Expert Report on Student Evaluations of Teaching (SET)

Richard L. Freishtat, Ph.D.

30 September 2016

Prepared for Ryerson Faculty Association ("RFA") and the Ontario Confederation of University Faculty Associations ("OCUFA")

I. **CURRENT POSITION AND AREAS OF EXPERTISE** (CV attached)

A. I currently serve as Director of UC Berkeley's Center for Teaching and Learning (CTL). In this capacity, I create, lead, and facilitate a variety of faculty development programs. Such programs include the Teaching Excellence Colloquium for new faculty, and the Presidential Chair Fellows Curriculum Enrichment Grant program, which aims to develop, improve, transform, and examine core areas of the undergraduate curriculum. Having consulted within each School and College across the campus, I provide individual and small group expert consultations with faculty on course-level pedagogy and assessment, oftentimes coupling consultations with classroom observations of teaching and the interpretation of student course ratings. In my 7+ years of professional experience in this field, I have reviewed as part of consultation thousands of SETs - seeing firsthand both their utility and misuse.

B. I am frequently invited to deliver custom workshops at UC Berkeley to faculty groups and departments on teaching and learning topics. Active in efforts to help faculty improve and innovate their pedagogy, and spotlight their successes, I launched and continually write for the Berkeley Teaching Blog, and author content for the Teaching@Berkeley newsletter.

C. I have been teaching courses at both the undergraduate and graduate levels since 2001 at institutions such as Pennsylvania State University, Arizona State University, and now at UC Berkeley. These courses have ranged across the areas of rhetoric and communication to education. The courses in education have focused on today's college student, and factors related to teaching and learning including evaluation.

D. I have professional experience in the evaluation of teaching, often sought out as a consultant and resource to inform the process and improve it, and within this a focus on the use and administration of student evaluations of teaching. I have published and presented widely on topics such as: An evaluation of student evaluations (2014, 2016), how social media impacts learning and student use of technology in the classroom (2010), ways to leverage faculty enrichment efforts to broaden participation and impact (2010, 2014), and fostering teaching excellence (2011, 2012, 2015, 2016). I also have served as a reviewer of conference proposals and scholarship of teaching and learning for national organizations and leading journals in the areas of pedagogy, student learning, and teaching improvement and evaluation, including for PODNetwork (most prominent international organization for professionals in the field of faculty development, teaching and learning).

E. As an ex officio staff representative to UC Berkeley's Academic Senate Committee on Teaching, I co-authored a white paper articulating ways to systematically improve the evaluation of teaching process and lay out practical means to change the current practices that fall outside existing policy.

F. I have been an invited speaker and leader of international programs on faculty development, teaching and learning, and the evaluation of teaching. I have conducted such programs and delivered talks at the Kuwait Foundation for the Advancement of Society, the UC Berkeley Center for Studies in Higher Education, the University of Toronto and have been invited to do so at Tokyo Tech University and in Beijing, China for the Chinese Ministry of Education's National Association of Education Administrators later in 2016.

G. In my role as Director of the CTL at Berkeley, I am often asked to write letters of assessment in support of personnel cases up for merit and promotion, and often asked by the faculty member and the review committee for an expert interpretation of the student evaluations, within the context of additional sources of evidence.

H. Appendix 1 is my current Curriculum Vitae.

I. All opinions set forth below are my own.

II. SUMMARY OF OPINIONS

A. I understand that student evaluations of teaching (SET) are called Faculty Course Surveys (FCS) at Ryerson University. I use the term SET below.

B. Aspects of teaching SETs do and do not measure.

1. There is little consensus on what SET do measure. SET ratings have been shown to have high correlation with students' grade expectations (Marsh & Cooper, 1980; Short et al., 2012; Worthington, 2002), reaction to instructor attractiveness (Ambady & Rosenthal, 1993), along with many other unrelated characteristics (these and others will be addressed in detail later in the report).

2. Student Evaluations of Teaching (SETs) are primarily measures of student satisfaction with their experience in a course. SETs do not accurately measure a faculty member's teaching effectiveness as a single source and method. While there is some debate in the literature, there is no compelling correlation between student learning and more highly rated instructors. In fact, whether a student is satisfied with their experience in a course depends on many confounding factors that have nothing to do with the instructor's teaching effectiveness. Many of the factors that affect SETs are not what should be affecting SETs (e.g., Was the student earning the grades they thought they deserved throughout the course? Was this a required course on a topic the student did not wish to take? Did the student find the instructor's accent or appearance to be pleasant or unpleasant?). More so, while the debate in the literature persists, a recent well-documented

review and analysis exposed flaws in the findings from many seminal studies advocating SETs as a measure of teaching effectiveness because they are reflective of student learning. In Uttl et al's (2016) review and re-analyses of previous analyses they found that the moderate SET-learning correlations reported were an artifact of small study size effects. When the SET-learning correlations were re-analyzed taking into account the small study size effects, the estimated SET-learning correlations dropped to near zero for nearly all of the SET-learning correlations reported in the previous analyses. As a result, SET surveys are often known as a measure of "customer satisfaction" (Beecham, 2009, p. 135).

3. Students should not be used to rate the adequacy, relevance, and timeliness of the course content nor the breadth of the instructor's knowledge and scholarship (Scriven, 1995). Most students lack the expertise needed to comment on whether the teaching methods used were appropriate for the course, if the content covered was appropriate for the course, if the content covered was up-to-date, if methods of student engagement used were appropriate to the level and content of the course, if the assignments were appropriate for promoting and assessing their own student learning, if what they learned has real world application, if what they learned will help them in future classes, if the type of assistance, help or support given to students was appropriate to the learning goals of the class, if the difficulty level of the course material was at an appropriate level, and if the course or the instructor was excellent, average or poor overall.

4. A point of clarification: SETs are not technically an instrument of measurement - although they are too often inappropriately used as such. Even proponents and advocates for SETs make it a clear point that they are not a tool to evaluate teaching, but rather a ratings method. As SET advocate and scholar Nira Hativa wrote in a 5/31/2016 post to the *Forum for Teaching and Learning in Higher Education*: "[W]hat SETs can contribute is insight into the experience of students who are in the class throughout a course. First, indeed and agreed upon by all experts, SETs do not measure teaching effectiveness (or teaching quality)! Rather than that, SETs do present students' perceptions/opinions of their teachers. Second, SETs do not measure - ratings and measuring are two different concepts."

5. Current students are well positioned to comment on their own experience of the class and inputs like: instructor's ability to communicate clearly, enjoyment, difficulty or ease, engagement or boredom, if an elective then whether they plan to take a sequel course, favorite/least favorite part of a course, whether they would recommend this course to other students, hours spent per week outside of class, and background information about pre-requisites taken or other courses in

the field. Even within these areas of student ratings, student comments can be valuable, but interpreting them is troublesome.

C. Effect of response rates on reliability of SETs

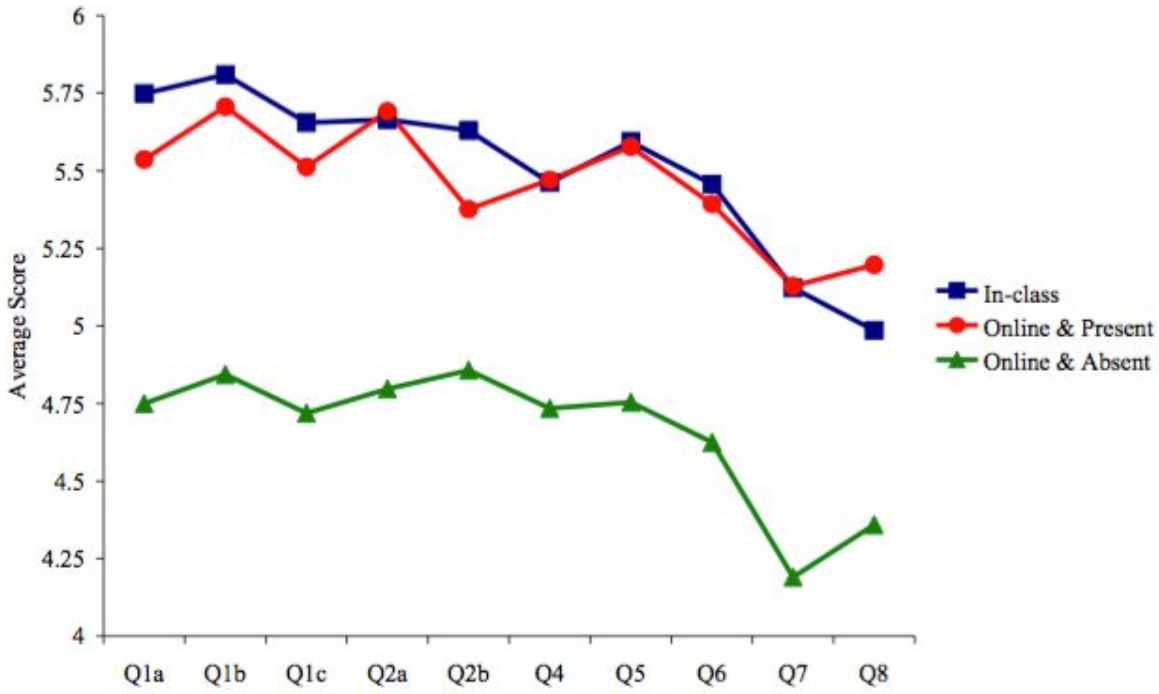
1. Response rates and who responds matter overall, and particularly in course evaluations administered online. Generally, two-thirds response rate is a minimum standard to inform the ability to present the spread of ratings as adequately representative of the class (Davis, 2009). This means the data can be used, but should be done so with caution and as part of a larger composite from multiple sources and methods of teaching effectiveness. Anything less than 100% response rate can be misleading in regard to the data and its interpretation (Davis, 2009), and even at 100%, SETs should never be used as the sole source of evaluating teaching effectiveness. Experts recommend further that when the minimum standard response rate is not met, the data should be interpreted cautiously for personnel decisions (Davis, 2009; Cashin, 1999; Theall and Franklin, 1990).

2. Responders are not a random sample and there is no reason their responses should be representative of the class as a whole. It is inappropriate to extrapolate, and therefore anything less than a 100% response rate can begin to impact the reliability of the spread of ratings.

3. Response rates and who responds significantly impact evaluations administered online - not necessarily because of the evaluation medium itself, but because of the students captured through the method. In a pilot study conducted at UC Berkeley, we found that the evaluations administered online capture a third group of students who traditionally were not responding to SETs - students who are enrolled, but do not attend class. Findings showed that not-in-class (absent) students tend to rate instructors and the course systematically lower on every survey item than those who attend class regularly (see graph below from UC Berkeley report which compares average ratings on question items across SETs administered on paper in-class in blue, online by students who attend class regularly in red, and online by students who do not regularly attend class in green). On measures related to academic performance, the absent group stands out. Compared to their counterparts enrolled in the course, the absent group tends to be less far along in their studies (though not necessarily younger), had a lower prior-term GPA, and earned lower grades in the course on average. Absent students who can now complete SET online disproportionately affect average question ratings and it brings up questions about the reliability of these responses if the students have not been actively attending class. These students tend to provide the most negative evaluations, producing a tendency for online

evaluations to be more negative than evaluations administered in-class on the whole.

Figure 4. Evaluation Responses by Mode and Attendance



4. Because of the need for instructors to garner high ratings *and* high response rates, which will inevitably include students who do not have a positive experience in a course for reasons outside of teaching effectiveness, teaching to SETs occurs. Instructors are disincentivized to improve and innovate teaching, and are instead incentivized to focus on approaches not driven by increasing student learning (e.g., lower course rigor) that are highly correlated to increased student ratings.

D. How personal characteristics of the instructor and/or student affect SETs.

1. Ambady and Rosenthal (1993) found that students' opinions about teachers are formed within seconds of being exposed to the nonverbal behavior and physical attractiveness of these teachers. Since that initial work was done, bias studies have also focused on other personal traits that are considered strongly related to SETs (Spooren et al., 2013). Examples include not only physical attractiveness (Campbell et al., 2005; Gurung & Vespia, 2007; Hamermesch & Parker, 2005; Riniolo et al., 2006), but also instructor fairness (Wendorf & Alexander, 2005), professor attitude (Kim et al., 2000), image compatibility as a subjective assessment of how well various images align (e.g., image of current teacher as compared to subjective image of ideal teacher; Dunegan & Hrivnak, 2003),

instructor likability (Delucchi, 2000), and initial impressions of a teacher (Tom et al., 2010). At the same time, bias studies have also examined student characteristics such as level of motivation, prior ability, and prior education in the field of the course (Langbein, 2008).

2. Gender matters (instructor and student). SET are affected by gender biases and stereotypes. In an example of how the gender of the student *and* instructor both matter, Boring (2015) demonstrated how male first-year undergraduate students favor male instructors with higher ratings, even though there is no difference between the academic performance of male students of male and of female instructors. MacNell et al. (2015) exposed similar bias based on instructor gender, showing when students think an instructor is female, students rate the instructor lower on every aspect of teaching, including putatively objective measures such as the timeliness with which instructors return assignments. Arbuckle and Williams (2003) demonstrated bias based on instructor gender finding that when students were told an instructor was young and male, students rated the instructor higher than for the other three combinations of young female, old male, old female, especially on "enthusiasm," "showed interest in subject," and "using a meaningful voice tone." Ottoboni et al. (2016) further revealed that the association between SET and (perceived) instructor gender is large and statistically significant, and that the bias based on gender of both instructors and students varies by factors such as discipline. Instructors whom students believed were male received significantly higher average SETs. Using data sets from both the U.S. and France, they concluded that "SET primarily do not measure teaching effectiveness, that they are strongly and non-uniformly biased by factors including the genders of the instructor and student, that they disadvantage female instructors, and that it is impossible to adjust for these biases" (p. 2).

3. Ethnicity and Race matter. Instructors of color tend to receive SET ratings that are biased downward (Huston, 2005; Basow et al., 2013; Boatright- Horowitz & Soeung, 2009; McPherson & Jewell, 2007; Smith, 2007; Smith and Anderson, 2005). Scholars have revealed a number of contributing insights and factors, including how Blacks and Asians were evaluated more negatively than White faculty in terms of overall quality, helpfulness, and clarity (Reid, 2010), and how the adjectives used to describe faculty in evaluations are less favorable as compared to White (male) faculty (Storage, 2016), how complaints from students in evaluations against Nonnative English speakers stem from a broader project of social exclusion and bias against accents (Subtirelu, 2015), and how racial minority faculty are rated lower across ratings of the quality of instruction (Babad, Darley, & Kaplowitz, 1999; Fortson & Brown, 1998; Ogier, 2005).

4. Age matters. Age has been found to negatively impact teaching evaluations. Bianchini et al. (2013), showed that student evaluations fluctuate based on

inherent characteristics of aged appearance, which are not changeable by an individual instructor.

5. Attractiveness matters. Instructors who appear attractive receive better student ratings (Hamermesh & Parker, 2005; Riniolo et al., 2006; Wolbring and Riordan, 2016). Wolbring and Riordan (2016) showed study participants a three-minute engineering lecture presented by a computer-animated professor who varied by gender and race (Black, White). They found that more attractive instructors receive better ratings. In fact, attractiveness ratings significantly correlated with each of the teaching dimensions.

6. What students read from other students matters. One study by Legg and Wilson (2012), found that comments on RateMyProfessors.com influence the attitude of initially unbiased students that then impacts the course SET negatively.

E. How course characteristics affect SETs.

1. Electives. Students tend to rate courses in their major field and elective courses higher than required courses outside their major (Marsh and Dunkin, 1992; Marsh and Roche, 1997; McKeachie, 1997).

2. Class size and discipline. Instructors of very small classes tend to receive higher ratings (e.g., <40). Gifford's (2007) findings suggest that crowding adds environmental stress and thus negative outcomes in a classroom and on SETs. Furthermore, findings show that quantitative courses are more negatively rated than non-quantitative courses because of student interest (Uttl et al., 2013). Humanities instructors tend to receive higher ratings than instructors in the physical sciences, with social and behavioral sciences in between (d'Apollonia and Abrami, 1997; Marsh and Dunkin, 1992; Marsh and Roche, 1997; Ory, 2001; Monks & Schmidt, 2010).

3. Innovation. In my experience, students negatively associate innovation in teaching as beneficial to their course experience. New or revised courses frequently get lower-than-expected ratings the first time out, and if the pedagogical approach used is novel to the students, the ratings will continue to be poor in subsequent iterations. I have worked with instructors teaching two sections of the same course, and conduct one in a traditional lecture format, and the other using active learning. With all other things being equal, the active learning course section is rated systematically lower. Additionally, despite evidence of increased learning gains, students report that they learned less in an active learning course versus a lecture course (Lake, 2001). This negatively impacted student ratings of course and instructor effectiveness, and open-ended responses from studies and in my review of SETs commonly share the incorrect

belief that: “The instructor did not teach me anything.” An over-reliance on SETs as a measure of faculty performance in teaching serves to deter pedagogical improvement and innovation.

F. Effect of subject area on SETs.

1. In certain subject and topical areas, it is necessary to address sensitive, challenging, and controversial topics. In these courses, I have worked with countless high quality instructors (those whose students have performed at high levels, and have garnered high SET ratings in non-controversial topical courses) who receive poor SETs. The explanation stems from the comments, which often speak to a student’s anger or resistance to confronting diverse perspectives and viewpoints. Courses addressing these kinds of topics may include discussions about anything from Evolution to Race to Society and more. Williams and Ceci (1997) and Schueler (1988) both found that professors may feel inhibited from discussing controversial ideas or challenging students' beliefs, for fear that some students will express their disagreement through the course evaluation form. SETs have been described as "opinion polls," suggesting that SET require professors to seek to avoid giving offense and putting style before substance if high student ratings are the aim.

G. Reliability and timing of SET administration.

1. Since grades and grade expectations have a strong influence on student experience, timing of the administration of the SET can significantly impact the outcome. The stress and anxiety level of the student fluctuates throughout the semester, usually at its highest around/during the final exam. Does the student feel prepared for the exam and therefore less anxious as they complete the SET immediately prior? This may prompt a more positive evaluation. Is the student lacking confidence with high anxiety while completing the SET after feeling as if they struggled on the exam? This could result in lower SET ratings.

2. The mean averages for all questions tend to be lower when administered late in the semester. Witt and Burdalski (2003) found that evaluations administered on the last day of the semester were lower than at the 11-week mark of a 14-week semester despite self-report data from the student sample that stated opinions of instructor effectiveness were the same or higher than when tested previously. Other findings come from Aleamoni (1981) and Braskamp et al. (1984), who suggest that evaluation results may be affected if administered before or after an examination. Braskamp et al. (1984) reported that student ratings collected during the final examination are lower than ratings collected during the semester and recommend administering the student evaluation instrument during the final two weeks of the semester.

H. Reliability of anonymous responses to open-ended questions.

1. Anonymous responses to open-ended SET questions should be a source of information which provides additional context to the numerical ratings, but are as reliable as SET's overall in terms of their inherent biases. In other words, they are not reliable as an indicator of teaching effectiveness and good pedagogy. The fact that the comments are anonymous further hinders their reliability because of how that anonymity enables students to provide information that raise concerns about their own credibility as a source of information. This is exacerbated in evaluations administered online where students appear to feel even more anonymous without identifiable handwriting, and we are seeing the emergence of "trolling" in online evaluations.

2. For example, in a course offered at UC Berkeley this past spring in the Social Sciences, a student responded anonymously that, "The only strength she [the professor] has is she's attractive, and the only reason why my review is 4/7 instead of 3/7 is because I like the subject." Neither of these sentiments has anything at all to do with the teacher's effectiveness or the course quality, and instead reflect gender bias and sexism.

3. Comments can be extremely valuable, helpful and informative to provide additional context to a summative review of teaching, but are not measures to rely upon as there is no appropriate way to use the qualitative data and account for the inherent biases. They are a source of subjective impressions of an experience. They help paint a picture and add depth to the numerical scores, at times. But even then, I find it common to see internally conflicting SETs where rating numbers are high and open-ended responses negative, and vice versa - even on matching pair question items from likert to open-ended. Comments can also illuminate where bias explicitly taints a student's rating. As a result, certain SETs could be disqualified from being included in reporting based on explicitly biased comments and therefore ratings (e.g., see comment in #2 above).

4. Not surprisingly, signed ratings are more positive than anonymous ratings (Blunt, 1991; Braskamp & Ory, 1994; Centra, 1993). This may be due to fear of retribution (Feldman, 1979).

I. Given the limitations of SETs, there are better ways to assess teaching effectiveness.

1. Agreed by both proponents and opponents of SETs, no single source or method of teaching evaluation should be used on its own. The consensus is that a teaching dossier is the ideal tool for assessing teaching effectiveness, incorporating SETs as part of a larger composite of one's teaching. Davis (2009)

explains that while student rating forms are administered at the end of the term to survey students' opinions about a course, a substantial body of research has concluded that administering well-crafted questionnaires to students makes sense as only one source of information for evaluating teaching. Research has also shown that reviewing end-of-course questionnaires alone tends not to help instructors improve their teaching (Hampton and Reiser, 2004; Kember et al., 2002; Marincovich, 1999; Nasser and Fresko, 2002; Schmelkin et al., 1997).

2. Evaluation of teaching plays a significant role in decisions regarding advancement and promotion. Thus, it is imperative that clear documentation of teaching ability and teaching contribution be included in advancement and promotion cases. Research clearly demonstrates that while any single source of reliable information about a faculty member's teaching is valuable, SETs alone do not provide a complete or reliable picture for evaluation (Davis, 2009). Indeed, for the reasons already described in this report, SETs can be biased and misleading. Therefore, multiple sources provide complementary perspectives on various aspects of teaching and together comprise a more comprehensive and accurate portrait of teaching as a complex and scholarly activity.

3. Teaching dossiers provide documented evidence of teaching and context for that evidence, drawn from a variety of sources. Dossiers provide the opportunity to evaluate teaching longitudinally, situating teaching as an ongoing process of inquiry, experimentation, and reflection (*e.g.,* Braskamp and Ory, 1994; Murray, 1997; O'Neil and Wright, 1995; Mues and Sorcinelli, 2000; Knapper and Wright, 2001; Seldin, et al., 2010). A teaching dossier would typically include the following items:
    a) Departmental letter summarizing the candidate's teaching - An effective letter will describe departmental teaching evaluation procedures, the nature and quality of a candidate's teaching, and the evidence upon which this evaluation is based;
    b) Candidate's statement - It is helpful if candidates provide a written statement of their teaching approach, including the goals of specific courses and choices of teaching strategies, along with their efforts to improve instruction and respond to criticisms of their teaching performance made by students on end-of-course evaluations;
    c) Description of courses taught - A list of courses and enrollment should be included; Description of student research directed - Candidates may want to describe their role in directing senior theses, masters and doctoral studies, and postdoctoral scholars;
    d) Peer evaluation - Reports or letters about the candidate's teaching performance from faculty colleagues familiar with the content could be included in the dossier while the letters should cite the basis and

evidence for judgments made (observation, review of instructional materials, and so on);

e) Student ratings - Student rating data (distributions and response rates; no averages) for each different course taught in the period under review should be presented. In addition, the dossier could include letters from current students or summaries of interviews;

f) Alumni evaluation - Former students, as well as teaching assistants, can provide information about a candidate's teaching performance in the form of group interviews, or summaries of surveys of alumni that specifically address the candidate's teaching.

4. Colleges and universities have found the dossier to provide excellent documentation for both formative and summative purposes (Edgerton et al, 1991). Root (1987) conducted one of the few studies that investigated colleagues' evaluations of teaching dossiers and concluded that a committee of colleagues could provide sufficiently reliable assessments of a complete dossier. The dossiers included course outlines, syllabi, teaching materials, student evaluations, and curriculum development documentation—much of what is generally prescribed for a teaching dossier with the exception of teacher reflections and evidence of student learning. Ultimately, the best way to get a valid summative evaluation of teaching is to base it on a dossier containing data from multiple sources—ratings from students, peers, administrators, self-ratings, and learning outcomes—that reflect every aspect of teaching including course design, classroom instruction, assessment of learning, advising, and mentoring (*e.g.,* Weimer, et al., 1988; Chism, 1999; Hoyt & Pallett, 1999; National Research Council, 2003).

5. Seldin (1993), who has done some of the seminal research work on teaching dossiers, notes that professors "stand to benefit by providing tenure and promotion committees with their teaching dossiers. It provides evaluators with hard-to-ignore information on what they do in the classroom and why they do it. And by so doing, it avoids looking at teaching performance as a derivative of student ratings" (p. 8).

6. There are some common pitfalls in the evaluation of dossiers for personnel decisions. Here they are offered as two don'ts:

   a) *Don't assume that everyone must teach in the same way*. It is better to allow individual differences in teaching styles and techniques as long as they can be tolerated by department and institutional goals. In general, it is best to develop criteria within the smallest practical unit: the department level (Seldin et al., 2010).

   b) *Don't assume that standards and ratings will be the same across academic disciplines*. Standards and ratings tend to fluctuate--sometimes

wildly and even unfairly. The same variation in standards and rating exists in all methods used to evaluate teaching. This is a very strong argument for the institution of a teaching dossier which allows a more comprehensive evaluation of teaching performance. Although popular and extensively used, appraisals of teaching based almost exclusively on student ratings is hardly the answer. It is better to install a teaching dossier program that has the advantage of documenting both the complexity and individuality of teaching and then refine the process of dossier evaluation so that it is accurate, fair, and complete (Seldin et al., 2010).

7. In terms of how to weight SETs and other dossier items for the purpose of evaluation, see Appendix 2 for an example of a teaching dossier evaluation form that helps to articulate what aspects of teaching should be considered in an evaluation, and how different sources, including SETs, could reasonably contribute to it. What we know about SETs is that they give us insight into the student experience, and this could reliably inform questions about whether classes are met on time, missed classes made up, and trends in student experience across courses. However, one should note the driving questions to prompt the review of various aspects of teaching in Appendix 2, and that most aspects of course design (e.g., whether materials and course content are appropriate for the course level, and if they represent the best work in the field), teaching methodologies (e.g., evidence of meaningful curricular development), content knowledge (e.g., currency of teaching materials, and whether they represent the best work in the field), student learning (e.g., if the grading philosophy is appropriate for the course/s taught, and evidence of real cognitive or affective student learning), and departmental responsibility (e.g., whether the faculty member seeks feedback about teaching performance, explores alternative teaching methods, and makes changes to increase student learning) cannot reasonably be addressed, or even informed, by SETs.

J. What to provide readers of SET ratings when they are interpreting SETs.

1. Distributions of SET scores should be reported, along with response rates. Response rates are important because the lower the response rate, the less representative the responses might be: there is no reason nonresponders should be like responders. Nonresponse produces uncertainty about the data, and the lower the response rate, the greater the uncertainty. Distributions are important because they can reveal insights into the student experience that can be obscured by diluting reporting to ratings' averages. For example, there is a difference between the instructor who receives mostly mid-range scores indicating a common, mediocre student experience versus a bipolar distribution where students seem to either love or hate the course/instructor. This is very important information to be

considered when interpreting SET data, particularly in relation to the nature of the course (e.g., a course covering sensitive or challenging subjects may elicit negative student experience regardless of teaching effectiveness, which a bipolar distribution would help illuminate in context).

2. Averages of scores should not reported. Personnel reviews routinely compare instructors' average scores to departmental averages. Such comparisons make no sense. They presume that the difference between a 3 and 4 means the same thing as the difference between a 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 (McCullough and Radson, 2011). 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. Even if averaging made sense, the mere fact that one instructor's average rating is above or below the departmental average says little. The distribution of scores for instructors and for courses should be reported, along with the percentage of ratings in each numerical category.

K. The standardization of SET questions.

1. The common use of SET by means of administering standard questionnaires to be completed (in most cases, anonymously) by all students across disciplines in a college/university is very problematic. Administering SET in a standardized way across an institution depersonalizes and ignores the complexity of teaching. It also ignores the contexts previously discussed in this report that affect ratings (e.g., qualitative versus quantitative course, required versus elective course). Instead, it asserts that everyone must teach in the same way to be rated well on standardized items, and that all instructors have equal opportunity to garner high ratings regardless of context. It is misleading to standardize SET across an institution because of what it will necessarily further obscure (e.g., biases that affect ratings based on course subject, type, format, level, content, etc.), and that the obscuring penalizes and ultimately discourages pedagogical experimentation and innovation - both things we want to encourage in order to promote increased student learning. This is why Seldin (2010) recommends that criteria for evaluation, and therefore the aligned SET questions for that criteria, be formulated at the department-level, and not beyond it.

2. The driving purpose of standardizing SET questions across any institution is to make comparisons of teaching effectiveness easier in merit and promotion decisions. The concept makes sense, but under scrutiny does not live up to the promise. The desire to compare teachers using a single source of ratings and quantifiable number/s does not align with actual teaching practice.

3. First, teaching is a highly complex activity and the result is that teaching can be a lot of different things while being equally effective (and in many cases students still rate the experience poorly). To ask the same exact questions across large swaths of a campus, disciplines, etc. invites the removal of any context to the analysis of the results. For example, the faculty member teaching a course with critical subjects, without any additional context from an SET will invariably score lower than an elective major course - even if all other things are equal. Unique questions about how the course impacted one's stance on an issue/topic, or how well the instructor addressed diverse viewpoints and included multiple perspectives in the classroom, make sense in these courses, but not others. Using SETs to compare faculty teaching effectiveness and performance does not make sense.

4. Second, the desire to compare using SETs is a flawed approach. SETs have both practical and statistical issues that arise from trying to compare instructors and standardizing questions only exacerbates those issues. SETs were never designed as a tool to inform comparison, but as a way to capture student experience. There is no reason to think that we should be able to easily and quickly compare an instructor's teaching effectiveness across different course levels, types, sizes, or disciplines. There are simply too many confounding, contextual, and unique factors at play that are lost by over-reliance on SETs, and particularly if they are standardized to only capture certain aspects privileged by an evaluating unit for the purpose of easy, quantifiable evaluation - which is itself a flawed proposition. With all of the confounding factors that exist within SETs, it does not make sense to assume that the same exact questions would be interpreted or answered the same across levels, sizes, types and institutions. The issue of who is composing the questions also raises concerns. Who gets to determine the questions? The nature of the questions make explicit a view of what teaching is and should be. Who gets to name it? And in that naming, students and faculty would have a justifiably wide variation in answering what teaching is and should be - making standardized questions unhelpful.

5. Additionally, students' interest in courses varies by course type (e.g., prerequisite or major elective). The nature of the interaction between students and faculty varies with the type and size of courses. These variations are large and may be confounded with SET (Cranton and Smith, 1986; Feldman, 1978; Feldman, 1984). Ultimately, it is not possible to make fair comparisons of SET across

seminars, studios, labs, prerequisites, large lower-division courses, required major courses, etc. (McKeachie, 1997), let alone across an institution or institutions.

References

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*, 153–166.

Aleamoni, L. M. (1981). Student ratings of instruction. *Handbook of teacher evaluation*, *110*, 145.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology*, *64*(3), 431..

Arbuckle, J., & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles*, *49*(9-10), 507-516.

Babad, E., Darley, J. M., & Kaplowitz, H. (1999). Developmental aspects in students' course selection. *Journal of Educational Psychology, 91,* 157– 168.

Basow, S., Codos, S., & Martin, J. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student Journal*, *47*(2), 352-363.

Beecham, R. (2009). Teaching quality and student satisfaction: nexus or simulacrum?. *London Review of Education*, *7*(2), 135-146.

Bianchini, S., Lissoni, F., & Pezzoni, M. (2013). Instructor characteristics and students' evaluation of teaching effectiveness: evidence from an Italian engineering school. *European Journal of Engineering Education*, *38*(1), 38-57.

Blunt, A. (1991). The effects of anonymity and manipulated grades on student ratings of instructors. *Community College Review,* 18, pp. 48-54.

Boatright-Horowitz, S., & Soeung, S. (2009). Teaching white privilege to white students can mean saying good-bye to positive student evaluations. *American Psychologist, 64,* 574–575.

Boring, A. (2015). Gender biases in student evaluations of teachers. *Document de travail OFCE*, *13*.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing Faculty Work: Enhancing Individual and Institutional Performance. Jossey-Bass Higher and Adult Education Series*. Jossey-Bass Inc., 350 Sansome Street, San Francisco, CA 94104.

Braskamp, L. A., Brandenburg, D. C., & Ory, J. C. (1984). *Evaluating teaching effectiveness: A practical guide*. Corwin.

Campbell, H., Gerdes, K., & Steiner, S. (2005). What's looks got to do with it? Instructor appearance and student evaluations of teaching. *Journal of Policy Analysis and Management*, *24*, 611–620.

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions*, 25-44.

Centra, J.A. (1993) *Reflective faculty evaluation.* San Francisco; Jossey-Bass.

Chism, N.V.N. (1999). *Peer Review of Teaching*, Bolton, MA, Anker Publishing.

Cranton, P. A., & Smith, R. A. (1986). A new look at the effect of course characteristics on student ratings of instruction. *American Educational Research Journal*, *23*(1), 117-128.

Davis, B. G. (2009). *Tools for teaching*. John Wiley & Sons.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American psychologist*, *52*(11), 1198.

Delucchi, M. (2000). Don't worry, be happy: Instructor likability, student perceptions of learning, and teacher ratings in upper-level sociology courses. *Teaching Sociology*, *28*, 220–231.

Dunegan, K. J., & Hrivnak, M. W. (2003). Characteristics of mindless teaching evaluations and the moderating effects of image compatibility. *Journal of Management Education*, *27*, 280–303.

Edgerton, R., Hutchings, P., and Quinlan, K. (1991). *The Teaching Portfolio: Capturing the Scholarship in Teaching*. American Association for Higher Education, Washington, D.C.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education*, *9*(3), 199-242.

Feldman, K.A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education,* 10, 149-172.

Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education*, *21*(1), 45-116..

Fortson, S. B., & Brown, W. E. (1998). Best and worst university instructors: The opinions of graduate students. *College Student Journal, 32,* 572– 576.

Gifford, R. (2007). *Environmental psychology: Principles and practice* (p. 372). Colville, WA: Optimal books.

Gurung, R., & Vespia, K. (2007). Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology*, *34*, 5–10.

Hampton, S. E., & Reiser, R. A. (2004). Effects of a theory-based feedback and consultation process on instruction and learning in college classrooms. *Research in Higher Education*, *45*(5), 497-527.

Hamermesch, D. S., & Parker, A. (2005). Beauty in the classroom: Instructor's pulchritude and putative pedagogical productivity. *Economics of Education Review*, *24*, 369–376.

Hoyt, D.P., & Pallett, W.H. (1999). "*Appraising teaching effectiveness: Beyond student ratings"*. IDEA Paper No. 36, Kansas State University Center for Faculty Evaluation and Development <*www.idea.ksu.edu*>.

Huston, T. A. (2005). Race and gender bias in higher education: Could faculty course evaluations impede further progress toward parity. *Seattle J. Soc. Just.*, *4*, 591.

Kember, D., Leung, D.Y., & Kwan, K. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching?. *Assessment & Evaluation in Higher Education*, *27*(5), 411-425.

Kim, C., Damewood, E., & Hodge, N. (2000). Professor attitude: Its effect on teaching evaluations. *Journal of Management Education*, *24*, 458–473.

Knapper, C., & Wright, W. A. (2001). Using portfolios to document good teaching: Premises, purposes, practices. *New Directions for Teaching and Learning*, *2001*(88), 19-29.

Lake, D. A. (2001). Student performance and perceptions of a lecture-based course compared with the same course utilizing group discussion. *Physical Therapy*, *81*(3), 896-902.

Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review*, *27*(4), 417-428.

Legg, A. M., & Wilson, J. H. (2012). RateMyProfessors. com offers biased evaluations. *Assessment & Evaluation in Higher Education*, *37*(1), 89-97.

Marincovich, M. (1999). Using student feedback to improve teaching. In P. Seldin and Associates (Eds.). *Changing practices in evaluating teaching: A practical guide to improve faculty performance and promotion/tenure decisions* (pp. 45-69). Bolton, MA: Anker.

Marsh, H. W., & Cooper, T. L. (1981). Prior subject interest, students' evaluations, and instructional effectiveness. *Multivariate Behavioral Research*, *16*(1), 83-104.

Marsh, H.W., and Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J.C. Smart (Ed.), *Higher Education: Handbook of Theory and Research* (Vol. 8 pp. 143-233) New York: Agathon.

Marsh, H.W., & Roche, L.A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*(11), 1187.

McCullough, B. D., & Radson, D. (2011). Analysing student evaluations of teaching: Comparing means and proportions. *Evaluation & Research in Education*, *24*(3), 183-202.

McKeachie, W.J., (1997). Student ratings: The validity of use. *Am Psychol., 52*(11), 1218–1225.

McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly, 88,* 868– 881.

Monks, J. & Schmidt, R. (2010). *The impact of class size and number of students on outcomes in higher education* [Electronic version]. Retrieved [9/29/2016], from Cornell University, School of Industrial and Labor Relations site: http://digitalcommons.ilr.cornell.edu/workingpapers/114/

Mues, F., & Sorcinelli, M. D. (2000). Preparing a teaching portfolio. *Amherst Mass.: University of Massachusetts, The Centre for Teaching*.

Murray, J. P. (1997). *Successful Faculty Development and Evaluation: The Complete Teaching Portfolio. ASHE-ERIC Higher Education Report No. 8, 1995*. ERIC Clearinghouse on Higher Education, Graduate School of Education and Human Development, George Washington University, One Dupont Circle, Suite 630, Washington, DC 10036-1183.

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, *27*(2), 187-198.

National Research Council. (2003). *Evaluating and improving undergraduate teaching in Science, Technology, Engineering, and Mathematics*. Washington, DC: National Academies Press.

Ogier, J. (2005). Evaluating the effect of a lecturer's language background on a student rating of teaching form. *Assessment & Evaluation in Higher Education, 30,* 477–488.

O'Neil, C., & Wright, A. (1995). *Recording teaching accomplishment: A Dalhousie guide to the teaching dossier*. Halifax, Canada: Office of Instructional Development and Technology, Dalhousie University.

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. *New directions for teaching and learning*, *2001*(87), 3-15.

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors. Com. *Journal of Diversity in Higher Education*, *3*(3), 137.

Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *Journal of General Psychology*, *133*, 19–35.

Root, L.S. (1987). Faculty Evaluation: Reliability of Peer Assessments of Research, Teaching, and Service. *Research in Higher Education, 26*, 71–84.

Schmelkin, L. P., Spencer, K. J., & Gellman, E. S. (1997). Faculty perspectives on course and teacher evaluations. *Research in Higher Education*, *38*(5), 575-592.

Schueler, G. F. (1988). The evaluation of teaching in philosophy. *Teaching Philosophy*, *11*(4), 345-348.

Scriven, Michael (1995). Student ratings offer useful input to teacher evaluations. *Practical Assessment, Research & Evaluation*, 4(7).

Seldin, P., Miller, J.E., & Seldin, C.A. (2010). The teaching portfolio: A practical guide to improved

performance and promotion/tenure decisions (4th Ed.). San Francisco, CA: Jossey-Bass.
Seldin, P. (1993). *Successful Use of Teaching Portfolios*. Anker Publishing Co.
Short, H., Boyle, R., Braithwaite, R., Brookes, M., Mustard, J., & Saundage, D. (2008, July). A comparison of student evaluation of teaching with student performance. In *OZCOTS 2008: Proceedings of the 6th Australian Conference on Teaching Statistics* (pp. 1-10). OZCOTS.
Smith, B. P. (2007). Student ratings of teacher effectiveness: An analysis of end-of-course faculty evaluations. *College Student Journal, 41,* 788– 800.
Smith, G., & Anderson, K. J. (2005). Students' ratings of professors: The teaching style contingency for Latino/a professors. *Journal of Latinos and Education, 4,* 115–136.
Spooren, P., Brock, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching the state of the art. *Review of Educational Research*, *83*(4), 598-642.
Storage, D., Horne, Z., Cimpian, A., & Leslie, S. J. (2016). The Frequency of "Brilliant" and "Genius" in Teaching Evaluations Predicts the Representation of Women and African Americans across Fields. *PloS one*, *11*(3), e0150194.
Subtirelu, N. C. (2015). "She does have an accent but…": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society*, *44*(01), 35-62.
Ottoboni, K., Boring, A., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*.
Theall, M., & Franklin, J. (Eds.). (1990). *Student ratings of instruction: Issues for improving practice* (No. 43). Jossey-Bass Inc Pub.
Tom, G., Tong, S. T., & Hesse, C. (2010). Thick slice and thin slice teaching evaluations. *Social Psychology of Education*, *13*, 129–136.
Uttl, B., White, C. A., & Morin, A. (2013). The numbers tell it all: students don't like numbers!. *PloS one*, *8*(12), e83443.
Uttl, B., White, C., & Gonzalez, D. (2016). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, http://dx.doi.org/10.1016/j.stueduc.2016.08.007
Weimer, M., Parrett, J.L., & Kerns, M. (1988). *How am I teaching?* Madison, WI: Magna Publications.
Wendorf, C. A., & Alexander, S. (2005). The influence of individual- and class-level fairness-related perceptions on student satisfaction. *Contemporary Educational Psychology*, *30*, 190–206.
Williams, W. M., & Ceci, S. J. (1997). "How'm I doing?" Problems with student ratings of instructors and courses. *Change: the magazine of higher learning*, *29*(5), 12-23.
Witt, J., & Burdalski, D. (2003). Regarding the Timing of Student Course/Instructor Evaluations. *Journal of the Academy of Business Education. Proceedings Issue. Available at http://www. abe.villanova.edu/proc2003/witt.pdf.*
Wolbring, T., & Riordan, P. (2016). How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of teaching. *Social science research*, *57*, 253-272.
Worthington, A. C. (2002). The impact of student perceptions and characteristics on teaching evaluations: a case study in finance education. *Assessment & Evaluation in Higher Education*, *27*(1), 49-64.

APPENDIX 2: TEACHING DOSSIER EVALUATION FORM

Adapted from Seldin, 2010

PART I: COMPOSITE EVALUATION

Teaching Dossier Component- Suggested Focus in Examining Materials
Course Design ● Are materials and course content appropriate for the course level? ● Do they represent the best work in the field? ● Are they appropriately challenging? ● What level of performance do the students achieve? ● Do course requirements appropriately address critical thinking development? Writing skill development? ● Are the teaching materials consistent with the course's expected contribution to the department curriculum?
Teaching Methodologies ● How do this faculty member's student ratings compare with others teaching similar courses? ● What trends are apparent across courses? ● What are this faculty member's teaching strengths? Weaknesses? ● Is there evidence of teaching improvement over time? ● Is there evidence of meaningful curricular development? ● Does the faculty member engage in team teaching? Interdisciplinary teaching?
Content Knowledge ● Are the teaching materials current? ● Is the best work in the field represented? ● Is the faculty member sought out as a resource in the discipline area by peers or students? ● Does he or she seek opportunities to learn more about the subject? ● Is there evidence that the professor uses expertise in settings outside the department? ● Does the faculty member actively involve students in scholarship?
Student Learning ● Is the grading philosophy appropriate for the courses taught? ● How suitable is the professor's grade distribution? ● Is there evidence of real cognitive or affective student learning? ● Are the professor's comments on student work appropriate? Thorough? Motivating? ● Is there evidence of assistance provided by the professor to students who are preparing publications or conference presentations? ● Do student essays, creative work, or fieldwork reports indicate deep, reflective thinking and

learning?
Departmental Responsibility ● Is this faculty member a "good citizen" with regard to teaching responsibilities? ● Are classes met on time? Missed classes made up? ● Does the professor instruct an appropriate number of students? ● Does he or she take an active role in the improvement of instruction in the department? ● Does the faculty member seek feedback about teaching performance, explore alternative teaching methods, make changes to increase student learning? ● Does he or she make an appropriate contribution as a student advisor?

PART II: COMMENTS AND OVERALL EVALUATION

Please comment here on your overall evaluation of this faculty member as a teacher.