Professional Leave Report Cover Sheet

Name: Kathleen Dyer

Department: Child & Family Science

College: Social Sciences

Leave taken: ◉ Sabbatical   ◯ Difference in Pay   ◯ Professional Leave without Pay

Time Period: ◉   Fall   2023

◯   Spring

◯   Academic Year

◯   Other

Your report will be sent to your Dean for your PAF and to the Library Archives.

Fall 2023 Sabbatical Report
Kathleen D. Dyer
Submitted January 2024

Success of the Leave

I had proposed two projects for my sabbatical leave, both allowing me to explore large systemic issues in higher education (Appendix A). First, I proposed to work on a proposal to write a book about my discipline of Human Development and Family Science (HDFS). Second, I proposed to compare shared governance on the 23 campuses of the California State University system. By the end of my sabbatical leaves, both of these objectives had been revised somewhat, but were ultimately successful.

First, the book proposal. I did submit the book proposal about my discipline, specifically profiles of historically important research in HDFS. I reviewed current textbooks to identify studies for inclusion, a project that I am currently writing into a manuscript for submission to a journal because it was such a revealing process. As a result of my proposal, my editor at Routledge has asked me to serve as a Series Editor to coordinate a series of textbooks for HDFS because I have such broad knowledge of the discipline, and have argued so compellingly for greater availability of disciplinary-relevant textbooks. However, I have decided to postpone writing the proposed book. The main reason is that another, different book started to seem a higher priority.

Near the beginning of my sabbatical semester, I finished copy edits to a manuscript that I wrote last year in which I used IPEDS data from the federal government to identify the scope of HDFS academic departments across the country. This paper was published in September 2023 (Dyer et al, <u>Overview of undergraduate degree programs related to human development and family science</u>). It immediately attracted a lot of interest from HDFS leadership in higher education, and I spent a lot of time conversing with colleagues across the country about the leadership of the discipline. I was asked to write a theory piece about the discipline for a special issue of a family science journal. I wrote that article and have completed a round of requested revisions. It will be published in December 2024. That invited article allowed me to immerse myself in the history and current status of HDFS, and I decided to continue work in that area. I launched a follow-up study that will add a department chair survey, a faculty survey, and an analysis of degree requirements to the data extracted from IPEDS data. I spent much of my sabbatical collecting that data. My professional organization, the National Council on Family Relations, has agreed (Appendix B) to publish that work as a research monograph. This research monograph will directly follow up on many of the issues raised by the original empirical article and the theory piece. It is much anticipated by disciplinary leadership.

Second, the system-wide analysis. By necessity, I narrowed the scope a bit and looked exclusively at the state of student ratings of instruction (SRI) on each campus of the CSU system. In a few cases, I drove to the campus and met with folks working on SRI. Mostly, I met with folks

over zoom and corresponded by email. It was deeply enjoyable, as I met several hard-working and thoughtful employees, ranging from administrators and faculty to staff. I compiled a report, comparing 1) SRI instruments, 2) practices related to administration of surveys, and 3) policies about the use of SRI data. My report includes some recommendations, both for our campus and for other CSU campuses. I have been invited to report my findings to a meeting of the AVPs of Faculty Affairs from all CSU campuses, as well as to the student ratings task force at CSU San Jose in February. In addition, I will be part of a panel at the 2024 CSU Teaching and Learning Symposium in February to discuss the use of SRI to improve teaching, and I am hosting a CSU system-wide meeting of those who work on SRI in March.

Finally, I would like to report that my sabbatical allowed me time to pursue multiple projects that were not listed as part of my formal proposal. These include:

- **Common Sense**: During my sabbatical, I finished revisions and copy-edits for a paper about <u>belief in common sense as a barrier to learning</u> based on data collected during online instruction during the COVID19 pandemic. It was published in October. When I went to a professional conference in November, I actually met someone who told me about exciting new research – and it was about my own paper! That had never happened to me before. ☺ But this is a topic that has not ever been addressed in our discipline. The excitement that this paper generated inspired me to launch a replication study that will significantly extend the hypotheses tested, and in a more normal setting. Along with a colleague, I collected data in several classes in my department in fall 2023. After data analysis, we will launch the second part of data collection in fall 2024.
- **History of Parenting Research**: I had been invited to write a book chapter for a book about the history of HDFS. During my sabbatical, therefore, I spent a full month researching the history of parenting research. This lined up with my growing interest in the history of my discipline. The chapter I've written is the first of its kind. It has been favorably reviewed, and will be published in 2025 in a book designed as a textbook for graduate school coursework in HDFS.
- **Lilienfeld Alliance**: Along with some colleagues at Columbia University, the State University of New York, and another from Fresno State, I have launched a new organization for professors in higher education – the <u>Lilienfled Alliance for the Teaching of Rational Skepticism in Higher Education</u>. We met at a national conference in October, and are planning that next year at the same conference we will host a half-day meeting focused on the scholarship of teaching and learning in the area of critical thinking/rational skepticism.

<u>Benefit to Faculty Member</u>
As anticipated, I needed the semester to transition out of my role as department chair. I was able to spend a lot of time off-campus, only popping in occasionally. I was able to advise our new department chair as often as she requested it, but the distance allowed me to put a hard stop at the end of my chair service, and psychologically transition to new possibilities. During the summer and fall, I had three papers published, which helped to jump start my return to more active scholarship.

The nature of the work I did during sabbatical allowed me to create meaningful relationships with disciplinary leaders on other campuses. I believe that these connections and the traction that my work has accomplished will allow me to participate in choosing the direction of my professional discipline.

Benefit to the University
My own department benefits from my sabbatical in that my expertise is more developed, and my department is primarily where I use that expertise.

Fresno State as a whole benefits from increased professional exposure and prestige as a result of my publications. For instance, the theory paper that will be published this paper tells the story of the history of Fresno State to illustrate how departments such as mine have evolved on university campuses.

My report on student ratings of instruction has already been read by people on most CSU campuses. About 30 of those people have RSVP'd for my system-wide meeting in March. That is a lot exposure, and I have received messages of gratitude from several recipients. Incidentally, one finding of my report is that Fresno State is the *only* CSU campus that is using an SRI instrument with demonstrated reliability and validity. A few people on other campuses have reached out to me to ask for guidance about how they might accomplish the same thing on their campus.

Appendices
Appendix A: Original Sabbatical Proposal
Appendix B: MOU with the National Council on Family Relations regarding a research monograph
Appendix C: Report on Student Ratings in the CSU System

Sabbatical Proposal

Faculty Member: Kathleen Dyer, Professor
College: College of Social Sciences
Department: Child & Family Science
Proposed Date of Sabbatical: Fall 2023
Submitted: September 2022

Section 1: The Proposal

As I enter the third decade of my professional career, my perspective on both my academic discipline of Human Development and Family Science (HDFS) and the system of higher education within which I do my work, has changed tremendously. In both cases, my view has expanded considerably, such that my interests are more systemic. As I complete my second term (my 8th year) as department chair, I have successfully achieved many significant goals of my academic department. At this point, I am eager to hand over the reins of local leadership to a colleague, and to pursue other work.

Thus, I propose a sabbatical that will allow me to address some large systemic questions about my profession in higher education. I propose to use my sabbatical to pursue two projects:

1) Write a book about the most impactful and important studies in family science, and
2) Compare how shared governance operates on multiple campuses of the California State University.

Book Proposal

In February 2022 my most recent book was published by Routledge. It is an introductory undergraduate-level research methods book for the discipline of Human Development and Family Science (HDFS). https://www.routledge.com/Research-Foundations-of-Human-Development-and-Family-Science-Science-versus/Dyer/p/book/9781032015576

While I was working on that book, I discussed with my editor the possibility of writing a smaller companion book that an instructor might choose to assign along with the textbook. My teaching strategy for the research methods class is to offer profiles of well-known and important studies that have shaped the discipline of HDFS, and to link those profiles to the research methods described throughout the semester. Therefore, I have been compiling a list (of sorts) of such studies for several years. Some are described in the text of my research methods book to illustrate principles of research. But not all were included, and I would like students to be able to analyze some such studies independently in order to apply what they have learned, so I didn't want all to be included. Therefore, my editor and I hatched the idea of a companion reader that could be published later, and could stand on its own as a description of seminal and impactful studies in family science.

I envision a book like "Twenty studies that revolutionized child psychology" by Wallace Dixon, first published in 2002 and then updated in 2015, but focused on family science rather than developmental science. The Dixon book is ranked #137 of all Developmental Psychology books on Amazon.com. The author identified "revolutionary" studies through a process of surveying developmental psychologists. Then he described each study in his own words, explaining the context in which the question was asked, the methods used by the research, the results, and the impact of that study on the discipline.

I propose using my sabbatical to write up profiles, in the same fashion as the Dixon book, of the studies that have landed on my list. At this point, the list has been populated by my personal understanding of HDFS, but also by talking to many of my family science colleagues and by casually perusing introductory texts. I propose a to conduct a more systematic analysis of citation records to identify the most widely cited studies in HDFS research, and a systematic review of the best-selling introductory textbooks to identify the most widely-cited studies described to students in the field. These projects could certainly yield journal article publications, but they would also help me identify the final list of studies to include in the book I want to write.

Shared Governance

I have significant university-wide leadership experience on the Fresno State campus as a result of my involvement in three activities:

1) Participation in various levels of the Academic Senate, currently by serving as a member of the Executive Committee.
2) Leading the transition to a new system for student ratings of instruction, and
3) Revitalizing and re-visioning the university-wide Council of Chairs.

All three of these activities represent shared governance on a university campus. The Academic Senate is the most visible and most formal manifestation of shared governance, in that the faculty literally make the policy by which the university functions. The Senate does so in the context of regular communication with upper administration, sometimes even at the request of upper administration, and policies must be approved by upper administration before they are enacted.

Student ratings (and other forms of evaluation of teaching effectiveness) also represent shared governance in two ways. First, the process is controlled by the Academic Senate. The transition to a new student ratings system was requested by the Senate, and the committee that I chaired was explicitly a Senate committee. The Senate supervises the administration of student ratings and maintains records of them. Furthermore, student ratings are used primarily by Personnel Committees, composed of faculty, for the purpose sharing the responsibility for governing faculty performance.

The Council of Chairs is not part of the formal structure of sharing governance for the university. However, department chairs are literally the first point of contact between the faculty and the administration. Chairs are the faculty members who are tasked with communicating between these two levels of governance: faculty and administration. Therefore, they play a pivotal role in how the university actually operates on a day-to-day basis. Therefore, the Council of Chairs has the capacity to share in the

governance of the university. At Fresno State, that capacity has rarely been realized, although the council is active and offers feedback to the administration even when not formally consulted. I have created an email discussion group with the chairs of other HDFS departments on other CSU campuses, and through that group have learned that chairs' councils are structured differently on other campuses. Some are formally recognized and officially incorporated into shared governance.

My involvement in Senate, student ratings, and the Council of Chairs during havoc wreaked in higher education by the COVID19 pandemic highlighted for me the ways in which multiple forms of shared governance either worked or did not work when under a pronounced strain. Some Senate policies had to be "suspended" during the COVID crisis, something that had never been done before and for which there was no allowance in our bylaws. It was declared that feedback from student ratings and peer evaluations were not to be used against faculty due to the burdens of the crisis, something that had no precedent. Student rating response rates dropped below reliability thresholds, thereby rendering them useless for personnel decisions. No policy response has yet been enacted. The Council of Chairs was strengthened because the chairs were an invaluable source of information about how curriculum was playing out in real time.

I am left wondering what is happening on other CSU campuses. How is shared governance working elsewhere? Would a comparison, even a casual one, help to identify some best practices? To that end, I would like to talk to some of my HDFS chair colleagues on other campuses. I would like to talk to some department chair leaders, some Academic Senate chairs, and to those who administer student ratings on other campuses. I will start by following up with my fellow chairs. From there, I will select campus visits based where I identify practices that are different from ours. When I visit another campus, I will start by talking to department chair leaders and Senate chairs. I will follow where my interviews take me. I expect that these conversations will lead me to meetings with union representatives, student body leaders, and other grass-roots campus leaders as well as reviews of formal policies and discussions with academic leaders such as deans and provosts.

Section 2: Benefits to the Faculty Member

I anticipate several benefits of the proposed sabbatical for me and my career. First, I need a little time to transition out of my long-term chair service, and to identify the next stage of my career. Since I do not yet know what that next phase will look like, this break in the patterns of my professional work will help me identify the next step.

Second, writing another book, and one contributes to the professional dialogue within my discipline, is a step I would like to take for my profession. It represents a step up in my professional maturity. I am now a member of the senior, most established generation of scholars in my profession. I want to make a contribution to my profession. I believe that the proposed book is a way that we in HDFS can look back at our history, and use that perspective to chart our future. Over the past two years, I have met occasionally with an ad hoc group of HDFS scholars to discuss the problems we see, as well as the promise of our interdisciplinary field. I find myself in a position of "in-between-ness" in my career that I

can use to contribute to this discussion, and broaden it beyond our small group. I want to grab the opportunity.

Finally, I am interested in pursuing professional work that allows me to contribute to Fresno State and to the CSU system as a whole, perhaps in an administrative capacity. I do not yet know what that will consist of, but I believe that the proposed sabbatical activities will broaden my skill sets and allow me to cultivate the required expertise to promote meaningful shared governance. This will help me identify and obtain my next job in my career trajectory.

Section 3: Benefit to the University

This sabbatical request is timed to facilitate my transition out of the role of department chair, and to help our next chair transition into that role. As I have been a long-term chair and I was involved in the hiring/training/promotion of virtually every other member of the department, it may be difficult for the next chair to assume the mantle of leadership if I am present, casting a shadow on that process. While I will make myself available to help, I think a little space will help the new chair have the freedom to establish their term. I believe that the whole department will benefit from facilitating a smooth transition.

The book and journal articles that I will produce as part of my disciplinary research will contribute to the academic reputation of Fresno State.

A more direct benefit to Fresno State will be the report I will produce regarding best practices in shared governance. This will be produced directly for the benefit of Fresno State. I will share the report with Provost Fu, the Council of Chairs, and the Academic Senate. My hope is that it will identify best practices with regard to shared governance, and that I will be able to make recommendations for specific changes we might consider on the Fresno State campus.

Section 4: Previous Leaves

I was granted a sabbatical in Spring 2016. I used that sabbatical to write a parenting education textbook that is still used in our CFS 135 (Parent Education) class, a required course for all majors in my department.

See attached for more details.

Nov. 20, 2023

To Whom It May Concern:

During August 2023, the National Council on Family Relations (NCFR) began working with Katie Dyer, Ph.D., of the Department of Child and Family Science at California State University, Fresno, on a research project regarding academic programs related to the discipline of Family Science, with understanding of the following:

- Dr. Dyer had begun and would continue to conduct a comprehensive review of U.S. academic programs that offer degrees related to Family Science; survey the department chairs and faculty of those academic programs; and review the undergraduate curricula of the programs.

- Following data collection, Dr. Dyer will write the results as a research monograph.

- NCFR has begun and will continue to provide consultation during Dr. Dyer's research process; provide a mechanism for peer review of the monograph; publish the final monograph; make the monograph available for sale; and advertise the monograph to NCFR members.

Sincerely,

Diane L. Cushman

Diane L. Cushman
Executive Director, National Council on Family Relations

About the *National Council on Family Relations* (NCFR): *Since 1938, the nonprofit, nonpartisan National Council on Family Relations has been the premier professional association for understanding and strengthening families through interdisciplinary research, theory, and practice. NCFR publishes three leading Family Science research journals, hosts an annual conference, certifies Family Life Education professionals, and provides numerous other publications and professional resources to its members. NCFR represents Family Science scholars and professionals by establishing standards and advocating for the Family Science discipline.*

The State of Student Ratings of Instruction in the California State University System

Kathleen Dyer, PhD
kdyer@mail.fresnostate.edu
January 2024

EXECUTIVE SUMMARY

Background: The use of Student Ratings of Instruction (SRI) became ubiquitous in higher education by 1990 as a result of pressure from both students and faculty. They are required by the collective bargaining agreement, and are used on every campus of the California State University (CSU) system. However, the practice remains controversial.

Objective: To describe the current use of SRI on the 23 campuses of the California State University (CSU) system. What is the quality of instruments being used? How are SRI administered? What policies govern the use of SRI results?

Methods: Information about SRI for each campus was identified via the campus website, an interview with at least one staff member who administered the system, and at least one faculty member or administrator who oversaw the process. Preliminary results were tabulated and checked for accuracy.

Results/Instruments: Campuses vary wildly in what name they give to the process of collecting student feedback about classes. The word "evaluation" is being removed and replaced with words like: feedback, opinions, ratings, reflections, and perceptions. Twelve campuses either use a single common instrument across campus, or have common instruments for a few types of classes (e.g., lectures and labs). The rest allow multiple instruments, which does not allow the possibility for testing for reliability and validity. Only one campus has explicitly tested its instrument for reliability and validity. Eight campuses are currently working on revising their system. This process generally occurs in the Academic Senate.

Results/Administration: There is no consensus about which office on campus administers SRI. It is being done by: Technology Services, Institutional Research, Faculty Affairs, deans offices, Academic Senate, and Center for Teaching and Learning. All campuses use online administration, but some also allow paper administration. All but three use a vendor for administration, with the most commonly used platforms being Scantron Class Climate, Anthology, and Explorance Blue. Response rates are alarmingly low across the system. Surveys are typically open for two weeks at the end of the semester, excluding final exams.

Results/Policy: All campuses collect qualitative comments from students, but four prevent those comments from becoming part of the personnel file and several others allow a mechanism for certain comments to be removed. Most campuses require that virtually all classes be rated with exceptions for supervision and low-enrolled classes. Most campuses do not have a policy about the use of incentives to improve response rates.

Results/Other Issues: Other issues that arose include a widespread interest in improving the potential for formative assessment to improve instruction, and the lack of guidance for personnel committees about appropriate use of SRI data.

Recommendations:
1. Improve validity and reduce bias by using expertise on campus to implement testing of instruments for reliability and validity. Include those with survey construction and statistical expertise in addition to representatives from multiple disciplines and class types. Revise instruments until they are theoretically based and demonstrably scientifically sound. On-campus experts should be compensated for this professional work. Task forces may need to be in place for longer than one year, as the process generally takes more than one year. This process could be facilitated centrally so that the burden does not rest entirely on each campus.
2. Reduce bias by using written feedback for formative assessment but excluding it from summative assessments that go in instructor personnel files. Each campus should carefully consider the use of comments in the process.
3. Prioritize student voice by maintaining the requirement that virtually all classes be rated without allowing individual faculty to selectively exclude classes. Communicate to students that SRI is an important and safe mechanism for them to be heard by their campus leaders.
4. Address the problem of low response rates. This issue should be investigated to identify evidence-based solutions. Currently, the best evidence is that requiring in-class administration is the most impactful practice. The use of incentives should be explored, particularly those that operate at the level of the institution rather than at the level of individual classes.
5. The window for administration of SRI surveys can safely be restricted to two weeks that should not include finals week. Longer windows increase work and annoyance without improving response rates.
6. Improve guidance offered to administrators and personnel committees about the use of SRI scores, especially when response rates are low and with regard to written comments.
7. Establish system-wide communication and collaboration about SRI practices in order to share expertise and experiences. The issues are the same on all campuses, yet currently each campus is addressing the issue alone. All could potentially benefit from an established network for those working on these reforms.

INTRODUCTION

Student ratings of instruction (SRI) refers to systematic feedback from students about their current classes that is solicited directly from students, via self-report, by the institution. SRI are used both to help instructors improve their teaching, and to provide administrators with a summative evaluation to use as the basis of personnel decisions. Furthermore, SRI are the only mechanism by which all students on a campus can have their voice heard by instructors and administrators alike, and they can do so in relative safety due to the anonymity of SRI.

The current report describes the current use of SRI on the 23 campuses of the California State University (CSU) system. The CSU system is the largest four-year university system in the country, and is thus a national leader in higher education. The policies and practices of the CSU system may therefore serve as an indicator of the status of SRI in higher education more generally.

Historical Context

SRI have been ubiquitous in higher education since the late 1980s (Seldin, 1998). This was the result of a push from both students and faculty. Both sides got what they wanted, student ratings administered by the institution with results considered in personnel decisions, and yet neither side currently seems happy about it.

In the early 20th Century, student feedback about classes was first solicited, compiled, and distributed by student groups (Canelos, 1985; Purdue University, 1972). These guidebooks about which faculty to seek and which to avoid were sometimes opposed by university administrators for the subversive nature of the reviews. They functioned, essentially, as Yelp reviews (or RateMyProfessor reviews) with no verification or quality controls. But these early reviews were sporadic because student groups have a lot of turnover and very rarely had the resources or infrastructure to regularly survey large samples of students. Students needed help to collect the desired information, and they wanted the institution to act to rectify the situation when poor instructors were identified. So student groups started to push for formal feedback to be collected and reviewed by the institution in order to maintain a high standard for teaching (Gelber, 2020).

Meanwhile, by mid-century, the social sciences had developed the expertise to create scientifically sound self-report instruments, and faculty interested in improving their teaching began to establish a scholarship of teaching and learning. Some scholars began collaborating with the student groups to systematize data collection and testing of instruments (e.g., Downie, 1952). Scientific results were promising. Advances were made in our understanding of effective college teaching (Canelos, 1985), and the instruments designed produced strong evidence of reliability and validity (e.g., Smalzried & Remmers, 1943; Spencer & Aleamoni, 1969). At this point, faculty started pushing for official recognition of their teaching responsibilities. They wanted retention, tenure, and promotion decisions to be based on their teaching as well as their scholarship (Canelos, 1985; Elbe, 1972; Guthrie, 1953). One way to do that was to use student ratings of instruction as part of personnel decisions (Seldin, 1980).

By 1990, the dual pressure from students and faculty had accomplished its goal. SRI were standard operating procedures on virtually all college and university campuses (Seldin, 1998). In

the CSU system, SRI are required by the collective bargaining agreement between the faculty union and each campus' academic policy.

Validity and Bias

Despite being actively pursued by both students and faculty, SRI have been controversial from their beginning (Canelos, 1985; Purdue University, 1974). In addition to the political pressures related to the sometimes-competing interests of students, faculty, and institutions, the social sciences were newly developing the science of self-report measurement when SRI were first created, and so the meaning of such ratings were contested at the beginning. As methodological and statistical innovations changed standards for educational research, investigation of SRI has continued unabated.

Popular news media regularly publish very critical opinion pieces, asserting that SRI do not reflect the quality of teaching, are fundamentally biased by race and gender, and should be banned (e.g., Flaherty, 2020; Kamenetz, 2014). The assertion is that SRI are biased by race and gender, and that they do not reflect teaching quality but just consumer satisfaction. All of this is said to drive lenient grading and grade inflation, and to harm the most vulnerable instructors doing the most difficult work, particularly young, female, scholars of color (e.g., Gutkin, 2023; Stroebe, 2020).

While the popular higher education media state that these negative findings are the uncontested truth, the scientific scholarship is not nearly so definitive. Research on the validity of SRI is perhaps one of the largest bodies of research literature in higher education (Linse, 2017), and it is profoundly split. In truth, no clear consensus can be declared about the value and utility of SRI. Some studies of student ratings find evidence of bias and a lack of validity while some studies find the opposite. Furthermore, proponents of SRI assert that they can be useful despite their limitations because they give a voice to students who are otherwise powerless in academia, and because better alternative methods for getting student input on instruction have not yet been identified (Burt, 2015; Gannon, 2018).

There is very little evidence in the published research that scholars have worked to explain the contradictory findings. Each side simply asserts that they are right and the other side is wrong. It has become a shouting match, with everyone apparently hoping to drown out the other side by simply making more noise.

Higher education needs to do better than this. If we are using invalid and biased SRI, then we need to find a better way of assessing teaching effectiveness. In short, we need to find out why we are uncovering contradictory findings. I have proposed (Dyer, 2021; Dyer & Donnelly-Hermosillo, under review) that the difficulty in arriving at consensus is produced by the use of different SRI instruments. The conflicting findings may be a result of a measurement problem. Some high-quality instruments may be more valid and less biased than other, lower-quality instruments. Perhaps this is why the research produces contradictory conclusions. Student ratings, per se, are not valid or invalid, but rather some measures are valid and other measures are invalid.

Specifically, it appears that a few instruments have been created carefully, based on theory, conceptualize teaching quality as a complex and multidimensional construct, and are well-tested before use. These instruments produce strong evidence of reliability and validity, and only rarely produce evidence of gender and race bias (e.g., Cohen, 1981; Costin, Greenough, & Menges, 1971; Marsh, 1982; Marsh, 1984). But some studies rely on less well-developed instruments, in fact, often just a single item asking for a global rating reflecting consumer satisfaction (e.g., "overall, how do you rate this class?"). These studies report dubious validity, and evidence of gender and race bias as well (e.g., Hamermesh & Parker, 2005; Kornell & Hausman, 2016; Marsh & Roche, 1997; Mau & Opengart, 2012; Rosen, 2018). It is not about SRI in general, it is about the quality of the SRI measurement tool.

The possibility that contradictory research findings is the result of a measurement problem inspires me to investigate the quality of SRI instruments being used in higher education. Are we using carefully constructed instruments that are theory-driven, multi-dimensional, and with evidence of reliability and validity? Or are we using instruments based on a consumer satisfaction model of education that rely heavily on single-item "satisfaction" items that have not ever been tested empirically?

Administration of Surveys
In addition to consideration of instruments used to measure student perceptions of instruction, a few other issues also weigh heavily on the use of SRIs in higher education. One is the delivery of the instrument. In the past, students were given paper surveys in class on one particular day of the semester. Response rates tended to be high, but students absent that day, or who needed a little more time to complete the survey, were out of luck. And of course, paper surveys are practically impossible in online classes. They can be mailed out to students and returned by mail, but that would be the only option.

As online classes became a staple of course offerings, universities found a way to deliver some surveys online. Gradually, online administration was adopted for some face-to-face classes as well for various reasons. The COVID-19 pandemic forced all classes online, and thus, all SRI were forced online as well. Research suggests that rating scores tend to be the same when surveys are completed online as compared to on paper, but response rates tend to drop, and the number of written comments tends to increase (e.g., Guder & Malliaris, 2010).

As we settle into post-pandemic modes of operation, presumably SRI remain primarily online. But this raises questions about the administration of SRI. Which office on campus is best equipped to manage that administration? What response rates are being achieved? What is the window for administration, and is the length of time that SRI are available for completion related to response rates?

Policy
Finally, there are important policy issues related to the implementation of SRI that have a direct bearing on the ethical use of ratings for personnel decisions. Advocates of SRI have always warned that SRI can be useful in personnel decisions only if they are "wisely-formulated, wisely-administered, and wisely-used," acknowledging that not all SRI instruments are equally valuable

and that there are limits on their use (Eble, 1971, p. vii). More contemporary scholars also suggest that the *way* SRI are used, rather than their existence, is the critical issue (e.g., Linse, 2017).

We here explore three issues related to the use of SRI. First is the use of open-ended written comments. Most instructors find comments invaluable when they use SRI to improve their classes because comments provide context and explanations for numerical ratings. Suggestions might be offered, and the nature of a critique can be explained. So comments are helpful to the instructor. But comments are also very likely to be riddled with gender and race bias.

Extensive evidence of gender and race bias is detected in the written comments on SRI instruments (e.g., Gelber et al., 2022; Hamermesh & Parker, 2005; Wallace, Lewis & Allen, 2019), and this has become the focal point of controversies related to SRI (e.g., Kreitzer & Sweet-Cushman, 2022). The Rate My Professor (RMP) online rating system, for instance, has appalling levels of gender bias that are easily detected in the written comments of RMP (Jaschick, 2015). As comments reflect simple consumer satisfaction, they can be provided to the instructor as constructive feedback, but should not be used for evaluative purposes because of the bias (Kreitzer & Sweet-Cushman, 2022).

The second issue investigated here is that of exclusions to the requirement to have all classes rated by students. The relevant issues include burden on students (where rating more classes is more of a burden, and thus may suppress response rates), burden on faculty (whereby administration of SRI may reduce class time and make them vulnerable to unfair attack) and the potential benefit to faculty (collecting evidence of high-quality teaching) and students (as an opportunity to have a voice, including the ability to report egregious behavior to administrators with low risk of retaliation). How are these competing concerns balanced by each campus?

Finally, we explore policies with regard to the use of incentives. Incentives are widely used in survey research to improve response rates, and are generally seen as compensation for a respondent's time (Dillman et. al, 2014). In the context of SRI, even very small incentives tend to be quite effective at improving response rates (Donmeyer et al, 2011; Sundstrom et al., 2016; Wode & Keiser, 2011.) Arguments against the use of incentives include the perception of coercion and the possibility of inauthentic responses.

Incentives might be at the individual level (e.g., extra credit points to students who complete the survey), at the class level (e.g., extra credit points to everyone if the class achieves a target response rate), or at the level of the institution (e.g., final grades are released early to those students who have completed their assigned ratings). Are campuses allowing the use of any of these incentives?

Prior Work and Current Objectives

A collaboration of the CSU Chancellor's Office with the California Faculty Association (CFA) and the CSU Statewide Academic Senate (ASCSU) documented the use of student ratings in the CSU system in 2008 (Joint Committee, 2008). That report found that the vast majority of CSU campuses used "homegrown" instruments developed by faculty committees, with only three

campuses using commercially-developed forms. The report did not indicate whether the so-called "home-grown" instruments had been tested for reliability and validity or not. The report seemed to assume that forms purchased from vendors had been tested for reliability and validity, but that has not been demonstrated and is likely to be a faulty assumption.

As of 2008, more than half of CSU campuses (n=16) used multiple different instruments. All but four campuses administered some surveys online, although only one campus relied entirely on online administration of surveys. All 23 campuses reported allowing written comments, and only one campus (Pomona) excluded comments from personnel files. The current report provides a status update on these and other issues.

In fall of 2020, the CFA (the faculty union) proposed changes to the language in the bargaining agreement about "student evaluations". Essentially, they proposed to remove the word "evaluation" and replace it with the word "opinions". They also proposed that faculty should have the right to respond to student opinion data that they believe to be biased by race or gender. As far as I know, these changes have not been incorporated into the collective bargaining agreement. And yet, it is good to know the intentions of the faculty body about student ratings, and this gives us some perspective.

METHODS

I identified the 23 campuses of the CSU system, and searched each university's website for information about their course evaluations/student ratings/student evaluations of teaching. I collected what information I could from these websites. Then, I used the website to identify someone on each campus to direct my questions to. When I could not find a website describing the process and identifying a contact person, I contacted someone in Faculty Affairs or the Academic Senate until I was directed to a person who could help.

On each campus, I was able to set up a meeting with one or two people who know enough about their SRI system to answer my questions. If the person I was talking to couldn't answer a question, I asked them to put me in touch with someone who could. Typically, it was a staff member who could tell me about the administration of surveys and a faculty member or administrator who could tell me about policy. Sometimes, one person on campus could answer all such questions. All data were collected in Fall 2023.

I was advised by our Committee for the Protection of Human Subjects that this project does not qualify as human subjects research\ because the institution is the subject, not the interviewee. Nonetheless, before I asked questions, I advised each interviewee of my plans to share my results in a report that I would share with all participants.

Most interviews were conducted over Zoom and took about 30 minutes each. In a few cases, the respondent preferred to exchange a series of emails instead. In a few cases, I visited the campus and had face-to-face meetings with respondents. I asked a standard set of questions about the issues described above and a few additional issues that arose in my conversations. A preliminary draft of this report was sent to all respondents to verify that their campus information was presented accurately, and several campuses offered minor corrections.

RESULTS

A partial profile of the system on each campus is provided in the three tables below: Table 1 is about the instruments in use, table 2 is about the processes for administration, and table 3 is about policies.

Instruments
Name? I discovered truly staggering degree of diversity in the naming of the systems used to collect student feedback about their courses and instructors. The term "course evaluation" is used by six campuses with several variations on how it is used. "Student evaluations of teaching effectiveness" is used two times, and "student perception of teaching" is used twice. The rest are all used only once and are listed in Table 1 below.

This diversity seems to reflect efforts to remove the word "evaluation" but a lack of consensus about what should replace it. Opinions, ratings, feedback, reflections, perceptions...all of these are options. There is also variety about what exactly is being addressed: courses or teaching/instruction.

My reading of the published literature suggests that "course evaluations" and "student evaluations of teaching – SET" are the most commonly used terms. Both use the term "evaluation" which is problematic. The argument against the term "evaluation" is that students are not really qualified to evaluate instructors. Instead, they provide information (ratings, feedback, opinions, reflections, or perceptions) and then peers and supervisors use that information to actually *evaluate* instructors. Prior research is clear that students are, in fact, reliable and valid reporters of what happens in their classes (Murray, 1983; Albanese, Schuldt, Case & Brown, 1991). When instruments ask them only to report, and do not ask them to evaluate, they can do so. The evaluation comes later, by those qualified to do so.

This reframing of what used to be called "student evaluations" gets right to the point about bias and lack of validity. If we ask students to do something that they are not qualified to do (i.e., evaluate) then the product will not be valid. But perhaps if we ask them to give us directly observable information instead, the results will be more valid.

Common Form? Several campuses have multiple instruments, rather than having one uniform instrument for the campus. On this point, there has been very little change since the 2008 report. Five campuses have unique instruments produced by each department with revisions allowed at any time – these campuses typically have more than 100 different instruments in use. Six campuses use a common instrument but allow departments to add items. Three have a few specialized instruments (e.g., for lab classes) but otherwise use a common instrument. The remainder (n=9) use a single common instrument for the whole campus.

The use of multiple instruments on the same campus (especially those that vary by department rather than class type) suggests that the instruments are very unlikely to have been created by those with expertise in survey construction, and are therefore likely to produce more bias. They are also almost impossible to test for reliability and validity because large numbers are necessary for that kind of testing.

Reliability and Validity? Only one campus (Fresno) uses an instrument that has been explicitly tested for both reliability and validity. Another (San Jose) has tested for internal reliability but not validity. Three more (Long Beach, Maritime, San Bernardino) have taken substantial measures to either carefully select items or revise items, usually with an eye toward reducing potential bias.

A lack of demonstrated validity is a problem if SRIs are used in personnel decisions. The 2008 report recommended using instruments produced by vendors as a way of addressing this concern, but very few CSU campuses followed that advice. Validated measures are not really available from vendors at this point, 15 years later.

The IDEA instrument claims to be validated, but it is based on the principle of self-reported learning, which is dubious. They describe the validation process as simple convergent validity between IDEA ratings and a single global question ("overall, how would you rate this class?"). This is not adequate to claim demonstrated validity. Some vendors currently offer access to a bank of questions, but not an instrument with demonstrated reliability and validity. Reliability (consistency) and validity (accuracy) can only be determined at the level of the instrument, not at the level of individual items, so pools of items do not satisfy the need for reliability and validity.

One solution is to use validated instruments reported in the scientific literature, most of which were created decades ago. Another solution is to use faculty who have expertise in self-report measurement / scale construction (these can be found in most social sciences) to conduct the necessary testing on campus. This expertise is certainly available on all CSU campuses. I suspect that the reason this testing has not been done widely is that this requires serious professional labor. Such labor needs to be appropriately compensated if it is to be done professionally – it goes beyond typical committee service for the scientist who would oversee the project.

Revision? Eight campuses are currently in the process of revising their instrument or otherwise considering their process for student ratings. Very few campuses have a standing faculty committee for oversight and review of SRI, but most place this responsibility firmly within the scope of the Academic Senate, usually the Policy committee. Efforts to revise typically arise within the Senate, and if there is enough momentum, a task force is appointed. A task force typically runs for one academic year. It is very difficult to complete a major research project on volunteer labor in a 9-month period.

Table 1: Instruments

Campus	Name	Website	Common Form?		Testing	Under Revision?
			2008	2023		
Bakersfield	SOCI: Student Opinionnaire of	None	Yes	Yes	No	No - revised in 2015

	Courses and Instruction				
Channel Islands	Student Ratings of Teaching	https://www.csuci.edu/academics/facultyaffairs/evaluation.htm	Yes	Yes (but programs can add items)	No	No
Chico	SFOT: Student Feedback on Teaching and Learning	https://www.csuchico.edu/ir/sfot/index.shtml	Yes (but depts can create their own instead)	Yes (but depts can create their own instead)	No	No
Dominguez Hills	Perceived Teaching Effectiveness	https://www.csudh.edu/faculty-affairs/perceived-teaching-effectiveness-pte/	Yes (but depts can add questions)	Yes	No	No – task force last year
East Bay	Student Evaluations of Learning	https://www.csueastbay.edu/online/classes-and-academic-support/course-eval.html	Yes (but depts can add questions)	Yes (but depts can add questions)	No	No
Fresno	SRI: Student Ratings of Instruction	https://sites.google.com/mail.fresnostate.edu/fresno-state-sri/home	No - each dept or college has its own	Yes	Reliability and validity	No
Fullerton	SOQ: Student Opinion Questionnaires	https://www.fullerton.edu/far/soq/index.php	No – each dept has its own	No – each dept has its own	No	No
Humboldt	CEBS: Course	https://cebs.humboldt.edu/	Yes (but depts can	Yes (one for lectures, one for labs)	No	No

	Evaluations by Students		add questions)			
Long Beach	SPOT: Student Perceptions of Teaching	https://www.csulb.edu/academic-technology-services/student-perceptions-of-teaching-spot	Yes (but depts can add questions)	Yes	Items selected from national pools	Yes
Los Angeles	SOS: Student Opinion Surveys	https://www.calstatela.edu/registrar/student-course-evaluations	Yes	Yes	No	No
Maritime	Student Course Evaluation Surveys	None	Yes	Yes (with additional questions for lab classes)	Items revised to remove bias	Yes
Monterey Bay	Course Evaluations	https://csumb.edu/cat/course-evaluations/	Yes	Yes	No	No
Northridge	SEF: Student Evaluations of Faculty	https://www.csun.edu/institutional-research/sef-resources	No – each dept uses its own	No – each dept uses its own	No	Yes
Pomona	Course Evaluations	https://www.cpp.edu/data/course-evaluations.shtml	Yes (but depts can add questions)	No – each dept uses its own	No	No
Sacramento	Course Evaluations	https://csus.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=1d9086fc-	No – each dept uses its own	No – each dept uses its own	No	No

		ccbb-47f0-9c90-ae6e010ea747				
San Bernardino	SOTE: Student Opinions of Teaching Effectiveness	https://www.csusb.edu/trc/resources/sotes	Yes (a couple of exceptions)	Yes	Items revised to improve value of formative feedback	Yes
San Diego	Student Feedback Surveys	https://fa.sdsu.edu/_resources/files/tenuretrack_evaluations/guidelinesforassessingteachingeffectiveness.pdf	No (by dept or college)	Yes (but depts can add questions)	No	Yes
San Francisco	SETE: Student Evaluations of Teaching Effectiveness	https://sete.sfsu.edu/	Yes (but depts can add questions)	Yes (but depts can add questions)	No	No
San Jose	SOTE: Student Opinions of Teaching Effectiveness	https://www.sjsu.edu/teachingeval/	Yes	Yes	Reliability	No – 2017 was most recent review
San Luis Obispo	Online Evaluations of Faculty	https://academic-personnel.calpoly.edu/content/course-evals-student-faq	No (by dept or college)	No – each dept uses its own, with a few common questions	No	No
San Marcos	Course Evaluations	https://www.csusm.edu/course-evals/	Yes, with variants by class type	Yes, with variants by class type	No	Yes – task force just disbanded but did not

						create new instrument
Sonoma	ROSE: Reflections on Students Experience	https://data.sonoma.edu/sete	Yes (but depts can add questions)	Yes (but depts can add questions)	Not yet	Yes
Stanislaus	SPOT: Student Perceptions of Teaching	https://www.csustan.edu/academic-senate/course-evaluation	Yes	Yes	No	Yes

<u>Administration of Questionnaires</u>
The administration of SRIs is mostly a matter of practice rather than policy. A summary of these practices is provided in Table 2. Campus websites related to SRI are listed for general information.

Responsible Office? Campuses vary widely in the location of student ratings of instruction on campus. About 8 campuses locate the administration of student ratings somewhere in their Technology Services division. Another 5 locate it in Institutional Research, and 5 locate it in the Faculty Affairs office or the Provost's Office. Two campuses place the responsibility with multiple analysts in deans' offices, 2 in the Academic Senate office, and 1 campus administers surveys through their Center for Teaching and Learning.

I spoke with many of the administrators of the system about how they judge the arrangement on their own campus. They recognized limitations of every possible placement. Both Tech Services and Institutional Research have the expertise to administer the large online systems, but they are perceived as being disconnected from instruction, so they simply do what they are told with very little room for assessment of how the system is working. When it's run through faculty affairs, there is a perception that it is disconnected from instruction and it seems to emphasize the punitive use of SRI rather than the potential for formative feedback. Everyone agreed, however, that Faculty Affairs has more right to see the information than does any other office on campus. Centers for teaching and learning offer an opportunity to capitalize on information about what could be improved, but most campuses want to retain faculty autonomy about when to seek help, so they do not want information from student ratings shared with staff automatically.

I spoke with a couple of staff members who are responsible for administration of ratings who have taken the job with them when they moved from one office on campus to another office on campus. The job seems to be attached to a competent person, and that determines what office houses the task. It is not the case that administrators are making thoughtful decisions about where it OUGHT to be, it is simply placed wherever a person is available and willing to do it.

Online or On Paper? The transition from paper to online surveys was just beginning in 2008, when one campus (San Diego) had moved entirely to online and 16 were rolling it out on a limited basis. By 2023, the transition to online had been very nearly fully accomplished. Very few campuses allow a paper option anymore. Two campuses (Chico and Sacramento) are obligated by policy to allow a paper option. Four more (Bakersfield, Fresno, Monterey Bay, and Stanislaus) allow paper as an option without the policy obligation to do so. Many others noted that the COVID-19 pandemic solidified the switch to online administration.

Vendor? Now that student ratings are mostly administered online, most campuses contract with an outside vendor to administer those surveys. Three campuses (Dominguez Hills, Los Angeles, San Jose) use an in-house system, and one (Dominguez Hills) is actively looking for a replacement.

The most widely-used vendor is Scantron Class Climate (n=12, soon reduced to 11). Some of those using Scantron are actively looking to replace it. Shortly before the pandemic, Scantron announced changes in the program, increasing the cost and reducing the amount of customization that would be supported. The main complaint by those who still use it is that customer service has declined precipitously. A second complaint is that custom solutions previously put in place will no longer be supported. But several users say it perfectly meets their needs, and does not require a huge investment of time to operate. In fact, I asked administrators to estimate how much time per semester they spend on student ratings. While those estimates were difficult to land on, those using Scantron Class Climate reported the least amount of time per semester on the task.

Four campuses use an Anthology product. Anthology has recently announced that it is moving clients away from the CourseEval product to a new system. That new system seems to be Campus Labs and IDEA, which Anthology recently acquired. They are trying to sell the IDEA product (a statistically a reliable instrument that is based on the problematic assumption that students can accurately report how much they learned).

Three campuses (soon to be four) use Explorance Blue. It seems to be the most powerful system, but also the most expensive and with the most difficult user interface. San Diego has recently started to use their ability to send survey invitations through text message, so we should all stay tuned to find out if this improves response rates for them.

A few campuses are looking at the Course Evaluation system offered by Qualtrics, and Long Beach has just started using it. No one has moved far enough along in the process to report back.

Response Rate? The main issue that every campus is facing is that of low response rates. When response rates are low, faculty fear disproportionate impact from a few disgruntled students. Low response rates also create a situation where ratings could not possibly be used in personnel decisions (one way or the other) because they are not legally defensible. But online rating systems always produce a drop in response rates. And the COVID pandemic seems to have created survey fatigue (or online work fatigue) that dealt a serious second blow to response rates for student ratings. Across the CSU system, and in higher education more generally, response rates are currently very low and everyone is scrambling to figure out what to do about it.

I asked each campus for their response rates last year. A few campuses could provide an exact number because they track it carefully, but many provided an estimate. Therefore, the numbers I offer here should not be taken as definitive. All respondents noted that the response rate varies considerably from course to course, department to department, and many have put significant effort into trying to identify the relevant factors.

The highest response rates were reported by Chico, Humboldt, Pomona, and San Francisco, but some of these are estimates, the reliability of which is unclear. It may be valuable to identify the strategies used on those campuses to see if they can be replicated elsewhere, but that was not within the scope of this report.

Response Window / Finals Week? Most campuses have a two-week response window for surveys to be open for students to complete, and all but one campus (Fresno) close that window before final exam week. Longer windows are not associated with higher response rates. And I know from my own experience that a very long window (such as what we have at Fresno) creates a lot of work for the administrator of the system, and a lot of frustrated angry emails from students who feel harassed by reminder emails when they persist for several weeks.

The rationale for closing ratings before final exams is two-fold: 1) During finals week, students should be focused on preparing for their exams. That should be their focus, and so it is wrong to ask them to do something unrelated during that time. 2) Grades may be finalized during that week, and there should be no link between ratings and grades. There should be no opportunity to "thank" an instructor for lenient grading, and no opportunity to "punish" an instructor for harsh grading.

Table 2: Administration

| Campus | Responsible Office | Online | | Vendor | Response Rate | Response Window | Finals |
		2008	2023				
Bakersfield	Enterprise Applications	Only for online classes	By instructor preference – 65% online	Scantron Class Climate	27% Online; 50-60% Paper	4-5 weeks	No
Channel Islands	Faculty Affairs	None	Entirely	Anthology Course Evals	37%	2 weeks	No
Chico	Institutional Research	Only for WebCT classes	By instructor preference –	Scantron Class Climate	50%	2 weeks	No

			80% online				
Dominguez Hills	Faculty Affairs	Some, moving that way	Entirely	In-house system (but looking)	30%	4 weeks	No
East Bay	Online Campus	Only in MA in Online Teaching program	Entirely (since COVID)	Scantron Class Climate	30%	2 weeks	No
Fresno	Academic Senate	Only for online classes, some dept preference	By instructor preference (98% online)	Explorance Blue	45%	7 weeks	Yes
Fullerton	Faculty Affairs	Only for online classes, some dept preference	Entirely	Scantron Class Climate (but looking)	40%	2 weeks	No
Humboldt	Administrative staff in college dean's offices	Only for online classes	Entirely	Scantron Class Climate	50-70%	2 weeks	No
Long Beach	Academic Technology	Only for some online classes	Entirely	Qualtrics Classroom	42%	3 weeks	No
Los Angeles	Instructional Technology	None	Entirely	In-house system	10%ish	2 weeks	No
Maritime	Administrative staff in college dean's offices	None	Entirely	Scantron Class Climate	36%	2 weeks	No

Monterey Bay	Academic Technologies	Only for online classes, some indiv preference	By instructor preference (about 75% online)	Anthololgy (but looking)	42% online; ? paper	17 days	No
Northridge	Institutional Research	Only for online classes, some dept preference	Entirely	Anthology Course Evals	30-35%	4 weeks	No
Pomona	Institutional Research	None	Entirely	Explorance Blue	60%	2 weeks	No
Sacramento	Tech Support	Only in Nursing	By instructor preference (80% online)	Scantron Class Climate	33% online; 65% paper	3 weeks	No
San Bernardino	Information Technology	Piloting	Entirely	Scantron Class Climate (but looking)	30%	2 weeks	No
San Diego	Office of Faculty Affairs and Student Success	Entirely	Entirely	Explorance Blue	35%	2 weeks	No
San Francisco	Academic Technology	Piloting in School of Business	Entirely	Scantron Class Climate	50-60%	12 days	No
San Jose	Institutional Effectiveness	Piloting	Entirely	In-house system	48%	2 weeks	No
San Luis Obispo	Faculty Affairs	None	Entirely	Scantron Class Climate	30-40%	1 week	No

				(but looking)			
San Marcos	Vice-Provost's Office	Piloting	Entirely	Scantron Class Climate - changing to Exploranc e Blue	Less than 30%	2 weeks	No
Sonoma	Institutional Effectivenes s	None	Entirely	Anthology	37%	2 weeks	No
Stanislaus	Academic Senate	Piloting	By instructo r preferen ce (about 90% online)	Scantron Class Climate	30-40%	2 weeks	No

Policy

A few policy issues came up in my conversations. While I do not have complete data on these questions, I present here (Table 3) what I learned about policies related to student ratings.

Comments in Personnel Files? All campuses include student comments (responses to the open-ended questions) as part of the SRI process. Most include these comments in the instructor's personnel file, but some campuses do not. Many campuses reported controversy over these comments. Some have created a mechanism by which faculty can request to have certain comments (those that reflect race or gender bias, or that are sexually explicit) removed from their record. Some campuses (Fresno, Fullerton, Long Beach, and Pomona) prevent all comments from going into personnel committees, on the grounds that they are provided to faculty for the purpose of improving instruction, but that they are very likely to reflect bias and so need not be part of a personnel file.

Classes Excluded? The collective bargaining agreement states that "all classes" should be rated, unless the campus president approves of a reason for that not to be done. Some campuses have built exclusions into their policy, which is signed off by the campus president.

Many campuses exclude supervision classes, or classes with low enrollment (although the operational definition of "low enrollment" varies from <10 students to < 3 students). A few allow departments to decide which courses should be excluded, sometimes by committee and sometimes by the department chair. Seven campuses (Chico, Dominguez Hills, Fresno, Los Angeles, Northridge, San Diego, Stanislaus) only require instructors to rate a specific number of courses, and allow faculty to decide which ones those will be.

The rationale for allowing faculty to exclude classes is not totally clear. At Fresno State, this policy was put in place when we used a vendor (IDEA) that charged by the class, so there was a financial incentive to reduce the number of courses rates. Another potential reason to exclude classes is to reduce the number of surveys students are expected to complete, and thereby potentially increase the response rates. However, faculty at Fresno State routinely solicit more student ratings than they are required to obtain. Faculty report that they want the feedback, and they want students to have the ability to share their perspective. It seems that offering exclusions may not actually reduce numbers very much, if at all.

On the other hand, several of my respondents noted that allowing faculty to opt out of student ratings seems counterproductive in that faculty will decline feedback in precisely the classes for which they might need it the most.

Incentives allowed? As a result of low response rates, many campuses are considering whether or not to allow (or even encourage) faculty use incentives to encourage participation. Policies on incentives vary widely in the CSU. Two campuses explicitly allow class-level incentives and disallow individual-level incentives. Most campuses have no policy about the issue of incentives.

Several of my respondents mentioned that institution-level incentives (e.g., early release of grades for those who have completed SRIs) would be desirable. This type of incentive removes the burden from individual instructors, and reduces any perception of a trade of good ratings for good grades. No one knows why this is not allowed, but Many are curious about whether it could be pursued.

Table 3: Policy

Campus	Comments in PAF?	Classes Excluded?	Incentives Allowed?
Bakersfield	?	?	Yes
Channel Islands	Yes	No- all classes rated	Yes
Chico	Yes	Tenured faculty only rate two classes per year. Classes with <10 not rated.	Yes – No policy
Dominguez Hills	Yes	Lecturers rate two classes per year, tenured faculty rate one class per year	No policy, but discouraged
East Bay	Yes	All classes except independent study/internship/student teaching. No restrictions by class size.	Yes
Fresno	No	Rate two classes per year for tenure track faculty, exclude if < 3 students	Policy is ambiguous

Fullerton	No	Exclude if enrollment is under 3; other exclusions per department requirements	No individual incentives, but group incentives are okay
Humboldt	?	Exclude <4 students, exclude supervision classes	Yes, recommended
Long Beach	No	must be c-classification and exclude classes with less than 6 students. No instructor choice.	No
Los Angeles	?	Tenured faculty rate 4 classes per year	Yes
Maritime	Yes, but faculty can request to have specific comments removed	No - All classes rated	Yes
Monterey Bay		No – all classes rated	Yes, recommended
Northridge	Yes	Yes – faculty select which to rate	Yes
Pomona	No	exclude classes with less than 6 students, departments maintain list of excluded classes	No individual incentives, but group incentives are okay
Sacramento	Yes	exclude independent study and classes with 5 or fewer students	No policy, but discouraged
San Bernardino	Yes	No – all classes rated	Yes
San Diego	Yes	all classes with 2+ students rated	No policy. No forbidden or allowed.
San Francisco	?	Exclude supervision classes	Yes – no policy
San Jose	yes, but faculty can request to have specific comments removed	exclude supervision classes, and those with 5 or fewer students. But not all have to go in the personnel file. Exclude 1 class per year if you taught 15+ units.	No
San Luis Obispo	Yes	exclude independent study and classes with fewer than 5 students	Yes
San Marcos	Yes	Yes, determined at college level	Yes
Sonoma	Yes	exclude independent study/thesis/supervision. Exclude classes with fewer than 6 students.	Yes – no policy

Stanislaus	No	Faculty rate 50% of their classes	No

Note: Data are not available for all of these questions because I started asking these questions only after noticing variability. I did not ask every campus these questions.

Campus Reports

Through my contacts, I was able to collect some written reports about SRI that had been written by various campuses. I share these below in case they are useful as references.

Table 4: Campus Reports

Year	Source	Report Title and Hyperlink
2008	Joint Committee of CSU, CFA, and ASCSU	Report on Student Evaluations of Teaching
2019	Fullerton	Student Opinion Questionnaire Committee Final Report
2019	Fresno	Task Force Report on the Creation and Validation of the Instrument
2022	Fresno	Response Rates for Student Ratings of Instruction
2022	San Jose	An Interpretation Guide for the Student Opinion of Teaching Effectiveness Surveys (SOTES)
	San Bernardino	https://docs.google.com/document/d/1QTr9Rwh39QKMhfPbF42qP3ery7QSl9kmhlcSjny7oC4/edit?usp=sharing
2023	Dominguez Hills	Evaluation of Current CSUDH Perceived Teaching Effectiveness (PTE) Practices – Subcommittee Report – Spring 2023 (not available online)
2023	San Marcos	Senate Task Force to Overhaul Student Opinion Survyes on Teaching (SOST) Final Report

Other Issues

Two other issues came up frequently in my interviews.

First, many campuses have identified the need for better guidance to personnel committees and administrators about how to use information from SRI in personnel decisions. What can be done when there are low response rates? When standard deviations are large? How should comments be considered (if at all)? What is considered a "low score" that should trigger concern? What else can be considered as evidence of teaching effectiveness so that SRI are not the only indicators?

One campus (San Jose) regularly updates a guide for the use of student ratings (link provided in Table 4). Other campuses might consider constructing a similar document for their campus to clarify how SRIs might be used appropriately.

Second, faculty on several campuses are concerned about how to better use SRI for formative assessment that directly contributes to improvement of teaching, rather than just as summative assessment used by administrators. On campus (San Bernardino) is experimenting with an optional early-semester feedback option that can be optional, in addition to the final summative rating. Another (Sonoma) is discussing a change in policy to require improvement in ratings over time, rather than a single static threshold for adequate quality.

RECOMMENDATIONS

Instruments

All campuses in the CSU are facing some similar challenges with regard to SRI. While some of the details are lost to history, it appears that most are still using SRI instruments that were created back in the 1980s when SRIs were first required, and none of those instruments were created according to scientific standards that we would hope for.

Many campuses are re-considering the name of their process. Generally, they are removing the word "evaluation" from the name to reflect a different approach to the system. Rather than asking students to evaluate instruction, they are asking students to provide information that will allow experts to evaluate instruction. We have not yet landed on an alternative word. I have used "student ratings of instruction" because that is what is used on my own campus. I like that the word "ratings" conveys that we are asking for numerical scores and that this is the core of what students are offering. I also like that it avoids words (such as "opinions") that suggest students are offering something very personal and subjective. However, I'm not sure that we need to agree on common language for this artificially. Until something naturally emerges as common language, it is slightly inconvenient that we do not share a name, but I think we can live with this inconvenience.

Consideration of what name to use, however, reflects something much more important that is happening on several campuses. Because we are generally rejecting "evaluation" we are reconsidering what we are asking students to do. If not "evaluate", then what? Many campuses have created task forces in recent years, typically through the academic senate, to explore issues related to SRI. Most of the work in this area has been in the form of reviewing the literature about sources of bias in SRI. Without exception, the campuses involved in such work have reported that they agree that there are problems with validity and bias in SRI, but that SRI must be maintained in order that students have some voice.

These senate task forces are interested in improving the quality of their instruments. To that end, several campuses are also investing in revisions to instruments based on item analysis to reduce explicit bias in items. If students are not evaluating their instructors, then they are reporting what

is happening in their class. That requires us to think carefully about what we want them to report. A lot of thoughtful work is being done in this area. In particular, task forces are recommending that we stop asking students to report things like friendliness and professional dress, because these are not essential to quality instruction and are often biased by gender and/or race.

Other recommendations that can be found in the literature include:
1. Identify the dimensions of quality teaching. Folks might be good at one thing (e.g., giving timely and meaningful feedback) but not great at another (e.g., allowing active participation in classroom activities). A good instrument should be multidimensional. There is a lot of published research about empirically validated dimensions of teaching.
2. Avoid double-barreled questions, those that ask two things in one question. For instance, when we ask "was class fun and interesting?" how should a student respond if it was interesting, but not fun?
3. Avoid consumer-satisfaction questions, such as "overall, how do you rate this class?" or "what is your overall opinion of this instructor?" These unfocused questions are just invitations to insert one's own bias rather than reporting on the instructor's teaching practices.
4. Avoid questions that ask students how much they learned. Human beings are very poor reporters of their own learning or skills.
5. Focus on so-called "low-inference" teaching behaviors. These are specific behaviors that can be directly observed or experienced. Students need not make complicated inferences about things that are invisible. For instance, a student would have to infer whether or a not a teacher is "knowledgeable" or "kind" because that cannot be directly observed. Low-inference items would include things like whether the instructor answered student questions, or listened to student points of view.

While progress is being made in the area of reducing bias in survey items, CSU campuses have not yet taken the step of subjecting their new or revised instruments to formal testing for reliability and validity. It is strongly recommended that they solicit the involvement of social scientists on their own campus to conduct such testing. This process requires explicit support, in the form of funding, from either the Chancellor's Office or Academic Affairs on each campus. While academics with expertise in survey construction are certainly available on each campus, they need to be compensated for their professional efforts to do this work. Scholars who devote their time and skills to validating SRI on their campus do so at the cost of their own scholarship and opportunities for publication. Therefore, they should be professionally compensated for that.

In addition, task forces created by the Senate generally last for one academic year. After appointing members and setting a meeting time, that generally leaves about 8 months for the work of the task force. While a task force may be the most appropriate way to get this work off the ground, a review of current practices, review of the literature, revision of items, and testing for reliability and validity simply cannot be done on an 8-month timetable. This may explain why the final step in the process does not usually get completed. Campuses may need to identify a scholar who will spearhead the validation and a small team to help with that, and plan to compensate that person (or the whole team) for professional work that will take two years.

Administration

There is a lot of variation in the administration of SRIs on CSU campuses. I could find no justification for recommending that one office or another should be used uniformly in the administration of SRI, or for the use of one vendor over another. Each campus settles on a process that meets their particular needs, which are quite varied across the system.

However, I strongly recommend a short window (2 weeks, maybe 3) for administration of surveys that does not include finals week. This conforms to the most widely used practices in the system, and the justification (above in the Results section) is convincing for both practices.

Finally, there may be some benefit to a system-wide discussion of response rates, and how to improve them. All campuses are struggling with this issue. Sharing resources and information might be helpful. See the section below on policies about incentives for more on attempts to increase response rates.

Policies
The inclusion of student comments in faculty personnel records remains a sore point without clear direction. I am convinced that they are deeply problematic as the likeliest place for gender and race bias to distort student ratings. There are rampant stories circulating of their misuse by personnel committees and administrators. For instance, I heard of instances where an administrator selected one atypical comment and used it to justify a negative tenure decision, despite dozens of comments that offered a completely different assessment. I've also heard of departmental staff members tasked with reading all comments and identifying representative ones to go on the formal record, which is not something they are qualified to do, nor should they be allowed access to that confidential data. Faculty also resent blatantly sexist or racist comments living forever in their personnel file. It feels demeaning. I personally think comments have no place in personnel decisions because of the risk of misuse and bias.

However, some respondents reported to me that the comments are considered by some chairs and administrators to be far more useful than the numbers. Faculty sometimes want to include comments in their file because the comments help contextualize scores. For instance, scores on the low side accompanied by comments such as "she's a really tough grader....you have to read the book for this class" might provide some justification and therefore protection to the instructor. Similarly, high scores accompanied by comments that describe the use of high-impact practices and timely responsiveness help to paint a picture of the instructor's good work. I also heard that committee members and administrators sometimes want the comments in order to understand the nature of low scores, which can help suggest solutions.

At the very least, policy-makers on each campus should consider how their policies might mitigate the risks of bias that are built into comments. Some campuses have procedures in place by which offensive and/or biased comments can be removed from the personnel file. This is one option that might reduce the risk of such comments. However, it creates another layer of work for many people, and even draws salacious attention to comments that may be counterproductive. But this issue is something that the Senate on each campus should consider.

Secondly, policy on each campus clarifies which classes should be exempt from student ratings. The CBA states that this is what is expected, but it allowed for exceptions to be decided on each

campus. If SRI exist to help improve instruction, then such feedback should be solicited for all classes. If SRI exist to provide a voice to students, then they should have the right to provide feedback about all of their classes.

Some classes must be exempt, of course, because ratings cannot practically be anonymous in a class with only one or two students. And there may be other legitimate reasons for excluding certain classes. But I recommend that this decision *not* be left to individual faculty members because of the risk of abuse.

Finally, each campus is entitled to develop its own policy about the use of incentives to promote student participation in SRI. Currently, most campuses have no such policy and do not have any idea whether their faculty are using incentives. My recommendation is that campuses explore whether incentives are being used, and if so, whether that use is appropriate. Campus policies should probably include some guidance about what is appropriate and inappropriate use of incentives. Furthermore, we might discuss with the Chancellor's Office whether it is possible to implement institution-level incentives. On my own campus, I've been told that we are prohibited from adjusting the release of grades on the word of the Chancellor's Office. Perhaps we should explore options.

CONCLUSION

In sum, there is room for reform of the use of SRI in the CSU system. Most of the instruments we use for SRI, and many of our policies, likely exacerbate the risk of invalid and biased results.

We ought to prioritize reform that maintains and respects student voice, reduces the potential for gender and race bias against instructors, demonstrates the reliability and validity of instruments used for this purpose, and maximizes the potential for feedback that helps to improve instruction.

References

Albanese, M. A., Schuldt, S. S., Case, D. E., & Brown, D. (1991). The validity of lecturer ratings by students and trained observers. *Academic Medicine, 66*(1), 26–28. https://doi.org/10.1097/00001888-199101000-00008

Burt, S. (2015). Why not get rid of student evaluations? *Slate, May 15, 2015.* https://slate.com/human-interest/2015/05/a-defense-of-student-evaluations-theyre-biased-misleading-and-extremely-useful.html

California State University (2008). Report on student evaluations of teaching. Produced by a joint committee of the California State University, the California Faculty Association, and the Academic Senate of the CSU. https://www.calstate.edu/csu-system/faculty-staff/academic-senate/Documents/reports/Report_on_Student_Evaluations_of_Teaching.pdf

Canelos, J. (1985). Teaching and course evaluation procedures: A literature of current research. *Journal of Instructional Psychology, 12, 4,* 187-195.

Dillman, D. A., Smyth, J. D., & Melani Christian, L. (2014). Internet, phone, mail, and mixed-mode surveys: The Tailored Design Method, 4th edition. Hoboken, NJ: Wiley.

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. Assessment & Evaluation in Higher Education, 29(5), 611-623.

Downie, N.M. (1952). With the Technicians: Student evaluations of faculty. *Journal of Higher Education, 23, 9,* 495-503.

Dyer, K. D. (October 14, 2021). Student ratings of instruction: Current controversies and future directions. Provost Award Lecture Series, California State University, Fresno. https://youtu.be/2t4PDRRtBoc

Dyer, K.D. & Donnelly-Hermosillo, D. (under review). Student ratings of instruction: Updating measures to reflect recent scholarship.

Eble, K.E. (1972). *The recognition and evaluation of teaching.* Project to Improve College Teaching, American Association of University Professors.

Flaherty, C. (2020). Even 'valid' student evaluations are 'unfair'. *Inside Higher Ed, Feb 26, 2020.* https://www.insidehighered.com/news/2020/02/27/study-student-evaluations-teaching-are-deeply-flawed

Gannon, K. (2018). In defense (sort of) of student evaluations of teaching. *The Chronicle of Higher Education, May 6, 2018,* https://www.chronicle.com/article/in-defense-sort-of-of-student-evaluations-of-teaching/

Gelber, S.M. (2020). *Grading the College: A History of Evaluating Teaching and Learning.* Johns Hopkins University.

Gelber, K., Brennan, K., Duriesmith, D., & Fenton, E. (2022). Gendered mundanities: gender bias in student evaluations of teaching in political science. *Australian Journal of Political Science*, 1–22. https://doi.org/10.1080/10361146.2022.2043241

Guder, F. & Malliaris, M. (2010). Online and paper course evaluations. *American Journal of Business Education, 3,* 131-138. https://files.eric.ed.gov/fulltext/EJ1060301.pdf

Guthrie, E.R. (1953). The Evaluation of teaching. *American Journal of Nursing, 53, 2,* 220-221.

Gutkin, L. (2023). The Review: Course evaluations are garbage science. *The Chronicle of Higher Education, December 4, 2023.* https://www.chronicle.com/newsletter/the-review/2023-12-04?utm_campaign=che-social-20231205&utm_medium=o-soc&utm_source=fb&fbclid=IwAR2QlrHpaIMR9BvyHQba6l3q-shnxCfwZ8H-izEADj3aasEAkzzYM9Nrwho

Hamermesh, D. S., & Parker, A. (2005). Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*(4), 369–376. https://doi.org/10.1016/j.econedurev.2004.07.013

Jaschick, S. (2015, Feb 9). Rate my word choice. *Inside Higher Ed.* https://www.insidehighered.com/news/2015/02/09/new-analysis-rate-my-professors-finds-patterns-words-used-describe-men-and-women

Joint Committee of the CSU, CFA, and CSU Academic Senate (2008). Report on student evaluations of teaching. California State University. https://www.calstate.edu/csu-system/faculty-staff/academic-senate/Documents/reports/Report_on_Student_Evaluations_of_Teaching.pdf

Kamenetz, A. (2014). Student course evaluations get an 'F'. NPRed, Sept 26, 2014. https://www.npr.org/sections/ed/2014/09/26/345515451/student-course-evaluations-get-an-f

Kreitzer, R. J., & Sweet-Cushman, J. (2022). Evaluating student evaluations of teaching: A review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics, 20*(1), 73–84. https://doi.org/10.1007/s10805-021-09400-w

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation, 54,* 94–106. https://doi.org/10.1016/j.stueduc.2016.12.004

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college teaching effectiveness. *Journal of Educational Psychology, 75*(1), 138–149. https://doi.org/10.1037/0022-0663.75.1.138

Purdue University, Measurement Research Center (1974). Institutional Research Bulletin, 74, 1,

Seldin, P. (1980). *Successful faculty evaluation programs: A practical guide to improve faculty performance and promotion/tenure decisions*. Coventry Press.

Seldin, P. (1993). The use and abuse of student ratings of instruction. *The Chronicle of Higher Education*, A-40.

Seldin, P. (1998). How colleges evaluate teaching: 1988 vs.1998.American Association of Higher Education Bulletin,50,3–7.

Smalzried, N.T. & Remmers, H.H. (1943). A factor analysis of the Purdue Rating Scale for instructors. *Journal of Educational Psychology, 34, 6,* 363-367.

Spencer, R.E. & Aleamoni, L.M. (1969). The Illinois Course Evaluation Questionnaire: Description of its development and a report of some of its results. Illinois University, Urbana, Office of Instructional Resources.

Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology, 42,* 276-294.

Sundstrom, E.D., Hardin, E.E., & Shaffer, M.J. (2016). Extra credit micro-incentives and response rates for online course evaluations: Two quasi-experiments. *Teaching of Psychology, 43,* 276-284.

Wallace, S. L., Lewis, A. K., & Allen, M. D. (2019). The state of the literature on student evaluations of teaching and an exploratory analysis of written comments: Who benefits most? *College Teaching, 67*(1), 1–14. https://doi.org/10.1080/87567555.2018.1483317

Wode, J., & Keiser, J. (2011). Online course evaluation literature review findings. Academic Affairs. Columbia College Chicago. Retrieved from website: http://about.colum.edu/academicaffairs/evaluation-and-assessment/pdf/Course%20Evaluation%20Literature%20Review.pdf