MUFA ad-hoc committee on student evaluation of teaching
Report
November 25, 2019

Composition of the Committee:
Michel Grignon (Chair, MUFA, Social sciences), Joe Kim (Faculty, Science), Lynn Martin (Faculty, Health Sciences), Kris Knorr (MacPherson Institute), Tiffanny van Lieshout (GSA, Science), Stephanie Bertolo (MSU, VP Education, Art & Science), Tasneem Warwani (MSU, AVP University, Art & Science).

Summary and recommendations:

Based on responses to a survey sent to Deans, interviews with stakeholders (at McMaster and on other campus) as well as a review of the literature, we found the following:

1. Students at McMaster fill out a relatively long Student Evaluation of Teaching (SET) questionnaire at the end of each course yet administrators use only one question (the one on the effectiveness of the instructor) to evaluate the instructor's performance. There is a feeling among students (and instructors) that the feedback from these questionnaires is not used, making the exercise useless.

2. Response rates for SETs are approximately 20%[1] which may be explained by the lack of confidence students have in how the information is being used. This low participation rate renders ratings meaningless in most cases (administrators in Faculties think that participation should be at least 60%, and preferably 70% for SET scores to be of help)

3. No simple way to increase participation rates was identified, and administrators are reluctant to allow incentives for students to participate (at least with grades). A concern is that students may trade participation or even good ratings for good grades in the course. More broadly, there is a sense that framing student evaluation as a way to reward or penalize the instructor raises issues of manipulation (by instructors, attempting to secure favourable scores with lenient grading or easy content; by students, attempting to secure favourable marks with the threat of low ratings). This introduces a negative dynamic in the teacher-learner relationship.

4. The literature is clear that no single question can accurately measure teaching effectiveness. There are good reasons to believe that responses to such a question will be biased (e.g., against women and minorities) and may lack validity in measuring teaching effectiveness with respect to learning outcomes (i.e., higher evaluation scores do not correlate with improved learning outcomes).

[1] This low level of participation also reflects the switch to online questionnaire (from paper, in class).

5. Questions on teaching effectiveness should never be summarized by their mean: ratings using Likert scales produce ordinal variables (rather than cardinal) and arithmetic operations on ordinal variables are meaningless. All that could be done is to measure the proportion of respondents who gave a rating above or below a threshold deemed to be meaningful (e.g., excellence or failure).

Our first conclusion is therefore that the current questionnaire used at McMaster lacks validity, relies in practice on the result of a single question (which may be biased and meaningless) and these perceptions may be a factor that leads to low participation in completion of the SET.

This is a critical opportunity to define a strategy for using student input on teaching in a way that serves the teaching and learning mission of the University. This must be based on an assessment by the McMaster community of what we see as high quality teaching and the strategy will be specific to McMaster (there is no ready-made tool or suite of instruments available on the market that would serve our purpose).

We can use student input about their learning experience for three different purposes:

1. Formative: to help instructors and programs provide better quality teaching suited to the needs of their students.

2. Summative: to help administrators assess the quality and effort provided by individual instructors (to provide support for faculty who may need it, and as a component of the assessment of teaching for tenure, permanence, and merit purposes).

3. Accountability: to help instructors, program managers, Deans and the VP Academic assess the experience of students at McMaster and respond to perceived weaknesses.

No single instrument can answer these three questions and we suggest three different approaches, one for each of these purposes:

1. Formative. Students can provide actionable feedback during the course (rather than at the end of the course), by commenting on their learning experiences and how it could be improved. At the mid-point of the term, instructors will administer a variant of a tool known as "Start/Stop/Continue" survey. In this easy-to-administer survey, students provide open-ended answers on what they think could be done to improve their learning experience in the second half of the course:
    a. Start: what can be done that is not currently done? (e.g., start posting lecture slides before class);
    b. Stop: what is currently done that hampers learning? (e.g., stop speaking so quickly and using complex terms);
    c. and Continue: what is currently done that positively adds to the learning experience? (e.g., demonstrations that involve members of the class).

The survey could start with one open-ended question on what the student thinks they are doing that contributes to their learning experience in the course (self-reflection), to clarify this is not about "evaluating" the instructor, but rather about perceptions of the learning experience. Students would fill out these surveys online anonymously (ideally through Avenue). Instructors would then provide feedback on the emerging trends, and adjustments to be made in the second half of the course. This report (including the response from the instructor) should be made public (including to students), but would not be used as such for evaluation purposes, except when included in the Report of Activity (SPS B1, item #6, evidence of incorporation of some form of formative evaluation of courses and response to concerns of students). When used, these short surveys usually have excellent participation rates.

2. Summative. Student feedback on their perception of their learning experience in the course is linked to the instructor's performance and effort:
    a. Although beyond the scope of the mandate for this committee, we believe that feedback from students on their learning experience and perception of the role of the instructor's effort and performance in that experience should be one of the components used by Chairs and Deans when they evaluate performance. Our committee was not mandated to weigh in on the relative weights of the different components of evaluation performance, we were simply of the opinion that student assessment of learning experience should not be given a weight of 0.
    b. We suggest that the subjective nature of this feedback should be recognized and that the method to collect such feedback should minimize the risk of biases. Thus, we recommend to stop using the current SET questionnaire at the end of the course or any quantitative (close-ended questions using Likert scale or any number) questionnaire for that purpose. In essence, ratings give the illusion of objectivity to a process which is inherently subjective: we try to capture what students perceive, which is subjective, knowing that everything objective can be captured otherwise (e.g., based on the course portfolio). Even more detrimental, the fact of using numbers, quantities, ratings on a scale is likely what induces the behaviours that we want to prevent, such as biases based on preconceived ideas (about gender roles or ethnicity) or manipulation (trading grades for ratings). Thus, we recommend replacing the current questionnaire, based on Likert scales, with a collection of qualitative feedback on learning experiences. There are different ways to conduct qualitative evaluations, either through open-ended questionnaires or focus groups. Our preferred version is the focus group: information on student learning experience in each course would be collected during a focus group[2] of designated students in the class (participation being mandatory) led by a team of trained evaluators. What we have in mind for trained evaluators is a pair, comprised of a trained undergraduate and an emeritus

[2] Note that focus groups do not have to be in person but could be conducted online (using a tool such as Top Hat (https://tophat.com/), already used at McMaster. This tool would provide anonymity to participants.

professor or a volunteer from the community. The professor or volunteer would moderate and provide an "external" perspective, while the trained evaluator would be closer to the members of the focus group. As we see it, these student evaluators would be mostly undergraduate students, who would receive credits or payments for conducting the evaluations. Importantly, these focus groups should be led according to a document provided by each program (or Faculty?) stating what teaching excellence means in the context of this particular program. The document could be based on the IQAP reviews. Each session would result in a report, written by the evaluation team and shared with the instructor for feedback. The report and feedback would then be shared with the administrator, and summarized in three broad categories (excellent, good, problematic) and would be one of the seven dimensions used to assess instructors in a given year (CP/M) or for T&P purposes. The detailed report could be used by Chairs and Deans to provide resources to help instructors improve student learning experience in their courses. We acknowledge this approach is exploratory, and McMaster would be a a leading innovator if it followed that path. As a result, there is not much evidence on whether qualitative feedback collected through focus groups is indeed less biased than quantitative feedback collected via close-ended questionnaires, or to support the feasibility or acceptability of such a qualitative approach to summative evaluation. Such a tool raises logistical issues and it would be unrealistic to produce such a report on all course taught in one academic year. We propose in this report a rotating formula where courses are evaluated once every two to three years, as a pilot and we recommend running studies (e.g., focus groups in some courses versus open-ended questionnaires in other courses) and documenting what works and what does not.

3. Accountability. Student feedback can be useful in capturing their overall learning experience in their program. This is where we see the value of a close-ended questionnaire on the perceptions of students but programs could also run town-hall meetings with graduates (as some currently do for IQAP purposes) and collect qualitative information on these perceptions. It has to be made clear that the goal is not to reward or punish a given instructor, but rather to help the program adjust and improve. The level at which such an evaluation should be conducted (course, program, Faculty, University) can be discussed. We recommend conducting it at the program level, and including questions on the organization of the program, rather than taking a narrow focus on the quality of individual courses. We also recommend asking graduates, or graduating students, to complete a questionnaire similar to the Course Evaluation Questionnaire (CEQ) used in Australia. McMaster Office of Institutional Research and Analysis (IRA) has experience surveying graduands at convocation (response rate between 30 and 40%) as well as drop-outs (response rate of 10%). IRA is also involved in the analysis of raw data from a survey of our graduates through the Ontario University Graduate Survey (OUGS), which surveys graduates at six months and two years, with a participation rate of 40%. For now, OUGS focuses on employment but questions could be added on program accountability (OUGS is administered by the Ministry but the Council of Ontario Universities can

influence the questionnaire). This is also a way for programs to stay connected with their graduates. A key question is that of participation: it is less crucial than when SETs are used to evaluate an individual instructor for CP/M or T&P purposes, but it may still be an issue if graduates who respond are of a certain type only (e.g., those who live in the GTHA, or those who benefited from their degree and have a good job). Response rates of 30 to 40% at a program level would guarantee sample sizes that allow for statistical inference. The only issue would be selection biases (if taking the survey is systematically linked to relevant opinions regarding the program) but this can tested by comparing those who respond to the whole population of graduands and/or graduates (e.g., by gender, age, grades, GPA, type of student etc.) We discussed incentivizing responses, perhaps through a frame for their degree or some recognition.

We recommend using the following questions if at the course level:

a. McGill Q2: I learned a lot from this course
b. UWO1: The instructor motivates me
c. UWO2: Overall, I had a great experience in that course

However, at the program level, CEQ is a full questionnaire with some validation.

We recommend running experiments on and statistical analysis of each of these three tools over time, such as, for instance submitting subjects to different questionnaires applied to the same material (taped lecture or portfolio), or various ways to implement the qualitative summative assessment, or analysis of response rates to “stop-continue-start” formative questionnaires.

Report:

1. What is done at McMaster today:
This section is based on a meeting with John Bell, who administers the questionnaires and a survey sent to Deans of all Faculties (responses received from Engineering, Health Science, Science and Social sciences).

Students must fill out a questionnaire (SET) toward the end of each course, with a series of (approximately 20) questions on the course and the instructor(s). Quantitative questions are usually on a scale of 1 (worst) to 10 (best), to the exception of questionnaires used in the Faculty of Health Sciences, that use a scale of 1 to 7. The data are entered online by students and collected and de-identified (but not anonymized) by John Bell, who then calculates and sends statistical information as well as verbatim qualitative comments to Department and Program Chairs. If instructors opt-in, students have access to the distribution of scores plus statistical information (mean, median, standard deviation) on the first question in the questionnaire (overall teaching effectiveness). Not all instructors opt-in (some opt-out and some do not tell whether they opt-in or out, which amounts to not making the information available to students). Chairs use the mean of the question on overall effectiveness as one component of the evaluation of teaching in annual CP/M decisions and in T&P decisions: according to the Deans, no other question on the questionnaire is used for these annual merit assessments. Means are sometimes adjusted (for level mostly) and compared either to a departmental average or to pre-established thresholds, and Chairs tend to use it as a categorical variable (below, equal, or above departmental average or threshold). The practise in the Faculty of Social sciences is to average scores over several years and to dismiss outliers in the distribution of scores (not individual outliers to the distribution of ratings in one course, but outliers in the "distribution" of mean scores for a single instructor over 5 years.) Instructors who want to innovate on pedagogy and take risks can discuss with their Chair and have the result of their evaluation waived on that particular course on which they want to innovate. Instructors are sent the full report prepared by John Bell for all the courses they teach (distributions of ratings on each quantitative question, plus statistical information, and verbatim open-ended comments after editing by the Chair in case of offensive comments – the latter being optional.)

Response rates are low on average (around 20%) and can be very low on some courses and it is deemed that, with such low response rates, SET scores are almost meaningless. What we heard about response rates was a real concern about the proportion of students who respond, not about the number (a 100% response rate in a class of 10 is still a small number) or the selection process (who are the students who respond?) Deans or Chairs are reluctant to allow instructors to incentivize responses to SETs in their courses, and very reluctant to the idea of incentives based on marks, for fear of launching a dynamic of quid-pro-quo around SETs (lenient marking or easy content in exchange for positive ratings). No Faculty or instructor at McMaster (as far as we know) has implemented a policy of advanced access to grades for students who complete their SETs (at Dauphine University, in Paris, France, as described below, students cannot see their final marks or transcripts until they have completed all their SETs). Students representatives on the committee were quite clear that these low response rates are the result of disillusionment on

the part of students, who do not feel their answers are used (recall they can only access to statistics on one question, and for some courses only) and that they are asked to provide many answers that are not even commented or analyzed.

Students are also asked for inputs in two other ways:

- Questionnaires at the university level, on student experience: the National Survey on Student Experience (NSSE), an international initiative to which McMaster takes part, and Canadian University Survey Consortium (CUSC).
    - The NSSE has close to 90 questions, on various aspects of student experience. It is sent to all 1st and 4th year students and the response rate is close to 25% (lower than the average in Ontario, at about 35%). Some sections (4, 5, 10, 13, 14, and 17) are directly related to teaching or instructors and McMaster publishes a statistical report but it is not clear whether any recommendations are derived from the report.
    - The CUSC is also a long questionnaire asked either to first year, middle years or graduating students (alternate years) in approximately 30 universities in Canada. The response rate is 29% (graduating students, 2018) but 18% only at McMaster. Sections on experience/satisfaction are 6.1 (perception of professors), 6.3 (satisfaction with quality of teaching), 7.0 (contribution to skills), and 8 (overall satisfaction, including value for money).
- Formative questionnaires (stop/start/continue) in some courses, at the initiative of the instructor (and for all instructors in the School of Business, according to information communicated during the focus groups).

2. What the literature tells us about the validity and quality of SET:

From now on, we use SET to mean "close-ended questionnaires with quantitative (Likert scales) questions on effectiveness of the instructor" and reserve the term "feedback on learning experience" for what we want to implement.
Until very recently, conventional wisdom (including published academic wisdom) was that questions on Student Evaluation of Teaching (SET), more precisely questions on overall effectiveness or quality of the course or instructor, were a decent indicator of the effectiveness of the instructor teaching a particular course and, as a result, could be used to assess the performance of an instructor either on a particular course or for their average performance in a given academic year. This is not to say that there were no dissenting voices and evidence to the contrary but, overall, reviews (Cohen 1981, 1982, 1983, Dowell and Neal, 1982, McCallum, 1984, Feldman 1989 and Clayson 2009) and reports based on the literature (e.g., Gravestock and Gregor-Greenleaf, 2008, for the Higher Education Quality Council of Ontario), while acknowledging imperfections, were of the opinion that SET could be used to infer something on the effort an instructor put into a course, or on the quality of what they were doing. The idea that such inference was robust enough to be used by administrators for hiring, firing and promoting was still controversial, and even proponents of it were quick to acknowledge it could not be the sole basis for making such high-stake decisions. However, the perception was that SET could be part of the summative evaluation of faculty.

Things changed dramatically around 2013 (a good description of this change can be found in Hornstein, 2017). A review by Spooren et al. (2013), though not comprehensive (see below), was the first to go beyond the narrow definition of "validity" used in previous reviews (which were based on results to final exams in multi-section courses with quasi-random assignment) and to add evidence on potential biases[3] (an instrument can be valid on average but biased). As a result, it cast doubt on the conventional wisdom that SET were valid and unbiased and could be used to infer something about the quality of an instructor's performance. On the heels of that extensive review, two prominent academics, who had never before written on the topic but came with strong credentials in statistics and education, wrote a scathing critique of SET (Stark and Freishtat, 2014):

- Firstly, they stated that SET findings are ordinal, not cardinal measures[4] (on a Likert scale, the difference between a 3 and a 4 has no reason to be the same as the difference between a 4 and a 5, these scales rank preferences but do not say anything about intensity of preferences) and, therefore, the average of a SET is meaningless. It is therefore meaningless to compare the average SET for one instructor for one course to the average SET for all instructors in the same department on all the courses they taught to infer something on the relative performance of instructors in a department. All that should be done is to comment on the distribution of scores received by a course on the Likert scale. McCullough and Radson, 2011, recommend commenting on binary variables (e.g., the proportion of students who gave the lowest or highest score or rate on the Likert scale for a particular course) with cut-off points determined by the administrators, after consulting with the faculty: does the university want to identify faculty in need of help -- focus on the lowest rate -- or excellent instructors, or instructors who are doing okay and better (use a cut-off lower than the highest rate)? And they recommend the use of a five point Likert scale – strongly agree, agree, neutral, disagree, strongly disagree with a statement on the course being good quality.)

- Secondly, they dismissed all the evidence on validity published in reviews of multi-section courses, stating that the randomization was not credible or that marks at the end of the course did not reflect learning. Turning to evidence based on follow-on courses (two studies only, one in a US military academy, Carrell & West, 2008 and one in an

[3] There does not seem to be any systematic review of biases in SET, nor even of one specific type of bias such as, for instance, gender bias, since Freedman, 1993 (which found no gender bias, but does not correct for size effects, and mixes studies that control for confounders, such as class size or level and programs, and studies that do not – it might be worth replicating that meta-analysis.) Empirical studies of gender biases in SET have been plagued by one major issue: the difficulty to control for true quality of the instructor. We were able to find 40 studies of gender biases in SET and will conduct a review of their findings. Spooren et al. (2013) list only four studies of gender bias (Basow and Montgomery, 2005, Smith et al. 2007, McPherson et al. 2009 and McPherson and Too Jewell, 2007), two of which find pro-women bias, one pro-men and one finds no bias. However, we know there are more than four such studies of gender bias. Savonick and Davidson, 2017 provide an "annotated bibliography of important recent studies [of] gender bias in Academe" that we can use as a starting point for a review.

[4] Technically, a cardinal measure is a measure that is unique up to an affine transformation.

Italian university, Braga et al. 2011) they found a lack of any correlation between SET in the pre-requisite course and performance to the follow-on course.

- Lastly, they cited eight studies (Anderson & Miller 1997, Basow 1995, Cramer & Alexitch 2000, Marsh & Dunkin, 1992, Wachtel, 1998, Weinberg et al. 2007, and Worthington 2002, none of which are cited in Spooren et al. 2013) showing clear biases, mostly against women instructors.

At the same time or shortly after, two empirical studies were published that used strong quasi-experimental settings (multi-sections with strict randomization or follow-on settings) and showed a lack of correlation between SET scores and learning outcomes: Braga et al. 2014. which is the published version of Braga et al. 2011, cited by Stark and Freishtat, 2014, using a follow-on setting at Bocconi University in Milan, Italy, and Boring et al. 2016, using a multi-section environment at Dauphine University in Paris, France. Stroebe, 2016, reviewing empirical studies linking grades and grade expectations to ratings of instructors by students with the goal to test whether grade leniency could be used by instructors to get better ratings, found that most studies using the follow-on setting demonstrated a lack of correlation between rating and grades in upper-level courses, and, when a correlation could be detected it was negative (suggesting that instructors at lower levels "bought" better ratings with easier, less demanding teaching).

Moreover, two studies, using sound methodology (quasi-experimental settings), strongly suggested gender biases in SETs:

- MacNell et al. 2014, exploited the existence of online discussion groups in a 1st level course to randomly attribute a gender to instructors (two instructors, one man and one woman, and four groups, each instructor led two groups, one with their own gender/identity and one with the gender/identity of the other instructor.) There were no objective differences in what the instructors were doing (e.g., identical turnaround for students assignments) and the difference between what the instructor received as an identified "male" and identified "female" was 0.47 out of a five-point scale, higher for males.
- Boring et al. 2016, using random assignments to multiple sections in a course with a single, anonymously graded exam, showed that, controlling for teacher's effectiveness (grades in final exam), female instructors were less well rated by their students, again suggesting a strong gender bias in SET.

These publications created a dissonance between what reviews were saying (SET are valid to some extent and correlate with one measure of the quality of teaching) and what most recent evidence was demonstrating.

Uttl et al. (2017), re-analyzing the studies in all published reviews, found that these reviews suffered from a substantial size-effect bias: some of the correlations used in the meta-analyses were based on a small number of sections and those correlations based on small numbers were usually very large (up to implausible correlations of .8 and higher, known in the field as "voodoo

correlations"), likely driving the average correlation (across all studies). Once corrected for the size effect (number of sections per study), the average level of correlation is actually zero, and the idea that SET questionnaires could tell something about the quality of instruction was severely damaged.

If Spooren et al., 2013 was the end of the beginning, Uttl et al. 2017, was the beginning of the end for SET as they were practised and used by administrators to make personnel decisions and perceived by non specialists (outside of the cottage industry of education science on students evaluation): for instance Elizabeth Barre published a blog (Rice University) in 2015 that was overall positive about SET (stating they have their weaknesses but can still be used) but she updated it in 2018, citing Uttl (2017) as the rationale for changing her mind and, after all, giving SET an F for their lack of validity (Barre, 2015 and 2018). The Ryerson decision, based on that new set of evidence, confirms what is now the state of knowledge in the field, that a single question on teaching effectiveness is not a valid assessment of teaching quality and there are chances that it is a biased way to assess instructors' performance. Not only is it bad practise to calculate the average of the summative question on teacher effectiveness and compare it to averages of all other instructors in the Department, but even a careful examination of the distributions of scores across instructors might lead to a biased and invalid assessment of teaching performance (see Recommendation 1).

To say that Uttl et al. 2017 was the beginning of the end does not mean that the discussion is over and we can still find rear-guard skirmishes in the academic literature. For instance, Linse, 2017 is a recent attempt at salvaging SET after the turning point, and we will spend some time analyzing it to show how the debate has changed and what seems to be the current defence line of proponents of SET (some of whom appear to have financial or professional interest in promoting the use of SET). First, Linse 2017 claims that the "lack of validity" of SET does not really matter, since SET were never meant to measure the quality of teaching but only the experience of students in the class. Administrators are then supposed to magically infer quality of teaching from experience of students (combined with other measures), even though experience of students have nothing to do (according to that line of defence) with the quality of teaching. It seems important here to reiterate that an instrument that does not measure quality of teaching cannot contribute in any way shape or form to an overall assessment of it, and that combining it with other, potentially better sources, will do nothing but pollute these other sources. Another line of defence in Linse 2017 is to claim that bias, and especially gender bias is not a weakness of SET. It is part of the same line of defence that claims SET can still be used, "in combination with other", more valid measures. We already explained why we don't think that line of defence makes sense, but it is crucial for it that SET are only invalid, but not biased, because if they were biased, combining them with other measures would make the overall measure biased. The way Linse tries to dismiss evidence on bias is particularly revealing of a partisan, rather than scientific, approach:

- First, she claims that evidence on bias comes from two recent studies, Braga et al. 2014 and McNell et al. 2015, suggesting that all previous studies could not find any bias. This implication that the finding of bias is recent and limited to few studies is factually wrong

as we already explained. Moreover, it is hard to understand why Linse mentions Braga et al. 2014 as this study is not about bias, but about validity.

- Linse then claims that these two studies "have methodological issues and significantly overstate the case", based on a publication by Ryalls et al. (2016)

- Reading Ryalls et al., it appears that the two publications in question are Boring et al. 2016 and McNell et al. 2015. The publication by Ryalls et al. (2016) is a newsletter published by The Idea (ideaedu.org), an organization selling SET to universities, and the authors are employees of the firm.

- Ryalls et al. 2016, which is the only source in Linse 2017 to dismiss the idea of any gender bias in SETs, is itself a partisan attack on the two studies, far from standard academic analysis.
    - Ironically, given their own conflict of interest, they start by attacking the quality of the outlet of Boring et al. 2016 (it becomes even more ironic in hindsight, since Anne Boring published a slightly different version of the 2016 paper, but with similar conclusions, in the Journal of Public Economics in 2017.)

    - They criticize the method used on the basis that the 2016 publication does not provide enough detail on critical empirical aspects (on the SET question or on how random the assignment process was); this is clearly in bad faith since Boring et al. 2016 cites a 2015 working paper (Boring, 2015) that gives all these details and is convincing on all methodological points (as, apparently, referees from JPE thought in 2017).

    - They then claim that the correlation found between gender of the instructor and rating is not very strong (0.09 all disciplines pooled together) and that, as a result, gender does not explain much of the variance in ratings. This is of course a mistaken reading of the results: when we study bias, we are not concerned by how much of variance it explains, rather whether it exists or not (it is clear that gender explains but a small portion of the variance in earnings, but any difference in earnings due purely to gender is wrong and points toward systematic discrimination.) In the 2017 paper, Boring shows a difference of around 0.1 to 0.13 points over a 4 point scale, a difference anyone would agree matters when high stake personnel decisions are made.

    - Last, they claim that the gender difference could be explained by many perfectly acceptable factors, such as experience; again, this is plain wrong since Boring et al. 2016 show that female instructors in their sample are better instructors than their male counterparts (students in sections taught by female instructors do better on the final exam.) Of course, Ryalls et al. 2016, assert that scores to the final exam is not a "valid" measure of quality of teaching, but this sounds more like desperation than anything else.

- Ryalls et al. 2016 also dismiss the results from MacNell et al. 2014, on the basis that instructors who participated in the experiment were fully aware of when they were identified as "male" and when they were identified as "female" and may have altered their behaviour with students in order to generate a difference in ratings penalizing their "female" identity. However, because interactions between instructors and students were tightly controlled (instructors worked together to make sure they were grading in a consistent manner and were returning assignments simultaneously), it is hard to imagine how they could have altered their behaviours significantly.

To summarize: even though the debate will linger on for a few more years, and even though we still need a meta-analysis of studies on biases in SET, it is quite clear that the bulk of the evidence shows that one single quantitative question on teaching effectiveness is a non valid and biased way to measure quality of teaching. As the Ryerson decision establishes it, it should not be used for personnel decisions by administrators.

3. Our discussions on how to use students' voices in assessing teaching and learning at McMaster:

This section is based on meetings of the committee, and discussions with John Bell (Director, Information Technology, Faculty of Humanities), Arig al Shaibah (AVP, Equity and Inclusion), Kimberly Dej (AVP Faculty), Pilar Michaud (Director, Human Rights and Dispute Resolution, Equity Office), and Tanya Martini (Psychology, Brock University) and members of the McMaster Teaching and Learning Advisory Board.

The first issue we discuss is: how can we use students' voices to help administrators make merit or tenure and permanence decisions?

The second issue is: what else can students' voices be used for (formative or accountability purposes)?

We discuss below what we know and what we need to know about these three purposes (summative, formative, accountability). We organize this broad discussion around the three following sub-discussions:

Discussion #1: Should we give up entirely on using student feedback on their learning experience in a way that can be used (together with other components) to assess instructor performance and effort or are there ways to do so that are valid and unbiased?

The question here is: Is there anything about an instructor's effectiveness (performance) and/or effort put into teaching that we need to learn through students because there is no way we can learn it through objective means (such as, e.g., teaching portfolios, turnaround for papers, tests, and assignments, ability to respond to emails)? Perhaps a better formulation would be: is the cost of running SET worth the value of the information collected? The answer to that question will come from the literature, documented practises in other institutions as well as institutional

introspection on what it is exactly that we call "effective teaching" at McMaster. Such an introspective study could be based on IQAP reports, that should have been completed by all departments as of now. There would then be a need to discuss which components of this McMaster effective teaching can be captured objectively and which ones would necessitate a subjective assessment by students.

Discussion #2: If we answer the latter to Discussion #1, i.e., there are ways to use student feedback on their learning experience to contribute to the measurement of teaching effectiveness that are valid and unbiased, and the cost is worth the information collected, what is the best way? We identified two main options: questionnaire using detailed questions to assess what instructors do from the perspective of students because it cannot be assessed from another perspective, but making sure subjectivity does not mean biased; or focus-group qualitative assessments led by a moderator. This discussion leads to two questions:

Discussion #2.1: Would using multiple questions on the course or the instructor's performance be helpful as "material" (rather than data) in the assessment of instructor merit by administrators? Which means: would it be more valid and less biased than the overall teaching effectiveness question or the composite index built on several questions?

Discussion #2.2: If the answer to 2.1 is no (it will not be more valid and/or less biased) or even "we don't know but it might be biased as well", it might not make sense to rely on close-ended questionnaires (we could experiment with questionnaires to get better evidence on question 2.1 but we would not use questionnaires to assess merit before we have that evidence) and the question becomes: would the cost of running focus-groups to extract subjective information on the instructor's performance be worth the value of that subjective information to assess merit?

Discussion #3: If we answer the former to Discussion #1, i.e., we should give up on student feedback on learning experience to learn anything on teachers' performance, such feedback could still be collected and used to help instructors improve their teaching (as is done, for instance, at the University of Lund) or to guarantee accountability from institutions or programs. In that case, the goal of collecting information from students would not be to capture performance but rather student experience (and this should be the test for validity: do we capture experience?), and bias may not matter that much (as long as it is not unanimous rejection of some characteristics of instructors by students, and/or that the information is not used to compare and rank programs or instructors, but rather to measure changes over time within the same program or institution, or for the same instructor).
The question is: what is the best tool to capture student experience, in order to help instructors and programs improve their teaching and/or help students select classes (match students to courses and instructors on the basis of epistemological beliefs of students and teaching practices of instructors)?
Best (2008) notes that the same questionnaire cannot be used to capture teaching effectiveness and assess student experience: questions to capture teaching effectiveness are centred on the instructor and their performance (e.g., "The course was well prepared and organized") whereas questions on student experience will represent teaching and learning as, in his words, "a

collaborative enterprise with work to be done by both instructors and students" (e.g., "I understood the professor's organization of the course").
It is interesting to note that SET instruments focused on teaching effectiveness implicitly endorse a vision of passive learning, where the instructor provides all the material and feeds it to the students. If McMaster wants to encourage active learning, it might be another reason for moving away from SET as teaching effectiveness and toward collecting feedback on student experience of learning. It is not clear either whether allowing students to shop around for courses and instructors on the basis of their learning style should be encouraged by McMaster: do we want to allow students who prefer passive learning and regurgitation to navigate their undergraduate years without ever having to take an active learning class or a class challenging their epistemological beliefs? Wouldn't a good definition of higher education precisely the challenge of one's such beliefs and perceptions?

Overall, student input can be used in three different ways to assess teaching and we detail below how this can be done:

Keep student feedback to evaluate teaching (teacher performance and effectiveness) but, instead of using one summative question, capture teaching quality based on a series of questions on the many different dimensions of what instructors do when they teach.
Keep student feedback to evaluate teaching, but, instead of quantitative surveys, use a qualitative approach with small groups of randomly selected students to discuss with a moderator their learning experience and how it relates to the instructor performance and effort course.

Use student feedback to capture student experience, without any pretense to measure teaching effectiveness or quality. This can be used in two ways:

> It would provide a way for students to voice their concerns and for the University, its Faculties, Departments and Programs, to listen and show accountability and responsiveness to these concerns. This type of student-experience oriented questionnaires could also be used as a way for students to select courses, based on previous descriptions or assessments (the information would not be used to rank instructors but rather to help students select instructors and teaching practises matching their learning style, so as to maximize effectiveness and experience).
>
> It could allow instructors to use student feedback to improve their teaching (formative purpose.)

We now provide the answers we reached (and, more importantly, the questions still unanswered) to these questions:

3. 1. Discussion #1: Are there ways to use SET to evaluate teaching effectiveness?

The question to be discussed is: Are there components of teaching effectiveness that cannot be assessed other than through students?
Our conclusion is that there exist elements of teaching effectiveness or quality that cannot be assessed "objectively" and we need the voice of students on their subjective assessment of their experience for the summative as well as formative assessment of teaching. However, what these dimensions are depends on how we define "quality of teaching" as an institution. To fully answer that question, we need to conduct an introspective exercise on our definition of good teaching.

3.2. Discussion #2: If yes to Discussion #1:

Discussion 2.1: Should we use quantitative questions, and a composite index based on multiple questions (one for each dimension of effective teaching)?

To illustrate, we provide a subset of questions asked at U of Michigan that we thought were the most objective ones, as well as two "control" questions][5]:

> [Q3 at Michigan] I knew what was expected of me
> [Q6 at Michigan] The instructor was well prepared for lectures
> [Q7 at Michigan] The instructor explained the material clearly {can be considered somewhat subjective, as your definition of clarity may vary with what you think of the person}
> [Q8 at Michigan] The instructor treated students with respect
> Control questions:
> [Q4 at Michigan] I had strong desire to take that course
> [Q5 at Michigan] The workload in that course compared to other courses

We should discuss the validity of these questions (do they capture what they claim they want to capture - readiness for lectures, clarity of expectations and explanations, respectful treatment?), which is not easy since we don't really have a way to measure these things otherwise. We could run experiments, changing the wording slightly, but that would tell us something about reliability or stability, not validity. We could also evaluate based on peer reviews and compare the results from the peer reviews to those of the questionnaire. Last, we will need to check these questions are not biased (ex post statistical analysis). We can ask Michigan if they ran those tests.

Note: we did not start from the six institutional questions suggested in the McMaster CTEC report 2017 because some of these questions were too subjective. These six institutional questions were:

1. The instructor made the subject interesting [we deemed it too subjective: interesting is linked to what a student thinks of the personality of the instructor and can be tainted by the charisma effect, as shown in the Dr. Fox's experiment]

[5] All questions are answered on a five item scale (from strongly disagree to strongly agree, with a neutral item).

2. Overall, I found that the quality of instructor's teaching in this course was excellent. [Similarly, very subjective and the literature agrees that students are not necessarily the best judges of "quality of teaching"]
3. Course projects, assignments and tests and/or exams provided an opportunity for me to demonstrate my understanding of the course content. [Somewhat similar to Q3 at Michigan, but we feel Michigan's wording is more objective: if a student failed an exam, they will be tempted to state they were not given an opportunity to demonstrate their understanding, but they might still agree they were told in advance what to expect]
4. Compared to other courses, I found the workload for this course to be challenging. [Q5 at Michigan]
5. Overall the quality of my learning experience in this course was excellent. [This is a good question, but captures experience - our question is: can a questionnaire capture both subjective experience and objective assessments of dimensions of teaching?]
6. I would recommend this course to another student. [Same comment as for 5]

The literature on validity as reviewed by Uttl et al. 2017 includes studies of uni-dimensional SET (one question) as well as of composite indexes. Therefore, it is likely that the latter (composite index) is as invalid as the former (single question). We cannot tell for sure, though, as the meta-analysis does not distinguish between the two. Similarly, the literature on bias does not seem to treat differently the two types of SET. We would need to review the literature with that particular focus in mind but it is very likely that composite index will suffer from the same fatal flaws as single question index.

> For instance, the University of Toronto recently released a report congratulating itself on the validity and absence of bias of its composite index on quality of teaching. The study uses 277,498 questionnaires (from 11,919 sections, or 23.3 questionnaires per section on average) collected over two academic years (2015/16 and 2016/17) at the four main academic units (Applied science and Engineering, Art and Science, Mississauga, and Scarborough) to assess the reliability and validity of the Institutional Composite Mean (ICM).
> What is the ICM?
> Toronto doesn't have a summary question on teaching effectiveness, but rather five items (reflecting dimensions) that must be part of any questionnaire in the cascaded course evaluation framework (CCEF) that has been in place since 2012. These five items are as follows:
>
> 1. Students are engaged: Item 1, "I found the course intellectually stimulating."
> 2. Students gain knowledge: Item 2, "The course provided me with a deeper understanding of the subject matter."
> 3. Atmosphere promotes learning: Item 3, "The instructor created a course atmosphere that was conducive to my learning."
> 4. Components improve understanding: Item 4, "Course projects, assignments, tests, and/or exams improved my understanding of the course material."

5. Students have an opportunity to demonstrate understanding: Item 5, “Course projects, assignments, tests, and/or exams provided opportunity for me to demonstrate an understanding of the course material.”

Each item is rated on a scale of 1 (not at all) to 5 (a great deal), and the ICM is simply the average of the five item scores. A course receiving 1 on all dimensions ends up with an ICM of 1, a course receiving 5 on all dimensions ends up with an ICM of 5.
The core questionnaire also has a summative question on overall experience in the course, from poor to excellent. Last, students are invited to provide two qualitative comments, one on the overall quality of the instruction in the course and one on the assistance received during the course.

The average response rate across courses was 42% (we don’t know how Toronto gets a higher response rate than McMaster – their surveys are administered online as well). Because RR are lower in large courses, the overall response rate was 36% (proportion of students invited who took the survey). There is no discussion of the selection bias (who are the student who take the survey?) and the report treats the sample of respondents as if it were a random sample of the population. Response rates are not influenced by the number of questions on the questionnaire (faculties, departments and instructors can add up to 11 questions to the 9 core ones), the number of course evaluations the student was invited to fill. The main determinants of response rate seem to be class size (negative correlation) and number of interactions between instructor and students (not expected grades or workload). Last, the correlation between RR and score is positive, but low (around .15).
Reliability: intra-class (raters within the same section) correlation is strong on each of the five items and for the mean and the five items are highly correlated (for the same student, across items), suggesting they all measure the same thing (“good experience”). The latter is somewhat problematic: if all items measure the same thing, isn’t it some kind of highly subjective impression rather than informed and rational assessment of each separate dimension and are we not led back to the issues plaguing the one question on overall effectiveness? What is, then, the difference between a composite index and a single question?
Test-retest reliability: the Intra-class correlation is .6 for the same student rating different course, .7 for the same instructor teaching different courses, and .8 for the same instructor teaching the same course. The difference between .8 and .6 can be read as evidence that the ICM measures something specific to the instructor and the course (the report’s reading) but the .6 can also be read as indication that student characteristics affects ratings (independent of what instructors do) to a large extent, which is problematic.
Construct validity: they test the correlation between the ICM and other components of the student questionnaire, some that have been used to calculate the ICM (and the ICM correlates strongly with them, which is not a surprise) and some not, such as reported attendance, interest in the topic, expected grade or workload (and the ICM does not correlate well with those). They conclude this validates the ICM, but this is not a validation. There is no attempt at validating the ICM on a measure of the quality of

teaching (grades to anonymous exams for instance, as in Boring et al. 2016 or in follow-on courses).
Biases: Class size matters: from 4.3 under 25 to 3.8 for 200+. Format matters: Online courses receive poorer evaluations (3.7 versus 4.1). The report does not find any gender difference in scores, but, since it does not control for size or other factors (as well as actual quality of teaching), it may be the fact that biases still exist: female instructors can still be under-rated given the level of effort or actual quality of their teaching.

A reason why any quantitative questionnaire to assess quality of teaching for summative purposes would suffer from lack of validity and biases (be "a popularity contest" rather than a rational evaluation) is that quantitative assessments using Likert scales appeal to the intuitive system rather than the rational one (see below). Therefore, any quantitative questionnaire for summative use may provide the wrong incentives to students: Students may manipulate SET if they know they are used for summative purposes and can affect faculty's careers, in order to get grading leniency and/or lower workload. There is a debate in the literature on biases due to students expectations, grades, and workload. Until recently, the literature was of the view that none of these existed but it might have changed recently as it was the same literature stating that SET were valid representations of teaching effectiveness - we recommend a re-evaluation of that literature, to test whether "objective" SET (questions on select components of teaching quality) can be manipulated by students and whether this manipulation is more likely when students know that SET are used by administrators to make high-stake decisions on career and merit. In the same vein, Hornstein, 2017, suggests that summative evaluations of teaching may lead to de-moralization of instructors (those below the median) and, paradoxically, lower quality of teaching.

For these reasons, we do not believe that a quantitative questionnaire and composite index would be immune from the weaknesses of the single question on teaching effectiveness if used for summative purpose and to make high-stake personnel decisions.

Discussion 2.2: Or should we use qualitative approaches?

The main argument in favour of qualitative approaches to collecting student input instead of quantitative questionnaires is developed by Merritt (2008): ratings appeal to the intuitive (system 1) rather than rational (system 2: deliberative, accountable, accurate, specific and evidence-based) decision-making processes. As a result, style trumps content. Another argument has been put forward by Lattuca and Donagal-Goldman (2007), that teaching needs to be assessed holistically rather than analytically (a course is a coherent project that needs to be understood in its general design and intentions, and cannot be decomposed into individual components) but they do not provide much empirical evidence for this and their main recommendations are to use non student input (they recommend peer-reviews, assessment of the portfolio, interviews with graduates or their employers). We fully agree with using non-student input as well as student input, but, because this committee was about student input in teaching assessment, we focus on Merritt's argument and conclusion.

Merritt's argument relies on experimental evidence conducted in the lab. The seminal and best known such experiment is the so-called "Dr. Fox's lecture" conducted in 1973 (Naftulin, Ware & Donnelley, 1973, experiment which suggested that attendants to a presentation (lecture), even experts in the field, could be fooled to praise a meaningless lecture (on "game theory applied to physician education") when the presenting style was impressive enough (delivered by a professional actor, showing warmth toward the audience, a lively demeanour and humour). The Dr. Fox experiment was replicated several times (Williams and Ware, 1977, Abrami et al. 1982, Marsh and Ware, 1982) with more nuanced findings[6]. Another such experiment, closer to the topic of SET is known as "the thin slice effect": Ambady and Rosenthal (1993) show that students asked to make judgements based on short snapshots of lectures, as short as 6 seconds, silent, or in a foreign language, predict actual SET ratings quite well (the correlation between subjects' ratings and actual SET ratings was .7). This study was replicated by Clayson and Sheffet, 2006, with the same results.

Biases come from the fact that style is based on nonverbal behaviour (nvb): some nvb can be altered and is part of what it is to be a good lecturer, but a lot is beyond the reach of the instructor, because it is filtered through pre-established stereotypes (e.g., black faces are seen as more aggressive; a woman seeking eye contact is seen as having an agenda or being authoritarian while a man doing the same is seen as relax and in control).
To help system 2 take over, Merritt recommends the following (see also Munro, 2000, Abbott, 1990):

- stay away from numerical ratings,
- give more time and more mental space to students
- add accountability (third party)
- create small-group discussions.
- Because all this is resource-intensive, conduct one every three years.

An alternative, suggested by Brockx, van Roy and Martelmans (2012), would be to use questionnaires with open-ended questions only, but it is not known what proportion of students currently use the opportunity of open-ended questions in the SET at McMaster and analyzing open-ended answers on a whole class can be very resource intensive.

We appreciate such a qualitative approach is resource intensive and may have its own issues (e.g., how to select participants to the focus group, how to make sure selected students participate, and how to train evaluators?) and we recommend a working group to discuss implementation. Another issue to be discussed is the fact that instructors rotate across courses and, if the evaluation is conducted every 3 to 4 years, an instructor may never be assessed

[6] Theall and Franklin, 2001, in an effort to dismiss all experimental findings on biases, assert that the Dr. Fox experiment could never be replicated. It is true that the conclusion that "content did not matter at all" was certainly too strong, but replications still found that style mattered as much, if not more, than content.

summatively on a given course they gave several times. Also, an instructor may be unfairly assessed summatively on a poor year whereas their experience is usually favourable.

Discussion #3: What are the objectives of measuring student experience or satisfaction?

Student voice is important, they are the experts in describing their experiences and it is important for instructors and administrators to know about their students' experience (and to be accountable to their students). Student experience can be used by instructors for formative purposes (to improve the course) and/or by administrators and the institution to improve the organization of teaching and learning within programs or at the Faculty or University level.

Beran et al. (2005) find that, in a large Canadian university, 50% of students use SET scores and 60% of those who use scores (hence, 30% of students) use them to select courses or instructors. Spencer and Schmelkin (2002) interviewing students at a large US college, find that students think SET should be used mostly for formative purposes to help instructors improve their courses (score of 4.7 over 5), then for course selection purposes (as a way to provide information to future students, score of 4.1 over 5), and, last, for summative purposes (tenure: 3.8, promotion: 3.5, and merit: 3.3). In their study, students respond they fill out SET (more than 70% of respondents do) but they actually don't use them for course or instructor selection.

Also, the simple fact for students to voice their concerns or approval can change the dynamic (motivation, involvement in learning), but the Committee did not discuss that question (if there is any evidence on it, which is not proven).

Should we measure experience or satisfaction?
Satisfaction results from an interaction between experience and expectations (based on past and anticipated outcomes). Satisfaction is often added to questionnaires on experience, sometimes as a way to validate questions on experience (see below, literature on CEQ).

In Australia, the objective is to compare the same programs across universities, to help university administrators improve the quality of what they deliver. In Sweden (Lund), the goal is to report back to students, mostly on how the instructor reacted to their feedback and what is going to be implemented.

What instruments are used to measure experience/satisfaction?
There are many questionnaires explicitly designed to capture students experience, such as the National Student Survey in the UK, the NSSE in Canada and the US, the Course Experience Questionnaire in Australia, the older SEEQ and the most recent (2007) SCEQ. These surveys on experience are usually at the program or institution level, not at the course/instructor level. One reason is their length, that would preclude using them for each and every course a student takes. Another reason is that these surveys want to capture the experience with learning as a whole, including organization of courses within a program (redundancies, gaps between pre-requisites and follow-on courses, areas not covered at all etc.)

The Course Experience Questionnaire is a popular instrument, used in Australia, the UK, and some universities in Europe (e.g., Lund) to assess how students perceive the quality of the course. It is used to measure performance in Australia and the UK, but as formative feedback to the instructor, administrators and students, after some deliberation, at Lund. In Australia, it is filled by graduates of programs, a few months post graduation.

The Committee did not find much literature on how to capture student experience within a course for formative purpose.
One interesting experience is that of Lund, in Sweden, that uses the CEQ to collect information on student experience in each course for formative purposes. CEQ scores are treated as material, not data: the instructor meets with student representatives and an expert in quality of teaching to discuss the scores and each participant to the meeting writes a short report. The expert then writes a final report that is distributed to students. The reasons they treat this as material rather than data are as follows: scores are contextual (course level, class size, required or optional course, time in the day, or quality of the room) and scores depend on the epistemological views of students. Students who believe that knowledge is true or false will rate a course that challenges them or makes uncomfortable as poorly as a course that is not well organized (they will see the former as not well organized even though it is). Our understanding is that, at the meeting, the instructor can (and is asked to) provide objective evidence on how well organized their course was (for instance). Kember and Wong (2000) suggest adding a question on student belief about learning: do they see learning as a passive or active mechanism and do they prefer deep or surface learning? Their study is a qualitative one, based on semi-structured interviews but there is a literature (cited in Kember and Wong) on quantitative measures of attitudes and beliefs toward teaching and learning. It must be noted, though, that promoters of the CEQ recommend not to use it at the course level (Wilson et al. 1997).
An alternative, much shorter, questionnaire, could be built on the basis of similar questions on student experience found in a variety of PSE institutions:

1. McGill Q2: I learned a lot from this course
2. UWO: The instructor motivates me
3. UWO2: Overall, I had a great experience in that course

These options (CEQ or ad hoc, short questionnaire) could be explored, but we could also rely instead on current practice at McMaster, based on what MacPherson does and what some of us (Joe Kim) have successfully experimented with: a mid-term and end-of-term open-ended questionnaire on what works, what does not work and what could work (also known as "Stop/Start/Continue") in the course. The instructor asks students to answer those questions anonymously, collects the answers, writes a report summarizing what was in the answers and what actions they will take in response (if any), with justifications for why no action is taken on some aspects. The Committee felt it is worth investigating whether such formative questionnaire could be generalized (made mandatory), with the report being included in the annual Record of Activity of the instructor. The Committee recommends that most of the RoA be automated (as is currently done in the Faculty of Engineering), which would give instructors enough time to write and present these narratives on how they responded to formative surveys on student experience

in their courses. This raises issues of implementation: if instructors run the questionnaire but do not act upon it, students might be even more disillusioned with SETs than they are now; also, in big classes, administering such a questionnaire will require resources beyond current allocations of TA's.

Turning to using surveys to capture student experience for accountability purposes, we summarize what we learn about the CEQ, which is the only questionnaire used country-wide (Australia and UK) for comparative purposes (across universities):

The current version of CEQ is comprised of 22 items (not counting two items on skills that were added later, see below), each scored 1 to 5, from definitely disagree to definitely agree, some with reverse scales):

Good Teaching:

1. The staff motivate students to do their best
2. The staff make a real effort understanding difficulties students may have with their work
3. Teaching staff usually give good feedback on where you are going
4. Lecturers are extremely good at explaining things
5. Teaching staff work hard to make subjects interesting
6. Staff never show interest in what students have to say (reverse scale)

Clear Goals and Standards:

1. It is always easy to know the standards expected
2. You usually have a clear idea of where you are going and what is expected of you
3. It is often hard to discover what is expected of you in this course (reverse scale)
4. The aims and objectives are not made clear (reverse scale)
5. The staff make it clear from the start what is expected from students

Appropriate Assessment:

1. Lectures give the impression they have nothing to learn from students (?) (reverse scale)
2. To do well, all you need is a good memory (reverse scale)

3. Staff seem more interested in testing what you've memorized than what you've understood (reverse scale)

4. Too many staff ask questions just about facts (reverse scale)

5. Feedback is provided to students only as marks and grades (reverse scale)

6. It would be possible to get through this course just by working hard around exam time (reverse scale)

Appropriate Workload:

1. The workload is too heavy (reverse scale)

2. It seems that the syllabus tries to cover too many topics (reverse scale)

3. We are generally given enough time to understand the things we have to learn

4. There is a lot of pressure on you as a student (reverse scale)

5. The sheer volume of work to go through in this course means you cannot comprehend it all thoroughly

The CEQ never states that it is about measuring teaching effectiveness, and has no summary question on it. It decomposes learning experience into components (good teaching, clear goals and standards, appropriate workload, and appropriate assessment) and it does not look like a composite score is calculated based on these components. An initial version of the CEQ had 30 items in total, but this number was reduced to 18 after the first rounds, based on a Principal Components Analysis, excluding items with lower loading factors. This suggests that promoters of the CEQ are looking for a latent construct (course experience) that all components capture in some way. The current version of the CEQ, as used in Australia, has 24 items, because a new component on skills was added (the perception by students that the course allowed them to develop new skills or to improve their skills).
We could not find any literature on how items in the CEQ were developed in the first place and it looks like a mixture of psychometrics (the Principal Component Analysis) and political (the skills component) motivations.
The literature on CEQ (and SEEQ, a Canadian contribution) is mostly about the link between Experience and Satisfaction (or Satisfaction and Skills). The Canadian study finds that SEEQ scores explain two thirds of the variance in satisfaction scores (one question on satisfaction with the course, one on satisfaction with the instructor) and one Australian study finds that the only components to explain Satisfaction and Skills are Good Teaching and Goals and Standards (therefore, not appropriate workload or appropriate assessment). Satisfaction correlates strongly with Skills (perceived).

Recommendations:

Based on the three discussions summarized above, we can use student input on teaching and learning for three different purposes:

1. Formative: to help instructors and programs provide better quality teaching, better suited to the needs of their students.
2. Summative: to help administrators assess the quality and effort provided by individual instructors (for tenure, permanence, and merit purposes).
3. Accountability: to help instructors, program managers, Deans and the VP Academic assess the experience of students at McMaster and take action in response to perceived weaknesses.

No single instrument can answer these three questions and we suggest three different approaches, one for each of these objectives:

1. **Formative**: Students can provide actionable feedback during the course (rather than at the end of the course), by commenting on their learning experiences and how it could be improved. At the mid-point of the term, instructors will administer a variant of a tool known as "Start/Stop/Continue" survey. In this easy-to-administer survey, students provide open-ended answers on what they think could be done to improve their learning experience in the second half of the course:
    1. Start: what can be done that is not currently done? (e.g., start posting lecture slides before class);
    2. Stop: what is currently done that hampers learning? (e.g., stop speaking so quickly and using complex terms);
    3. and Continue: what is currently done that positively adds to the learning experience? (e.g., demonstrations that involve members of the class).

    The survey could start with one open-ended question on what the student thinks they are doing that contributes to their learning experience in the course (self-reflection), to clarify this is not about "evaluating" the instructor, but rather about perceptions of the learning experience. Students would fill out these surveys online anonymously (ideally through Avenue). Instructors would then provide feedback on the emerging trends, and adjustments to be made in the second half of the course. This report (including the response from the instructor) should be made public (including to students), but would not be used as such for evaluation purposes, except when included in the Report of Activity (SPS B1, item #6, evidence of incorporation of some form of formative evaluation of courses and response to concerns of students). When used, these short surveys usually have excellent participation rates.all instructors will conduct a start/stop/continue survey mid-term and, optionally, end-of-term. These are easy-to-administer surveys in which students simply state (open-ended answers) what they think the instructor should do that they don't do, should

stop doing, and should continue doing (what works) to improve their learning experience. It is not about, and should not be framed as, students telling instructors what to do, rather constructive comments on how students think their learning experience could be enhanced in the course. Students would fill out these surveys online anonymously. Instructors would then provide feedback on what they would change (if anything) to the course in response to the survey, either in the second half of the course or in the coming year. This report should be made public (including to students), but would not be used as such for evaluation purposes, except when included in the Report of Activity (SPS B1, item #6, evidence of incorporation of some form of formative evaluation of courses and response to concerns of students). When used, these short surveys have usually excellent participation rates. We acknowledge they may raise a workload issue (even though one of us currently runs the start, stop, continue survey in a large – 2,400 students – class and reports back to the class on the emerging trends and actionable changes, and it takes a reasonable amount of time): it is certainly costly to collect and analyze these data and a recommendation of the committee is for the administration to provide additional TA resources to conduct this work. Some instructors who use these questionnaires at McMaster use TA resources to analyze and summarize feedback and are happy with it. The DeGroote School of Business has made them mandatory and we should be able to learn collectively from their experience. Beside workload, this method for collecting feedback from students requires buy-in from all instructors, otherwise students will, again, become disillusioned about the whole process and may disengage from it. We suggest that, by asking that the report (including responses from the instructor) be included in the annual Record of Activity, instructors will be motivated to take these seriously; another suggestion is to pair reluctant instructors with instructors who have conducted these data collection exercises successfully, so that they can learn about the benefits of conducting them as well as how to conduct them without inflating workload. An issue raised in focus groups is that 1st year students might need a different set of questions: questions on what could be done, what should be changed and what works well for learning experience are easy to understand by students who can compare their experience in the course to their experience in other university courses, but less so for students who do not have a comparator (except high school courses, which are quite different). The questionnaire could be less prescriptive and more descriptive (how is my learning experience in the course) in 1st year classes. The experience of the DSB might be helpful on all these questions (workload, disengaged instructors, and 1st year classes).

2. **Summative:** Student feedback on their perception of their learning experience in the course is linked to the instructor's performance and effort:
   We recommend replacing the current questionnaire, based on Likert scales, with a collection of qualitative feedback on learning experiences. We do not believe any quantitative questionnaire (even a 4-5 questions multi-dimensional one) would be of any help. In essence, ratings give the illusion of objectivity to a process which is inherently subjective: we try to capture what students perceive, which is subjective,

knowing that everything objective can be captured otherwise (e.g., based on the course portfolio). Even more detrimental, the fact of using numbers, quantities, ratings on a scale is likely what induces the behaviours that we want to prevent, such as biases based on preconceived ideas (about gender roles or ethnicity) or manipulation (trading grades for ratings).

There are different ways to conduct qualitative evaluations, either through open-ended questionnaires or focus groups. Our preferred version is the focus group[7]: information on student learning experience in each course would be collected during a focus group of designated students in the class (participation being mandatory) led by a team of trained evaluators. What we have in mind for trained evaluators is a pair, comprised of a trained undergraduate and an emeritus professor or a volunteer from the community. The professor or volunteer would moderate and provide an "external" perspective, while the trained evaluator would be closer to the members of the focus group. As we see it, these student evaluators would be mostly undergraduate students, who would receive credits or payments for conducting the evaluations. Importantly, these focus groups should be led according to a document provided by each program (or Faculty?) stating what teaching excellence means in the context of this particular program. The document could be based on the IQAP reviews. Each session would result in a report, written by the evaluation team and shared with the instructor for feedback. The report and feedback would then be shared with the administrator, and summarized in three broad categories (excellent, good, problematic) and would be one of the seven dimensions used to assess instructors in a given year (CP/M) or for T&P purposes. The detailed report could be used by Chairs and Deans to provide resources to help instructors improve student learning experience in their courses. We acknowledge this approach is exploratory, and McMaster would be a a leading innovator if it followed that path. As a result, there is not much evidence on whether qualitative feedback collected through focus groups is indeed less biased than quantitative feedback collected via close-ended questionnaires, or to support the feasibility or acceptability of such a qualitative approach to summative evaluation. Such a tool raises logistical issues and it would be unrealistic to produce such a report on all course taught in one academic year. We propose in this report a rotating formula where courses are evaluated once every two to three years[8], as a pilot and we recommend running studies (e.g., focus groups in some courses versus open-ended questionnaires in other courses) and documenting what works and what does not. We acknowledge that such a solution raises many issues that are not addressed in the literature we read:

- Collecting feedback every three year can be seen as unfair to instructors: what if the year on which feedback is collected was a "bad" year? Our sense is that the

[7] Note that focus groups do not have to be in person but could be conducted online (using a tool such as Top Hat (https://tophat.com/), already used at McMaster. This tool would provide anonymity to participants.

[8] This can be seen as a long gap between evaluations. However, for an instructor teaching four courses a year, there would be on average one evaluation report every year.

risk of a "bad" crop is much lower in a qualitative assessment environment than through quantitative questionnaires with a response rate of 20%. In a focus group, students will be invited to reflect on their own contribution to their learning experience, to see the course as a collaboration between themselves and the instructor (rather than as a product they consume), and to highlight the main features of the course rather than the small details that may not have worked as well as the instructor wanted them to work. Overall, we believe (but agree we don't have much evidence to support this claim) that focus groups will focus on the signal rather than the noise and will not penalize an instructor meaning and doing well for small glitches in a course that otherwise worked well. This is why we do not think variance in assessment will be as much of an issue as in current SET (again, especially with response rates as low as 20%)

- Questions were raised in the focus groups about the selection of students who would then participate in the discussion groups. We don't have any definitive answers to that question and would defer to experts in qualitative research and MacPherson institute who run such exercises on a regular basis and know how to select informants. It must be noted that response rate is not as much of an issue in focus groups as in questionnaires: focus groups do not attempt to sample the population and be representative, but rather to get some variation in characteristics within a group of 5 to 10 people (e.g., not all males or females, not all students with an F or A+ on the mid-term). As a result, if one contacted participant refuses to participate, they can be replaced with someone in the same category .
- What we have in mind for trained evaluators is a pair, comprised of a trained undergraduate and an emeritus professor or a volunteer from the community. The professor or volunteer would moderate and provide an "external" perspective, while the trained evaluator would be closer to the members of the focus group. As we see it, these student evaluators would be mostly undergraduate students, who would receive credits for conducting the evaluations. It is also important that these evaluations would be conducted on the basis of a document provided by each program (or Faculty?) and stating what teaching excellence means in the context of this particular program. The document could be based on the IQAP reviews. Each "evaluation" would result in a report, written by the evaluation team and shared with the instructor for feedback. The report and feedback would then be shared with the administrator, and summarized in three broad categories (excellent, good, problematic). It would be one of the seven dimensions used to assess instructors in a given year (CP/M) or for T&P purposes.

3. **Accountability:** Student feedback can be useful in capturing their overall learning experience in their program. This is where we see the value of a close-ended questionnaire on the perceptions of students but programs could also run town-hall meetings with graduates (as some currently do for IQAP purposes) and collect qualitative information on these perceptions. It has to be made clear that the goal is not to reward or punish a given instructor, but rather to help the program adjust and

improve. The level at which such an evaluation should be conducted (course, program, Faculty, University) can be discussed. We recommend conducting it at the program level, and including questions on the organization of the program, rather than taking a narrow focus on the quality of individual courses. We also recommend asking graduates, or graduating students, to complete a questionnaire similar to the Course Evaluation Questionnaire (CEQ) used in Australia. McMaster Office of Institutional Research and Analysis (IRA) has experience surveying graduands at convocation (response rate between 30 and 40%) as well as drop-outs (response rate of 10%). IRA is also involved in the analysis of raw data from a survey of our graduates through the Ontario University Graduate Survey (OUGS), which surveys graduates at six months and two years, with a participation rate of 40%. For now, OUGS focuses on employment but questions could be added on program accountability (OUGS is administered by the Ministry but the Council of Ontario Universities can influence the questionnaire). This is also a way for programs to stay connected with their graduates. A key question is that of participation: it is less crucial than when SETs are used to evaluate an individual instructor for CP/M or T&P purposes, but it may still be an issue if graduates who respond are of a certain type only (e.g., those who live in the GTHA, or those who benefited from their degree and have a good job). Response rates of 30 to 40% at a program level would guarantee sample sizes that allow for statistical inference. The only issue would be selection biases (if taking the survey is systematically linked to relevant opinions regarding the program) but this can tested by comparing those who respond to the whole population of graduands and/or graduates (e.g., by gender, age, grades, GPA, type of student etc.) We discussed incentivizing responses, perhaps through a frame for their degree or some recognition.

We do not recommend using this tool at the course level. One main reason is recall issues: graduates will have more to say (and their recollections will be more accurate) about their 3rd and 4th year classes than about their 1st and 2nd year ones. However, they will be able to provide an assessment of the program as a whole, from 1st to 4th year, having navigated all years in the program. However, would McMaster want to develop a tool to collect information on quality at the course level from graduates or alumni, we recommend using the following questions:

1. McGill Q2: I learned a lot from this course
2. UWO1: The instructor motivates me
3. UWO2: Overall, I had a great experience in that course

At the program level, CEQ is a full questionnaire with some validation.(captures the experience of students): this is where we see the value of a questionnaire on the perceptions of students. It has to be made clear that the goal is not to reward or punish a given instructor, but rather to help the program to adjust and improve. The level at which such an evaluation should be conducted (course, program, Faculty, University) can be discussed. We recommend conducting it at the program level, and including questions on the organization of the program, not only on the quality of individual courses. We also recommend asking graduates, rather than current students to

respond, on the model used in Australia. This is also a way for programs to stay connected with their graduates. A key question is that of participation: it is less crucial than when SET are used to evaluate an individual instructor for CP/M or T&P purposes, but it may still be an issue if graduates who respond are of a certain type only (e.g., those who live in the GTHA, or those who benefited from their degree and have a good job). This can be tested, by comparing the profile of respondents to the profile of graduates in a program, and corrected by reaching out to under-represented categories of graduates. We discussed incentivizing responses, perhaps through a frame for their degree or some recognition.

In the short run, and in order to implement these three approaches successfully, we recommend the following actions:

1. Neutralize answers to the summative question on overall effectiveness in the current McMaster's SET: if the department or program uses an algorithm to calculate merit, put a weight of zero on that variable or set all faculty members at the department average when calculating their merit score; if the department or program does not use an algorithm, do not use the score on that question to justify merit assessment on teaching. Other questions are not mandatory and we recommend to drop them from the questionnaire (this amounts to suspending the administration of SETs in the short run.)

In the longer run, revise the yellow document to drop the question on overall effectiveness from the questionnaire and to make the end-of-term questionnaire (SET) optional or, even better, replace it with a better solution.

1. Run an institutional (McMaster) introspective study of the definition of effective teaching at McMaster and what components of it should be assessed based on information collected through subjective assessment by students. That study could build on the series of IQAP reports from all departments and an internal debate (at JC or Senate or McPherson?) on the cost-effectiveness of collecting information on teaching effectiveness through SET. We could then assess which dimensions of effective teaching should be assessed subjectively and which ones can be assessed objectively, based on the teaching portfolio.

2. Run experiments (try different questionnaires, administration modes, or incentives) on how best to capture student experience at the course, program, or institutional level (and how to improve response rates, that are notoriously low at McMaster compared to other Canadian or even Ontarian institutions). The McMaster CTEC 2017 report recommends better implication from instructors and a strengthened communication as a way to improve response rates.

3. Establish a working group on how to implement qualitative assessment of teaching for summative purposes, using trained evaluators (students and "adults" from the community or emeriti), preferably based on focus groups or discussions. The following should be

discussed: frequency, how to control bias, representativeness. The working group would also discuss how to phase it in (e.g., start with 1st year courses or prioritize teaching professors, for whom evaluations of teaching represent 80% of their career assessment).

4. Establish a working group on how to implement formative "stop/start/continue" assessments during courses on a mandatory basis.

5. If we collectively think we have the resources to commit to a public good effort, McMaster could contribute the larger literature on SETs and their biases in the two following ways:

    - Conduct a statistical analysis of all SET collected at McMaster, using as many years of data as possible to assess biases due to instructors or courses characteristics as well as rater's characteristics, and characteristics of the SET itself (response rate, time trend). Because questionnaires are de-identified rather than fully anonymised, we can link responses to raters' ids and, as a result, characteristics (such as GPA, level, programs, majors, minors, honour or not etc.).

    - Conduct two surveys of the literature: one on the effect of conducting SET on students' expectations (including quid-pro-quo between grading leniency and/or workload and instructor ratings) and one on gender bias in SET.

List of References:

Abbott, R. D., Wulff, D. H., Nyquist, J. D., Ropp, V. A. & Hess, C. W. (1990) Satisfaction With Processes of Collecting Student Opinions About Instruction: The Student Perspective. Journal of Educational Psychology 82(2): 201-206.

Abrami, P.C., Leventhal, L. & Perry R.P. (1982) Educational Seduction. Review of Educational Research, 52(3): 446-64, https://doi.org/10.3102/00346543052003446

Ambady, N., Bernieri F.J. & Richeson J.A. (2000) Toward a Histology of Social Behavior: Judgmental Accuracy from Thin Slices of the Behavioral Stream. Advances in Experimental Social Psychology, 32: 201-271, https://doi.org/10.1016/S0065-2601(00)80006-4

Anderson, K., & Miller, E.D. (1997). Gender and student evaluations of teaching. PS: Political Science and Politics, 30(2), 216-219.

Barre, E. (2015) Do Student Evaluations of Teaching Really Get an F? Reflections on Teaching & Learning, Rice University Centre for Teaching Excellence blog, posted July 9.

Barre, E. (2018) Research on Student Ratings Continues to Evolve. We Should, Too. Reflections on Teaching & Learning, Rice University Centre for Teaching Excellence blog, posted February 22.

Basow, S.A. (1995). Student evaluations of college professors: When gender matters. Journal of Educational Psychology, 87(4), 656-665.

Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. Journal of Personnel Evaluation in Education, 18, 91–106. doi:10.1007/s11092-006-9001-8

Béran, T., Violato, C., Kline, D. & Frideres, J. (2005) The Utility of Student Ratings of Instruction for Students, Faculty, and Administrators: A "Consequential Validity" Study. The Canadian Journal of Higher Education – La Revue Canadienne de l'Enseignement Supérieur, 35(2): 49-70

Best, A. (2008) Student evaluations of law teaching work well: strongly agree, agree, neutral, disagree, strongly disagree. Southern Law Review 38 Sw. L. Rev. 1

Boring, A. (2015). Gender biases in student evaluations of teachers (working paper). OFCE-PRESAGE-SCIENCES PO and LEDa-DIAL.

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. Retrieved from Science Open Research. doi:10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Boring, A. (2017) Gender biases in student evaluations of teaching. Journal of Public Economics 145: 27-41, http://dx.doi.org/10.1016/j.jpubeco.2016.11.006

Braga, M., Paccagnella, M., & Pellizzari, M. (2011). Evaluating students' evaluations of professors. Bank of Italy Temi di Discussione (Working Paper) No, 825.

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. Economics of Education Review, 41, 71–88. http://dx.doi.org/10.1016/j.econedurev.2014.04.002

Brockx, B., Van Roy, K. & Mortelmans D. (2012) The student as commentator: students' comments in student evaluations of teaching. Procedia – Social and Behavioral Sciences 69: 1122-33

Carrell, S.E., & West, J.E. (2008). Does professor quality matter? Evidence from random assignment of students to professors (No. w14081). National Bureau of Economic Research

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. Journal of Marketing Education, 31(1), 16–30. http://dx.doi.org/10.1177/0273475308324086.

Clayson, D.E. & Sheffet, M. J. (2006) Personality and the Student Evaluation of Teaching. Journal of Marketing Education, 28(2): 149-160 DOI: 10.1177/0273475306288402

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, 51(3), 281–309. http://dx.doi.org/10.2307/1170209.

Cramer, K.M. & Alexitch, L.R. (2000). Student evaluations of college professors: identifying sources of bias. Canadian Journal of Higher Education, 30(2), 143-64

Dowell, D. A., & Neal, J. A. (1982). A selective review of the validity of student ratings of teachings. The Journal of Higher Education, 53(1), 51–62. http://dx.doi.org/10.2307/1981538.

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. Research in Higher Education, 30(6), 583–645.

Gravestock, P. and Gregor-Greenleaf, E. (2008) Student Course Evaluations: Research,

Models and Trends. Toronto: Higher Education Quality Council of Ontario.

Horstein, H.H. (2017) Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. Cogent Education, 4, 1304016 http://dx.doi.org/10.1080/2331186X.2017.1304016

Kember, D. & Wong, A. (2000) Implications for Evaluation from a Study of Student's Perceptions of Good and Poor Teaching. Higher Education 40(1): 69-97, Stable URL: https://www.jstor.org/stable/3447952

Lattuca, L., & Domagal-Goldman, J. (2007). Using qualitative methods to assess teaching effectiveness. New Directions for Institutional Research, 136, 81–93. doi:0.1002/ir.233

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, http://dx.doi.org/10.1016/j.stueduc.2016.12.004

MacNell, L., Driscoll, A. & Hunt, A. N. (2014) Whats in a name: exposing gender bias in student ratings of teaching. Innov. High. Educ. 1–13.

Marsh, H. W. (1987). Student's evaluations of university teaching: Research findings, methodological issues, and directions for further research. International Journal of Educational Research, 11, 253–388. doi:10.1016/0883-0355(87)90001-2

Marsh, H. W., & Ware, J. E. Effects of expressiveness, content coverage and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. Journal of Educational Psychology, 1982, 74, 126-134.

Marsh, H.W., & Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), Higher education: Handbook of theory and research, Vol. 8. New York: Agathon Press.

McCallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. Research in Higher Education, 21(2), 150–158. http://dx.doi.org/10.1007/BF00975102.

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

McPherson, M. A. (2006). Determinants of how students evaluate teachers. Journal of Economic Education, 37, 3–20. doi:10.3200/JECE.37.1.3-20

McPherson, M. A., & Todd Jewell, R. (2007). Leveling the playing field: Should student evaluation scores be adjusted? Social Science Quarterly, 88, 868–881 doi:10.1111/j.1540-6237.2007.00487.x

Merritt, D. J. (2008) Bias, the Brain, and Student Evaluations of Teaching. St John's Law Review, 82(1), Article 6, Available at: https://scholarship.law.stjohns.edu/lawreview/vol82/iss1/6

Munro, G. S. (2000) Outcomes Assessment for Law Schools. Institute for Law School Teaching, Gonzaga University School of Law, Spokane, Washington.

Naftulin, D. H., Ware, J. E., and Donnelly, F. A. "The Doctor Fox Lecture: A Paradigm of Educational Seduction." Journal of Medical Education, 1973, 48, 630–635.

Ryalls, K., Benton, S., Barr, J., & Li, D. (2016). Response to bias against female instructors. IDEA Research and Papers. Editorial Notes.

Savonick, Danica and Davidson, Cathy, Gender Bias in Academe: An Annotated Bibliography of Important Recent Studies (2017). CUNY Academic Works.
http://academicworks.cuny.edu/qc_pubs/163

Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. Women's Studies in Communication, 30, 64–77.
doi:10.1080/07491409.2007.10162505

Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. Assessment & Evaluation in Higher Education, 27, 397–409.
doi:10.1080/0260293022000009285

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. Review of Educational Research, 83, 598–642.
http://dx.doi.org/10.3102/0034654313496870

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. doi:10.14293/S2199-1006.1.SQR-EDU. AOFRQA.v1Stark. Retrieved from Science Open:
https://www.scienceopen.com/document/vid/42e6aae5-246b-4900-8015-dc99b467b6e4?0

Stroebe, W. (2016) Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations. Perspectives on Psychological Science, Vol. 11(6) 800–816

Theall, M. & Franklin, J. (2001) Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction? New Directions for Institutional Research, 109: 45-56, https://doi.org/10.1002/ir.3

Uttl, B., White, C.A., & Wong Gonzalez D. (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, http://dx.doi.org/10.1016/j.stueduc.2016.08.007

Weinberg, B.A., Fleisher, B.M., & Hashimoto, M. (2007). Evaluating methods for evaluating instruction: The case of higher education (NBER Working Paper No. 12844). Retrieved 5 August 2013 from http://www.nber.org/papers/w12844http://www.nber.org/papers/w12844

Williams, R. G., & Ware, J. E., Jr. (1977) An extended visit with Dr. Fox: Validity of student ratings of instruction after repeated exposures to a lecturer. American Educational Research Journal, 14, 449-457.

Wilson, K. L., Lizzio, A. & Ramsden, P. (1997) The development, validation and application of the Course Experience Questionnaire. Studies in Higher Education 22(1): 33-53, https://doi.org/10.1080/03075079712331381121

Worthington, A.C. (2002). The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education, Assessment and Evaluation in Higher Education, 27(1), 49–64.

Appendices:

Appendix 1: Terms of references:

MUFA Ad-Hoc Committee on Using Students Assessments in Evaluation of Teaching.

Terms of Reference

Preamble

The procedures for the assessment of teaching are described in SPS B1. This policy is clear that "effective teaching" is a condition for promotion, the granting of tenure, salary increments based on merit, and teaching awards. SPS B1 II is clear that a student assessment of teaching must be conducted toward the end of each undergraduate course; this assessment must be conducted by questionnaire, the "first question" of which must be a summative question on the effectiveness of the instructor for the course. Other questions can be tailored to the need of each faculty, department, or even instructor. Information from the questionnaire must be "consolidated" by the department in a report to the instructor; for numerical questions (including the summative one), consolidation does not imply the mean or median, it must only be a "tabulation of data" (that would include frequency tables). The report must include contextual evidence (same tabulation summarizing information on all courses taught in the department). Section II states that the report "will be used by the department as input", but Section III of the same policy does not constrain departments to take it into consideration: "To the extent that the students' ratings feature in the consideration by departments…". It can therefore be said that the policy is clear on how to conduct student assessment of teaching effectiveness, but much less so on how the information should be used.
Moreover, a recent decision at Ryerson University, based on two reviews of the empirical literature on the measurement of teaching effectiveness through summative questions in student questionnaires, has made it clear that central values (mean or median) of the distributions of answers to such questions **should not be used** for tenure and promotion purposes. Not only are these central values biased in ways that are not possible to correct, but they are meaningless as they treat ordinal variables as if they were cardinal. It is safe to assume that they should not be used either for salary increments based on merit (as most universities in Ontario, Ryerson does not have a CPM scheme, which is why the decision weighed on T&P processes only).

MUFA is interested in updating the procedures for the assessment of teaching in SPS B1, more precisely section II, and any other section where the student survey is mentioned.

MUFA proposes an Ad-Hoc committee tasked with providing recommendations on the best methods to include students' assessments in the evaluation of teaching (including not taking them into account), starting in fall 2018. The Ad-Hoc Committee is not tasked with re-writing SPS B1, but rather with providing recommendations to a future drafting committee.

Mandate

The Ad-Hoc Committee ("the Committee") is tasked with answering two questions:

Should students **evaluate** their professors for tenure/promotion and CP/M?
If yes, how should that evaluation be conducted? If not, how should the **assessment of student experience** be part of the evaluation of teaching?

Deliverables: The committee will write a report answering questions 1 and 2 above by spring 2019.

Responsibilities

As part of its responsibilities, the Committee will consider and make recommendations concerning:

Whether students should evaluate instructors, or whether teaching evaluation should be based on the assessment of students experience in the course, or whether the summative evaluation of teaching should not use any student input, keeping those for formative evaluation purposes only. These questions can be answered differently for T&P and CP/M purposes.

What are the best processes/methods to collect information on the evaluation of teaching effectiveness by students or on the assessment of students experience in a course. The Committee will discuss the value of quantitative versus qualitative methods, and, for each type, the best practices to produce unbiased and meaningful information on teaching quality. The committee will make recommendations on best practices for various types of courses.

Last, should the information collected (or part of it) be made public and, if so, how?

Research and Consultation

In order to fulfill its mandate, the Committee shall:

Review existing practices across faculties and departments at McMaster

Review existing practices and processes in comparator universities in Ontario to identify best practices

Consult with stakeholders (students, chairs, Deans) and experts in the field of assessment of teaching effectiveness

Review any other documents the Committee deems helpful to the fulfilment of its mandate.

Membership

4 MUFA members, including the President, one member of the exec, one teaching professor and one tenure-track, one representing the Faculties of Humanities and Social Sciences, one representing the Faculties of Science and Engineering, one representing the Faculty of Health Sciences, and one representing the DeGroote School of Business.
2 MSU members.
1 Graduate Student Association member (they are students and TAs).
1 research staff paid equally by MUFA and the administration (possibly a McPherson staff.)
Total: 8 members.

Reporting

The Committee will keep MUFA Exec informed of its progress through its mandate. The Committee will invite MUFA exec to comment on the report and approve it.

Appendix 2: Meeting with John Bell, February 27.

Students fill the questionnaire on a dedicated portal, populated with courses based on information pulled from Mosaic (courses taught on a given term). A first issue is data quality: some courses are not listed on Mosaic, that should be, and some courses listed are actually not offered or, when they are, there is a discrepancy between the name of the instructor on Mosaic and in reality. This is why each Department has to have an administrator in charge of checking that the information provided by Mosaic to the portal is accurate. This can be problematic in the case of cross-listed courses, because it is not clear what department is in charge (or, the course can be listed as two separate courses, with two different class lists). Another issue is with team-taught courses: some departments will create one "shell" for each instructor, allowing J. Bell to create as many SETs as are needed for that course [what is not clear to me is whether the department would create different courses on Mosaic. If that is the case, how are final grades entered for the course?] Despite all precautions, there is still a rate of error of 1% in the data he receives from Mosaic.

Versions: all questionnaires have to have the question on the effectiveness of the instructor (in the exact same form: 10 point scale) but, other than that Faculties, Departments or even instructors can have any questions they see fit (including open-ended or scales different than 10-point). There are between 40 to 45 different instruments on campus now, not counting those used for Teaching Assistants or Lab Assistants. We asked to get access to theses different instruments, to see how much variation there is across campus but these data belong to Deans and we will add a question to our survey to require access to the questionnaires used in the Faculty.

J. Bell receives a series of flat files (txt) from Mosaic: they are an SQL extraction from a view of the Oracle database, created nightly every day. As a result, it is not a real time view but a 24 hour lagged one, which does not allow J. Bell to do much data cleaning or monitoring before the end of the evaluation period. This should change in the future and he should have access to the database in real time.

Delegates can then generate a PDF for each course with the distribution (frequencies), mean, median and standard deviation of all quantitative variables, plus full answers to the open-ended questions. This is sent to department Chairs who can then share with their faculty members, once all grades have been entered on Mosaic.

A data file is then provided to UTS who can import it into MOSAIC. The file contains the following information: the mean, median, standard deviation and distribution of scores for the first question of the SET only, if there are at least 5 responses and if the instructor opted-in to have their teaching evaluations shared with students. The issue is that there are actually three categories of instructors: opted-in, opted-out (they show on the portal with the mention that they did not want to share information on their courses) and no option (the instructor has neither opted in nor opted out). Stephanie and Tasneem confirmed that, as a result of this incomplete

information, students at Mac don't really use the teaching evaluation results to select courses. UTS should have a distribution of faculty by status.

Not all programs are part of this evaluation tool: for instance, the BHSc Program is the only one in FHS to use it. It is also not mandatory for graduate courses.

Q&A:

Why a 10-point scale?

He doesn't know. His sense is that 7-point scales work much better (there is consensus around this in Sociology)

Are data anonymized?

Student IDs are encrypted but this can be reverse engineered if need be. This could be changed but it was decided to go anonymous to all users (including administrators) in 2008 at the time the system was established in Humanities, out of fear that simple de-identification would make students reluctant to respond. [I am in favour of de-identification, to link responses to respondent's characteristics, but would be interested in reading the literature that led to the decision to anonymize.]

Can the system be used for formative evaluations?

Yes, and he would strongly encourage that. The IT system can do anything but what is lacking are clear instructions about the purpose(s) of these SETs (what is the instrument hoping to achieve?) [our comment: from a student perspective, the purpose can be: to penalize or reward the instructor, to provide review to facilitate shopping around, or to improve the quality of teaching. If the latter, evaluations could be provided at several points during the course, e.g., mid-term, 9 weeks and 12 weeks, and John confirmed the system could handle that]

Can we get data on response rates by course, Faculty, and year?

Yes, I need to send John a request and he will ask the AVP Academic to release the data.

Appendix 3: Survey on how Faculties administer SET at McMaster:

Summary of the responses received from the Deans (responses are provided below):

Four responses were received, from Engineering (one at the Faculty level and one for each Department), Health Sciences (one Department), and Social Sciences (Faculty level), Science (one for each Department).

These four faculties use the "standard" McMaster questionnaire, without any modification. There does not seem to be much alteration to the questionnaire at the Departmental level. All questionnaires are completed online.

In Engineering, six Departments of Science (Geography and Earth Sciences, Psychology and Neurobiology, Mathematics and Statistics, Kinesiology, Chemistry and School of Interdisciplinary Sciences) and Health Sciences (Biochemistry and Biomedical Sciences), instructors are encouraged to allow some time for students to complete the SET in class (15 minutes, the instructor leaving the room), but not in Social Sciences. It is interesting to note that, in Engineering, one Department (Chemical engineering) is not aware of that encouragement and does not think it is usual practice. The danger of "unsupervised" (as opposed to in class) completion of SETs for reliability has been pointed out (Kinesiology).

Poor response rates are seen as a major issue in Science and Health Sciences, to the point that the response states that "current participation rates make SET data meaningless in most cases" (Biochemistry and Biomedical Sciences). All respondents indicated minimal response rates that are well above what we tend to get on average at McMaster: 50% in Social Sciences, 65% in Health Sciences (75% for graduate courses), and between 20 and 50% (depending on class size) in Engineering and in Science according to some Departments, but 70-80% at the Faculty level for Engineering. It is interesting to note that some of our respondents indicated they had no expertise on that (which suggests they assume somebody made those decisions upstream).

The question of incentives (to increase participation) brought interesting responses: there is a general reluctance to allow instructors to incentivize on grades (e.g., give x% to all students if participation rate reaches a given level) because students learn how to inflate those incentives, playing instructors against each other. There is even mention of students threatening to lower their rates on the questionnaire if the incentive is not large enough (in essence trading evaluation scores, rather than participation, for grades).

All Departments use the summative question on the questionnaire, except Physics and Astronomy, which neutralizes it (because it is unreliable) and uses other metrics (peer reviews, portfolio etc.) and they do not use any other question, with two minor exceptions: when rates are very low, other questions are used to understand what happened, and written comments are used systematically by Kinesiology. Administrators use the mean, without much adjustment (outliers are not removed), except, in some instances, for class size and "difficulty". In Social sciences, evaluation ratings are averaged for the instructor over several years (SPS B1 explicitly mentions 5 years) and inconsistent scores are dismissed (an instructor doing well on average but who failed once will not be penalized). In Material Science and Engineering, scores are compared to a Departmental average which is weighted by class size.

Graduate courses are treated the same as undergraduate courses in Engineering and Health Sciences, but not in Social Sciences (where Departments can do what they want for graduate courses) and Science (some Departments do not evaluate graduate courses at all due to small class sizes, and Kinesiology uses MacPherson to run evaluations of graduate courses).
Dan Centea (Engineering) states that all seven dimensions of good teaching listed in SPS B1 should be scored quantitatively. As of now, SET is the only dimension to receive a quantitative assessment, making it easier to use than the qualitative assessments, and Chairs naturally give it preeminence.
Last, the School of Interdisciplinary Sciences recommends we conduct a study of why our students are so few to respond to the SETs.
In an ideal world, quality of teaching should be measured by learning (either quiz in and out of term or performance in follow-on courses), or at least by experts such as MacPherson (suggested, respectively, by Physics and Psychology).