exist, but the challenge is often related to culture change within the organization (Arreola, 2007). Frameworks and processes need to be developed internally by the faculty themselves to be fully adopted and sustained. Deeply held beliefs about teaching, suspicion about evaluation practices, and a history of misuse and overreliance on SRIs, have created strong emotional barriers to change (Hativa, 2013). A change in teaching evaluation is inevitably linked to questioning and reshaping the institution's cultural values around teaching and faculty work.

The University of Denver has demonstrated commitment to high quality teaching. At DU, aspirations of a teacher-scholar model predominate, where faculty members find mutual benefits from the teaching and scholarship aspects of their work. The establishment of a Teaching Professorial Series in 2015 showed strong commitment to validating and sustaining the important role of teaching. DU's Faculty Senate Policies and Procedures Relating to Faculty Appointment, Promotion and Tenure follow national recommendations in stating that quality of teaching, for *all* faculty lines, should be judged using multiple sources of evidence. Yet in practice, according to years of discussions and DU surveys, teaching evaluation is often underemphasized, unsystematic, and anecdotal. DU suffers from the same barriers and roadblocks as other research institutions. Additionally, the highly decentralized model at DU creates a lack of consistency across campus as well as a lack of centralized support.

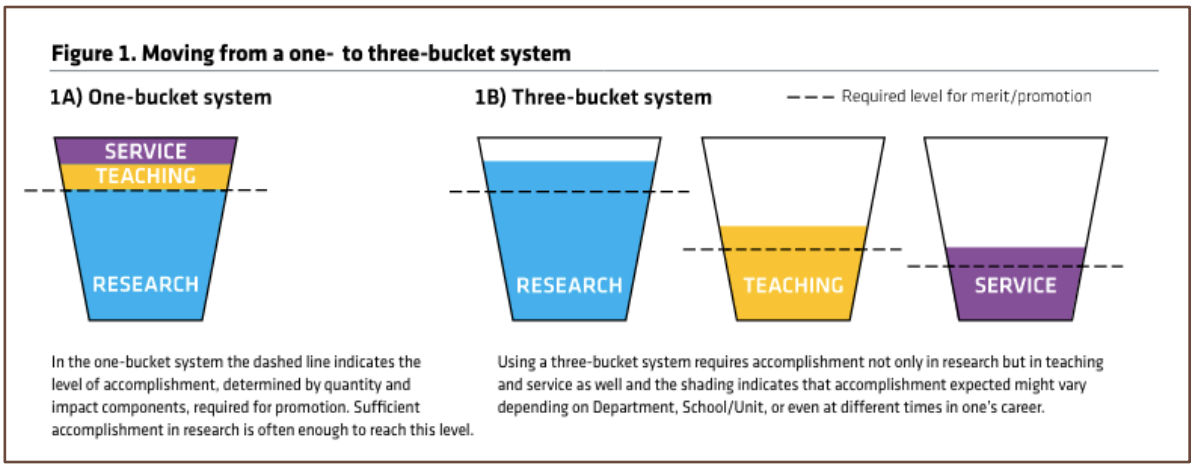
The purpose of this report is to act as a reference point for current initiatives relevant to changing the reward structure for teaching. Although this field is surprisingly vast, and no brief summary could capture all the literature or proposed frameworks, this is an attempt to pull together what the author considers to be the most vetted literature in the field, as well as useful examples from other institutions relevant to the DU context.

Limitations of Current Practices

> *"Research universities need to create an environment where the continuous improvement of teaching is valued, assessed, and rewarded at various stages of a faculty member's career and aligned across department, college, and university levels."*
>
> - *Association of American Universities, Aligning Practice to Policies: Changing the Culture to Recognize and Reward Teaching at Research Universities (2017)*

It is no secret that the current reward structure in higher education focuses faculty time and effort towards research. Teaching is touted as a strong value and a strength for most universities, yet common practices typically encourage faculty to spend their limited time and energy on efforts other than teaching. It is very often the case that faculty can be promoted without any serious focus on teaching.

The visual below from the University of California at Irvine demonstrates this current reality. The visual depicts a common "one-bucket system," where teaching, service and research are all expectations for faculty, yet it is possible for someone to demonstrate satisfactory achievement with only or primarily research contributions. The desired shift is toward a "three-bucket system," where research still may constitute the majority of someone's role, but a faculty member will be expected to achieve a certain level of accomplishment in teaching and service.



University of California Irvine - Figure 1 from Dennin, et al. 2017. "Aligning Practice to Policies: Changing the Culture to Recognize and Reward Teaching at Research Universities." American Association of Universities (p. 7).

Expectations for teaching, research, and service may vary. Faculty in DU's new Teaching Track Series have teaching expectations that constitute 90% of their position. For these individuals, a comprehensive and balanced teaching evaluation is vital. However, in all institutions that value student learning, and especially those with a teacher-scholar model, a true expectation for teaching is not only for those in a teaching track series. All faculty that have a teaching "bucket" should be expected to demonstrate accomplishment in teaching, albeit with different level of expectations.

Teaching evaluation is a difficult topic that does not inspire lighthearted discussion, yet without a comprehensive and fair system of rewarding, and thus evaluating, teaching, it remains as a lesser-valued aspect of faculty work. One solid indicator of how much an institution values teaching is the existence, or lack of existence, of a serious, balanced, and supported teaching reward framework (Bergquist & Pawlak, 2008). For institutions that value outstanding student learning, such a reward structure is essential. Examples of institutional practices working towards this goal are highlighted at the end as well as throughout the report.

> An Association of American Universities survey of 1000 faculty from 50 research universities "showed there is some disconnect between what is publicly supported within colleges and universities and what actually happens in day-to-day processes" to evaluate teaching. (p. 4)
>
> Only 15 % of those surveyed rated the evidence used to measure teaching as high quality. (Dennin, et. al., 2017)

Part of the difficulty of creating a solid teaching evaluation framework can be attributed to the competing goals of evaluation. Evidence that supports teaching evaluation can be used for different purposes. Summative evaluation is used to make a final judgment about progress or performance often compared to a benchmark or standard, whereas formative evaluation provides information during the process for the purpose of ongoing improvement. Summative purposes – such as comparisons and ranking, determining baseline levels of acceptable performance, or making decisions about merit, appointment, promotion and tenure – can be at odds with formative purposes – such as ongoing teaching development or students' ability to provide necessary feedback about their learning. It is possible, but delicate, to use data for both purposes. When evidence is used to make high-stakes decisions, the data integrity is questioned in ways not necessary for formative purposes.

Student Ratings of Instruction

> *"Using student feedback to inform teaching but not to assess teaching is important progress." - Philip B. Stark, UC Berkeley*

Student Ratings of Instruction (SRIs) are feedback forms designed to gather information about a course or learning experience from those for whom the experience is designed – students. There is widespread debate in the literature about SRIs, however all scholars on SRIs agree that they were designed to provide feedback about teaching but were never designed to serve as the evaluation of a faculty member. All scholars, organizational recommendations and institutional policies argue that SRIs should be only one source of information within an overall teaching evaluation, never the sole or primary source (Arreola, 2000; Benton & Ryalls, 2016; Berk, 2005). Yet even though they are

routinely administered in every college course with the intent to provide both formative and summative information, in practice they are often only used for summative purposes.

Berk suggested that there are four major limitations to using only student ratings for decision making:

(1) **Students' limited qualifications as raters** – Although students are uniquely qualified to rate their experience in a course, there are teaching behaviors and skills that students are not qualified to rate, such as content expertise, learning outcomes or quality of course materials.
(2) **Technical inadequacy and bias** – Professionally-developed forms are routinely tested for validity and reliability, but many universities use homegrown instruments, often created by faculty committees.
(3) **Misuse of scales and misinterpretation of ratings** – Although guidelines, instructions, and manuals usually accompany student rating scales, they are still administered at many institutions under uncontrolled, unstandardized, and/or inappropriate conditions which can significantly decrease the response rate or even render the answers invalid.
(4) **Inadequate source of evidence for decision making** – SRIs are the were never intended to be used as an evaluation of teaching, but rather a source of feedback that is used by a chair or committee to make an evaluation. It is the misuse of SRIs that is compounds their problematic nature. (Berk, 2014)

Teaching is a complex process that cannot be distilled into a single metric. The IDEA Center, a publisher of SRIs and advocate for the use of reliable SRIs, has long recommended that SRIs are a necessary but insufficient measure of teaching effectiveness. They conclude that SRIs should never constitute more than 30% to 50% of a faculty member's overall teaching evaluation (Hoyt & Pallett, 1999). The evaluation policies of the American Association of University Professors propose using "various measures of the effectiveness of [teaching] efforts" (AAUP, 2006). Yet policies typically do not reflect actual practice and overreliance on SRIs for determining a faculty member's teaching performance continues to exist (Seldin, 1999; Pallett, 2006).

Practice to Consider – Renamed Student Feedback Forms

Many universities have changed the name of their SRI forms, from the commonly used "Student Evaluations of Teaching (SETs)", to a name that more accurately reflects their intended use. Students are not providing a final evaluation of a faculty member. Rather, these forms provide information to be used by a committee or chair to conduct an evaluation. Some examples include:

- Student Ratings of Instruction
- Student Perceptions of Teaching
- Student Experience Questionnaire
- Student Response to Instruction
- Course Experience Questionnaire
- Learning Experience Survey

Teaching is a complex phenomenon without prevalent agreed-upon criteria, and alternative methods for collecting teaching evidence are time consuming and subjective. SRIs, although subjective in

their own right, are comparatively easy to administer and report than other methods. All stakeholders in the process want a fair and balanced way to get feedback on teaching, but because there are no easily-acquired, quick and objective alternatives, and decisions about teaching evaluation can be emotionally and politically charged, SRIs remain the default method. The search for the perfect has remained the enemy of the good.

Issues of Concern with SRIs

Student Ratings of Instruction are the most studied aspect of teaching in higher education, with well over 1000 studies conducted in almost 100 years. There is more research on and experience with student ratings than all of the other measures of teaching effectiveness combined (Berk, 2018). Because of its misuse as a high-stakes evaluation instrument, some studies about SRIs have garnered a great deal of attention, yet critics often lament the vast variability in controls and levels of rigor. There is much debate in the literature, which can be quite complex. Nearly everyone has strong opinions on SRIs, although relatively few scholars have been able to take the time to truly explore the literature in depth.

One very accessible source through which to understand the complexity of SRIs is Rice University Center for Teaching Excellence Blogs where Elizabeth Barre (a former Religious Studies faculty member and Teaching and Learning Center Director) conducted a deep dive into the literature as she co-chaired Rice's Teaching's Subcommittee on Teaching and Course Evaluations. She created widely-read blog posts and an extensive bibliography. Among much complexity, she concludes:

- Individual student ratings forms have low reliability, but class averages have higher reliability, relative to class size. The larger the class, the higher the reliability (thus the recommendation not to use SRIs for a class smaller than 10).
- Validity is more difficult to determine because we do not often agree about what exactly we are trying to measure (student attitudes, teaching practices, student learning?), or which alternative measures are the most useful for comparison.
- There are studies that have shown no correlation (or even inverse correlations) between the results of student evaluations and student learning. Yet, there are just as many, and in fact many more, that show just the opposite.
- She explains why issues of bias are more complicated than they appear. For example, grades and workload are correlated with SRI results, but are not considered to be a problematic bias. Gender is a point of bias but may also be related to the gender of students (in some disciplines females rate female faculty higher). Discipline does create some bias (humanities courses are often rated higher than social sciences, which are higher than STEM).

Similar to all other aspects of the literature around SRIs, issues of bias are hotly debated. At DU, reports such as the AHSS Committee report about Current Evaluation Measures and Recommendations explored some of the SRI literature and recommended, among many other things, that the questionable validity and potential bias of SRIs against women and minority faculty be clearly communicated to all those involved in tenure and promotion decisions, and that AHSS cease relying on SRIs as a significant factor in tenure, promotion, and merit-award decisions because of its systemic biases.

This paper cannot provide an adequate summary of the literature about bias in SRIs. Some of the most widely shared studies with shocking results about bias have also been the most highly criticized from a methodological perspective. One reasonable conclusion can be that for SRIs overall, when looking at large data sets and tightly controlled studies, issues of bias are not as simple as they may appear. Regardless, when a faculty member receives a personal insult through SRIs, or is unfairly judged due to issues unrelated to teaching, the damage done cannot be measured or erased.

Practice to Consider – McGill University

In an effort to uphold principles of academic freedom, integrity, responsibility, equity and inclusiveness, McGill developed a protocol for SRIs to be completely removed if students use hateful or discriminatory language. According to their policy, "This protocol is not intended to silence students or minimize the value of student feedback, which is crucial to curricular and pedagogical development at the University. Students do have real concerns about their instructors and courses; however, it is imperative that these concerns be expressed in a constructive and respectful manner." To support this effort, they developed guidelines for students on providing constructive feedback.

There can also be an effect on SRIs based on pedagogical practices. Many SRI questions were developed and validated before active learning methods became commonly used. Some have argued that certain questions could contradict modern teaching practices ("The formats for the lectures and the lecture presentations were informative" is on a DU Biology SRI form). When new teaching methods are used or faculty members engage in innovative practices, there can be a predictable dip in SRI scores. It is possible that an overemphasis on SRIs can disincentivize pedagogical experimentation and innovation.

It should also be noted that students may not view SRIs as a venue for giving useful feedback. In the author's many discussions with students collecting data on teaching, students very often comment that they feel no one reads the SRIs, or that they have given constructive feedback in the past but have not seen it acted upon. Students also need to see the value of giving feedback if we expect them to be constructive and take the process seriously.

Although there is wide agreement to use SRIs as just one source of feedback within a larger context, in practice this is still not often the case. Barre's final conclusions about SRIs echo these sentiments. In her own words:

- "Overall, student ratings of teaching effectiveness are a useful, *but ultimately imperfect*, measure of teaching effectiveness. Despite this, we have not yet been able to find an alternative measure of teaching effectiveness that correlates as strongly with student learning. In other words, they may be imperfect measures, but they are also our *best* measures.

- Yet, there is simply too much room for error with individual comparisons, so we should put a moratorium on using student ratings results to rank and compare individual faculty to one another."
- "While comparing faculty to one another is dangerous, the quantitative scores can still be valuable if used to chart growth of a single instructor over time." (Barre, 2018)

Practice to Consider – University of Southern California
As of Spring 2018, the Provost at University of Southern California mandated that SRIS would no longer be used in tenure and promotion decisions. Newly adjusted SRIs will still be administered, and faculty will be expected to explain how they use student feedback to improve instruction in their teaching statements, but SRIs will be taken out of high-stakes personnel decisions. A peer review model will be used instead along with additional initiatives to support teaching development and reward teaching excellence.

Revising Student Ratings of Instruction

In general, gathering student feedback on courses and teaching is valuable and necessary. We do not want to take the student voice out of the teaching and learning equation. As SRIs are baked into the higher education culture and will likely continue to be used, it is in everyone's interest to make them as useful and objective as possible, and to foster supportive use. Suggestions and current initiatives focus on two areas: changing the nature of the questions and using them appropriately.

Change the nature of the questions

Many universities have revised their SRIs questions in recent years in part to address concerns over bias and misuse. An immense scholarship on the psychometric qualities and standards of creating reliable and valid SRI questions exists (Arreola, 2000; Benton & Cashin 2012; Hativa, 2013) and is well beyond the scope of this report. But some recommendations and ideas about current initiatives will be shared.

Because there is a general sense that questions that ask students to rate courses or instructors in a generic way are more susceptible to bias, some institutions (e.g., University of Kansas, University of Southern California, Stanford), have changed their questions to focus on observable behaviors or feedback on teaching methods, rather than general attributes. Many of these revisions have resulted in more descriptive and criterion-based questions, as opposed to evaluative and norm-based. Some questions explore student contributions to the course to help interpret results (interest and motivation, time and effort put into a course), or include questions that ask for a self-report of student learning rather.

For example, some of the universities explored for this report have redesigned SRIs with questions similar to the approach found in the Student Assessment of Learning Gains (SALG) instrument (See Stanford example below). SALG is a free instrument developed 20-years ago as part of an NSF grant to focus questions exclusively on the degree to which a course has enabled student learning. The questions ask students to report the extent to which they made gains on a host of learning objectives as a result of their course and the extent to which course features and faculty interactions facilitated their learning (http://www.salgsite.org/).

Practice to Consider – Stanford University

Stanford redesigned their evaluation process in 2015 to focus on:

- A redesigned SRI focused on learning, with customizable questions
- An emphasis on mid-semester feedback
- Supporting ways to use SRI results to make changes in teaching
- An overall multi-modal evaluation

Stanford's redesigned SRI includes:

- Nine standard questions (3 about student participation, 5 about the course - 3 of which are open-ended, and a question for future students considering this course which is public to all students)
- Customizable questions from the instructor about learning goals (can add 4-10 learning goals), usefulness of course elements (can add up to 5), and up to 3 other questions

If the SRI is truly to be formative in some way, narrative comments should continue to be included and could be focused specifically on course design and instructional practices rather than general statements (such as eliminating, "please comment on the weaknesses of this instructor.") Instructors should also be allowed to add their own questions.

The use of global questions (evaluative broad-stroke questions such as, "overall, this is an excellent instructor") are still hotly debated in the literature. Some argue that they provide a summary of students' perceptions about their learning experience, especially when they appear at the end of the instrument. However, others argue they are more subject to student bias, and their simplicity allows them to be more easily mistaken for objective information in summative decisions.

DU Pilot – New SRI Questions

In the Spring quarter of 2018, a working group of the APC and SRC in Faculty Senate, along with OTL staff, conducted focus groups with students and piloted new SRI questions designed to include questions about student learning and student contributions to the course. Survey questions also asked volunteer faculty participants and students about their reactions to these questions. Results of this pilot are not yet available, but important faculty discussions have begun about revisiting the goal and purpose of SRIs at DU.

Appropriate use of SRI instruments

The recommendation remains, that no matter what questions are used, it is the misuse of SRIs that is most problematic. Effective use not only means using multiple sources of evidence in addition to SRIs in a comprehensive evaluation, but also appropriate interpretations of the SRI instrument itself.

In interpreting SRI results, considerations of potential bias and use of multiple measures should be taken into account. Nearly all reports about teaching evaluation call for chair and committee training on the limitations and appropriate use of SRI data. The Office of Teaching and Learning (OTL) created a handout and sponsored four Academic Chairs roundtable discussions over the last four years to share best practice in using SRIs, and some training may occur within some departments, but DU, like many institutions, does not routinely provide formal support in this area.

How data is presented to faculty and chairs is also a topic of discussion in many of the reports reviewed (e.g., presenting data in context with courses that have similar characteristics such as size or topic area, controlling for known biases or other relevant factors, issues of who has access and for how long, minimum response rates for reporting, etc.). Currently, SRI data is presented at DU with comparisons to other courses in the department, the unit, and the university. Other reporting recommendations exist but are outside the scope of this report and would need to come through Faculty Senate. Given the overwhelming concerns over bias and misuse of the data, this often-overlooked component could be further explored.

Practice to Consider – Yale University
In an effort to change the culture of SRI use, Yale University added a "CLICK HERE TO NOMINATE FOR A PRIZE" button to screens that contain SRI forms to solicit positive feedback about teaching.

DU's Current Use of SRIs

Currently, DU has a very complex SRI system that allows decisions about questions and policies to be developed at the unit, and sometimes department level, reflecting the decentralized nature of the campus. The exact number changes each quarter depending upon what classes are offered, but for Spring quarter 2017, there were over 50 different SRI forms used for 17+ units/programs, containing a total of 243 unique questions (504 questions in all).

While the graduate programs typically use one form for the entire unit (including GSSW, GSSP, DCB, MCE, SCOL, KSIS and UCOL), there is much more variety in the undergraduate programs. For example, there are 11 forms used in Natural Sciences and Mathematics (NSM) and 12 in Arts, Humanities and Social Sciences (AHSS). Although disciplinary differences in teaching exist and more specific questions allow for more useful feedback, the current system is known internally to be difficult to manage and is externally viewed by vendors and those from other institutions as extraordinarily complex.

DU does not have a consistent record of when SRI questions have been changed, but according to current institutional knowledge, many of the current questions were created during an intensive modeling study undertaken in 2005-2006 by Faculty Senate. A committee spearheaded by the head of Institutional Research at the time conducted an extensive exploration into peer institution practices and the current literature. Deans and the faculty were involved in the process and a large part of the discussions at the time revolved around the scale used in the form (strongly agree to strongly disagree, with or without a midpoint).

Calls for revisiting the SRI instrument periodically appears to be good practice although there does not appear to be a known practical recommendation for how often this should occur. University College changed their form in 2017 and GSSW changed their form in Fall 2018. The Morgridge College of Education piloted a new form about a decade ago, but it was not officially changed. In recent years, there have been proposals to add a question about course rigor and about inclusive excellence, and Faculty Senate piloted separately a new set of student-focused questions, but none of these have been formalized.

For most units, it appears the forms have remained the same since the 2005-2006 revisions. At that time, each unit or department decided on the standard questions for their particular group, and it was decided that 3 global questions would be asked on every form across campus. The global questions are:

- Overall, this is an excellent course
- Overall, this is an effective instructor
- I learned a great deal in this course

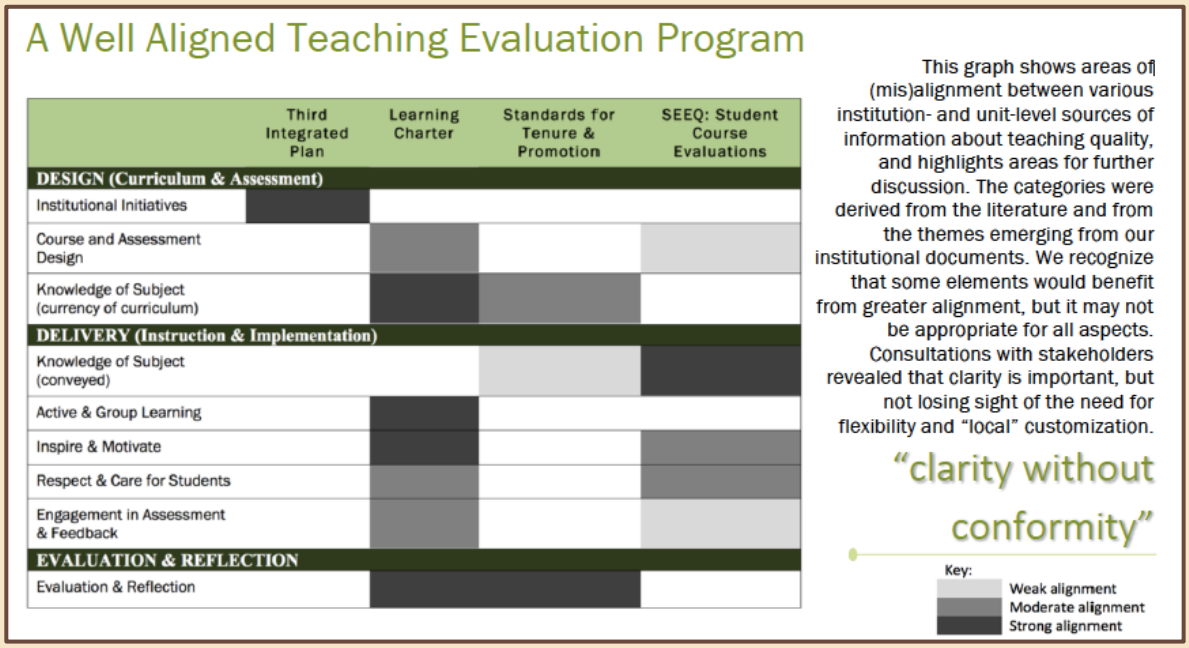
DU used a homegrown online software system to administer SRIs starting in 2006, moved from mostly paper evaluations to fully online course evaluations in 2011, and is currently transitioning to a new vender called EvaluationKIT which should allow even more flexibility and options. According to the Office of Institutional Research, in the fall of 2018, the average response rate was 75%. This was up from an average of about 60-62% for several years before this. As of 2011, faculty have been able to add 2 questions of their own to each form.

Throughout the author's involvement in many committees over the years, it appears most faculty members do not seem to know who created or decided upon the questions used on SRIs. Many faculty members seem to think that Institutional Research owns the process and the questions, while IR staff reiterate that the faculty owns the process and Faculty Senate must approve any changes.

Although policies for managing and changing SRIs are not formally established in writing, according to discussions with Institutional Research staff, Deans must approve any changes to the forms at the college level. Changes at the University level need to be approved by Faculty Senate and the Provost. Faculty members, and perhaps students, should be consulted at all stages of these changes.

Practice to Consider - University of Saskatchewan

In preparation to develop a shared teaching quality framework, a cross-campus committee at the University of Saskatchewan undertook a year-long project to map out and determine the alignment between stated goals and policies for teaching and current practices. The resulting heat-map showed how little alignment was occurring between the universities stated goals and current practices.



**From Building a Shared Framework for Teaching Quality at the University of Saskatchewan: Phase I Final Project Report (p. 29)*

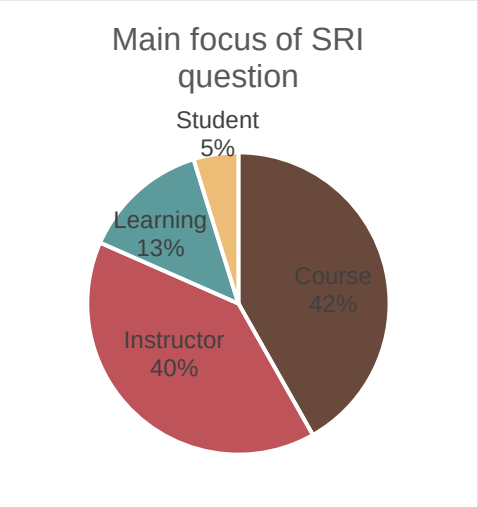
In 2017-2018, a working group made up of two committees on Faculty Senate (APC and SRC) and members of the OTL began to address the issue of changing SRIs at DU. Based on the University of Saskatchewan's model shown above, an attempt was made to determine the level of alignment between DU's stated values and goals for teaching and its ongoing practices. Bridget Arend from the OTL conducted an analysis, comparing Teaching Aspirations created by a group of DU faculty and staff (goals and values for teaching) with the current questions asked in DU's 246 SRI questions. Additional efforts to expand this approach to other methods of teaching evaluation have not yet happened.

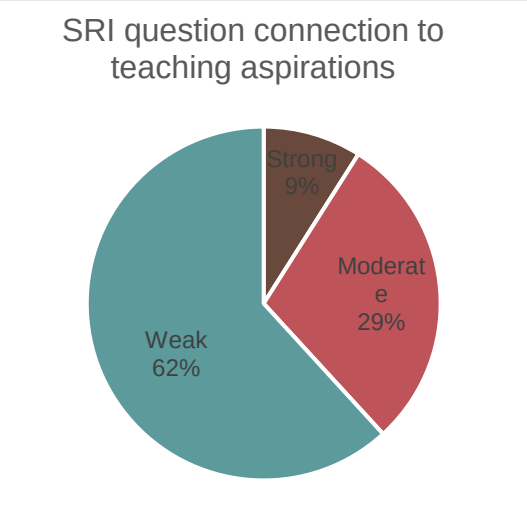
DU Pilot - Alignment of SRIs with Teaching Aspirations

Three rounds of inductive and deductive coding of DU's 504 SRI questions was conducted to explore SRI questions through the lens of significant, learning-centered, inclusive, and reflective teaching practices.

Preliminary Findings:

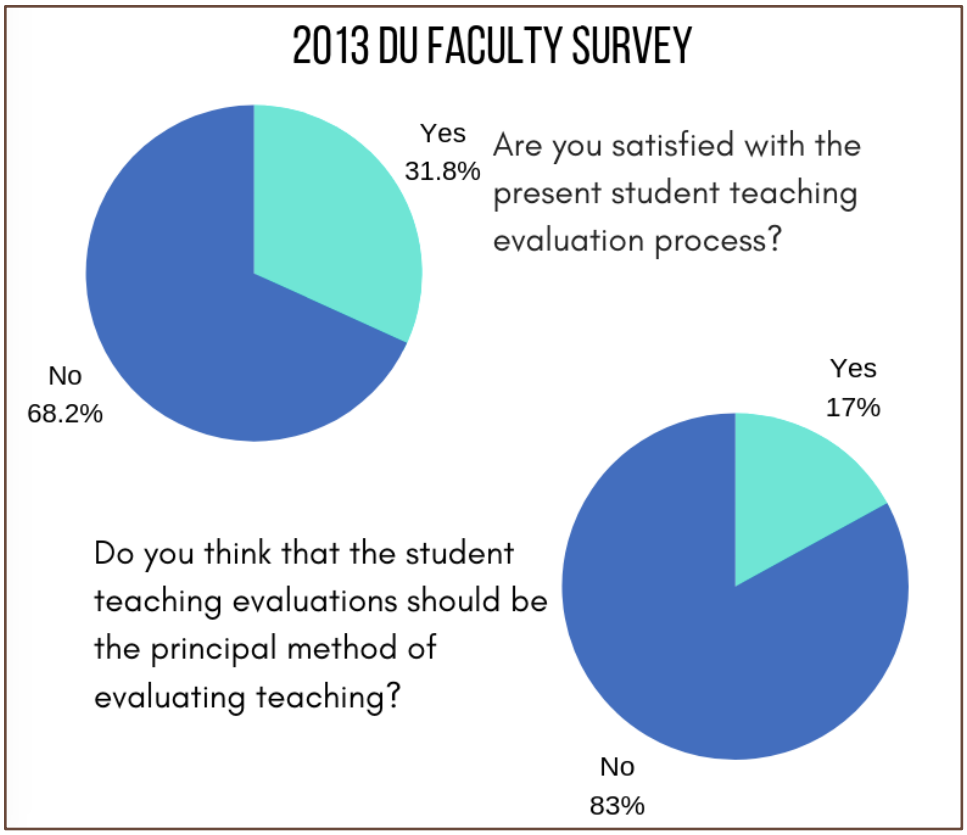
- The vast majority (82%) of questions focused on instructor traits and course format. (e.g., The instructor was organized, knowledgeable, strengths of this course)
- Less than 10% of the questions had a strong/direct connection to the type of teaching described in our teaching aspirations. (Where a response to the question was considered very likely to relate to one of the articulated teaching aspirations.)
- The more abstract the question, the more difficult it is to determine what is being measured (e.g., challenging, organized). The more specific the question, the more it is likely to inform and improve teaching and learning, but the less applicable it is to different disciplines or types of courses.





Teaching Evaluation: Can we measure what really matters? (Arend, 2017)

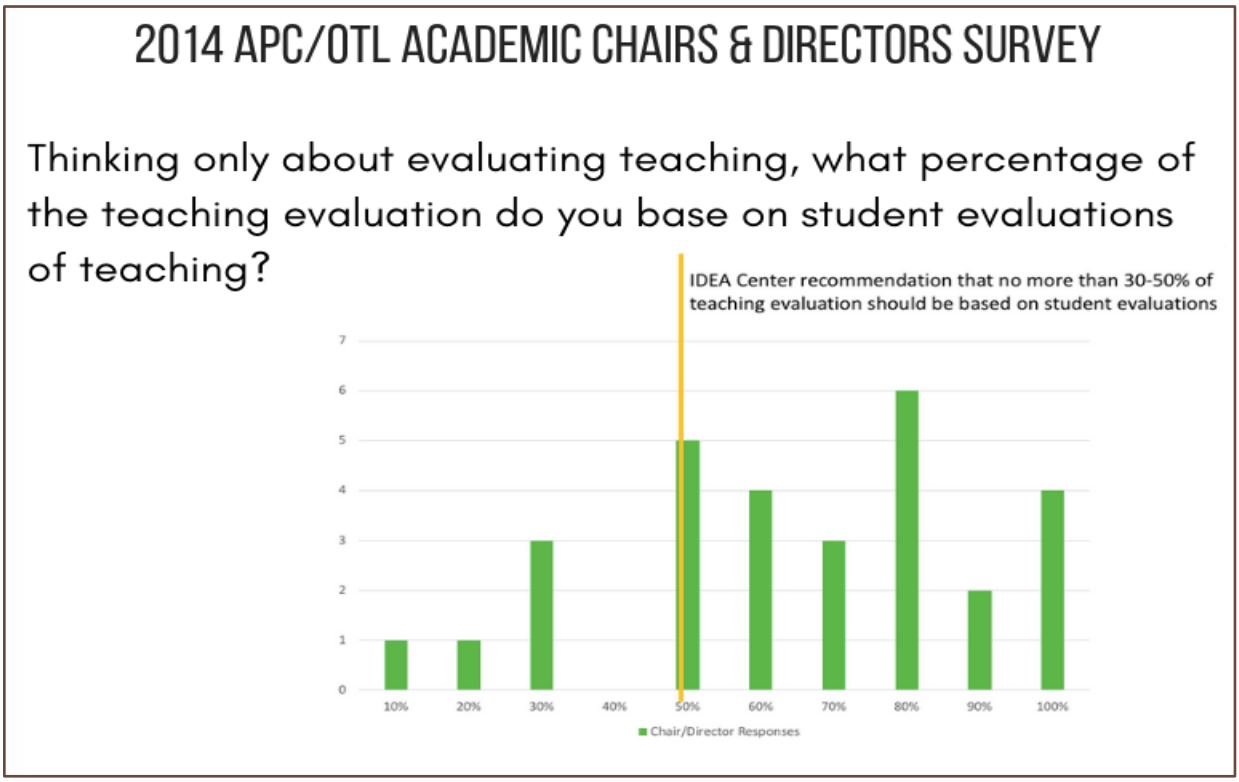
Not surprisingly, and similar to national sentiments, the majority of faculty members at DU are not satisfied with the use of SRIs to evaluate teaching performance. In a 2013 Faculty Senate survey (n=240), only 47% of faculty respondents said they were satisfied with the standardized questions that appear on the present student teaching evaluation. Other questions on the survey showed similar dissatisfaction on the overall process of using SRIs, as shown in the image below.



2013 Faculty Senate Survey of Student Teaching Evaluations

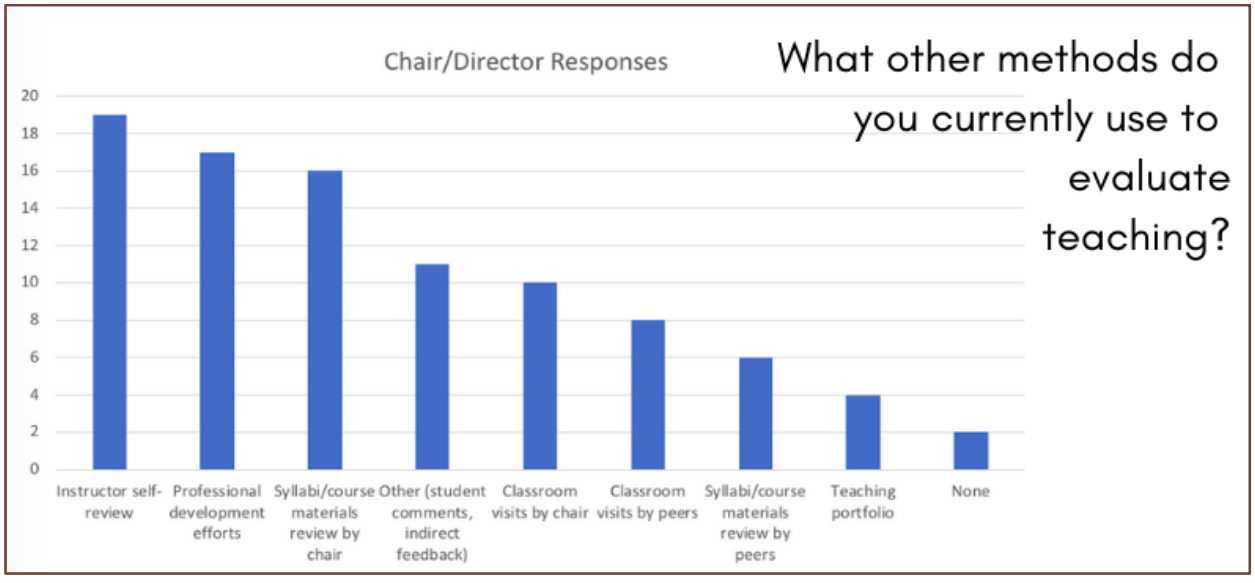
DU's Faculty Senate Policies and Procedures Relating to Faculty Appointment, Promotion and Tenure, state that quality of teaching should be judged using multiple sources of evidence. "The evidence to be used, though not to the exclusion of other evidence available, shall come at a minimum from each of the following three areas: (A) self-analysis, (B) teaching observations and/or the review of course material, and (C) empirical data on teaching effectiveness" (2015, p. 26).

The policy is in place to align DU with best practice recommendations for using multiple methods of evaluating teaching. Yet in practice, SRIs still appear to predominate. Anecdotally, this is the narrative across campus. And, in a 2014 OTL/APC survey to Academic Chairs and Directors, only 5 of the 29 departmental chairs who responded said they based less than 50% of teaching evaluation on SRIs. The IDEA Center recommendation is that SRIs should count for no more than 30-50% of an overall evaluation of teaching.



2014 Faculty Senate APC and OTL Survey of Academic Chairs and Directors

In that same survey, when Chairs were asked what other methods were use to evaluate teaching, many options were cited (see image below). There are certainly many ways to demonstrate evidence about teaching effort and performance, and faculty members can and should be able to use unique and non-standard methods of evidence. However, without commonly agreed-upon standards and criteria across campus, individual faculty members can be subject to very different realities.

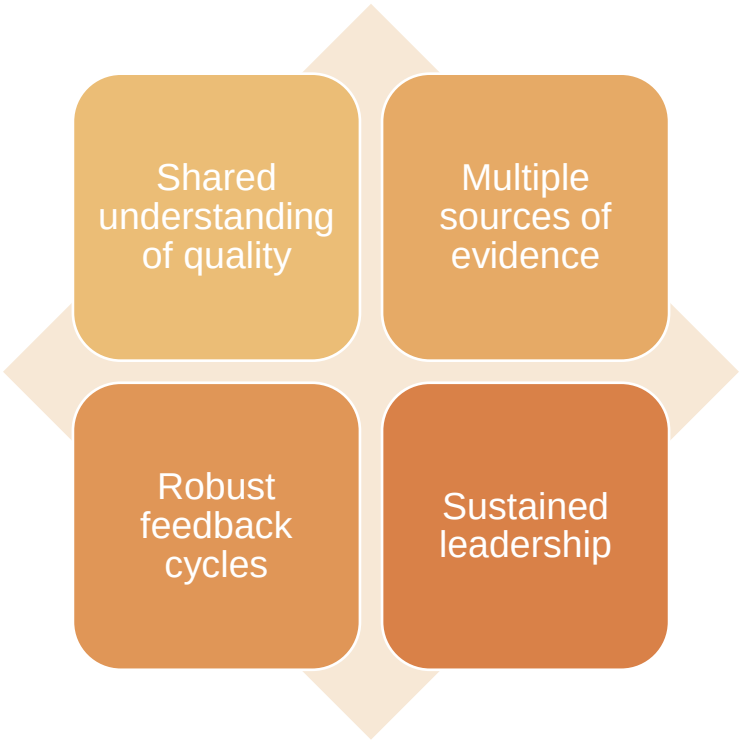


2014 Faculty Senate APC and OTL Survey of Academic Chairs and Directors

Comprehensive Teaching Evaluation Frameworks

In nearly all the literature about best practice for teaching evaluation, there are calls for developing a comprehensive model, where values and goals align with policies and practices that are supported at all levels across campus. These models go well beyond changing SRI questions. They involve culture change. Any focus on improving the reward structure for teaching should also include focus on supporting teaching practice and acknowledging the time and effort that is spent on teaching. Although this report focuses on the evaluation of teaching, these pieces are all interrelated and the success of any changes in the teaching evaluation process would be better fostered with an overall focus on supporting teaching excellence in all areas.

An extensive literature review created for an International Forum on Teaching Evaluation in Canada in 2017 identified four recurring themes required for effective teaching evaluation; a shared understanding of quality, multi-faceted data and evaluation, robust feedback cycles and sustained leadership for education, engagement, and change (Wright et al., 2014). Examples and ideas about these four areas will be shared in the rest of this paper.



Shared Understandings of Quality

Many comprehensive models of teaching evaluation have been proposed over the years. A critical component in all of them is the articulation of the type of teaching desired. This is typically the first step in any evaluation process – it is difficult to measure something if you do not have a clear definition of what is to be measured – yet this important step has been historically overlooked in teaching evaluation. The rationale may be that we know good teaching when we see it, or that teaching is too personal and complex to define. Certainly, teaching is complex and varied, and there is not one "right" way to be an effective teacher. Yet if we are to judge teaching performance, even on the basis of effort and development, we need to articulate the kind of teaching we are aiming to reward.

Practice to Consider – University of Oregon

UO has defined teaching excellence in three main areas:

- Inclusive Teaching
- Engaged Teaching
- Research-led Teaching

Oregon uses these three areas as the foundation for their teaching evaluation. The questions that are asked in the SRIs are framed around these three areas, as are the elements expected in instructor reflections, and the criteria present in faculty teaching evaluation rubrics. Professionalism is considered a fourth area but is used only as a baseline competence rather than criteria for ongoing development.

Many of the comprehensive models for teaching evaluation suggest a set of categories or criteria for defining teaching quality. Similarly, most institutions actively working on creating their own model are defining their own categories or elements. Frameworks and models already exist. However, each institution needs to choose a framework that resonates with their particular culture and context. Some examples of categories to define teaching quality are shown in the table below. (Rubrics and details for these examples as well as others are found in the TEI Portfolio site under the Teaching Evaluation tab).

Table 1: Examples of Teaching Quality Frameworks from other Institutions

Institution	**Teaching Quality Dimensions**	**Source**
University of Kansas	• Goals, content and alignment • Teaching practices • Achievement of learning outcomes • Classroom climate and student perceptions • Reflection and iterative growth • Mentoring and advising • Involvement in teaching service, scholarship or community	Boyer (1990), Glassick, et al., (1997)
CU Boulder	• Goals, content and alignment • Preparation for teaching • Methods and teaching practices • Presentation and student interaction • Student (and other) outcomes • Mentorship and advising • Reflection, development and teaching service/scholarship	Based off of the KU rubric
University of Saskatchewan	• Design (course design) • Delivery (feedback on instruction, SRIs) • Evaluate & Reflect (self-reflection, plan for growth) • Develop, Share, Lead (also professional development and SoTL)	Kolb's (1984) experiential learning cycle
Florida International University	• Learning centered teaching • Evidence based teaching • Culturally responsive teaching	Developed internally
United States Air Force Academy	• Evidence-based, learning-focused teaching practices • Respectful, engaging learning environments • Sustained professional development and improvement in their teaching and course design	Developed internally
Boise State University	• Design course materials in alignment with course learning outcomes • Implements evidence-based practices • Uses an inclusive, student-centered approach • Practices reflective teaching to drive continuous improvement of teaching	Developed internally
University of Oregon	• Inclusive Teaching • Engaged Teaching • Research-led Teaching • Professionalism	Developed internally
University of Southern California	• Respectful and professional • Challenging and supportive • Inclusive and diverse • Relevant and engaging • Prepared and purposeful • Fair and equitable • Evidence-based	Developed internally

Teaching Aspirations Work

Work towards a campus-wide understanding of teaching quality has already begun at DU.

In 2015-16 the Office of Teaching and Learning sponsored a faculty learning community of 14 active members across disciplines to tackle the question, *What kind of teaching do we aspire to at DU?* The group engaged in a modified design thinking process to gather evidence from their own experiences, the educational literature, students, rapid prototypes, and interactive feedback sessions with at least 5 faculty groups (Arend & Pitts, 2015).

Thirty statements about teaching emerged in four main categories that closely align with the teaching ideals espoused in DU's strategic planning documents (found on the TEI Portfolio site). The categories and statements are considered a continually ongoing work-in-progress, meant to inform discussions and support of teaching rather than to define or summarize. But could form the basis of a content framework for understanding teaching quality.

At DU, we aspire to teach in a way that is…

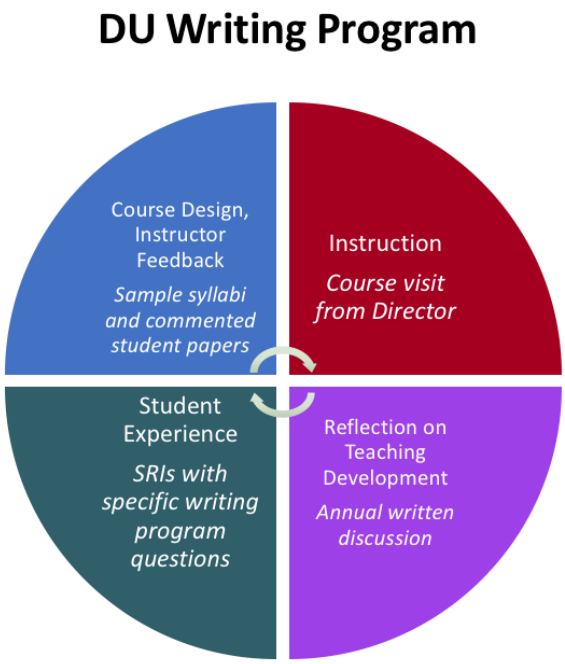
- **Significant and Impactful**
- **Learning-Centered, Meaningful and Active**
- **Inclusive, Inviting and Empathetic**
- **Reflective and Evidence-Based**

At DU, some units and departments have taken steps to develop a common understanding and articulated process for teaching evaluation, a few in conjunction with the Teaching Aspirations work. Three examples that have already been in use are shared below. (Details and contact information for these DU examples are found in the TEI Portoflio site.)

Writing Program

The Writing Program has utilized an extensive process for evaluating teaching for many years. Teaching evaluation includes:

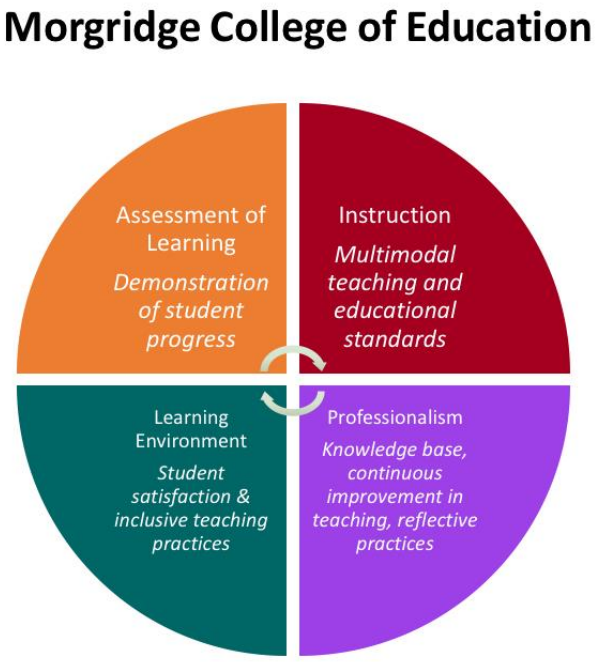
- **Reflection on teaching development** - 2-4 page written discussion of the past year's teaching
- **Course design/Instructor feedback** - upload syllabi and 3 commented student papers showing instructor feedback from each writing course taught
- **Instruction** - course visit from the Director of Writing and submission of artifacts of teaching
- **Student Experience** - SRIs including questions specific to the writing program



Morgridge College of Education

The Morgridge College of Education recently approved a new approach to teaching evaluation based on four areas:

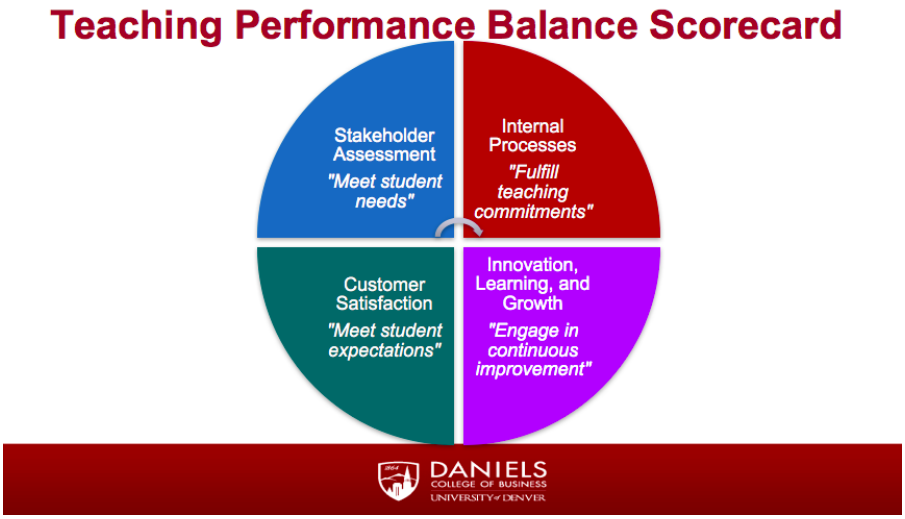
- **Professionalism -** demonstration of knowledge base, continuous improvement in teaching, reflective practices
- **Instruction** - demonstration of multimodal teaching and educational standards reflected in content
- **Assessment of Learning** - demonstration of student progress
- **Learning Environment** - demonstration of student satisfaction and inclusive teaching practices



Daniels College of Business

The Daniels College of Business has been piloting an approach to evaluating teaching based on four main categories of teaching performance:

- **Meeting student needs** - course design, challenge, alignment, teaching strategies, inclusivity, etc.
- **Meeting student expectations** - student satisfaction through course rating forms
- **Engaging in continuous improvement** - course and curricular innovation, professional development around teaching, etc.
- **Fulfilling teaching commitments** - course preps, number of students, types of courses taught, student engagement outside classroom, etc.



Common Aspects

Three units/programs at DU have created teaching evaluation frameworks unique to their particular culture and disciplinary norms. All are variations of these general categories of teaching evidence.

Table 2: Possible Teaching Evaluation Categories

Possible Teaching Evaluation Categories
Course design
Instruction/Facilitation
Student experience/Satisfaction
Student learning
Continuous improvement in teaching
Teaching commitments

Clearly there are many variations of teaching quality definitions. Some are focused on the type of teaching desired (inclusive, evidence-based), whereas others are focused on the common aspects and roles of teaching (course design, teaching development). If the goal is for a faculty member to fairly demonstrate all aspects of their teaching within a given framework, it appears that, despite their differences, many of the examples explored could be successful.

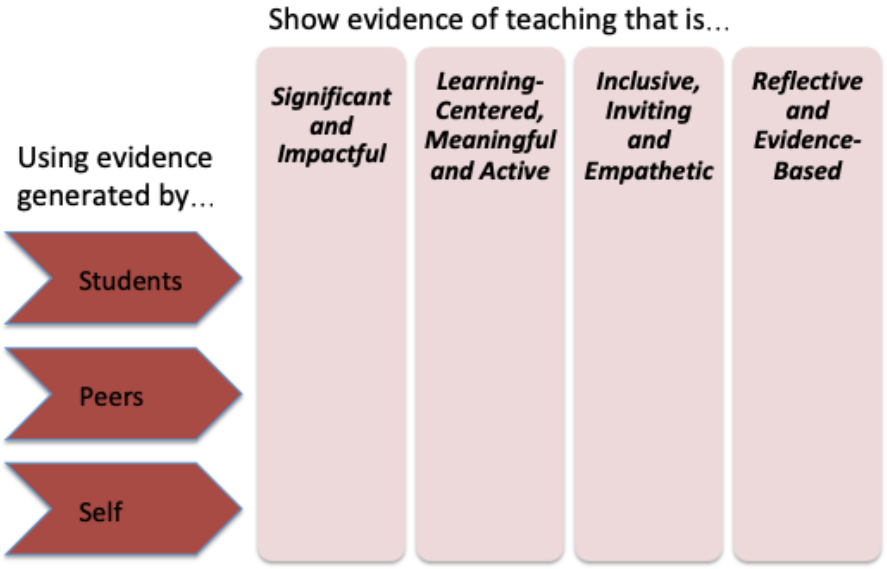
Most institutions explored for this report determine a framework at the university-level that is generic enough to be applicable to multiple disciplines. They then encourage units or departments to customize as appropriate. According to the CU Boulder model, this “provides the university with a common framework while preserving disciplinary identity and specificity” (TQF Handout, 2017). However, at DU some unit-level frameworks already exist.

Multiple Sources of Evidence

There are endless methods that could be used to demonstrate teaching effort and performance. In practice, teaching evaluation practices tend to prioritize a narrow dimension of teaching activity (the behavior of the instructor in the classroom) and a limited source of evidence (student evaluations).

Most recommendations suggest that methods to be used demonstrate evidence of the categories within a shared understanding of quality, as well as represent the many voices of the data – from students, from peers, and from the instructor.

For example, an ideal structure based on DU's Teaching Aspirations could include evidence from all three areas that support the desired teaching goals.



No alternative method will be a perfect measure of teaching. All sources of evidence will have subjectivity, and all will provide information about only some aspects of teaching. No one source or measure will provide an overall measure of teaching. Rather, it is the triangulation of various pieces of evidence, from different sources, that can be put together in context to create an accurate picture as to the ongoing development of a faculty member's teaching practice.

According to Miller & Seldin (2014) in repeated surveys of over 500 academic Deans, the most widely used sources for summative decisions are student end-of-course ratings (94.2%), chair and dean ratings (79%), self-evaluation (67.6%), dean evaluation (67.6%), classroom visits (60.4%), committee/peer evaluation (52.5%) colleagues' opinions (41%) and scholarly research/publications (28%). Of these top nine options, all except colleagues' opinions have increased in use since 2000, including SRIs, which increased in use from 88% in 2000 to 94% in 2010.

The more common alternative sources of evidence for teaching quality are briefly summarized according to their source: student feedback, peer review, and self-analysis.

Student Feedback

Clearly SRIs have been the most commonly used form of gathering student feedback about teaching. But given the current structure and limitations of SRIs, they are not ideal instruments to gather formative feedback to actually improve instruction. There are other methods for gathering student feedback about teaching. These could be considered in two categories: student perceptions about the learning experience, and direct measures of student learning.

Student Perceptions

Students are uniquely positioned to provide insights on teaching. They spend the greatest amount of time engaging with teaching and are the ultimate recipient of any teaching actions. Although students may not always be able to identify the long-term impact of a learning experience, only they can speak to their perceptions of the immediate impact. Student's perceptions about teaching could be collected in some of the following ways.

Faculty reflections on SRI data/themes – especially the open-ended questions. Instructors can reflect on changes made as a result of SRI feedback, self-reflection on student concerns or achievements, themes identified, or changes made over time.

Mid-quarter student feedback – formative feedback from students collected mid-term provides valuable information to instructors but also lets students see that their feedback is actually taken into consideration and acted upon. Gathering mid-quarter feedback can be acheived through anonymous surveys, through class discussions, through student representatives, or through Student Feedback Sessions offered through the OTL. The OTL also has a webpage with sample questions and ideas for gathering feedback. In any case, being careful to keep data confidential, simply documenting that this occurred, along with a few key themes that emerged, shows evidence of using student feedback for ongoing improvement.

Instructor-created surveys or student comments – instructors can ask for student feedback at any time about learning, teaching methods, resources used or class activities. Sometimes survey questions might be asked in conjunction with assignments, tests or papers that elicit feedback on the teaching and learning process. These data can be summarized and shared as a form of feedback on teaching.

Student letters – student letters are sometimes solicited for promotion and tenure packets, and student emails or notes can show evidence of an instructor's willingness to provide extra support or that a learning experience had made a significant impact on a student.

Alumni surveys or exit interviews – gathering feedback from exiting students or recent alumni can be a valuable source of seeing the long-term impact of a course or learning experience.

Student Learning

Students provide valuable information all the time through the work they produce in a course. Direct measures of student learning can provide information about teaching, but with some cautions. In the 2014 DU Chairs and Directors survey, ideas about using student learning data to show evidence of

teaching varied widely. Responses ranged from “we couldn’t possibly assess teaching without it” to “I would never use learning assessment data to punish faculty.”

Although learning is the ultimate goal of teaching, courses have varying levels of expectations and student preparedness and we would want to ensure that teachers are rewarded fairly given the contexts of their classrooms. Grades, given their subjectivity, would never be an accurate way to assess teaching. Some potential sources include the following.

Student artifacts – showing select examples of student work can provide insight into the goals and outcomes of teaching methods. In the Writing Program, faculty have been asked to share student papers with the instructor’s comments for samples that are considered both stronger and weaker. This demonstrates not only the student outcomes and the range of students the instructor is teaching, but also how the instructor provides feedback to students.

Summaries of student progress – Select student work could show the impact on specific teaching methods or changes in instruction. Pre- and post-results on a certain test question, progress in ongoing assignments, or student work that shows changes in conceptual understanding could demonstrate an impact on learning in a certain area of focus within the course.

Formative assessment – Collection of formative feedback on learning, such as popular Classroom Assessment Techniques (CATs), can show intended, or sometimes unintended changes in student learning as the result of instruction.

Peer Review

Peer colleagues can provide expert feedback to help improve teaching that students cannot provide. Peers bring content knowledge and context about students and are uniquely equipped to provide feedback on course materials and student work. The peer review process for teaching is often suggested because of its parallel process with peer review for other forms of scholarship (Berk, 2005).

Peer reviews are the most common choice for adding evidence about teaching beyond SRIs. But it should be noted that there is consensus among teaching experts that peer observation and review data should be used for formative rather than for summative decisions (Berk, 2005). When engaging in peer review or peer observation, the goals should be clearly stated at the beginning. Fletcher (2018) provides a distinction between three common uses of peer review: for evaluation, for development, or for collaboration, and how the dynamics, relationships, and benefits change with each purpose.

In this process, the definition of peer should be clarified. Weimer (2010) argues that peer evaluation programs can fail because they are based on the assumption that more experienced faculty are the most qualified to judge the teaching effectiveness of less experienced faculty. She suggests that if the goal is to improve teaching, faculty need a diverse collection of colleagues who can play a variety of roles. This may include peers in other departments, at other institutions, with similar goals, or someone with a background in teaching. She suggests seven possible peer roles, organized as a developmental sequence:

1. colleague as collaborator (working on a shared project such as designing a new assignment)
2. colleague as co-learner (of teaching scholarship, a new instructional practice or tool, etc.)
3. colleague as student (offering possible student reactions to course materials, exercises)
4. colleague as questioner (asking about pedagogical beliefs or course policies, for ex.)
5. colleague as critic (constructively disagreeing, identifying practices that may limit learning)
6. colleague as advocate (speaking publicly about policies that enhance or compromise learning)
7. colleague as confidant (listening to one's joys and struggles)

Nancy Van Note Chism's book, *Peer Review of Teaching: A Sourcebook* (2007), is considered to be an invaluable guide to structuring peer review, including pitfalls to avoid and many sample templates and forms. The follow suggestions of possible peer review activities all come with the same caveat that this type of evidence is best used for formative purposes, with faculty self-reflection about the reviews serving as evidence for summative evaluation purposes. Some possible peer review activities that can also create a culture of sharing and peer learning include the following.

Peer review of teaching artifacts – In contrast to reviewing teaching materials for the purpose of evaluation, peer activities that serve teaching development purposes can also act as peer review. Some known options include a syllabus exchange, annotated syllabi creation, classroom visits, or assignment reviews. From these activities, instructors could share their self-reflections, summaries or themes from the peer feedback, before and after versions of an artifact, or plans for future changes.

Peer collaboration on teaching projects – Peers can work together to tackle an issue identified within the department that crosses many courses, or to implement new teaching methods or focus on types of desired learning (self-directed learning, inclusive practices). The DU Writing Program asks faculty to work in small groups on teaching projects and write summary reflections of their work, which then can become evidence for promotion packets.

Peer observations – Although peer observations are common in higher education, they are also fraught with potential difficulties if done for evaluative purposes without following a systematic process. Ideally peer observations are conducted by multiple people and occur over multiple sessions. Recommendations for best practice include a process that includes:

- ***Faculty ownership.*** The concerns and goals of the faculty member under review should guide the process.
- ***Confidentiality.*** Discussions of teaching should be kept between the faculty and the peer reviewers and only shared with others by the reviewee.
- ***Relationship of equals.*** Colleagues need to work together as equals in a consultative, not evaluative process.
- ***Collegial feedback.*** Reviewers should provide feedback that is constructive and collegial rather than evaluative, with an aim towards collective learning.
- ***Open-ended process.*** Like all forms of scholarly endeavors, peer review should occur over time with cycles of practice, review, and application of feedback as part of a culture of ongoing improvement. (Arreola, 2007).

The OTL has a collection of class observation rubrics and observation instruments are also easily found online. Observation forms that include many open-ended questions and explore commonly agreed upon teaching aspirations, chosen by the instructor being reviewed, can feel more supportive than checklists or forms with a rating-scale.

Some new observation forms in STEM fields have been used in peer review and claim to focus on observable behaviors, as opposed to judgments about teaching performance. For example, the Classroom Observation Protocol for Undergraduate STEM (COPUS), involves some training on its use but looks for the use of evidence-based practice in the classroom. (Smith, et al., 2013). The Reformed Teaching Observation Protocol (RTOP) was designed to determine where the activities in a classroom fall on the continuum from teacher-centered to student-centered (Budd, et al., 2013).

Self-Analysis

Faculty members are uniquely positioned to understand the self-knowledge and beliefs that influence their interpretations of their teaching experiences. Although it not always articulated, faculty members themselves know what motivates them to use certain instructional methods or make specific changes, their true goals for student learning, and what kind of teaching they aspire to achieve.

Clearly any self-reported evidence is associated with some degree of bias. But the faculty voice is vital in any in an evaluation of teaching, and some might argue that some faculty context (for other's understanding) and reflection (for growth) should be a part of every source of evidence used. The difference between the busy work of collecting information and pulling together useful data to improve practice is the reflective element. If a true goal is ongoing improvement in learning, then purposeful, self-reflective processes are most likely to yield information that can facilitate improvements in teaching (Blumberg, 2014; Seldin, 1999). Some examples of self-analysis products are shared below.

Self-reflections on teaching – any self-reflection is useful, and some would say vital for ongoing development in one's teaching practice. Evidence could include after class reflections, journal excerpts or journal summaries or highlights, or any sources that provide insight into the motivations and intentions behind instructional choices or practices.

Self-reflections on teaching artifacts – reflective summaries could also be created about teaching consultations or video recordings of class sessions, sample lesson plans, annotated syllabi or annotated assignments, course alignment maps, or even a video-walk through of a Canvas course shell. If such reflections are to be used as evidence of teaching, they should ideally include the intended goals and rich description.

Teaching statements – teaching statements are often submitted as part of a teaching portfolio or larger promotion and tenure package. An articulation of one's teaching philosophy can help evaluators understand a faculty member's goals, choices and evidence in context.

Teaching Portfolios

Teaching portfolios are the ultimate form of self-reflection on teaching practice. Because of their time intensive nature, teaching portfolios are recommended for promotion and tenure decisions, although elements of the portfolio can and should be created and collected each year. Much like any portfolio for tenure and promotion purposes, a teaching portfolio ideally shows alignment between three elements:

- **Beliefs/Goals** – values, principles, what makes someone unique as a teacher
- **Actions** – what actions are taken to enact these goals, teaching responsibilities, experiences, practices, approaches to teaching, or contributions made
- **Evidence or Impact** – evidence or artifacts that demonstrate the actions taken, or evidence that actions have made a difference in some way

Faculty members may not be accustomed to collecting evidence about teaching, so putting together a teaching portfolio may seem daunting until a culture exists where evidence on teaching is generated on a regular basis.

Practices to Consider: University of California Irvine

In 2016, UC Irvine required faculty to upload at least one additional type of evidence to evaluate teaching (e.g., reflective teaching self-statement, syllabus, peer-evaluation, or measure of student achievement). This change was designed to be a first step toward conducting a more thorough evaluation of the contributions to teaching. It broadens the discussion of teaching by everyone involved in the review process, and thus has the potential to increase awareness of the innovative and effective teaching practices taking place on campus.

Ongoing teaching development

Teaching is considered an ongoing practice that is always in development. The path of one's teaching practice is never checked off as complete, just as a research or scholarship path is never checked off as complete. Teaching is an ever-evolving field. The methods are evolving, student needs and expectations are changing, and our content fields are always adapting. A focus on continual improvement should be part of the picture for teaching evaluation. Evidence about ongoing effort spent on teaching can be provided in many ways.

Professional development – faculty have many opportunities to participate in professional development around their teaching through campus sponsored (OTL and others) programming as well as programs in disciplinary fields or external organizations. The possibilities are numerous, from attending workshops and reading books, to participating in Faculty Learning Communities or multi-week intensive experiences, or leading discussions or workshops about teaching practices.

Externally-developed education rubrics – some professional development activities also result in vetted certifications or rubric completion. Quality Matters is one such source for online courses, or rubrics for syllabi construction (see the University of Virginia's syllabus rubric) or course design also exist.

Scholarship of teaching and learning (SoTL) – The ultimate form of reflection on teaching is to take a scholarly approach and conduct a study or produce a scholarly work about one's teaching practice. SoTL work could be considered the ideal form of ongoing development in teaching and often just engaging in the process makes tremendous advances in one's teaching practice.

A note about teaching awards – Teaching awards are often cited as evidence of teaching excellence, and in theory should demonstrate such, but they are not recommended as standard criteria for summative decisions about teaching. Faculty awareness of and participation in the award process is a form of professional development in itself and can be beneficial. However, literature about teaching awards cites concerns about their politicized nature, lack of alignment with direct criteria about good teaching, and the relatively small number awarded, as limiting factors for use in this process (Chism, 2007; Poproski & Greene, 2018*).*

Practice to Consider – Lund University

At Lund University in Sweden, an institution considered a research powerhouse with over 40,000 students, staff in their Division for Higher Education Development created a Pedagogical Academy to support faculty development and teaching evaluation. This is a voluntary but highly competitive program designed to foster a scholarly approach to teaching. Pedagogical Academies are career paths that result in something similar to a 'master teacher' recognition called an Excellent Teaching Practitioner (ETP). To achieve ETP designation, faculty members create a teaching portfolio that is peer-reviewed against rigorous standards. The portfolio must include evidence to support excellence in student learning, pedagogical content knowledge, professional development, scholarly approaches and pedagogical leadership. Most pieces of evidence take the form of SoTL papers or presentations. Much like a tenure and promotion package, the teaching portfolio takes years to assemble and some are not awarded designation the first time around. This is a similar but separate process to research and tenure. Those who achieve ETP designation receive a permanent salary increase.

Of course, there are many more types of evidence that could be used to demonstrate effort and outcomes in teaching. Ideally, all forms of evidence would support the others, resulting in a rich picture of someone's teaching practice.

An example of methods desired might look something like the sample chart below which was created as a result of meetings on this topic at DU in recent years. For the chart to be useful, it would be modified by individual departments and based on the content framework chosen by the university.

Table 3: Example of Recommended Evidence to Support Teaching

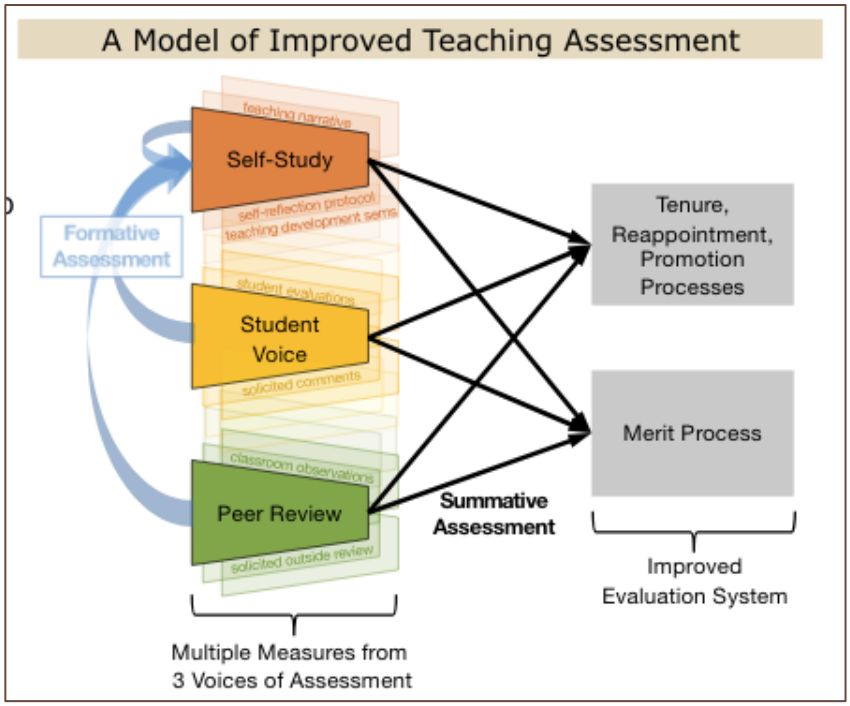
	Course Design and Instructional Strategies	Student Experience & Learning	Continuous Improvement
Minimum Annually	• Syllabi • Sample assignments • One other source of evidence	• Student Ratings of Instruction (with instructor reflection) • One other source of evidence	• Participation in one teaching-related professional development activity • Plan for ongoing development
Additional types of evidence to choose from	**Course Design** • Annotated syllabi • Course alignment map • Externally-reviewed course design rubrics **Teaching Strategies** • Description/examples of teaching methods **Assessment methods** • Sample rubric • Sample graded work w feedback provided **Learning impact** • Service learning, experiential learning, etc.	**Student experience/interaction** • Mid-quarter student feedback • Analysis of narrative/ qualitative comments from students • Course climate questionnaire • Student letters **Student Learning** • Samples of student work with or w/o instructor feedback • Evidence of learning based on 1 or 2 outcomes (not grades)	• Teaching consultation w changes/results reflection • Mid-quarter student feedback with changes/results reflection • Focused development on a specific topic • Intensive/multi-week workshop • Faculty learning community • Workshop attendance • Invitations to present • Mentoring participation • Scholarship of Teaching and Learning project - research on one's own teaching • Etc.

Feedback cycles

It is important to note that all the time and effort spent collecting evidence should provide valuable feedback to inform teaching and learning. In an ideal scenario, a large amount of rich formative data would be collected and select highlights and summaries could be passed along as summative assessment data. In that sense, the process of collecting evidence for teaching could be thought of as a layer of reflection built on to the already ongoing tasks and work of teaching. When we put a reflective lens on our everyday work, it becomes useful for informing our practice.

If the goal is that all these sources of evidence provide such feedback to improve teaching, this should be kept in mind as policies and process are created. The image below from the CU Boulder

Teaching Quality Framework shows how all three voices in teaching should support the instructor's self-analysis of teaching for continual improvement.



CU Boulder Teaching Quality Framework Initiative Summary Description (2018)

One could assume that the closer the evidence is to the raw data about our work, the more useful it is to improve teaching. Formative feedback collected for our own purposes is most useful but may not be appropriate to share with committees for confidentially or efficiency reasons. Evidence used in annual review may be more succinct and summarized, therefore losing some of the rich description, but can be useful for chairs to gain an accurate picture of an instructor's everyday teaching practice. For promotion and tenure purposes, evidence shared may be collected at a more abstract level, less useful for improvement, and more focused on showing evidence of good practice.

Useful to the instructor Useful to the institution

If we are to develop policies where evidence is collected in a way that is manageable and that provides useful feedback, it may be helpful to view these goals on a continuum, rather than serving completely separate purposes.

Sustained Leadership for Implementation

Changing a teaching evaluation structure is not an isolated endeavor but is about culture change. It involves creating a culture that values teaching at many levels and should be part of a larger effort.

Changing a university culture to value teaching results in changes in three main areas: supporting ongoing teaching development, creating fair and balanced evaluation of teaching, and rewarding teaching excellence. This takes investment from the university not only at the faculty level, but from the provost's office, from a teaching and learning center, as well as at the unit and department level.

Practice to Consider – University of Southern California

USC's new model for a comprehensive focus on teaching includes three areas: development, evaluation and reward. The initiative asks for support at all levels: the university, the schools, and individual faculty, to work on initiatives that support teaching development, evaluation and rewards. The Center for Excellence in Teaching is the central support source of the initiative. Evaluation changes are attracting the most outside attention, but the focus is meant to be on all aspects of supporting teaching excellence.

Deans and department chairs play a vital role as they implement practices at a grass-roots level and truly create the culture of teaching. It is at this level that the contextual differences of teaching exist, so it is important to allow the specifics of teaching evaluation to be determined at this local level. But this is also where the rubber meets the road, in essence, where written policies are interpreted and show how much the institution truly values teaching.

Creating a comprehensive model is perhaps the biggest struggle for DU. Many departments and individual faculty members already use alternative methods of teaching evidence. Yet without an overarching framework, the approach at DU remains piecemeal and of varying levels of quality.

The author has visited several committees at DU over the years, and most do not know the norms or policies in other units or departments. The OTL has taken a role in bringing these individuals together and sharing practices across campus, but the workload still remains in the individual departments, where faculty members are already pressed for valuable time. DU's highly decentralized structure and faculty-owned process has many benefits. There is the potential for a lot of local buy-in and contextual considerations. But it also means that each department or unit is spending valuable faculty time recreating a framework and process, often from scratch.

At a practical level, Berk suggests the following process to tackle this issue (personal communication, 2018):

1. Assemble a small, cross-campus ad hoc committee of dedicated individuals
2. Map out reasonable outcomes given the university context
3. Start with SRIs (his suggestion, although many universities choose to do this as a separate endeavor so as not to derail the committee with this overly-negative aspect)
4. Review other sources of evidence in turn (peer review, self-assessment, etc.)
5. Determine the combination of sources (at the university level is possible, at the unit/dept level if necessary)
6. Build out the other measures and get approval on each one

This process may work well for institutions that are not very decentralized. Perhaps the best way for DU to understand how to move forward is to look at other institutions that are taking a unit/department-level approach. The TEI Portfolio site contains contact information, promising practices, and links to reports and resources for 14 different institutions that are making progress in this area. The current paths of five of those institutions are described in more depth in the next section of this report.

Examples from Other Institutions

Any implementation of new initiatives that support a culture of teaching is dependent upon an institution's particular context and culture. A few institutions that are taking serious steps to implement a new teaching framework are highlighted below. All contact information and resources from these institutions, as well as nine others, are found in the Teaching Excellence Initiative (TEI) Portfolio site (http://portfolio.du.edu/teachingexcellence/).

CU-Boulder

The University of Colorado Boulder is one of four institutions involved in a five-year, $2.8 million grant from the National Science Foundation (NSF) for the development and promotion of evidence-based teaching practices. The other institutions are the University of Massachusetts Amherst and the University of Kansas. A researcher from Michigan State University (Ann Austin) is studying the process as the fourth institution (and would be a good person to talk with down the road).

At CU Boulder, this project is run through the Teaching Quality Framework Initiative (TQF). The initiative is supported by dedicated staff through its Center for STEM Learning, although the project is open to all disciplines. Half of the funding comes from the NSF grant and the other half from academic units.

The Teaching Quality Framework (TQF) is defined as consisting of two elements:

1) **A campus-wide, cross-disciplinary structure for defining teaching quality and associated tools for assessing teaching quality**

The TQF content framework is based on Boyer's *Scholarship Reconsidered: Priorities of the Professoriate* (1990), which has been adapted and operationalized by Glassick, Huber, & Maeroff in, *Scholarship Assessed: Evaluation of the Professoriate* (1997), then further adapted by Bernstein and colleagues (2010) at the University of Kansas. Although CU Boulder originally began using the KU model and rubric, after the first year they are adapting it further to better support the CU Boulder context. The latest version of their rubric was revised in December of 2018. The seven areas are:

- Goals, content and alignment
- Preparation for teaching
- Methods and teaching practices
- Presentation and student interaction
- Student (and other) outcomes
- Mentorship and advising
- Reflection, development and teaching/service scholarship

Within these areas, evidence should be gathered from all voices involved in teaching and learning: students, peers and faculty. Common examples so far are:

- From students – class interviews (like our student feedback sessions)
- From students – FCQs - Faculty Course Questionnaire (SRIs) (they are beginning to look at FCQs at a university-wide level)

- From peers – observations, letters
- From faculty – self reflections, professional development, Teaching Dossier for P&T

Their content framework was developed by a central team, but the implementation (tools and weighting) will occur at the departmental level, aiming for a common approach to assessment, while preserving disciplinary identity and specificity.

The goal is to create a culture where teaching is treated as a scholarly activity analogous to research. Also critical to this model is the hope that feedback to improve teaching should be built in throughout.

2) A process for contextualizing the structure to each disciplinary unit and enacting it across campus.

The CU Boulder model is also focused on necessary organizational change processes and elements. Their approach includes:

- Campus-wide acceptance of an overall framework
- Opt-in pilot of academic departments to contextualize the framework and develop local standards and processes
- Cross-department teams share resources
- Institutional support from above
- A centralized resource manages and facilitates the process

Their departmental approach utilizes a **Departmental Action Team (DAT)** process to implement change. This is a model whereby change happens at the departmental level through regularly occurring facilitated meetings of a departmental sub-committee. A DAT lead (chair) might get service credit or a course release, but a TQF research associate helps facilitate the meetings and process. Currently over 12 departments are at different phases of this process.

Those DATs at the beginning phases are working on discussions of value - understanding the framework, talking with other departments, and exploring their current context and practices. Those who are in “Phase III” of their work are involved in bi-weekly meetings with a TQF facilitator to go through implementation process and develop tools and rubrics. The process roughly involves:

- exploring current practices
- exploring priorities for change
- understanding the cultural and political context within the department (including any “no-fly zones”)
- Then, going after the low-hanging fruit – how to take practices that they already do and adjust or redesign so that they are evidence-based (observations is common example)
- Some current DATs are developing tools and instruments, but most are not yet at this point
- Next steps – pilot these instruments and work on new methods

Tensions, roadblocks, surprises, lessons learned

- It is difficult to find the right balance between letting faculty develop frameworks and tools themselves (can be too abstract and time consuming with little agreement), versus providing concrete tools or frameworks (can be too specific and lacks room for buy-in and contextual

concerns). For example, the rubric used in this initiative was too-specific at first, lacked full support of faculty, so now they are revising it based on feedback.
- A big roadblock is the time and resources needed to develop and pilot methods. It takes time to do assessment appropriately. They are trying to focus on current practices, making them more efficient and evidence-based.
- In addition, political roadblocks exist. Nothing in teaching evaluation is perfectly objective, there are no perfect measures, and this can be a high-stakes discussion. Additionally, some faculty believe that teaching cannot be assessed or feel they are being told how to teach.
- Within DATs, scheduling difficulties and competition with other demands, especially at certain times of the year, can present roadblocks.
- The current process is very hands-on. It would not be happening without centralized support of staff and resources, but in it's current form, it is not scalable.

University of Massachusetts Amherst

The University of Massachusetts Amherst is the lead institution for the multi-institution five-year, $2.8 million grant from the National Science Foundation (NSF) for the development and promotion of evidence-based teaching practices described above.

Gabriela Weaver, former vice provost for faculty development and director of the Institute for Teaching Excellence and Faculty Development, convened a cross-campus working group of faculty and staff and spent a year and half reviewing research, practices at other institutions and internal practices to create recommendations for change. These recommendations became the catalyst for the multi-institution NSF grant which is now supporting the process. The institutional partners allow for collaboration and idea sharing, and the funding adds credibility and helps the faculty take the process seriously.

The working group decided to use the generalizable rubric developed by the University of Kansas as the starting point. They also developed Principles for Evaluating Teaching to guide the process:

- ***Evaluation should include multiple dimensions*** of teaching: categories of activities that capture the teaching endeavor in its totality, including aspects that take place outside of the classroom.
- ***Evaluation should include multiple lenses***: multiple *sources* and *types* of data, including faculty self-report (e.g., course materials, evidence of student learning and reflections on it), peer input (e.g., class visits, review of course materials, discussions with the instructor), and student voices (e.g., course evaluations, alumni feedback).
- ***Evaluation should involve triangulation***: no measure should be used in isolation, and analysis and interpretation should include an acknowledgement of the ways in which these measures provide reinforcing and/or conflicting perspectives on an instructor's effectiveness.
- ***Both formative and summative*** uses of the data must be possible to maximize the impact on teaching effectiveness. In addition, the evidence should be useful in a longitudinal view (over courses, semesters, and years) so that improvement over time can be documented.
- There must be a ***balance between uniformity*** across departments ***and customization*** to different disciplines in order to maximize usefulness to the administration as well as faculty.

Departments have to apply to be part of the cohort to receive funding to support their work ($5000 the first year, $3000 the second year). Four departments are chosen each year to be part of a cohort. They are tasked with analyzing the proposed rubric, determining their own measures and assigning weights among the components of the rubric according to the context and culture of their department. The first cohort is still in the first year of discussions and are working through issues of meaning and trust. They have not developed anything yet, but the goal is to create their own measures and weights for the rubric by end of second year.

Their version of cultural change in institutions is based on the idea that buy-in is needed at multiple levels. The people most affected by the change should have a lot of voice, rather than taking a top-down approach. This takes much longer, and the focus is on the process. The staff from their offices of teaching development and of assessment are involved as consultants throughout the process. In some departments they have played a bigger role than in others.

According to the campus dissemination plan, after 3 years when 16 departments have piloted their own versions of the rubric, they will create a campus-wide rubric and set of tools. Although portfolios or peer review documents might be used at the department level, they are not sustainable to use at the administration level for a campus of their size. So instead, their focus is on a generalizable rubric as the eventual outcome. After they creation of a campus-wide solution, the outcomes would need to go through faculty contracts and an active union in order to implement anything beyond voluntary participation.

The working group also identified recommended changes to their SRIs (what they call STRI – Student Responses to Instruction). This is currently happening campus-wide but is a parallel process to the departmental grants. It is being spearheaded by their Associate Provost in the Office of Academic Planning and Assessment. She was also part of the working group and was considered vital to that process because she brought a very different perspective on SRI use across campus than the faculty members.

Tensions, roadblocks, surprises, lessons learned

- This is a very a long process that will likely take longer than expected.
- Sometimes there is strong defensiveness and assumptions that this work is coming from administration, even when it is meant to protect faculty and is coming from the faculty. It's a very hot-button issue.
- Departments do not have knowledge about or comfort in conducting peer review – both in a logistical sense and in a content sense (what to look for in someone's teaching). A lot of support is needed in this area.

University of Kansas

The University of Kansas is the third university involved in the 5-year NSF grant, along with CU Boulder and UMass Amherst. KU's rubric, called "Benchmarks for Teaching Effectiveness," had been in a pilot stages for a few years before the grant officially began in 2017, and has largely been adopted by the other institutions in the grant. The rubric has seven areas:

- Goals, content and alignment
- Teaching practices
- Achievement of learning outcomes
- Classroom climate and student perceptions
- Reflection and iterative growth
- Mentoring and advising
- Involvement in teaching service, scholarship or community

KU had been working on this rubric for peer review purposes, which are required for third-year, pre-tenure, and tenure review. They used the rubric to structure interviews and observations and the language was deemed useful for committees in providing a letter back to faculty. The guiding questions that underlie each of the seven criteria aspects are meant as reflective prompts that make the tool useful for ongoing development.

The levels of quality on the rubric mirror what already exists in the promotion and tenure model at KU. (CU Boulder initially did not include these levels but has since included a similar version). In an ideal world, the levels would be more formative and feel less evaluative, but they have found that most faculty members seem to like the specificity of the levels.

The NSF grant was designed to allow departments to experiment with best ways to use the rubric. KU's departmental approach is similar to that of UMass Amherst in that departments write a proposal to apply to a yearly cohort. Departments receive $5000 in the first year and $3000 in the second year to spend as they deem appropriate to help advance their work (summer salary, GTA, food for meetings). Departments have been chosen largely based on their readiness to begin and the level of buy-in within the department.

The 5 departments in the first cohort are submitting their plans by the end of December 2018, so the outcomes of this first cohort are not yet evident. The second cohort of 5 departments has just begun. A third and final cohort will be selected for 2020, and the hope is that after that time they will have strong examples with which to develop university-wide guidelines. Departments meet with each other and share ideas and resources throughout the process.

The approach at this time is to focus on departmental discussions. Departments are given the rubric and asked to adapt the language and build consensus about how to use it. Many departments are not adjusting the rubric very much, but it is the process that is important. They should have buy-in to truly own the process they create. Some groups have been given a template, which is essentially a blank version of the rubric, and asked to fill out the boxes themselves. Much of the first year is spent on this, but they do not want to rush this part.

Departments can choose how to use the rubric, but most are using it for formative purposes – for peer review triads or mentoring junior faculty. Some are using it for high-stakes decisions already. It is meant to be an iterative process where colleagues talk about the elements of the rubric, use it, and then revise as necessary.

Dea Follmer Greenhoot, who is the Director of KUs Center for Teaching Excellence, along with the center's associate director, are co-P.I.s on the grant and oversee most of the efforts. a faculty fellow and half of a staff member's time are also dedicated to support the initiative. As the center is seen as an arm of the faculty, rather than an arm of administration, the initiative has largely been seen as a faculty effort. Yet is has also been important to get administrative support throughout. This is a balance. While it would ruin the project for it to be perceived as a mandated top-down approach,

faculty also don't want to feel that they are creating these new tools and materials if they will not be used or valued by their Deans and by the administration.

KU plans to pull together lessons learned after the first cohort once they receive and are able to digest the reports. These should be posted on their website over the next year.

Tensions, roadblocks, surprises, lessons learned

- They are surprised at how little the departments have changed the rubric. There have been some adjustments – one group separated out mentoring and advising into two separate categories – but most departments largely adopt it as is. However, they feel the adaption discussions are still the most important part of the process.
- Finding the balance between a bottom-up and top-down approach. Faculty need to own the process, but the administration needs to be aware and visibly show support for the project through written memos or in meetings.
- They would recommend starting with low-stakes use first. It appears that departments can get behind it easier if they can start using it in a supportive way.
- The competitive aspect of the grant process – having to write a proposal and apply to be selected – has meant that departments are really prepared to begin. The grant approach holds them accountable and keeps the momentum and the NSF designation adds credibility.
- The next big challenge to tackle will be the balance between the summative and the formative aspects of this work. This process has seemed to tilt towards a formative focus, so they will need to be thoughtful about bringing in the summative elements.

Florida International University

At Florida International University (FIU) the process for developing a comprehensive teaching evaluation framework began with a directive from the Provost to develop new guidelines to evaluate and reward teaching. FIU is currently 2 years into their process.

They defined teaching excellence using the three areas below. This was developed by their center for teaching and learning staff based on years of knowing the literature and was generally supported and adopted by the working teams.

- **Learning-centered** refers to teaching models focused on optimizing student learning, growth, and development.
- **Evidence-based teaching** describes instructional practices informed by existing educational research, and/or both quantitative and qualitative evidence collected while teaching.
- **Culturally responsive teaching** recognizes the significance and value of students' cultural identities to their learning and strives for equitable outcomes.

In addition to using evidence that supports these three areas, evidence should come from students, self and peers.

FIU is only focusing on annual review at this time. This work will be helpful for P&T packets and future changes to P&T guidelines, but they are first trying to create a culture of gathering evidence about teaching.

After a planning team developed the parameters of the initiative to use multiple measures of teaching, a pilot of six departments (3 faculty representatives from each with modest stipends) reflected on their current practice and drafted policies and a pilot new process. Currently, this has been scaled up to the 3 largest colleges, 18 departments in all, who will be turning in their pilot plans at the end of December 2018.

A steering committee provides oversight on a more symbolic level, a core team (mostly center staff) develops resources and provides support, and faculty fellows act as facilitators for the departments going through the change process. The departments go through a course in Canvas, and the faculty fellows facilitate (who receive a stipend, overload pay or course release). Teaching center staff help develop tools and resources. The department groups are asked to reflect on their current practices, review resources, and develop pilot plans. They hope is to fully implement in Spring of 2020.

Isis Artze, who is heading up the project, feels they cannot compare faculty work to the vision yet. The language of learning centered, evidence based, and culturally responsive teaching is still too new to evaluate performance in these areas. Rather, they prefer to focus on demonstrating progress towards these aspirational goals as it is not in the culture yet for faculty, or for chairs. Their teaching center is conducting programming around these three areas for faculty and for chairs (going to chairs retreat, providing mock faculty work to evaluate).

There is also work happening on their campus to revise SRIs as a parallel effort. They have rebranded SRIs as SPOTs: Student Perceptions of Teaching surveys and are currently conducting focus groups to identify new questions to add that have value for faculty. In Florida, there are state-mandated questions they cannot change that are generally disliked. But they want to de-emphasize SRIs to focus on the collection of alternative methods.

Tensions, roadblocks, surprises, lessons learned

- Transparency – Focus on why this is happening, what problem it is solving, over and over. There is a lot of suspicion of motives in this process.
- Create a website for this work from the beginning to share the rationale, what is happening, and any relevant documents so that everyone can see. They are only now creating this resource but use their 2-page overview at every stage. They suggest the website be part of the Provosts office rather than T&L center.
- There are risks associated with teaching and learning center staff, who typically provide formative teaching support, being tied too closely with summative evaluation of teaching. This can also be the case for the faculty fellows who support the process.
- Even given the risks, there is a need for teaching and learning folks to be at the table to advocate for outcomes that reflect best practice in teaching and learning
- Those departments that had more interaction with the T&L center are moving along more quickly. Possibly because they have more familiarity with the concepts and more trust with center staff.
- The language is not in the culture yet, especially “culturally responsive.” There are many questions about what this means and what it looks like in practice.

University of Southern California

USC made national news this spring when the Provost mandated that SRIs would no longer be used in tenure and promotion decisions.

After convening multiple cross-campus tasks force committees to explore ways to improve teaching effectiveness over recent years, a set of recommendations for teaching excellence at USC was presented to the Academic Senate in May 2017 and submitted to the Provost. Faculty experts in education and curriculum from across disciplines were consulted to help develop a plan, that would include substantial resources, to enact the faculty's recommendations in the reports. The plan included:

1. Defining excellence in teaching at USC
2. Developing development opportunities (as a resource to schools only, not required)
3. Developing a peer-review process (with available resources and training) to evaluate best teaching practices (as a resource to schools only, not required)
4. Revising the student evaluation instrument to reduce bias and increase utility to faculty
5. Developing recommendations to strengthen rewards for teaching excellence.

The USC definition of excellence in teaching was created after consultation with scholars in their Rossier School of Education and the advisory boards, and discussions at an Academic Senate meetings. University guidance is that school-based definitions need to be broadly aligned with the USC definition and grounded in evidence-based best teaching practices. From there, faculty may decide to adopt or customize the USC definition as their school-based definition or decide to develop an entirely different definition.

The Provost agreed that schools may use SRI data as one of multiple sources of data about student engagement, one component of the teaching evaluation process, or SRIs can also be used for a number of other purposes. It was suggested that responsiveness to student feedback could be measured through a teaching statement. Aggregated SRI data can still provide information at the school or program level, or to assess a school's diversity and inclusion efforts. They can be used to identify faculty who are excelling at student engagement, or those who may need further support, or alert the university about problematic behavior that might require investigation.

The sense is that SRIs are important to understand the students' experience, but caution must be used when they are applied to faculty evaluation.

The student evaluation instrument was revised to:

1. Provide subscales that were more useful to faculty
2. Include items on students' experience of inclusive practices
3. Include student engagement items (intended to help students see their own role in their learning, and to give faculty information on engagement)
4. Decrease biased responses by removing vague items and giving guidance on how to give professional feedback

5. A new protocol was added to increase response rates to address faculty concerns about low response rates.

The faculty committee reports recommended a move to a peer-review system of teaching evaluation. The rationale for peer-review has many layers.

1. It promotes teaching development. Faculty will receive feedback on their strengths and where they can improve, their course design, assessments, inclusive practices, and other elements of teaching. Peer-review can make the evaluation process both meaningful and informative, providing both formative and summative evaluation.
2. The act of reviewing one's peers can also promote teaching development, allowing faculty to learn from each other's work, much like in peer-review of scholarship.
3. Learning to use peer-review tools is informative to one's own teaching, as it allows faculty to learn about best practices and to do critical self-reflection on their teaching.
4. Peer-review is the standard used in academia. We should be reviewing one another's work and providing both support and challenge.

A major part of the initiative includes a new teaching model that has involves investment from three levels: University, school, and faculty, and also endeavors to make change in three areas development, evaluation and reward.

1. The university level includes a university-wide definition of excellence in teaching, investment in training and resources for teaching development, peer-review tools and training, and consultation for alignment of reward structure with definition, development, and evaluation.
2. The school level will involve a faculty-led process to develop school-based customized plans to promote teaching excellence. This will include a school-based definition that reflects pedagogical best practices within the discipline, development opportunities, peer-review tools and processes--in addition to other discipline-endorsed evaluation methods, and a revised incentive structure that is sufficiently rewarding to motivate faculty to invest in teaching development, peer-review, and performance. Faculty will determine how each of these elements will play out.
3. The faculty level will involve individual or groups of faculty reflecting on their own teaching goals, and identifying resources and opportunities to engage in the new model.

The Provost's office is now working with school leaders to work on faculty-led school-based teaching plan development. They have asked that the evaluation part of the plan be submitted by the end of spring and plan development will continue in the next academic year. They expect many schools will opt for a phased implementation. This will give schools time to fully flesh out criteria, processes, and plans for implementation.

**Information about USC was acquired via online documents and email exchange. The author did not have the opportunity to ask about roadblocks, surprises or lessons learned to date.*

Final Thoughts

At DU, there have been countless discussions over many years to try and create a culture that values and rewards teaching excellence among all the faculty. DU has its own unique context and culture, and it is worth articulating some of the assumptions and potential guiding principles that have been shared and have emerged after years of discussion on this topic at DU.

- Learning to teach is not a one-time endeavor. Teaching is a lifelong practice that involves ongoing reflection and engagement in continuous improvement. The support, reward and evaluation of teaching should take a developmental approach.
- Instructors all have different developmental trajectories, different interests, strengths, and contexts for teaching. Therefore, plenty of choice and flexibility (a menu of options) should be provided within some required areas of development. The process should be individualized and self-directed.
- There is no one "right way" to teach. But at the same time, there are widely agreed-upon evidence-based standards of best practice.
- In order to improve teaching, reflection by the instructor is essential. The intellectual work of teaching - the deeper thinking about ones teaching practice - is harder to represent and evaluate than the more superficial aspects of teaching.
- Any summative judgment about an instructor's teaching should be based on multiple types of data points collected from multiple sources (students, instructor, chair, peer/colleagues, external experts). Qualitative or narrative sources of evidence are essential due to the complicated nature of teaching.
- The main focus of teaching support and evaluation should be on the process of continually enhancing the teaching practice.

The information summarized in this report is meant to support initiatives and efforts to change the reward structure for teaching at DU. No specific recommendations are made as no one person should decide on the framework, methods or process for such a complicated and sensitive topic. Rather, as much work has already occurred at DU as well as in institutions across the country, it is hoped that the information collected here will provide a solid foundation from which DU can move forward in this critical area. The DU faculty owe it to themselves to create a fair and balanced structure that encourages, values and rewards teaching excellence.

References

AAUP (2006). Statement on Teaching Evaluation. In *Policy Documents and Reports, 10th ed.* (pp. 200-203), Washington, D.C.: AAUP.

Allen, A., Caron, M., Forrest, A., Liddle, M., Sabourin, B. & Wright, A. (2015). *Evaluation Vignettes.* Invited panel discussion at Weighed in the Balance: Evaluating Teaching in Higher Education, University of Windsor, Windsor, ON.

Arend, B. (2017). *Teaching Evaluation: Can we measure what really matters?* Poster presented at POD Network Conference, Montreal, Canada.

Arend, B., & Pitts, V. (2015). *Reflecting on our Special Sauce: Developing Collective Teaching Aspirations*. Paper presented at Professional and Organizational Development (POD) Network Conference, San Francisco, CA.

Arreola, R. A. (2000). *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system.* Bolton, MA: Anker Publishing Co.

Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system: A guide to designing, building and operating large-scale faculty evaluation systems.* Bolton, MA: Anker Publishing Co.

Barre, E. (2018, February 22). Research on Student Ratings Continues to Evolve. We Should, Too [web log post]. Retrieved from http://cte.rice.edu/blogarchive/2018/2/20/studentratingsupdate

Benton, S. L., & Cashin, W. E. (2012). *Student Ratings of Teaching: A Summary of Research and Literature. IDEA Paper #50.* Manhattan, KS: The IDEA Center. Retrieved from https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_50.pdf

Benton, S. L., & Ryalls, K. R. (2016). *Challenging Misconceptions About Student Ratings of Instruction. IDEA Paper #58.* Manhattan, KS: The IDEA Center. Retrieved from https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_58.pdf

Berger, J. B. & Braxton, J. M. (1998). Revising Tinto's Interactionalist Theory of Student Departure Through Theory Elaboration: Examining the Role of Organizational Attributes in the Persistence Process. *Research in Higher Education, 39(2)*, 103-119.

Bergquist, W.H. & Pawlak, K. (2008). *Engaging the six cultures of the academy: Revised and expanded edition of the four cultures of the academy* (2nd ed.). San Francisco, CA: Jossey-Bass.

Bernstein, D. J., Addison, W., Altman, C., Hollister, D., Komarraju, M., Prieto, L., Rocheleau, C.A. & Shore, C. (2010). Toward a scientist-educator model of teaching psychology. In D. F.

Halpern (Ed.), Undergraduate education in psychology: A blueprint for the future of the discipline (pp. 29-45). Washington, DC: American Psychological Association.

Berk, R. A. (2018). Start spreading the news: Use multiple sources of evidence to evaluate teaching. *Journal of Faculty Development, 32*(1), 73-81.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17(1)*, 48-62.

Berk, R.A. (2009). Using the 360° multisource feedback model to evaluate teaching and professionalism. *Medical Teacher, 31(12)*, 1073-1080.

Berk, R.A. (2014). Should student outcomes be used to evaluate teaching? *Journal of Faculty Development, 28(2)*, 87-96.

Blumberg, P. (2014) *Assessing and Improving Your Teaching: Strategies and Rubrics for Faculty Growth and Student Learning (1st ed.).* San Francisco, CA: Jossey-Bass Inc.

Bolden, R., Petrov, G., & Gosling, J. (2009). Distributed leadership in higher education: Rhetoric and reality. *Educational Management Association and Leadership, 37(20)*, 257-277.

Boyer, E. L. (1990). *Scholarship Reconsidered: Priorities of the Professoriate*. Lawrenceville, NJ: Princeton University Press.

Bradforth, S. E., Miller, E. R., Dichtel, W. R., Leibovich, A. K., Feig, A. L., Martin, J. D., ... & Smith, T. L. (2015). University learning: Improve undergraduate science education. *Nature News*, *523*(7560), 282.

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.

Budd, D. A., Van der Hoeven Kraft, K. J., McConnell, D. A., & Vislova, T. (2013). Characterizing teaching in introductory geology courses: Measuring classroom practices. *Journal of Geoscience Education*, *61*(4), 461-475.

Buller, J. (2012). *Best practices in faculty evaluation: A practical guide for academic leaders*. San Francisco, CA: Jossey-Bass.

Canale, A. M., Herdklotz, C., & Wild, L. (2012). *Evaluation of Teaching Effectiveness.* Rochester, NY: The Wallace Center at RIT, Office of Faculty Career Development. Retrieved from http://www.rit.edu/academicaffairs/facultydevelopment/sites/rit.edu.academicaffairs.facultydevelopment/files/docs/Evaluation_of_Teaching_Effectiveness.pdf

Cashin, W. E. (1996). *Developing an Effective Faculty Evaluation System. IDEA Paper #33.* Manhattan, KS: The IDEA Center. Retrieved from http://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/Idea_Paper_33.pdf

Chalmers, D., Cumming, R., Elliot, S., Stoney, S., Tucker, B., Wicking, R., & Jorre de St Jorre, T. (2014). *Australian University Teaching Criteria and Standards Project*. Sydney: Office for Learning and Teaching.

Chism, N. (2007). *Peer Review of Teaching: A Sourcebook, 2nd Edition*. Bolton, MA: Anker Publishing Co.

Cohen, P. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13*(4), 321-341.

Cox, B.E., McIntosh, K.L., Reason, R.D., & Terenzini, P.T. (2011). A culture of teaching: Policy, perception, and practice in higher education. *Research in Higher Education, 52(8)*, 808-829.

Dennin, M., Schultz, Z. D., Feig, A., Finkelstein, N., Greenhoot, A. F., Hildreth, M., . . . 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).*

Ebert-May, D., Derting, T., Hodder, J., Momsen, J., Long, T., & Jardeleza, S. (2011). What We Say Is Not What We Do: Effective Evaluation of Faculty Professional Development Programs. *Bioscience, 61*(7), 550-558.

Fairweather, J. S. (2002). The ultimate faculty evaluation: Promotion and tenure decisions. *New directions for institutional research*, *2002*(114), 97-108.

Feldman, K.A., & Paulsen, M.B. (1999). Faculty motivation: The role of a supportive teaching culture. *New Directions for Teaching and Learning, 1999(78)*, 69-78.

Fink, L. D. (2008). Evaluating teaching: A new approach to an old problem. In Chadwick-Blossey, S., & Robertson, D. R. (Eds.), *To Improve the Academy: Resources for Faculty, Instructional, and Organizational Development, Volume 26* (pp. 3-21). San Francisco, CA: Jossey-Bass.

Fletcher, J.A. (2018). Peer observation of teaching: a practical tool in higher education. *Journal of Faculty Development* 32 (1), 1-14.

Flaherty, C. (2018, May 22). *Teaching Eval Shake-Up*. Retrieved from https://www.insidehighered.com/news/2018/05/22/most-institutions-say-they-value-teaching-how-they-assess-it-tells-different-story

Gibbs, G. & Coffey, M. (2004). The impact of training of university teachers on their teaching skills, their approach to teaching and the approach to learning of their students. *Active Learning in Higher Education, 5(1)*, 87-100.

Glassick, C. E., Huber, M. T., & Maeroff, G. I. (1997). *Scholarship Assessed: Evaluation of the Professoriate. Special Report.* San Francisco, CA: Jossey Bass Inc.

Graniero, P. & Hamilton, B. (2016). *Making Sense of Teaching Evaluations: The Value of Noise*. Paper presented at the 36th Annual Conference of the Society for Teaching and Learning in Higher Education, London, ON.

Gravestock, P. (2011). *Does teaching matter? The role of teaching evaluation in tenure policies at selected Canadian universities* (Unpublished doctoral dissertation). University of Toronto, Toronto, Ontario.

Gravestock, P. & Gregor-Greenleaf, E. (2008). *Student Course Evaluations: Research, Models and Trends*. Toronto: Higher Education Quality Council of Ontario.

Grayson, J. P., & Grayson, K. (2003). *Research on retention and attrition*. Montreal, Canada: The Canada Millennium Scholarship Foundation.

Hativa, N. (2013). *Student Ratings of Instruction: Recognizing Effective Teaching*. USA: Oron Publications.

Hénard, F. (2010). *Learning our lesson: Review of quality teaching in higher education*. Paris: Organization for Economic Cooperation and Development (OECD). Retrieved from: https://www.oecd.org/edu/imhe/44058352.pdf

Hénard, F. & Roseveare, D. (2012). Fostering quality teaching in higher education: Policies and practices. *An Institutional Management in Higher Education (IMHE) Guide for Higher Education Institutions*. Retrieved from http://www.oecd.org/edu/imhe/QT%20 policies%20and%20practices.pdf

Hines, S. R. (2011). 20: How mature teaching and learning centers evaluate their services. *To Improve the Academy, 30(1)*, 277-289.

Hoyt, D. P., & Pallett, W. H. (1999). *Appraising Teaching Effectiveness: Beyond Student Ratings. IDEA Paper #36.* Manhattan, KS: The IDEA Center. Retrieved from http://www.theideacenter.org/sites/default/files/Idea_Paper_36.pdf

Kolb, D.A. (1984). *Experiential Learning*. Englewood Cliffs, NJ: Prentice-Press.

Kreber, C., Brook, P., & Policy, E. (2001). Impact evaluation of educational development programmes. *International Journal for Academic Development, 6(2),* 96-108.

Kustra, E., Doci, F., Meadows, K., N., Dishke Honzel, C., Goff, L., Gabay, D., Wolf, P., Ellis, D., Grose, J., Borin, P., Hughes, S. (2014). *Teaching Culture Indicators: Enhancing Quality Teaching.* Report to the Ministry of Training, Colleges and Universities Productivity and Innovation Fund Program, University of Windsor, Ontario. Retrieved from http://www1.uwindsor.ca/ctl/system/files/Teaching_Culture_Indicators.pdf

Miller, J. E., & Seldin, P. (2014). Changing practices in faculty evaluation. *Self*, *58*, 67-6.

Mårtensson, K., & Roxå, T. (2016). Working with networks, microcultures and communities. In D. Baume & C. Popovic (Eds.), *Advancing practice in academic development* (pp. 174–187). London: Routledge.

Olsson, T., & Roxa, T. (2013). Assessing and rewarding excellent academic teachers for the benefit of an organization. *European Journal of Higher Education, 3(1),* 40-61.

Olsson, T. & Roxå, T. (2008) Evaluating rewards for excellent teaching – a cultural approach. In *Engaging Communities*, Proceedings of the 31st HERDSA Annual Conference, Rotorua, July, 261-272.

Pallett, W. (2006). Uses and abuses of student ratings. In P. Seldin & Associates (Ed.), *Evaluating faculty performance: A practical guide to assessing teaching, research, and service* (pp. 50-65). San Francisco, CA: Anker Publishing Co.

Piccinin, S. J. (2003). *Green Guide 4: Feedback key to learning.* Society for Teaching and Learning in Higher Education. London, ON.

Poproski, R. and Greene, R. (2018). *Metrics and Measures of Teaching Effectiveness*. Unpublished white paper, Georgia Institute of Technology.

Roxå, T. & Mårtensson, K. (2013). *Significant networks for educational development*. Retrieved at http://www.konferenslund.se/pp/CEQ_Roxa_Martensson.pdf

Seldin, P. (1993). *Successful use of teaching portfolios.* Bolton, MA: Anker Publishing.

Seldin, P. (1999). Current Practices–Good and Bad–Nationally. In P. Seldin & Associates (Ed.), *Changing Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions* (pp. 1-24). Bolton, MA: Anker Publishing.

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): A New Instrument to Characterize University STEM Classroom Practices. *CBE—Life Sciences Education*, *12*(4), 618-627.

Southwell, D., & Morgan, W. (2009). *Leadership and the impact of academic staff development and leadership development on student learning outcomes in higher education: A review of the literature: A report for the Australian Learning and Teaching Council (ALTC)*. Queensland University of Technology.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching. *Review of Educational Research, 83(4)*, 598-642.

Sterman, J.D. (2006). Learning from Evidence in a Complex World. *American Journal of Public Health, 96(3)*, 505-514.

Theall, M., Franklin, J., Abrami, Philip C., & Mets, Lisa A. (2001). Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction? *New Directions for Institutional Research, 2001(109)*, 45-56.

Trowler, P. (2008). *Cultures and Change in Higher Education: Theories and Practices.* Basingstoke, Hampshire: Palgrave Macmillan.

University of Denver (DU). (2015). *Policies and Procedures Relating to Faculty Appointment, Promotion, & Tenure.* Retrieved from https://www.du.edu/facsen/media/documents/apt_jan16_2015.pdf

University of Denver, College of Arts, Humanities and Social Sciences (AHSS). (2018). *AHSS Teaching Committee Report: Current Evaluation Measures and Recommendations.*

Retrieved from http://otl.du.edu/wp-content/uploads/2018/12/AHSS_Teaching_Committee_Report1.pdf

Van Note Chism, N. (2006). Teaching awards: What do they award? *The Journal of Higher Education, 77(4)*, 589-617.

Weimer, M. (2010). *Inspired college teaching: A career-long resource for professional growth*. John Wiley & Sons.

Wieman, C. (2015). A Better Way to Evaluate Undergraduate Teaching. *Change: The Magazine of Higher Learning, 47*(1), 6-15.

Willness, C., Turner, N., George, C., & McDougall, P. (2015). *Building A Shared Framework for Teaching Quality at the University of Saskatchewan: Phase One: Alignment and Comparison*. Retrieved from https://teaching.usask.ca/documents/gmctl/GMCTE%20Teaching%20Quality%20Framework%20Phase%20I.pdf

Wright, A. W., Hamilton, B., Mighty, J., Scott, J., & Muirhead, B. (2014). The Ontario Universities' Teaching Evaluation Toolkit: Feasibility Study. *Centre for Teaching and Learning Reports.* Retrieved from https://scholar.uwindsor.ca/ctlreports/4

Wright, A. W., Hamilton, B., Raffoul, J., & Marval, P. (2014). Leading the Leaders: Embedded Educational Leadership Initiatives at the University of Windsor. *Centre for Teaching and Learning Reports.* Retrieved from https://scholar.uwindsor.ca/ctlreports/3