RESOLUTION TO ELIMINATE QUANTITATIVE RANKING OF FACULTY

RATIONALE

Quantitative student ratings within the instrument commonly known as Student Evaluations of Teaching Effectiveness (SETEs) are used at San Francisco State University as a primary tool in evaluating instructional faculty for both ordinary and extraordinary employment decisions. However, these ratings fail to provide actionable data to improve teaching outcomes and student learning outcomes, are demonstrably biased, are statistically meaningless, and fail in numerous other dimensions. Scholars have amply documented the harm caused by SETs to women (Austin, 2021; Gelber et al, 2022; Hoorens et al., 2021) and BIPOC (Chavéz, 2020; Lazos, 2012; Wang & Gonzalez, 2020) faculty. Further, despite prior claims of a high correlation between positive student evaluations and student learning, recent studies found low or even zero correlation, meaning that students do not learn better from instructors who receive positive scores (Stroebe, 2020; Uttl, White & Gonzalez, 2017). While some scholars of faculty evaluation propose methods for extracting limited usable insight from SETs while minimizing bias (Kreitzer & Sweet-Cushman, 2021; Linse, 2017), the application of these methods requires considerable additional resources, training and labor, reducing the likelihood that they will be implemented. Functionally, quantitative ratings in SETEs constitute an "arbitrary classification," which violates the Equal Protection Clause (XIVth Amendment) of the U.S. Constitution. Further, the inappropriate use of quantitative ratings erodes faculty confidence in teaching effectiveness assessment systems and attendant faculty retention, promotion, and tenure processes. In light of these and other problems, there is no way to recuperate quantitative ratings for any legitimate purpose, and therefore, quantitative ratings must be eliminated from teaching effectiveness assessment practices.

WHEREAS, the numerical ratings component of the instrument known at San Francisco State University as Student Evaluations of Teaching Effectiveness (SETES) are used as a primary instrument by departments, programs, and colleges at San Francisco State University in the classification of instructional faculty work performance as satisfactory or unsatisfactory, which in turn determines in significant part both ordinary and extraordinary employment decisions (eg. which instructional faculty are retained or not retained, or are granted or denied range increases or promotions; which instructional tenure-track faculty are granted or denied tenure; and which instructional lecturer faculty are granted or denied contracts for subsequent semesters or years). Therefore, quantitative ratings in SETEs have a profound impact on employment actions related to instructional faculty; and

WHEREAS, numerous examinations of student evaluations designed similarly to San Francisco State University's SETES have shown that SETE do not accurately or reliably measure short-term or long-term teaching effectiveness (i.e. what they purport to measure), and in a 2017 "Meta-analysis of faculty's teaching effectiveness," Uttl, White, and Gonzalez argue, "The best evidence − the meta-analyses of SET/learning correlations when prior learning/ability are taken into account − indicates that the SET/learning correlation is zero," and conclude that there is no correlation between SETs to student achievement of learning outcomes (2017); and

WHEREAS, UC Berkeley Professor of Statistics Philip Stark and Richard Freishtat, Vice President of Curriculum at UC Berkeley Executive Education, expose the rating system of student evaluations as predicated on multiple errors of basic statistical science and debunk their apparent objectivity:

> Personnel reviews routinely compare instructors' average scores to departmental averages. Such comparisons make no sense, as a matter of Statistics. They presume that the difference between 3 and 4 means the same thing as the difference between 6 and 7. They presume that the difference between 3 and 4 means the same thing to different students. They presume that 5 means the same thing to different students and to students in different courses. They presume that a 3 "balances" a 7 to make two 5s. For teaching evaluations, there is no reason any of those things should be true [6]. SET scores are ordinal categorical variables: The ratings fall in categories that have a natural order, from worst (1) to best (7). But the numbers are labels, not values. We could replace the numbers with descriptions and no information would be lost: The ratings might as well be "not at all effective,"…, "extremely effective." It does not make sense to average labels. Relying on averages equates two ratings of 5 with ratings of 3 and 7, since both sets average to 5" (Stark and Freishtat, 2014; p . 2).

These findings lead to the conclusion that quantitative ratings in student evaluations cannot be recuperated for legitimate purposes; and

WHEREAS, the consensus of scholarship on SETs conclude that their numerical ratings conceal and amplify bias with respect to race, gender and other characteristics (Austin, 2021; Chavéz, 2020; Gelber et al, 2022; Hoorens et al., 2021; Lazos, 2012; Wang & Gonzalez, 2020), and that, while redesign and advanced training in statistical science and anti-bias may reduce such bias, it cannot be fully eliminated (Kreitzer & Sweet-Cushman, 2021; Linse, 2017); and

WHEREAS, the majority of those who use SETE ratings as part of employment decision processes receive little or no anti-bias training on how to appropriately interpret and apply SET quantitative ratings and comments for employment purposes;

WHEREAS, student ratings distributions are typically negatively skewed, giving more weight to students with biased outlier views: "In skewed distributions, means are sensitive to (influenced by) outlier ratings; in student ratings, these outliers are almost always low scores...While students with outlier views are not unimportant, they should not be given more weight than the views of most students" (Linse, 2017, p. 101-102) and further that "When results are summarized and only mean or median ratings are included in a dossier, negative scores ... are inadvertently awarded extra weight in a review" (Linse, p. 103), thus amplifying the harm of biases (whether implicit or explicit); and

WHEREAS, the majority of those who use SETE ratings as part of employment decision processes receive little or no statistical analysis training on how to appropriately interpret and apply SET quantitative ratings and comments for employment purposes; and such training, to be effective, would be onerous [eg. the training suggested by "A Guide for Making Valid Interpretations of Student Evaluation of Teaching (SET) Results":

> Relevant stakeholders [i.e. anyone involved in producing or analyzing data from SET ratings] should receive training on both basic survey research principles and

psychometric concepts such as validity and reliability. Survey research training should focus on key concepts, such as sample size, MOE, and confidence levels, and how each of these factors interacts in the context of SETs. Training on basic psychometric properties such as validity and reliability should focus on the notion that validity not only addresses the accuracy of a set of scores but also the appropriate interpretation and use of scores. (Royal, 2017)]

WHEREAS, such lack of training has been shown to exacerbate the tendency to "over-interpret small differences as indicative of a problem, a decrease in quality, or an indication that one faculty member is materially better than another" (Linse 2017, p. 100); and

WHEREAS, at SFSU, nearly every department, program, or school's retention, tenure, and promotion policy specifies that faculty SETE ratings must be "better than the department mean" or meet a number typically below 2 (on a scale from 1-5, in which 1 represents excellence); and according to Linse, "Unit means are not an appropriate cutoff or standard of comparison because there will always be some faculty members who are, by definition, "below the mean." This is particularly problematic in units with many excellent teachers" (2017; p. 102); and

WHEREAS, because poor ratings can be produced by multiple variables including factors beyond the control of the instructor (Hoben, Bedenhorst, & Picket, 2020; Linse 2017, p.100; Uttl & Smibert, 2017; Wolbring & Treischl, 2016), there is no evidence that these scores provide actionable data to instructional faculty to improve teaching outcomes; and

WHEREAS, summative ratings of teaching effectiveness, given at or near the close of the semester of teaching being evaluated provide no opportunity for formative professional development of faculty teaching effectiveness; and

WHEREAS, "Inappropriate use of student ratings breeds mistrust, fosters inequities and inconsistencies, and ultimately demoralizes the faculty" (Linse, 2017; p. 103); and

WHEREAS, San Francisco State University is a public university, an agency of the state of California, and as such subject to the guarantees of the equal protection clause of § 1 of the Fourteenth Amendment to the Constitution of the United States; and

WHEREAS, the Supreme Court of the United States has held that, in accordance with the equal protection clause, the rules, principles, or standards employed by a state and its agencies may not create classifications among individuals which are "essentially arbitrary" (*Lindsley v. Natural Carbonic Gas Co.*, 220 U.S. 61, at 79 (1911), quoted in *Morey v. Doud*, 354 U.S. 457, 464 (1957); a classification "must be reasonable, not arbitrary, and must rest upon some ground of difference having a fair and substantial relation to the object" of the rule, principle, or standard, "so that all persons similarly circumstanced shall be treated alike" (*F.S. Royster Guano Co. v. Virginia*, 235 U.S. 412, at 415 (1920)); and see also *Skinner v. Oklahoma ex rel. Williamson*, 316 U.S. 535 (1942), *Reed v. Reed*, 404 U.S. 71 (1971), and *Eisenstadt v. Baird*, 405 U.S 438 (1972); and

WHEREAS, the Student Evaluations of Teaching Effectiveness (SETE) currently employed to classify the quality of work of instructional faculty at San Francisco State University are essentially arbitrary, not a reasonable classification of the quality of teaching effectiveness, fail to treat similarly circumstanced persons alike, and are an illegitimate, unfair, and illegal

employment of the university's power to evaluate the quality of work by instructional faculty and to classify instructional faculty employees accordingly; so therefore be it

RESOLVED, that San Francisco State University shall expeditiously eliminate quantitative rating systems from student perspective gathering instruments and that such ratings be retroactively and henceforth excluded from the personnel files of all instructional faculty.

References

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. https://www.scienceopen.com/document?vid=818d8ec0-5908-47d8-86b4-5dc38f04b23e.

Boyle. M. & Schmierbach, M. (2021) Measurement in the classroom: Using student evaluations to explain research concepts and improve your teaching. Routledge.com. https://www.routledge.com/blog/article/using-student-evaluations-to-explain-research-concepts

Chávez, K. (2020). Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity. *PS, Political Science & Politics*, *53*(2), 270–274. https://doi.org/DOI:10.1017/S1049096519001744

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*, *57*(2), 199–220. https://doi.org/10.1080/10361146.2022.2043241

Gelber, S. (2020). G*rading the college: A history of evaluating teaching and learnin*g. Johns Hopkins University Press.

Hoben, J.L., Badenhorst, C. & Pickett, S. (2020). Student evaluations and the performance of university teaching: Teaching to the test. *LEARNing Landscapes, 13*: 161—172. https://files.eric.ed.gov/fulltext/EJ1261356.pdf.

Hoorens, V., Dekkers, G., & Deschrijver, E. (2021). Gender Bias in Student Evaluations of Teaching: Students' Self-Affirmation Reduces the Bias by Lowering Evaluations of Male Professors. *Sex Roles*, *84*(1–2), 34–48. https://doi.org/10.1007/s11199-020-01148-8

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*. https://doi.org/10.1007/s10805-021-09400-w

Stark, P. B. & Freishtat, R. (2014) An evaluation of course evaluations. *ScienceOpen Research*. doi.10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

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*(4):, 276—294. https://doi.10.1080/01973533.2020.1756817

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. *PeerJ*, *2017*(5), 1–13. https://doi.org/10.7717/peerj.3299

Uttl, B. White, C., & Gonzalez, D.W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*: 22—42. ISSN 0191-491X, https://doi.org/10.1016/j.stueduc.2016.08.007.

Wang, L., & Gonzalez, J. A. (2020). Racial/ethnic and national origin bias in SET. *International Journal of Organizational Analysis*, *28*(4), 843–855.

Wolbring, T., & Treischl, E. (2016). Selection Bias in Students' Evaluation of Teaching: Causes of Student Absenteeism and Its Consequences for Course Ratings and Rankings. *Research in Higher Education*, *57*(1), 51–71.

https://doi.org/10.1007/s11162-015-9378-7