Report of the Working Group on Student Course Feedback

OCTOBER, 2021

UNIVERSITY OF DELAWARE

The report of the Commission on Tenure-Track Faculty, which was submitted to the Provost and the Senate in Fall, 2017, identified UD's use of student course feedback in the evaluation of faculty members as problematic. The report noted the following:

Our current P&T guidelines privilege student course evaluations as a primary source of evidence of teaching quality. And departmental reviews typically place a very heavy emphasis on the numerical ratings of two common questions related to the overall quality of the instructor and the overall quality of the course. There are three problems with this. First, student course evaluations often measure student satisfaction, but do not necessarily address teaching quality. Second, there is a growing body of scholarship documenting the negative effect of bias on student evaluations. Third, there is no common course evaluation instrument in use, campus-wide. Therefore, we offer the following recommendations:

- *We recommend that the University P&T document contain cautionary language about the utility and reliability of student course evaluations as a measure of teaching quality.*
- *We recommend that the University consider establishing a common set of questions for all student course evaluations that are more focused on student learning opportunities than on student satisfaction.*

In the Fall of 2018, the Vice Provost for Faculty Affairs organized a working group, consisting of faculty and staff members from across campus to follow up on the commission's recommendations. The working group began by reviewing UD's current practices around soliciting and using student course feedback and then moved on to reviewing the scholarly literature and best practices around both student course feedback and more holistic frameworks for evaluating teaching quality. This work was suspended from March, 2020 to March, 2021, due to the COVID-19 pandemic. Having regrouped in Spring 2021, the working group resumed its work and now presents to the Provost and the Faculty Senate the following report and recommendations.

Members of the working group included:

Julianna Butler, *Associate Professor of Economics*

Carlton Cooper, *Assistant Professor of Biology*

Dave Costrini, *Director, IT Applications Development*

Erin Daix, *Library Assistant Director*

Laura Deschere, *IT Project Leader II*

Josh Enszer, *Associate Professor of Chemical & Biomolecular Engineering*

Alan Fox, *Professor of Philosophy*

Eric Greska, *Assistant Professor of Kinesiology and Applied Physiology*

Kevin R. Guidry, *Associate Director of Educational Assessment, Center for Teaching & Assessment of Learning*

Kim Isett, *Professor of Public Policy*

Matt Kinservik, *Vice Provost for Faculty Affairs*

Chrysanthi Leon, *Deputy Dean, Honors College*

Sharon Neal, *Associate Professor of Chemistry and Biochemistry*

Paul Rickards, *Director, Academic Technology Services*

Mark Serva, *Associate Professor of Accounting and Management Information Systems*

Matt Trevett-Smith, *Director, Center for Teaching and Assessment of Learning*

Shawna Vican, *UD-ADVANCE*

Chris Williams, *President of the Faculty Senate*

Writing and using student ratings of instruction: A practical summary of scholarly literature

This document summarizes three areas of scholarship. In the first section, we summarize Benton and Cashin's influential 2014 literature review with some brief notes about avenues of research that have been explored since it was published. The second section summarizes and synthesizes scholarly literature that describes how other colleges and universities have sought to improve these surveys including how the survey instruments are constructed and administered as well as how their results are responsibly interpreted and used. Finally, we summarize and compare multiple teaching quality frameworks that include SRIs as well other kinds of evidence. All three sections are written with a pragmatic perspective and a keen focus on the charge of this working group to make practical recommendations for improving how we collect and use information about teaching at the University of Delaware.

Writing a literature review of scholarly materials - studies, metastudies, case studies, critical analyses, etc. - about student ratings of instruction[1] (SRIs) is a mammoth undertaking. Even if the scope is limited to materials published even in just the past five years, separating the wheat from the chaff - including only materials focused on higher education, responsibly including or excluding the multitude of well-intentioned amateur publications, etc. - is extremely time consuming and challenging to do rigorously and consistently. Moreover, although there have been advances in this field of study during the previous decade it is unlikely that a review that only includes rigorous, well-informed work would produce results that differ significantly from the 2014 review by Benton and Cashin, the two most eminent scholars of this topic.

Rather than spend an immense amount of time and energy attempting to incrementally improve on that 2014 review, we instead focused on our efforts on locating and summarizing scholarly materials that describe how other colleges and universities have sought to improve the (a) surveys of students used to collect information about teaching quality, including how the results of those surveys are responsibly interpreted and used, and (b) larger teaching quality frameworks into which those survey results are placed alongside other evidence.

Evidence of rigor in student ratings of instruction

U.S. colleges and universities have been surveying students to gather information about teaching for about a century (Gelber, 2020; Murray, 2005). In large part because of their central role in evaluating teaching and faculty members, Benton and Cashin (2014) have written that "[this] topic has been studied more than any other in higher education" (p. 279). A thorough and formal review of this immense body of literature is beyond the scope of this working group. Instead, we summarize Benton and Cashin's 2014 chapter in Higher education: Handbook of theory and research, a relatively recent summary of this research. We also provide a brief summary of relevant research that has been published since 2014. To support the charge of this working group, these summaries focus strongly on practical implications for the use of these surveys at the University of Delaware.

Research published prior to 2014

A significant complication in any analysis of SRIs - creation and use of an instrument, analyses of the data, and ethical use of the data - is that they are inherently multidimensional. Different researchers have analyzed different instruments and drawn very different conclusions about how many underlying dimensions exist with the exact number ranging from 3 to 28.

Student ratings of instruction are surveys so this section focuses primarily on the two broad areas of trustworthiness and error in surveys, validity and reliability. Each concept is briefly introduced and defined before summarizing Benton and Cashin's review of the relevant literature. There are also significant contemporary and historical concerns about bias so a separate summary of that research is also provided.

Validity

In survey research, validity refers to the relationship between what is being measured and what is intended to be measured (Groves, et al., 2011). In other words, when we use this specific survey in this specific context do we collect information about the topic(s) in which we are interested? Do our respondents understand the survey questions and response options the way that we intended them to understand them? Critically, whether a survey is valid in a specific context depends not just on how the survey was written and how its respondents answer it but also on how others use the results.

[1] These surveys have many common names, including student evaluations of teaching. Scholars who study these instruments and their uses have mostly converged on the name "student ratings of instruction (SRI)" with the rationale that students are not qualified to "evaluate" teaching but they are more than qualified to provide their own ratings.

In the context of SRIs, the most basic way to frame validity is to determine if these ratings convincingly measure students' perceptions of teaching effectiveness. This is a challenging question to answer given the multidimensional nature of these instruments, the inherent complexity of teaching, and the lack of agreement on a definition of teaching effectiveness. Nevertheless, multiple studies have established scales to measure the following dimensions of teaching:

1. Organization—course materials and teacher well prepared; lessons linked to overall course framework
2. Clarity—simplified explanations; understandable; links to students' prior knowledge
3. Enthusiasm/expression—being enthusiastic about the subject or about teaching; making dynamic presentations; using humor
4. Rapport/interactions—encouraging students to ask questions, to discuss and share ideas, and to invite a variety of viewpoints (Benton & Cashin, p. 287)

Research into the correlation between the results of SRIs and direct measures of learning (e.g., grades) has been mixed. Many studies have found a positive correlation between these measures providing some support that these ratings are valid measures of teaching effectiveness. Others have not found that relationship. Benton and Cashin hypothesize that these differences may occur because some measures of student learning are themselves not well constructed; they place more weight on studies conducted in disciplines such as psychology and education where instructors likely have more training and experience in measurement and assessment and those studies tend to find a positive relationship between these different measures.

Researchers exploring the validity of these instruments have also compared their results to other measures and sources of information. Examining the relationship between these results and improvement in student grades in subsequent courses (i.e., value-added) has been inconclusive with weak correlations and many significant methodological challenges. There are positive relationships between the ratings of instruction given by students and ratings provided by (a) the same instructor (i.e., self-rating), (b) peers and administrators, and (c) alumni. Analysis of written comments provided by students in these surveys matches well with the quantitative results.

In summary, the literature suggests that SRIs can be a valid measure of students' perceptions of teaching effectiveness, but that developing a valid instrument for UD would require significant work, including the development and testing of possible instruments.

Reliability

In survey research, reliability refers to the degree that data are measured without error (Groves, et al., 2011). In other words, do questions on the same feature correlate higher than different features? Are the measures for a given faculty member consistent over time?

Methods to explore survey reliability are primarily quantitative in nature and generally seek to compare responses to multiple administrations of the survey. The research that has been conducted into the validity of specific instruments has indicated that SRIs can be valid approaches for measuring students' perceptions of teaching effectiveness. Some studies have focused on ratings collected from the same students in the same semester (e.g., split-halves, internal consistency coefficients for different subscales) and others have focused on comparing ratings of the same instructor at different points in time; the different methods had similar findings of high levels of reliability.

Sources of bias

In their review of literature published prior to 2014, a body of literature that encompasses many decades of research by many researchers in different contexts, Benton and Cashin conclude that "[d]espite widespread faculty concern, researchers have discovered relatively few variables that correlate with student ratings that are not also related to instructional effectiveness (i.e., student learning)" (p. 295). This is at odds with more recent research and that will be discussed in the next section.

Instructor characteristics

Research conducted prior to 2014 found relatively few instructor characteristics that may bias student ratings of instruction. Faculty who are more expressive - enthusiastic, friendly, and charismatic - may receive higher ratings than faculty who are less expressive. Many faculty may believe that these personality characteristics are unrelated to teaching effectiveness and consider this to be a bias. Faculty with higher rank tend to receive higher ratings however their rank is confounded with their greater experience..

In the research that they reviewed, instructor age and teaching experience were not typically correlated. In the studies that did find correlations, they were weak correlations that tended to be negative (i.e., older faculty received slightly lower ratings). Findings about biases related to gender were very mixed with only some studies finding (weak) correlations. The multidimensionality of SRIs is hypothesized to be a significant complication in this body of research as some studies found that students tended to rate male and female instructors differently on different dimensions

of teaching. Some studies also found a weak positive relationship between ratings that students gave to instructors who were of their same gender. Benton and Cashin found very few studies of possible biases related to race and ethnicity in the literature that was available when they conducted their literature review. More work needs to be done, therefore, to understand the effects of instructors' characteristics on SRI reponses.

Student characteristics
Students who have more interest in a course tend to provide higher ratings. This has been tied to the reason for the student enrolling in the course in the first place as students enrolled in required courses tend to provide lower ratings. Research into the relationship between these ratings and students' expected grade in a class is mixed with some studies reporting a small but positive correlation. In other words, there is some support for the idea that students provide higher ratings for courses in which they expect to receive higher grades. Interpreting these findings is very complex with multiple competing hypotheses, however, and multiple researchers have concluded that instructors can much more effectively improve their ratings by improving their teaching.

Many other student characteristics have been explored and found to have no meaningful relationship with SRIs. Student age, gender, level, grade-point average, and personality characteristics have not been found to be meaningfully correlated with these ratings.

Course characteristics
The characteristics of some courses impact SRIs. Some studies have found that smaller classes tend to receive slightly higher ratings. Interpreting these results is not straightforward, however, as plausible arguments are made that some aspects of effective teaching are more difficult or impossible to manifest in very large classes. Although the level of the student - first-year, sophomore, etc. - does not correlate with these ratings the level of the class - 100-level, 200-level, etc. - does have a weak correlation with upper-level courses being rated slightly higher. Courses in the humanities and arts tend to receive higher ratings than courses in the social sciences which in turn tend to be rated more highly than math and science courses. Researchers have suggested many possible explanations and some of those explanations are complicated by findings that courses that are more difficult and require more (productive and meaningful) work tend to receive higher ratings.

Surprisingly, the precise timing of the collection of these ratings does not seem to affect them. Although Benton and Cashin only cite several older studies (1979 is the most recent one), the studies consistently found that administering these surveys during the second half of the semester resulted in similar responses regardless of the exact timing even if they were conducted as late as the first week of the following semester.

Research published after 2014

Of course, research about SRIs continued after Benton and Cashin's literature review was published in 2014. The primary thrust of much of this research has been understanding the role of biases - students', faculty members', and administrators' - in collecting and using the information provided by these surveys. More specifically, there has been significant concern that these instruments that play a crucial role in the hiring, retention, promotion, and pay of faculty members are biased against people who have historically experienced bias and prejudice in their professional and personal lives e.g., women, Black people, queer people.

Much of the research conducted after 2014, particularly the research that has drawn significant attention from the media, argues that SRIs are biased and unreliable. SRIs are impacted by student, instructor and course characteristics (e.g., Kreitzer & Sweet-Cushman, 2021; Gourley & Madonia, 2021; Mitchell & Martin, 2018; Uttl & Smibert, 2017; Boring, Ottoboni & Stark, 2016). These studies found that student evaluations are impacted by instructor characteristics such as gender, race, age etc. (e.g,. Kreitzer & Sweet-Cushman, 2021; Mitchell & Martin, 2018; Boring, Ottoboni & Stark, 2016), student characteristics such as gender, grade etc. (e.g., Boring, Ottoboni & Stark, 2016), and discipline (e.g., Uttl & Smibert, 2017). For example, Mitchell and Martin (2018) found that student ratings are higher for male instructors than women even if they are teaching identical courses. Moreover, Uttl & Smibert (2017) suggest that student ratings tend to be especially "hazardous" to instructors in quantitative fields.

The findings of bias in many of the studies have motivated many institutions to take concrete steps to change their instruments and evaluation processes. Several of these are summarized in the following section that focuses on teaching quality frameworks developed by institutions. These concerns have also motivated several scholarly organizations to make specific recommendations to not rely solely on surveys of students to evaluate teaching quality. For example, the American Sociological Association issued a statement saying "student feedback should not be used alone as a measure of teaching quality" (p. 2) with a specific concern that this feedback is biased and disadvantages faculty from marginalized groups. These concerns were

also among the reasons that the National Academy of Sciences convened a workshop in late 2019 to "frame the national conversation around effective ways to evaluate undergraduate teaching, particularly in STEM areas, that can help drive adoption of evidence-based instructional approaches" (p. 2).

However, we must also recognize that much of this research has significant limitations. Most were conducted at only one institution, often with small numbers of students. Most were conducted with only one instrument (and not the same instrument as other studies). And most were not situated within the very large body of research that preceded them. Indeed, a rigorous metastudy of this body of research adds nuance to these broad findings of bias, noting that these instruments and their related processes and data do appear to be biased against women but that effect is conditional on other factors (Kreitzer & Sweet-Cushman, 2021).

This research leaves us in a very uncomfortable place. Methodologically, much of this work is so limited that it is difficult to draw definite conclusions based on the available empirical data. However, it would be ridiculous to believe that these instruments and their related processes, including how the results are used, are immune to the biases and prejudices that permeate our cultures and societies, including those at the University of Delaware. Even if the available evidence in this specific area of practice is not at the high levels of rigor we would like, the importance and sensitivity of these data require that we take these concerns seriously and minimize bias in our SRIs, including the questions that are asked, the ways in which the surveys are administered, and how the resulting data are interpreted and used to make decisions.

Responsible use of the results of student ratings of instruction

Benton and Young (2018) recommend several general principles that should be incorporated into the writing of SRI instruments. The following sections provide further definition and concrete recommendations about how to responsibly write, administer, and interpret these surveys and their findings to make them trustworthy and fair.

Writing the instrument

A fundamental requirement of any survey is that it produces information that is accurate: the instrument must be reliable and valid, elements that are well-known to survey researchers and psychometricians with many standard approaches to measure and improve. Benton and Young (2018) caution that this requires a balance between a too-simple approach that masks or undervalues important differences (e.g., between disciplines, instructor and student demographics) and a too-complex approach that makes using the results difficult and impractical.

Fundamental recommendations for writing survey questions and instruments are, of course, applicable for these surveys (e.g., Groves et al., 2011). Most importantly, respondents must be able to answer the questions they are asked. Questions must be written in a language that all students can readily understand. They must ask about phenomena for which students can provide accurate answers. If they are asked to make judgments, they must be judgments that they are qualified to make. If they are asked to report their experiences, they must be experiences that are memorable. Hativa (2013, p. 41) synthesises the literature to formulate similar recommendations specific to SRIs:

1. Students are able to answer questions "accurately based on their own experiences in the course"
2. Questions "reflect campus values of teaching and teaching effectiveness"
3. Questions are "most important for effective teaching"

In her 2013 book-length summary of literature and original research, Hativa also makes several recommendations regarding specific questions or kinds of questions:

1. Overall opinion of the course, instructor, and their own learning
2. Teacher's use of the most important "general teaching-behavior items"
3. Specific teaching behaviors (for formative evaluation)
4. Open-ended questions
5. Demographic questions (discussed below)

Interpreting these data fairly also requires some information that is not directly about students' experiences and judgments of teaching with implications for additional questions. Benton and Cashin's 2014 literature review lists several variables that are known to have a relationship with student's rating of instruction. More recent research has also uncovered or opened for discussion and investigation additional variables. These variables must be included in data collection so they can be controlled or accounted for in analysis or reporting. They do not have to be included as questions for respondents; in fact, some data elements are not suitable for survey questions but should be incorporated from other data sources e.g., institutional course data. These were all discussed in the previous section and are listed below in Table 1.

Dimension	Factor	Potential source(s)
Instructor	Rank Expressiveness	Institutional data ???
	Demographics (age, race/ ethnicity, gender)	Institutional data
Student	Motivation	Survey(s)
	Expected grade	Survey(s)
	Demographics (age, race/ ethnicity, gender)	Institutional data (if surveys are not anonymous) Survey(s) (if surveys are anonymous)
Course	Class size	Institutional data

Table 1: Factors known or suspected to be related to bias in student ratings of instruction

Hativa (2013) also recommends several kinds of questions **not** be included in these surveys. In general, these are topics that students are not qualified to judge.

1. Judgments about the disciplinary appropriateness or importance of the course
2. Suitability of readings
3. Instructor's knowledge of content
4. Instructor's competence in the discipline(s)
5. Judgments and opinions not relevant to the teaching of the course
6. Judgments and opinions that contradict campus values and purpose
7. Any questions unlikely to exhibit variability in responses i.e., questions that most or all students will answer the same way
8. Comparisons with other teachers or courses

Administering the instrument

Although some faculty conduct these surveys using physical, paper instruments, the majority are conducted online. Benton and Cashin (2014) and Hativa (2013) agree that the broad body of research has found no meaningful difference in the responses of students to online or paper surveys. The primary difference is a markedly lower response rate to online surveys, a difference that cannot always be attributed to the change in administration medium as this change has frequently been accompanied by other significant changes. In particular, many paper surveys were administered during class (usually with the instructor not present to reduce the chance that his or her presence would influence student responses) whereas students have often been asked to complete online surveys during other times. Online surveys administered during class - with students told beforehand to bring their laptops, tablets, or smartphones - have similar response rates as paper surveys administered in class.

As briefly mentioned in the section summarizing the large body of literature about student ratings of instruction, the timing of survey administration does not appear to change the responses. The studies cited by Benton and Cashin (2014) are all several decades old and focused exclusively on administering the survey some time in the second half of the semester (or very beginning of the following semester). However, Hativa (2013) recommends that surveys be administered prior to final exams to reduce the chance that student stress and worry about their exam performance will influence their survey responses.

Interpreting the results

One of the most common pitfalls of reports and analyses of the results of SRIs is the inappropriate use of statistical tests and results (e.g., Boysen, 2015; Boysen, 2016). The literature warns of two specific pitfalls. First, a failure to control for factors that we know affect these survey results can result in unfair comparisons. For example, we know that class size can affect these survey results so comparing unadjusted average or median scores between large and small classes could be unfair to the faculty who taught those classes. Second, an inappropriate level of precision can be attributed to these data. It is clear from decades of research and centuries of experience that teaching is extremely complex with multiple dimensions and cannot be responsibly summarized in scientific-appearing measurements with multiple decimal points of accuracy.

Benton and Young (2018) make three recommendations to guide appropriate interpretation of quantitative results of these surveys:

1. Do not focus solely on average scores. Examine other descriptive statistics e.g., frequencies, standard deviations.
2. Incorporate contextual factors (i.e., those listed in Table _: Factors known or suspected to be related to bias in student ratings of instruction). There are factors beyond the control of the faculty member and they should not be penalized for them.
3. Look for patterns for each instructor across the courses they teach. Remembering that different

courses have different contextual factors, it is helpful to determine what feedback students are consistently providing about a faculty member.

Across multiple studies and syntheses of relevant research, Boysen has made several specific recommendations to minimize misinterpretations of quantitative results of SRIs:

1. Always analyze these quantitative data systematically. This requires an initial investment of time and training and an ongoing commitment to apply this system instead of relying on quicker, simpler heuristics but results in more sound, fair judgements (Boysen, 2017).
2. Only make judgments based on multiple sets of data analyzed together. This tends to reduce measurement error and discourage overinterpretation of random fluctuations (Boysen, 2016).
3. Require statistical information whenever these data are presented. This may discourage inappropriate judgements that can result from impartial information. It may also result in increased statistical knowledge among those who regularly have to use these data (Boysen, 2015).
4. Provide professional development to those who must interpret these data. Although some faculty are familiar with quantitative data and statistical tests, many are not. Further, even those who have significant training and experience often fall prey to common pitfalls (Boysen, 2015).

Scholars also recommend that regular, formal interpretation and analysis of these data should not be limited to evaluators (e.g., chairs, deans) but should also be conducted by the faculty member who taught the class(es). As adopted by several teaching quality frameworks discussed in the next section, self-reflection can be a powerful tool for faculty in productively understanding and using these data. Boysen (2016) suggests that including some self-reflection and explanation of unique context can help evaluators (e.g., chairs) understand these data. Finally, Penny & Coe's 2004 meta analysis indicated that formally discussing these survey results with a peer and setting goals can be effective.

It is also critical to look beyond the immediate interpretation and use of these survey results. Even the most accurate survey results can present inaccurate and misleading characterizations of faculty teaching (Esarey & Valdes, 2020). Moreover, those who use these data to evaluate faculty - chairs, deans, committee members, etc. - must be able to not only interpret these data accurately and fairly but they must also be able to place them into a larger context and help faculty understand their teaching and, where appropriate, improve. This will be discussed further in the next section.

Teaching quality frameworks

As we aim to improve the use of student ratings of instruction relative to teaching quality at UD, it is of benefit to review current undertakings of peer institution teaching quality frameworks (TQF) [Appendix A]. These frameworks tend to do two things. First, they describe the elements of teaching quality that the institution believes are critical. These philosophical elements of teaching quality inform the second purpose of these frameworks, the operationalization of these elements.

TQFs provide two sets of definitions or lists of characteristics. First, they provide a philosophical orientation and claims about high quality teaching, particularly the kinds of evidence that can be collected and summarized to establish that high quality teaching has taken place. Our review of available TQFs at research universities identified several common claims about high quality teaching:

- The ultimate aim of teaching is student learning and success
- Teaching is a skill that can be observed, characterized, and improved
- Teaching is a complex set of skills that can only be characterized by multiple forms of evidence
- Students can provide invaluable information about teaching but there are critical elements that they are not qualified to evaluate
- Peer review is valued as necessary input relative to teaching quality and growth, not focused on subject content
- Self-reflection demonstrates accountability and helps determine modifications to teaching/mentoring practices
- Understanding and improving teaching is not solely the responsibility of individual faculty members; we have critical responsibilities to one another including peer review and high quality support for those who evaluate faculty teaching
- A highly functional system of characterizing high quality teaching must include extensive support for understanding and improving it

Second, TQFs also specifically define teaching quality and its characteristics. These elements are operationalized in rubrics, the student surveys, and many other operations. Common definitions or characteristics include:

Characteristics	Questions	Possible Evidences
Goals, content, and alignment	What are students expected to learn? Are course goals appropriate? Is content aligned with the curriculum?Does content represent diverse perspectives?	• Goals for student learning and skill-development are established and are at appropriate level for the course and the students expected to take it. • These learning goals are well-articulated to students. • The course goals are clearly connected to program or curricular goals. • Content is challenging and innovative or related to current issues and developments in the field. • Topics are of appropriate range and depth, with integration across topics. • The instructor includes high quality materials that are well-aligned with the learning and skill-development goals for the course. • Assessments are varied and well-aligned with learning goals.
Methods and teaching practices	How is in-class and out-of -class time used? What assignments, assessments, and learning activities are implemented to help students learn? Are students engaged in the learning process?	• Activities are well planned, integrated, and reflect commitment to providing meaningful assignments and assessments. • Use of effective, high-impact and/or innovative methods to improve students' understanding and support their learning. • In- and out-of-class activities provide opportunities for practice and feedback on important skills and concepts. • Efforts are demonstrated to support learning in all students. • Teaching practices result in high levels of student engagement.
Achievement of learning outcomes	What impact do courses have on learners? What is the evidence of student learning? Are there efforts to make achievement equitable?	• Standards for evaluating student understanding are connected to program or curriculum expectations. • Standards are well-communicated to students. • Multiple forms of effective assessment, aligned with course objectives, are used. • Level of learning supports success in other contexts (e.g., subsequent courses) and/or is increasing over successive offerings.
Class culture and student perceptions	What sort of climate for learning does the instructor create? What are students' views of their learning experience and how has this informed teaching?	• Evidence that class climate is respectful, cooperative, inclusive, and civil. • Evidence that class climate encourages motivation and engagement. • Instructor is accessible and interacts well with students. • Students perceive that they are learning important skills or knowledge.
Mentoring and advising	How effectively has the instructor worked individually with UG or grad students?	• Evidence of quality and time commitment to advising and mentoring (defined as appropriate for the discipline).
Reflection and iterative growth	How has the instructor's teaching changed over time? How has this been informed by student learning evidence?	• Evidence that instructor is responsive to, and reflective on, student feedback in the short- and long term. • Regularly makes adjustments to teaching/mentoring practice based on reflections on student learning, within or across semesters. • Re-examines student performance following adjustments. • Improved student achievement of learning goals based on modifications to teaching/mentoring practices.
Involvement in teaching service, scholarship, or community	How has the instructor contributed to the broader teaching community, both on and off campus?	• Engagement with peers on teaching (e.g., teaching-related presentations or workshops). • External presentations. • Publications to share practices or results of teaching or educational activities. • Scholarly publications or grant applications related to teaching.

Similarly, we identified several common elements of the operationalization of these frameworks:

- Student surveys
  - Mid-semester survey
  - End-of-semester survey
- Peer review
- Self-reflection
- (In a minority of cases) Additional evidence provided by faculty e.g., direct evidence of student learning, participation in professional development, scholarship and service related to teaching

Currently at UD, there is neither a clear, shared understanding of the philosophical elements of teaching quality nor of how those are operationalized. To the extent that these do exist, they exist primarily at the college and departmental level. This does not allow for a clear definition of teaching quality or accountability across the university, creating different standards in promotion and merit processes and many challenges for faculty who are required to collect and use this information to inform their own teaching, faculty and administrators who use this information to evaluate faculty (e.g., annual appraisal, contract renewal, promotion, tenure), and those who are charged to support all of these processes.

Practicality of adapting framework at UD

The TQFs reviewed from peer institutions are fairly robust, demonstrating commonalities to UD processes as well as differences that must be considered. In undertaking such an adaptation, it must be noted that the peer institutions selected did so in a rolling process, choosing certain departments to make the adaptation each semester. This allowed for dynamic changes to occur between semesters to allow for successful implementation in these large, complex organizations. A common theme across the TQFs is a continuous cyclic assessment, always building from the previous cycle. As well, most peer institutions have developed rubrics that allow for more consistency and transparency in the evaluation of teaching, defining different domains and proficiency. Moreover, these philosophical frameworks and definitions are frequently used to inform professional development at the institution e.g., the teaching center explicitly positions many of its workshops and services to develop specific skills or practices in the framework.

Many of the TQFs identify the use of SRIs to obtain student feedback on the class experience, as we do at UD. In most cases, the SRI instrument was explicitly constructed to address elements of the framework as a reflection of how the institution had defined teaching quality. For example, the available descriptions of the frameworks at the Universities of Oregon (UO) and Southern California (USC) explicitly provided the questions contained in the SRIs. Comparable to UD, USC utilized a four-point scale to define the student learning experience based on seventeen questions, splitting the questions into five specific domains: course design, instructional practices, inclusion practices, assessment practices, and course impact. USC also required the students to provide a written justification for any "disagree" or "strongly disagree" responses to the questions. UO utilizes SRIs at two points during the semester, midway and end, with the same survey presented during both assessments. The midway SRI serves to provide feedback to the instructor for adjustments and clarification of course expectations, with the end of semester SRI possibly demonstrating student perceptions of course alterations. Their SRI includes thirteen three-point scale questions, two "select the best answer" questions related to course improvement, and a student accountability question related to hours per week invested in the course. The three-point scale utilizes positive selection choices that include: "beneficial to my learning", "neutral", and "needs improvement to help my learning". Though the other peer-institutions did not provide specifics of their SRIs, they all utilize them as a necessary component, or "voice", required for a holistic assessment of teaching.

Another necessary component that was identified in the selected TQFs was self-reflection. The current implementation of this may vary at UD as it is a necessary part of the merit and P&T processes to a certain degree. The selected TQFs identify this as a necessary component ("voice") to demonstrate effective teaching and many require or recommend a teaching portfolio that should be updated each time a course is taught to demonstrate reflection or improvement. Some of the TQFs expand the concept of self-reflection further by considering professional development sought, as well as contributions to the teaching community; other TQFs separate these components into a teaching service category.

The other necessary component that was identified is peer review. Currently at UD, this process is commonly performed at the departmental-level with each department developing its own processes and standards. The concept addressed by the selected TQFs identifies a common rubric that is utilized across the university with its objectives focused on the evaluation of teaching, not content. USC allows for the formative portion of a peer review to remain confidential, but

the summative feedback is utilized during review. The other peer institutions allow for peer review to be used as supporting evidence for teaching quality, with the University of Saskatchewan also allowing for course curriculum peer review.

From this brief summary of peer TQFs, some of the specifics are currently being assessed at UD, but only at the unit level. In moving towards a campus-wide assessment, the following objectives can be accomplished:

A. To ensure sustainability and compliance, the implementation of a revised evaluation process can be rolled out over a set amount of time.

B. The ability to assess all faculty across different disciplines relative to teaching quality can be accomplished by adopting a university standard rubric. In our current position, there is no need to "reinvent the wheel", as the rubrics reviewed provide more than enough information for direct adoption or adaptation.

C. SRIs should continue to be utilized to inform the instructor of the student perception of their teaching, but self-reflection should become a required component to illuminate areas of strength, areas of improvement, and alignment with the university's definition of teaching quality.

D. The use of peer-review on teaching by experienced reviewers should be implemented. Currently, UD has a Faculty Peer Observation Program (FPOP) that accomplishes this. The framework of this program, which places faculty into a group of three with members from different departments, allows for a specific focus on teaching assessment and reflection. This framework is time-intensive but it or something very similar can be adapted and incrementally grown to fit the needs of the campus.

The Case for Change at UD

The impetus for change in how we evaluate teaching quality is spurred by two core values at UD: (1) our commitment to non-discrimination; and (2) the high value we place on teaching quality. As the foregoing survey of scholarship on SRIs (student ratings of instruction) shows, student course feedback can be influenced by biases on the lines of gender, race, age, and other categories and their application can exacerbate inequities among faculty members. Guided by research and examples of best practices, we can take steps now to address this concern. Additionally, current events provide additional urgency to maintaining our longstanding commitment to teaching quality. The necessary shift to remote instruction during the COVID-19 pandemic brought the university classroom into the homes of many students, resulting in a renewed focus on teaching quality on the part of students and their parents. This focus on teaching quality is not likely to fade anytime soon; rather, it will likely increase as our region approaches the so-called "demographic cliff" by the middle of this decade, when the number of high school seniors will fall precipitously. That will lead to greater competition for undergraduate students, for whom teaching quality will be a key criterion for college selection.

Student course feedback is an essential factor in evaluating teaching quality. This has long been recognized at UD and is the reason why student feedback is collected at the end of nearly every course and included in all faculty review processes. ***However, the way that feedback is collected, disseminated, and used is problematic and must be changed.*** The necessary changes range from the design of the SRI instrument; to how faculty incorporate the student feedback into their teaching; to the ways our promotion & tenure and merit review policies rely on student feedback. In light of the foregoing summary of the scholarly literature on SRIs, a review of our practices leads to the unavoidable conclusion that we have developed some problematic institutional practices. Knowing this, the time has come to change them.

The Importance of a Teaching Quality Framework

Currently, there is no university-wide consensus on what constitutes teaching quality or how to assess it. The evaluation of teaching quality at UD is guided by department-specific policies on promotion and tenure and merit pay review. There is also no requirement to collect student feedback and no single, campus-wide SRI. Instead, there is a wide variety of SRI's in use (many electronic, but some still on paper), and at least one department offers the option of using multiple different SRIs in a given semester. Disciplinary variation is inevitable and desirable. Teaching first-year writing is radically different from teaching a field-based upper-division course. But in the absence of a guiding Teaching Quality Framework, the variation in SRIs and the uses to which they are applied become problematic.

The chief problem with operating without a teaching quality framework is that the SRI becomes the predominant element in evaluating teaching quality. Student course feedback—especially the numerical ratings—is not meant to stand alone; rather, it should be one of many sources of evidence used in the evaluation of teaching quality. Numbers are powerful, and even when SRI's are presented as part of a broader set of evidential materials, the numerical ratings given by students can exert an oversized influence on faculty review processes. And as the subsequent analysis shows, departmental P&T and Merit Metric policies tend to privilege SRIs over other evidentiary materials--sometimes subtly, sometimes quite explicitly.

A related problem that results from the absence of a teaching quality framework is that classroom performance can become the be-all and the end-all of evaluating teaching quality. Every faculty member knows that the moment of instruction is just that: one moment in a longer process of course development, research, and assessment. And, indeed, most departmental policies indicate that evidential material related to course planning and assessment is relevant. But if there is no broader consensus or guidance on the nature or value of those elements of teaching, their role in faculty evaluations become secondary to the SRIs. Student ratings become an indirect (and inappropriate) measure of the quality of all the pedagogical labor that occurs outside of the classroom performance.

The most worrisome problem of evaluating teaching without a teaching quality framework is that the focus on SRIs can specifically disadvantage faculty members along the lines of race, gender, age, and other characteristics. A campus-wide teaching quality framework will both guide the evaluation of teaching in a more uniform way and inform the nature of the questions asked of students in the SRI instrument. As the foregoing literature review shows, the best practices focus SRI questions on student learning opportunities and not on general questions of satisfaction with the course or instructor, which tend to introduce bias. At UD, our current SRI instruments prioritize the latter over the former.

In the absence of a teaching quality framework, how do UD's policies handle student course feedback? How do SRIs figure into faculty review processes? Is there a de facto campus consensus on how to use this information? To answer these questions, we reviewed all departmental P&T and Merit Metric policy documents. The results of this review follow and they provide a strong argument to take corrective action now.

Department Promotion & Tenure Policies

In reviewing departmental P&T policies, we utilized multiple reviewers and consensus coding to answer the following questions:

1. Is student course feedback specifically mentioned in the P&T policy?
2. Is the inclusion of student course feedback required as part of the P&T dossier?
3. Does the policy require student course feedback to be presented in the dossier in a specific manner (e.g., a comparison of numerical scores to a departmental average)?

The answers to these questions demonstrate the centrality of SRIs to the promotion and tenure process—especially the numerical ratings. We found that 100% of department P&T policies specifically mention student course feedback. Of those, around 75% explicitly require the inclusion of that feedback as evidence of teaching quality. Finally, just over 40% of all departments specify how student feedback should be presented in the dossier.

Although not every departmental P&T policy explicitly requires material from SRIs, a careful reading of these policies suggests that, functionally, every department expects to see this material--especially numerical ratings--in the P&T dossier. For instance, SRI content might be included in a list of evidential material that is presented as optional, but later in the policy, there will be a description of how that material must be presented. One department policy suggests that the SRI material is optional, but later says that the department chair will collect the evidential material and provide it to the candidate. There are many such examples throughout the university, and in cases like these, the inclusion of student course feedback seems more compulsory than optional, and may seem especially so to new or junior colleagues trying to learn their departmental norms.

Department Merit Metric Policies

In reviewing departmental Merit Metric policies, we utilized multiple reviewers and consensus coding to answer the following questions:

1. Is student course feedback specifically mentioned in the Merit Metric policy?
2. Does the Merit Metric policy use student course ratings in a formulaic way to produce a numerical rating?

Student course feedback also figures in the annual review of faculty to determine merit pay. We found that 67% of departments specifically mention student course feedback in their Merit Metric policies. And around 25% of all departments have a provision that converts SRIs into a numerical rating on the annual appraisal's 1-9 scale for teaching. Those provisions are especially stark examples of how the SRIs (again, almost exclusively the numerical ratings) become the only measure of teaching quality. In other words, what is meant to be one piece of evidence in the evaluation of teaching quality becomes not just the predominant, but the sole determinant of the 1-9 rating that will determine merit pay.

Student input in the evaluation of teaching quality is indispensable. Nothing in this analysis is meant to suggest that student input has no place in the evaluation process. But this review shows us that we developed policies and practices that give student course feedback an outsized and sometimes unfiltered role in our faculty evaluation processes. The task before us is to correct this situation so that we are collecting more valid student feedback and using it as only a *part* of our evaluation process. To that end, we recommend the following:

1. Develop a common SRI instrument to collect student feedback that is focused more on learning opportunities and established best practices and less on student satisfaction with the course and instructor.
2. Develop a teaching quality framework that adequately and consistently accounts for all the work that goes into excellent teaching.
3. Encourage departments to review their P&T and Merit Metric policies in light of the issues identified in this report.
4. Explore the feasibility of adopting a third party SRI tool with powerful reporting features.

Appendix A: Teaching Quality Frameworks

The following teaching quality frameworks were included in our review:

- **Boise State University**
  Implementing a Framework for Assessing Teaching Effectiveness (IFATE)
- **Colorado State University**
  Teaching Effectiveness Framework
- **University of Colorado Boulder**
  Teaching Quality Framework (TQF) initiative
- **University of Kansas**
  Benchmarks for Teaching Effectiveness
- **University of Massachusetts Amherst**
  Transforming the Evaluation of Teaching: A Study of Institutional Change (TEval)
- **University of Oregon**
  Continuous Improvement and Evaluation of Teaching System
- **University of Saskatchewan**
  Teaching Quality Framework
- **University of Southern California**
  USC Excellence in Teaching Initiative

Appendix B: Literature Review Methodology

This appendix briefly describes how the bodies of literature summarized in this document were identified and compiled. Each of the three sections - evidence of rigor in student ratings of instruction, responsible use of the results of student ratings of instruction, and teaching quality frameworks - required a different approach with different results and considerations.

In the first section, the general literature review of rigor in student ratings of instruction, we dealt with the largest body of literature. Staff in the university's Center for Teaching & Assessment of Learning and the Library, Museums and Press have collaborated in the past to attempt to rigorously collect and evaluate this literature. These efforts were inconclusive but were invaluable in informing this work. More specifically, these efforts illustrated the size of this body of literature with even carefully constructed searches of relevant databases returning thousands of results. Each of these results would require further analysis to determine if it was suitable to be included in a literature review. Reviewing these documents would require many months of work by multiple trained, skilled researchers, resources that were not available for this working group.

Luckily, this body of literature has been systematically reviewed, synthesized, and summarized numerous times by highly qualified scholars. Focusing on a small number of high quality, recent summaries was a manageable task. Moreover, it seems highly unlikely that if we were to conduct our own extensive, rigorous literature review that we would reach significantly different conclusions if we took the time to write our own original summary. Hence we were intellectually comfortable with the decision that was required by our limited resources.

There is one significant limitation to this approach that relied heavily on two key documents, one published in 2013 and the other in 2014: Important research was published after these documents were published. Our general sense of this contemporary body of research is that the broad themes of the previous decades of research have been well supported. One particular area, however, has been the subject of very fierce and public debate: the degree to which responses to these surveys may exhibit a bias against faculty members whose identities commonly subject them to bias in other areas of life e.g., women, racial and ethnic minorities. This is a critical area of research that we believed must be included in our review and we did so by focusing primarily on the studies that received significant press. More specifically, we searched

the archives of *The Chronicle of Higher Education* and *Inside Higher Ed* to identify not only the studies that have been widely discussed but also how they have been discussed within the academy.

In writing the second section, the section that focused on the responsible use of the results of student ratings of instruction, we faced similar challenges as writing the first section. This body of literature is not as large but it is more complex as some of the relevant, useful material is not published in scholarly journals and proceedings. In particular, significant documents were written by other similar working groups and task forces at other colleges and universities.

Since much of this material is published in two different kinds of repositories, we employed two different approaches to locate it. Our initial attempts at locating the peer-reviewed materials used searches of the relevant databases. However, we quickly abandoned that approach as the number of results ballooned into a size that was not manageable by this working group. Instead, we began with a few key documents - Benton and Young's 2018 IDEA paper *Best Practices in the Evaluation of Teaching*, Boysen's 2015 *Uses and Misuses of Student Evaluations of Teaching: The Interpretation of Differences in Teaching Evaluation Means Irrespective of Statistical Information*, and Hativa's 2013 *Student Ratings of Instruction: A Practical Approach to Designing, Operating, and Reporting*. In particular, we used Google Scholar to conduct a reverse citation search for each of those documents to identify materials that cited one or more of those key documents. The results that appeared to be most relevant and appropriately rigorous were included in our reviewed materials.

Our second approach at identifying relevant documents about the responsible use of the results of student ratings of instruction focused on locating documents written by and for other colleges and universities engaged in work similar to that of our working group. This search was incomplete due to resource limitations and the time available for our work. It primarily focused on searches of the archives of the POD mailing list, a very active mailing list used by educational developers around the world (but with a large concentration and focus on the United States and Canada). This search identified relevant documents shared or described on the mailing list with some materials also identified as they were referenced in those original materials. We do not directly cite these documents, in large part because they relied on many of the same materials that we had already reviewed, but they informed and confirmed many of our conclusions.

The third section of this document, the summary and synthesis of teaching quality frameworks, followed a similar approach to that employed to find institutionally-authored materials for the second section. The search of the POD mailing list was significantly supplemented by a broad Google search for "teaching quality framework" as well as our own professional knowledge of institutions that have been working in this area e.g., we included the framework from the University of Saskatchewan as we knew of Nancy Turner's work from a paper presented at the AERA conference a few years ago. We also limited our focus to frameworks in use at or being developed by research universities.

References

American Sociological Association. (Sept. 2019). Statement on Student Evaluations of Teaching. https://www.asanet.org/sites/default/files/asa_statement_on_student_evaluations_of_teaching_feb132020.pdf

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. *In Higher education: Handbook of theory and research* (pp. 279-326). Springer, Dordrecht.

Benton, S. L., & Young, S. (2018). *Best Practices in the Evaluation of Teaching*. IDEA Paper# 69. IDEA Center, Inc.

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*.

Boysen, G. A. (2015). Uses and misuses of student evaluations of teaching: The interpretation of differences in teaching evaluation means irrespective of statistical information. *Teaching of Psychology*, 42(2), 109-118.

Boysen, G. A. (2016). Using student evaluations to improve teaching: evidence-based recommendations. *Scholarship of Teaching and Learning in Psychology*, 2(4), 273–284. https://doi.org/10.1037/stl0000069

Boysen, G. A. (2017). Statistical knowledge and the over-interpretation of student evaluations of teaching. *Assessment & Evaluation in Higher Education*, 42(7), 1095–1102. https://doi.org/10.1080/02602938.2016.1227958

Esarey, J., & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education*, 45(8), 1106–1120.

Gelber, S. (2000). *Grading the College: A History of Evaluating Teaching and Learning*. Baltimore, MD: Johns Hopkins Press.

Gourley, P., & Madonia, G. (2021). The impact of tenure on faculty course evaluations. *Education Economics*, 29(1), 73-104.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology* (Vol. 561). John Wiley & Sons.

Hativa, N. (2013). *Student ratings of instruction: A practical approach to designing, operating, and reporting*. Oron Publications.

Kreitzer, R. J., & Sweet-Cushman, J. (2021). Evaluating student evaluations of teaching: A review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics*, 1-12.

Mitchell, K. M., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, 51(3), 648-652.

Murray, H. G. (2005, June). Student evaluation of teaching: Has it made a difference. In *Annual Meeting of the Society for Teaching and learning in Higher Education* (pp. 1-15).

National Academies of Sciences, Engineering, and Medicine. 2020. *Recognizing and Evaluating Science Teaching in Higher Education: Proceedings of a Workshop–in Brief*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25685.

Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A metaanalysis. *Review of Educational Research, 74*, 215–253. http://dx.doi.org/10.3102/00346543074002215

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ*, 5, e3299.