Report of the OCUFA Student Questionnaires on Courses and Teaching Working Group

February 2019

OCUFA

Ontario Confederation of University Faculty Associations
Union des Associations des Professeurs des Universités de l'Ontario

Report of the OCUFA Student Questionnaires on Courses and Teaching Working Group

February 2019

Ontario Confederation of University Faculty Associations (OCUFA)

The Ontario Confederation of University Faculty Associations has been the provincial voice of university faculty since 1964. OCUFA represents over 17,000 professors and academic librarians in 29 faculty associations across Ontario.

17 Isabella Street, Toronto, Ontario M4Y 1M7
416-979-2117 | ocufa@ocufa.on.ca
www.ocufa.on.ca

Table of contents

Executive summary

Report of the OCUFA Student Questionnaires on Courses and Teaching Working Group

Background

End-of-term student questionnaires, sometimes called student evaluations of teaching (SETs) or student questionnaires on courses and teaching (SQCTs), have been common practice at Ontario universities for decades. Ontario faculty view SQCTs as important tools for student feedback, but they are often used inappropriately by university departments and administrations to make decisions about faculty pay, tenure and promotion, or the re-appointment of contract faculty.

Prompted by a rise in complaints about the misuse and inappropriate interpretation of questionnaire scores and reports of harassment, OCUFA established a working group to examine the use of student questionnaires at Ontario universities. Members included experts in the fields of methodology, research ethics, and human rights.

After examining student questionnaires along those three dimensions, the working group concludes that SQCTs are a useful mechanism to gather student feedback for "formative" purposes to inform the understanding of the teaching and learning experience, but they are counterproductive, even harmful, when used for "summative" purposes to evaluate the performance of faculty.

Using SQCTs for performance evaluation penalizes women, racialized and LGBTQ2S+ faculty, and faculty with disabilities. These faculty are also more likely to be the target of harassment in the anonymous comments sections of the questionnaires. Further, using SQCTs for performance evaluation risks undermining effective teaching and intellectual diversity.

SQCTs: The context of postsecondary education in Ontario

The working group started from the premise that student learning must be at the heart of why student questionnaires are used. SQCTs ought to provide valuable information that helps faculty improve their course design and teaching; they should be formative for courses and teaching. This is a very different matter from evaluating an instructor's performance for use in decisions that will affect their career.

The institutional context determines whether SQCTs are used to support quality teaching or are misused in ways that subvert the goals of teaching excellence and the academic character of university education. Trends associated with the reduction of provincial government support for universities, a shift to tie future funding to "accountability" and "performance indicators," expectations that higher education only exists to prepare students for the labour market, and an increased focus on ideas of consumer choice and satisfaction, each operate against the grain of teaching excellence and the academic learning experience.

For universities, SQCTs are a cheap substitute for qualitative assessments of teaching. When university credentials are presented as commodities exchanged for the price of admission, the priority of education is flipped from student learning to student satisfaction. In this context, using summative

SQCTs incentivizes teaching geared towards attaining higher scores rather than teaching excellence and improved student learning – undercutting the educational mission of Ontario universities.

The onus is on the advocates of summative questionnaires to show that they are valid instruments of teaching evaluation, that they and their administration conform with ethical standards for the treatment of students and instructors, and that they are consistent with the human rights and faculty agreement rights of instructors.

Findings

Methodology

There is a growing body of evidence that demonstrates that SQCTs are not a suitable method for evaluating teaching performance. The apparent simplicity of numerical scores contributes to pervasive misuse and misinterpretation, but there are deeper problems. For one, student questionnaires do not measure teaching effectiveness and student learning. For another, student questionnaires are characterized by incipient bias.

Many factors figure in student responses, making it impossible to determine the degree to which SQCT scores correlate with actual student learning. Some have nothing to do with the qualities of the teaching: class size, time of day, subject discipline, whether the course is required or an elective, etc. Research also shows that women, racialized, and LGBTQ2S+ faculty receive lower SQCT scores than their white male colleagues. There are obvious equity and human rights implications.

Even responses to questions about specific features – the length of time in returning assignments, for example – are affected by respondents' broader assessment of an instructor, and thereby impacted by unconscious bias. This "halo effect" makes it impossible to unscramble the effects of the multiple determining factors. Student questionnaires may still be useful for an instructor to compare responses for their own courses year after year in order to assess the effect of changes in form and content, but the working group concludes that SQCT scores should not be used to compare instructors under any circumstance.

For related reasons, the working group also finds that SQCT scores should not be aggregated and used as performance indicators for academic programs or universities.

Most importantly, the working group has determined that summative questionnaires do not contribute to student learning and academic achievement. In addition to incentivizing teaching strategies which do nothing to advance student learning, they can work to discourage classroom innovation or the study of challenging subjects. New or challenging approaches may yield lower SQCT scores due to student resistance, even when these improve student learning. Moreover, relying on SQCTs for evaluating faculty performance risks generating intellectual conformity which is not consistent with the spirit of academic freedom and inquiry.

Research ethics

Student questionnaires are not required to undergo the same research ethics review as other university-based research surveys, but the working group has concluded that SQCTs must be conducted in conformity with similarly high ethical standards. Ethical compliance is not a substitute or corrective for methodological shortcomings. However, there have been changes in ethical standards, as well as social and technological developments, since SQCTs were first introduced.

Consent from students should be *active* and *informed*, and sought immediately before completing the questionnaire. They should be aware of the many ways in which their responses might be used and who will have access, including, for example, institutional researchers and third party providers of SQCT services. Students and faculty should have confidence in the data security measures used to protect their privacy. Students should also be aware of the circumstances in which their identity is confidential but not anonymous, that is, in the event of an investigation of harassment via the anonymous comments section of SQCTs.

Consent is different for faculty, and turns more on how fully instructors are willing to embrace the process of SQCTs, and on their disposition towards sharing the results. Unlike summative SQCTs, which are externally imposed and entail potentially punitive consequences, formative questionnaires treat faculty as subjects who participate, differently perhaps, in the exercise with their students.

Except at the option of the instructor, and in a manner to ensure their rights, the working group can see no reason to require SQCT results to be shared or made public. Not only are the scores a poor guide for choosing courses, but publicizing them provides no information about whether an instructor has made any changes in response to previous results, much less what kind of changes or what the instructor might be planning for the next iteration of a course. This is the kind of information pertinent to faculty members' reports about how they have used SQCTs to develop their courses and teaching.

It is impossible to say whether the knowledge that instructors are willingly reporting on their use of SQCT results will be satisfactory to students who wish to know they have been heard. The working group firmly believes, however, that formative questionnaires do refocus attention on the faculty and student and their teaching and learning relationship.

Human rights

The standard of judgment for SQCTs as far as human rights are concerned is consistency with the law – the *Ontario Human Rights Code*, and the *Occupational Health and Safety Act* – and with faculty agreements. These feature prohibitions against discrimination, including *constructive* or *adverse effects* of discrimination, and provisions for addressing harassment. Not only do women, racialized, and LGBTQ2S+ faculty receive lower scores than their white, cis-gender, male colleagues, they are also more likely to be targets of abusive comments.

However unintentional, using SQCT scores to evaluate performance has effects on pay, career status, and career progress that are negative, iterative, and systemic. For those subject to intersecting biases, gender and race or ethnicity for example, the added effect can be even larger gaps in earnings.

Superficially neutral, SQCTs do not account for those factors that shape faculty members' capacity to meet teaching expectations. Women, racialized, and LGBTQ2S+ faculty pay a "cultural tax" in the form of higher workloads that follow from demands to represent diversity and to work with communities of which they are a part. To attain scores comparable to their white male counterparts, they also need to work more and harder to overcome perceptions of competence and intellectual qualities which are biased towards white males.

Moving through the ranks, including from contract to full-time, is challenging enough for women, racialized and LGBTQ2S+ instructors, without the unnecessary pressure of boosting SQCT scores. The effects on the demographic and intellectual diversity of universities are predictable. The working group adopts an equity perspective and takes the position that the simplest solution is not to use SQCTs for summative purposes. It is naïve at best to try to adjust scores as an attempt to offset the effects of bias.

Along with faculty organizations in Australia, the United Kingdom, and the United States, faculty associations in Ontario have reported a rise in the incidence of harassing comments that coincides with the increasing use of online SQCTs. The first consideration must be to protect instructors' rights to effective action against harassment. In addition to the psychological toll of harassment, there are practical questions about how these harassing comments are processed and how the accompanying scores should be excluded from the results.

Universities cannot contract out of their legal and contractual obligations: they should ensure students and faculty are aware of institutional human rights, workplace harassment, and sexual harassment policies and procedures.

Equity and safety are not incidental to the academic mission of universities and the quality of student learning. They are critical conditions for cultivating academic freedom and open inquiry, and an atmosphere that thrives on demographic and intellectual diversity.

Recommendations

To return student questionnaires to their original purpose – as sources of formative feedback – the working group recommends discontinuing their use for summative evaluation of teaching performance. Moreover, they insist the administration of SQCTs be consistent with faculty agreements, in accordance with ethical norms, and respect human rights.

The working group proposes seven principles for refocusing student questionnaires and placing faculty and students and their teaching and learning relationships at the forefront.

1. Limit the use of SQCTs to formative purposes

SQCTs are only suitable for informing faculty about students' understanding of their learning experience, and most valuable for the further development of courses and teaching. They are not equitable and not appropriate for determining pay, tenure, permanency, or promotion for full-time faculty, or appointment and renewal for contract faculty.

2. SQCTs should provide useful feedback for instructors

How different the design of formative questionnaires will be from summative end-of-course versions currently in use will vary, but summative questions do not have a place. Nor will a one-size-fits-all model provide instructive feedback if SQCTs are intended to shed light on different iterations of a course. Common questions follow from, rather than guide, the design of formative instruments.

3. SQCT results should be confidential except at the instructor's discretion

Results and scores should not be made public, or shared with anyone other than those whom the instructor chooses. They are dubious guides for students choosing courses. If the questionnaires are formative, the responses should matter to no more than the faculty member, and perhaps those competent to help interpret them and inform teaching strategies. Any departure from this default must be subject to the terms of faculty association agreements.

4. SQCTs must seek informed and active consent from students

If harassment is to be challenged wherever it appears, student comments on questionnaires cannot be an exception. Students must be advised of their institution's policy on harassment, and the scope of confidentiality in the event of an investigation of alleged harassment or threat of violence.

5. Surveys for other reviews should be separately administered

To avoid double counting, canvassing respondents not in the relevant population, and tainting of the results with bias endemic to SQCTs, surveys for program and institutional reviews should be administrated separately. Further, no other methods of teaching evaluation should be reduced to numeric scores and used as metrics for program or institutional performance.

6. Teaching evaluation requires a suite of tools

If SQCTs are included as part of teaching evaluations, they should be only one tool in a bigger toolkit. The principal methods are the careful examination of teaching dossiers and in-class observation by peers. If SQCT results feature, it is not the scores that are informative but the instructor's explanation of how the responses figure in the faculty member's own evaluation and development of their courses.

7. Peer evaluation should be the rule

No student graduates with a university credential having taken courses from only one professor: university education is a collective responsibility. Evaluating teaching is a collegial responsibility that should not be contracted out. There is no substitute for peer knowledge of the content, the nature and value of teaching activities outside the classroom, and the differences between courses and modes of delivery.

Each of these principles is encoded in provisions of one or more faculty agreements. So too are prohibitions against discrimination, as are commitments to natural justice and procedural fairness, reasonable and fair standards, and academic freedom. These are in keeping with the formative feedback that teaching evaluation is generally meant to provide. How these seven principles can be applied to the particulars of faculty agreements must be a matter negotiated with faculty associations.

Putting the principles into practice will require resources. A fully functioning peer review process for teaching evaluation requires time and investment, including training reviewers to recognize bias and the ways it manifests in the review process, and enabling contract faculty to engage as peers in the evaluation of their fellow contract instructors. Making a renewed commitment to teaching excellence and academic achievement will require more funding and less focus on metrics from the provincial government. It will also require the willingness of university administrations to allocate resources to support faculty, students, and teaching as vital to the academic mission.

Report of the OCUFA Student Questionnaires on Courses and Teaching Working Group

Introduction

Under one name or another, Student Questionnaires on Courses and Teaching (SQCT) have been in use at Ontario universities for decades.[1] In recent years SQCTs, or student questionnaires, as we shall call them in this report, have attracted considerable attention from various quarters. The interest is more than academic.

The attention given to student questionnaires is commonly framed as a concern about "teaching quality" or "teaching excellence." It is often accompanied by two other themes: preoccupations with "value for money" and "accountability." These discourses can no doubt be linked in some way to the retreat of the provincial government from a commitment to higher education as a public good, and to the expectation that it is instead a machine for preparing students for the labour market, or to the commodification of higher education and elevation of consumer choice and satisfaction as guiding values. In fact, it is hard to imagine that these trends have arisen solely within the academy, with no influence from such external forces.

Pressures such as these confront faculty members at the heart of what they do. Teaching is a core activity for virtually all faculty, and for many it is the primary responsibility, in addition to being a matter of professional integrity and a source of job satisfaction. Feedback from students through questionnaires on courses and teaching should be an important mechanism for instructors to gauge how well they are doing in that regard. However, what teaching quality and student learning mean for the instructors who are actually engaged with students is increasingly at odds with the meaning, logic, and use being applied to end-of-course student questionnaires by their advocates.

It is these cross-purposes which prompted OCUFA to establish a working group to review the use of student questionnaires at Ontario institutions of higher education. Uses that are made of these questionnaires have real consequences for faculty members and students, and they do not relate only to the quality of the teaching and learning experience. In addition to the methodological matters fueling much of the debate about student questionnaires, there are substantive ethics and human rights issues that get little if any attention. It is with these concerns in mind that the we undertook our review.

We conclude that student questionnaires should be used for "formative" purposes only; they should not be used for "summative" purposes, that is, for performance evaluation.[2] Questionnaires are not only inappropriate for use in career decisions for faculty members, they can also be self-defeating if their purpose is to support effective teaching and academic diversity. We are also persuaded that nothing positive is gained by making SQCT scores public. Quite apart from respect for the confidentiality of

---

[1] Names vary by institution, and over time. They include such terms as Student Opinion Surveys, Student Ratings of Instruction, Student Evaluations of Teaching, Faculty Course Evaluations, to list just a few.

[2] Summative is the adjective form of summation, i.e., it is end-of-course and for evaluation. Formative is developmental, and etymologically from the French *formation*, or the transfer of necessary knowledge.

information about faculty, the scores are misleading guides to course options for students. Because of their limitations, simply aggregating the scores from student questionnaires does not yield useful information about an academic program or faculty, much less offer meaningful, policy-relevant information about institutions.

Student questionnaires are not the only means students have to communicate to their instructors about their courses and teaching, but they are certainly the most inclusive avenue for getting the broadest possible feedback. In return, students should have confidence that, even if they cannot always see the effect themselves, that faculty members have a similar interest and investment in soliciting students' perspectives for the purpose of advancing learning. Lest it give students false hope, and end up alienating students and faculty alike, to lend integrity and credibility to student questionnaires and the process by which they are administered, we are convinced it is necessary to ensure that questionnaires are formative only.

We believe student questionnaires can and should be useful for formative, developmental purposes. A number of principles follow from that premise. It is not as simple as re-purposing existing student questionnaires. If they are to be respectful of the faculty member and the students, and to meet the purposes set for them, survey design and the process by which they are administered matter. Further, even though student questionnaires are mandated by university decision-making bodies, the primary audience should be the faculty member whose course the students are taking. Disclosure of the results should be at the instructor's option. Finally, questions to students about academic programs and the courses which comprise them should be administered separately.

Irrespective of the use to which student questionnaire results are put, if they continue to be administered, we believe a fundamental re-evaluation of consent procedures must be conducted for the protection of students and faculty members. We conclude not only that active, continuing, explicit and informed consent should be required, but also that the information provided must be considerably more than currently seems to be the case. This is especially pertinent for taking steps to curtail their use as "institutionally sanctioned instruments of harassment," as one informant described online student questionnaires.

This report summarizes our review of research and expert reports, responses from faculty associations, available documentation pertaining to the administration of student questionnaires and research ethics, our critical reflections on the methodological, ethical and human rights issues raised by the use of student questionnaires, and our recommendations for guidelines or principles to follow.[3]

These reports are prepared by the working group for OCUFA, and as such, they do not constitute an OCUFA policy statement. Nor do they represent or prescribe positions faculty associations might take

[3] A separate report on faculty collective agreements and student questionnaires as well as an extended bibliography are for exclusive use by OCUFA member associations.

on student questionnaires. However, it is our hope that the reports will provide useful analyses and navigational aids for each faculty association to use as it deems appropriate, in accordance with its own particular circumstances and priorities.

Report background, scope, and structure

The impetus for establishing a working group to examine the use of student questionnaires is twofold. First, faculty associations are dealing with the direct effects these questionnaires have on members. Complaints about the inappropriate interpretation and use of SQCT scores have been on the rise, as have reports of incidents of harassment – both problems amplified by online technology. There is a common set of issues faced by all faculty whenever scores are used for performance evaluation: members disciplined, pay increments withheld, tenure or continuing status denied, contracts not renewed.

We use the terms faculty member and instructor interchangeably throughout this report to reflect the shared features of these problems, but the difficulties are more acute and the consequences more severe for some faculty than others. There are teaching-intensive faculty and especially contract faculty for whom student questionnaires weigh disproportionately on pay, permanency, career progress, and/or the prospect of contract renewal. Women, racialized and LGBTQ2S+ faculty, and faculty with disabilities are also more likely to be penalized, sometimes in combination with their type of appointment, and to be the target of harassment.

Second, OCUFA faces a policy environment in which "accountability" and "performance indicators" are assuming a higher profile in policy discourse. The fixation with performance metrics predates the 2012 report of the Commission on the Reform of Ontario's Public Service (the Drummond Commission), but the Commission's recommendation to adopt SQCT scores as a performance measure has since been echoed by the Ontario Auditor General and in a Ministry of Training, Colleges and University consultation report on the funding model for universities. Most recently, the scores appeared as a suggested optional metric for the 2017 to 2020 round of Strategic Mandate Agreements (SMA) between institutions and the provincial government.[4]

Very little attention seems to have been paid to the actual effects of student questionnaires. The working group was put together to remedy that deficiency by asking questions and offering perspectives that tend to be overlooked by advocates of summative end-of-course student questionnaires. We were asked to examine three aspects of student questionnaires and their use: methodology, research ethics, and human rights. Each of these areas has corresponding vectors of harm. For faculty members, these are experienced in performance evaluation, appointment status and

[4] Commission on the Reform of Ontario's Public Services, 2012; Office of the Auditor General of Ontario (OAG), 2012; OAG, 2014; Ministry of Training, Colleges and Universities, 2015; Ministry of Advanced Education and Skills Development, 2017. Three universities – Carleton, McMaster, and Western – adopted student questionnaire scores as an institution-specific metric in their 2017-2020 SMAs.

career progress, compromised confidentiality of results, and harassment. For students, the hazards are to effective teaching and learning, informed consent, and intellectual diversity.

Experts in each of the three areas are members of the working group, respectively: Jay Michela of the University of Waterloo on methodology, Pierre Boulos of the University of Windsor on research ethics, and Terezia Zoric of the University of Toronto on human rights. The OCUFA Collective Bargaining and Grievance Committees are also represented on the working group, by Jeff Tennant of Western University and Hannah Scott of the University of Ontario Institute of Technology. Professors Tennant and Scott have taken the lead in authoring this and the collective agreement reports, with the support of OCUFA Senior Research Analyst Russell Janzen. OCUFA Policy Analyst Sandra Smele provided research assistance.

We were asked to review student questionnaires and the policies and practices regarding their use at Ontario institutions. Available institutional policies, administrative guidelines, and survey instruments were collected toward that end. Appendix A outlines the collection. As documents that constrain the use of student questionnaires, collective agreements and memoranda of agreement between faculty associations and their institutional employers are part of this collection also. Unlike institutional policies and guidelines, a complete collection of these faculty agreements was available.[5]

We were also asked to survey faculty associations regarding the problems experienced by their members and ways of preventing or mitigating potential harm. Not all associations responded, and the results are confidential. Even if the response rate had been 100 per cent, it would not yield data which could be used confidently to identify patterns. Differences in collective agreement provisions and in association practices for managing case files and maintaining records mean the information gathered is more a categorized set of referrals to unique data sets and anecdotal evidence than an exhaustive report. This review acknowledges these limitations and makes no pretense to being comprehensive in this regard.

The review has benefited considerably from the insights and references distilled in reports by four experts, retention of whom was supported by OCUFA in two separate proceedings. These include Drs. Richard Freishtat and Philip Stark in the case of an interest arbitration between the Ryerson Faculty Association and Ryerson University, and Drs. Susan Basow and Frances Henry in an unpublished proceeding.[6] Between the expert reports and the literature reviews they included, responses to the survey of faculty associations, and the documentation reviewed by the working group, there is strong evidence pointing to the need to rethink how student questionnaires are used. In light of the mounting and credible evidence of harm, it is the working group's view that the burden of proof is on those who

[5] "Faculty agreement" is used in this report as a generic term for both collective agreements and memoranda of agreement between faculty associations and their institutional employers.

[6] Basow, 2018; Freishtat, 2016a, 2016b; Henry, 2018; Stark, 2016.

advocate the use of student questionnaires for the evaluation of teaching performance to demonstrate that student questionnaires are valid instruments of evaluation.

To be clear, we have not concluded that student questionnaires have no value. Rather, we assert that they can and should be useful for formative purposes. When they are used for summative purposes, however, they are methodologically not up to their assigned task as performance assessments, and are harmful to faculty members, student learning, and the teaching mission of universities. For those reasons, such summative uses should be discontinued.

Why student questionnaires?

The working group starts from the premise that student learning must be at the centre of why student questionnaires are used in the first place. It is thus axiomatic that the purpose of the questionnaires should be the formative development of instructors and information for the design and conduct of courses they teach. Indeed, these questionnaires' original purpose when introduced decades ago was for formative evaluation.[7] While questionnaire responses can be said to provide a summation of students' reflections at the end of a course, this does not mean they should be considered as "summative evaluation" in the sense normally attributed to this term in educational discourse.

Students' expressions of their perceptions in response to a course survey is a very different matter from evaluating an instructor's teaching performance for career progress decisions, or comparing "performance" using indicators from a common, standardized set of questions. It is the use of student questionnaire responses for summative performance evaluation of instructors that the working group sees as a departure from the logic of seeking feedback from students, and which needs to be called into account.

Teaching effectiveness and student learning: a faculty perspective

The working group recognizes that, for faculty members, student learning goes beyond training and education for employment. It is framed in part by the very nature of teaching as one part of an integral academic enterprise that also comprises scholarship and service. That is true not only for faculty members whose responsibilities are the classic trio of teaching, research and service, but also for teaching-intensive faculty who are expected to engage in scholarship and service. It is also true for those teaching on a per-course basis and who, practically speaking, are undertaking unpaid scholarly work and professional development which they bring to their teaching.

"Teaching excellence" in this context entails being student-focused, developing deep learning and critical thinking, and challenging students, sometimes with very sensitive topics. It may incorporate innovative teaching methods, but these are means to the end of student learning. Considered thus,

---

[7] Algozzine et al, 2004.

teaching excellence is not about being perfect or popular.[8] A realistic and responsible understanding of teaching excellence recognizes it cannot guarantee success for each and every student, but it is about *enabling* students to achieve the outcomes set out for the course and the program.

Where these standards are reflected in faculty association agreements with their institutional counterparts, it is in the provisions on rights and responsibilities in the area of teaching. The language commonly articulates values about the nature of teaching and the learning environment, for example: "scholarly competence and effectiveness in teaching"; fostering scholarly learning; academic freedom and inquiry; fair and ethical treatment of students; conscientiousness and currency in devising and updating courses. Provisions may also address practical expectations regarding duties associated with conducting and managing a course. Together, the provisions speak to the role faculty members have in creating the *conditions* for student learning, which depends also on what students contribute to the process.

Provisions in faculty agreements addressing student questionnaires speak less to those conditions than to setting out the role student questionnaires may play in evaluating teaching, which is acknowledged to comprise a breadth of activities and responsibilities. As academic professionals, instructors are concerned about their teaching effectiveness. They want to know how they are doing, and they take steps to ensure their courses remain relevant. To that end, student questionnaires have the potential to provide instructors with valuable information about student perceptions of the course and how they see it contributing to their learning. But this does not make these questionnaires evaluation instruments.[9]

Once student questionnaires are given the mismatched task of performance evaluation, there is an inherent risk that teaching strategies will be geared to attaining higher SQCT scores rather than aspiring to teaching excellence and maximizing student learning. Certainly, when compensation and job security are tied to scores, there is an incentive for instructors to choose the path of least resistance. Jay Michela's review (Appendix B) and the section on methodology in this report discuss this further. Suffice to say, the evidence suggests that effective teaching strategies and student learning consistent with the academic mission of universities are the casualties. The sections on research ethics and human rights discuss parallel concerns.

Student priorities

Outside of faculty members, students have the greatest direct interest in student questionnaires, and the strongest reasons to be skeptical about or disgruntled with the status quo. Irrespective of

---

[8] Excellence in teaching is not a bar all are expected to meet, except for promotion to top-most ranks in some cases; even then, a faculty member's portfolio contains much more than scores from student questionnaires.

[9] One of the earliest discussions of the working group led to a decision not to use the term "student evaluation of teaching" because it is misleading about what students are actually providing as feedback about their experiences. Questionnaires, on the other hand, solicit responses which provide data about student perception.

differences between students and faculty about what the nature of academic learning and effective teaching might be, student expectations with respect to student questionnaires do not come from nowhere. In the best of circumstances, one would hope for a convergent interest in the formative purposes of the questionnaires. At present however, the mismatch between the purported and actual benefits of using student questionnaires for summative purposes is only one reason students' voices might appear not to be heard. In addition to paying ever-escalating tuition fees, students have been facing steadily larger class sizes taught more and more by contingently employed faculty members. In such unenviable circumstances, it is no surprise that student organizations would highlight the need for effective questionnaires.

The Canadian Federation of Students (CFS) and the Ontario Undergraduate Student Alliance (OUSA) express dissatisfaction, even downright frustration, with the current practice of student questionnaires. And both are strongly supportive of student questionnaires for formative purposes. CFS has declared in its policy on instructor evaluations that they should be formative only. While acknowledging response bias and placing less emphasis on summative use than previously, OUSA supports further development of more robust survey instruments with a view to including quantitative metrics for SMAs.[10]

Lest there be any confusion, let us clarify that the working group is working from the assumption that faculty members and students both approach student questionnaires in good faith and that the problem lies with the institutional context, in which questionnaires can either be used appropriately to support quality teaching, or misused in a way that subverts the goal of teaching excellence.

Student questionnaires: formative or summative

Where the working group parts ways with other interested parties is with respect to their use for summative purposes, especially in light of changing circumstances. For us, the changing composition of the student and faculty populations, evolving technologies and forms of social interaction, and the availability of resources require a rethink about the consequences of using student questionnaires for summative evaluation. For advocates of summative questionnaires, these developments seem to be cause for doubling down on and generalizing the use of summative questionnaires, if proposals by the Drummond Commission and others to establish a set of common questions and incorporate the responses as performance indicators are any evidence.

Whether the growing interest in student questionnaires mentioned at the outset amounts to a rethinking or a redoubling is perhaps a matter of perspective. It strikes us as no small matter that the energy put into research on student learning and educational quality, of which assessment of teaching is a part, coincides with increasing commodification of higher education and accountability

[10] CFS, 2014; OUSA, 2014, 2015a, 2015b, 2018. CFS-Ontario does not have a policy specifically on student questionnaires; their policy on University Quality (2014) does, however and unfortunately, endorses a set of questions common to all instruments across the province.

requirements. It is for others to debate whether and to what degree the conception of quality in this context is a function of a neoliberal agenda for the economy and education. Much of the work by practitioners and specialists in teaching and learning is undertaken in good faith. We are concerned, however, that there are pressures to adopt conceptions of student learning and teaching excellence that are at odds with the nature of university education, or even redefine it to conform to different purposes that substantially erase its *sui generis* character.

The fault line can be drawn with the reduction, by successive governments, of per-student provincial support for university operating costs and the transfer of a greater portion of the fiscal burden directly to students and their families. In return, the provincial government substituted accountability criteria and performance indicators. None of this is unique to Ontario; the main difference is probably the rate of shift to private, namely household, sources of revenue.

If reduced government support is a necessary feature of the commodification of higher education, it is really the emphasis on university education as a labour market credential that completes the transformation. The effect is one thing for private sector employers happy to shift the effort of training to universities, and the costs of developing "human capital" to public institutions and, directly or indirectly, to individuals.

The effect is another thing when university credentials are presented as commodities exchanged for the price of admission. For universities, new or modified programs are a consistent feature of bridging academic missions with "real world" challenges, but the temptation and the logic of treating education as a commodity and students as consumers is to flip priorities from student learning to student satisfaction. The combined impulse is for government funders, and institutions in turn, to prioritize summative evaluations and quantitative measures over formative questionnaires and qualitative assessments of teaching ability.[11]

Institutional priorities

It is, of course, at the institutional level that student questionnaires are operationalized, whatever their form and content. Even apart from conditions for provincial funding and accountability, changes in the student population and available technology alone would have been sufficient cause for a re-evaluation of student questionnaires. As branches of universities, the Teaching and Learning Centres (TLC) that have been present on Ontario campuses for some time have a predictable interest in the future of student questionnaires as well.[12] As a pedagogical interest, it is one the working group believes lends itself to engaging with faculty members to make the most of formative evaluations. They may have another role in educating evaluators about the nature and role of formative questionnaires, and about the suite of tools for evaluating teaching, but their role in teaching development is compromised if they

[11] See Rytmeister (2013) for a concise comparison of different conceptions of quality and their implications.

[12] As with student questionnaires, TLCs go by different names at different institutions.

advocate for student questionnaires as summative evaluations – becoming an adjunct to the faculty relations and human resources administrative arms of the university.

It is the use of student questionnaires as proxies for evaluation of teaching performance that the working group sees as the most in need of critical analysis. There is no question that universities have an interest in the formative value of student questionnaires to further their academic mission. Nor is there any surprise that universities, as employers, seek to monitor employee performance. That much is recognized in, and constrained by, faculty agreements. For employers, the attraction of student questionnaires as an expedient "cost-effective" substitute for teaching evaluation is obvious.

In the name of cost containment, universities also outsource various functions, which increasingly include the administration of student questionnaires. Some of the most spirited defenders and prolific writers on student questionnaires also have stakes in firms that provide such third-party services. We see no point in questioning the motives behind institutional or third-party advocacy of summative evaluations; their interests are what they are. We do not presume their interests coincide with the interests of faculty and students, but we do question the ends to which student questionnaires are put and whether these questionnaires are up to the tasks set out for them.

The next section assesses whether student questionnaires meet the test with respect to adequacy. The sections that follow review their appropriateness in light of the obligations universities owe to faculty members and students, as employers and as guardians of the rights to dignity and equity.

Student questionnaires and methodology

The working group did not set out to review the survey instruments in use at Ontario institutions to see how well they measured up to the task set out for them. Apart from the scale of such an exercise and the obstacles to obtaining a suitable cross-section of samples, the prior question is whether student questionnaires can provide a valid basis for evaluating teaching performance as such. Are student responses about teaching effectiveness and can they be accepted as neutral, or are they sources of incipient bias? The stakes are high for faculty members' careers.

Student questionnaires typically include questions intended to provide formative feedback on course aspects and global summative questions using such phrasing as "Overall, the course…" Summative scores (which may combine responses to specific and general questions) are used in teaching evaluations for a variety of career steps depending on the institution. They include salary increments, often based on annual reports which include a faculty member's scores. There also is pre-tenure renewal, tenure and promotion for tenure-stream faculty, similar steps for teaching stream, and hiring and renewal for contract faculty.

The most common problem identified by faculty associations is the use of scores to assess how individual faculty members compare to some benchmark (e.g., mean, median) that is based on the

scores of the department or another comparator group.[13] As we will return to in the section on student questionnaires and human rights, the problem goes beyond the particular way in which the results are tabulated and reported. Whether the metric is mean, median, mode, or distribution, women, racialized, and LGBTQ2S+ faculty disproportionately receive lower scores.

Little of the research we compiled addressed other sources of bias – difference in physical ability, for example – but the gist is the same nonetheless: the potential consequences for individual faculty members is such that the methodological bar must be very high for any instrument used to assess performance, teaching included. We conclude that student questionnaires do not meet that test. We believe also that, in as much as their use inhibits the adoption of effective pedagogical models and the presentation of challenging content, summative student questionnaires do a disservice to students and their learning.

Student questionnaires: use and consequence

One level of contention about the methodology of student questionnaires takes the questionnaires as a given. The use of SQCT scores for comparative purposes currently is circumscribed in various ways at different institutions – prescribing which questions are to be used, specifying the scope, or defining the method of comparison. Within this frame, a typical problem is the misuse and over-interpretation of "averages" which, although colloquially conventional and superficially informative, are based on contested calculations because of the type of scales used.[14] The decline in response rates resulting from the transition to online surveys raises more questions about the reliability of the results. As if to underscore the point, a recent arbitration award found that the use of averages has no place in tenure and promotion decisions.[15]

As important as this award is, neither a commitment to valid forms of aggregation nor better response rates solve problems arising from the nature of the instruments themselves – nor can asking "better" questions. A different, and deeper, level of concern is reflected in debates about whether the student questionnaires used measure what their champions purport, and whether the scores for any faculty members are free of bias against the member. If student questionnaires are not effective instruments of evaluating teaching performance, they are not valid for comparative or summative purposes. Whether they are valid for formative purposes is a different question.

---

[13] Practice varies regarding whether or not student comments provided anonymously to open-ended questions on surveys are included in the file of the member being evaluated. These too can prove problematic.

[14] Student questionnaires typically use categorical-ordinal scales for which averaging, and comparing unlike items, is problematic. See, e.g., Boysen (2015) on administrative propensities in this respect. Whether it is averaging a single omnibus summative question or trying to produce a single, global expression out of multi-dimensional attributes across different questions, neither is appropriate. Stark and Freishtat (2014), Freishtat (2016a) and Stark (2015, 2016) offer pointed methodological criticisms of the practice.

[15] Ryerson University v Ryerson Faculty Association, 2018 CanLII 58446 (ON LA), <http://canlii.ca/t/hsqkz>, retrieved on 2018-06-28.

One of the principal criticisms of student questionnaires is that students, not being teachers themselves, are not qualified to evaluate teaching. Students may be experts on their *own* experience in the classroom and with the course, and well-designed questionnaires conceivably can provide valuable feedback for formative purposes. There is no question about that. But students are not situated to assess how appropriate the materials are, how well the subject is being covered, or if course objectives are being met.[16] The evidence suggests that student questionnaires do not measure teaching effectiveness.[17] Even some of the more enthusiastic advocates of summative end-of-course student questionnaires acknowledge they are registers of satisfaction rather than direct indicators of student learning or teaching effectiveness.[18] Not surprisingly, student grade expectations are strongly correlated with scores.[19]

As Professor Michela notes in his review, there is a wide array of other factors which have nothing to do with teaching effectiveness but so confound the results that it is essentially impossible to determine the degree to which SQCT scores do correlate with actual student learning.[20] Influences include those that have nothing to do with the qualities of the instructor – the time of day, the physical environment, subject discipline, whether the course is required or an elective, etc.[21] That alone throws doubt on the veracity and value of the scores, but there are also the effects of bias based on instructors' gender, race, age, sexual orientation, and other characteristics. These biases may themselves vary by differences within the student population.[22] This alone should be sufficient cause to relieve student questionnaires of their summative role.

Defenders of student questionnaires claim that *some* more or less significant correlation between scores and teaching effectiveness and student learning can be found amidst the data noise. They might

---

16 Bjork et al, 2013.

17 Reviews making this point (e.g., Nilson, 2012; Stark and Freishtat, 2014; Stark, 2015; Stroebe, 2016; Uttl et al, 2017) cite studies examining the relationship between questionnaire scores and student performance (e.g., Boring et al, 2016; Braga et al, 2014; Carrell and West, 2010; Clayson, 2009; Weinberg et al, 2009).

18 Hativa, 2013, 2015.

19 Boring et al, 2016; Short et al, 2008; Sinclair and Kunda, 2000; Vaillancourt, 2017; Vasta and Sarmiento, 1979; Worthington, 2002. On perception of instructor "fairness" also, Wendorf and Alexander, 2005.

20 Expert reports from Basow (2018), Freishtat (2016a), and Stark (2016) provide very instructive summaries and references supplementing Professor Michela's review, a number of which are cited in this and the next section.

21 Basow, 1995; Basow and Montgomery, 2005; Bedard and Kuhn, 2005; Boring et al, 2016; Cranton and Smith, 1986; Feldman 1978, 1984; Hill and Epps 2010; Monks and Schmidt 2010; Uttl et al 2013; Uttl and Smibert 2017. Cf. d'Apollonia and Abrami, 1997; Marsh and Roche, 1997; McKeachie 1997.

22 On gender: Arbuckle and Williams 2003; Basow, 1995, 2000; Basow et al, 2006; Boring, 2015; Boring et al, 2016; MacNell et al, 2015. Ethnicity and race: Boatright-Horowitz and Soeung, 2009; McPherson and Jewell, 2007; Reid, 2010; Smith, 2007, Smith and Anderson, 2005. Gender and race: Basow et al, 2013; Huston, 2005; Storage et al, 2016. Sexual orientation: Ewing et al, 2003. First language and accent: Ogier, 2005; Subtirelu 2016. Age; Levin, 1998. Gender and age: Bianchini et al, 2013. One study (Ambady and Rosenthal, 1993) records just how quickly impressions can be made and registered in questionnaire responses; others highlight the role of perceived physical attractiveness: Campbell et al, 2005; Gurung and Vespia, 2007; Hamermesch and Parker, 2005; Rinolio et al, 2006; Wolbring and Riordan, 2016.

also concede there is evidence of gendered or racialized bias, yet contend it is not a significant one. Setting aside questions of statistical significance versus effect size, at best these can only be arguments that student questionnaires might be useful for formative purposes for individuals. With respect to bias, wherever the bar might be set to argue there is a significant statistical correlation between SQCT scores for measures treated as proxies for teaching effectiveness, the standard we must apply with respect to bias is quite different: *any* correlation with bias should invalidate the use of such scores for summative purposes.

Even supposing the SQCT scores of women or racialized faculty are not discernibly different from those of their white male colleagues, the absence of evidence of bias is not evidence of the absence of bias. Without considerably more data than is typically collected in the course of student questionnaires, there is simply no way of knowing whether those same women or racialized faculty would not have received higher scores had they not been women or racialized. Thanks to an experiment based on online teaching, however, we do know that the same instructor in the same course can receive very different scores depending on the gender they identify to their students.[23] We return to the issue of bias in the section on student questionnaires and human rights.

Second, as far as any correlation between SQCT scores and teaching effectiveness is concerned, sorting out the bias is difficult enough even for formative purposes, and certainly beyond what would be required for comparative use. The "halo effect" – when a global evaluation about an instructor drives responses to more particular questions, despite having sufficient and relevant information to assess or comment on the specific question – also skews scores.[24] Whether the halo effect in any given set of responses is due to bias (unconscious or not), or simple like or dislike, it affects results even for questions on such ostensibly neutral matters as length of time in returning assignments.[25] Answers on notionally specific questions essentially stand in for a more general statement about students' satisfaction.

There is no guarantee that responses even to questionnaires intended for strictly formative use will be free of the various biases or the halo effect. And the halo effect may be magnified by expecting student questionnaires to do double duty as formative and summative instruments, depending on the survey items selected. Questions soliciting responses on students' "overall" assessment, which are often favoured by administrators for summative purposes, may be most susceptible to bias.[26]

As Professor Michela demonstrates, there is no way to know if the particular set of scores for a specific individual is accurate, and no way to correct for bias even knowing it is present. There are simply too

---

[23] MacNell et al, 2015; Boring et al, 2016.

[24] Nisbett and Wilson, 1977; see also Feeley, 2002, and Keeley et al, 2013.

[25] Boring et al, 2016; Sproule, 2000.

[26] Worthington, 2002.

many variables to be taken into account for the results to be comparable across a department, let alone a faculty or broader unit. He describes the effect of the biasing factors as one of "scrambling" the relative positions, "best" to "worst," of the individual instructors where student ratings are concerned. It is these scrambled results that are seen by review committees. In the circumstances, the only suitable comparator for an individual is their self in each of their own courses over time.

Eliminating summative use for end-of-course questionnaires does not spell the end of formative versions. Student questionnaires will still be useful, and necessary, for instructors to assess changes in course content or delivery responses by comparing whether students' responses to pertinent questions are more or less favourable from one year to the next. For example, someone's courses may use different pedagogical methods that students are generally inclined to like to a greater or lesser degree. One could seek to enhance aspects of a particular course and hope to see a change in students' ratings *in that course*, *for that aspect*, after implementing a change in course delivery.

Student questionnaires and student learning

The damage caused by student questionnaires used for summative purposes is not simply to faculty members' career trajectories; it is also to teaching strategies and student learning. That the effect of biases can yield a negative correlation between SQCT scores and teaching effectiveness should be concern enough in itself.[27] Yet there is also reason to think twice about the implications for teaching innovation, intellectual diversity, and academic standards.

Some time ago, the provincial government adopted a public policy objective of encouraging teaching innovation. On the face of it, this was to take advantage of new technologies and evolving pedagogical techniques to improve and sustain teaching quality. As Professor Michela outlines, the prospect that instructors will receive lower scores in classes that incorporate teaching methods shown to be more effective for student learning is a disincentive to innovation of this type. On the one hand, students may believe they are learning less in a hitherto non-traditional model.[28] On the other hand, students can be resistant to innovation and express their opinion in or through student questionnaires.[29]

Student resistance, not just to instructional form but to course content, represents a different kind of challenge. To reiterate, in addition to the cultivation of deep learning, teaching excellence entails the development of critical thinking and often addresses challenging subjects. Even if the fear of being punished via student questionnaires does not lead faculty members to compromise their academic freedom and decide against raising contentious subjects in their classrooms, apprehension about

---

[27] Boring et al, 2016.

[28] Lake, 2001.

[29] Ellis, 2013, 2014.

the possible fallout exists nonetheless.[30] Resistance or hostility is manifest most obviously in student comments which attribute bias or a non-academic "agenda" to the instructor.[31] It is not a stretch to infer that the accompanying scores will reflect those sentiments.

Considering the weight that simple likability carries in student questionnaires, there are reasons for faculty members to find ways of boosting their SQCT scores when student questionnaires are used for summative evaluation. The incentive is greater for full-time faculty with teaching-intensive appointments and for contract faculty whose contract status is heavily dependent on the scores, but it is present for any faculty member with teaching responsibilities. Some strategies to put students in a positive and generous state of mind do not necessarily undercut teaching.[32] However, grade inflation, sometimes abetted by the behaviour of academic administrators keen to mollify disgruntled students, is a different matter.[33] If the objective is to dispense credentials as expeditiously as possible, grade inflation is one strategy, but it is contrary to the sense of academic education and teaching excellence set out in the section on teaching effectiveness and student learning from a faculty perspective.

Other uses of student questionnaire scores

An advocate of end-of-course summative student questionnaires might concede each of these points, and still defend their use for other purposes. SQCT scores are reported in course, curriculum, and program reviews, for example. Scores also might be published for members of the university community at large to read, the level of aggregation, type of data, and breadth of access varying by institution. For good reason, SQCT scores do not figure prominently in program reviews undertaken in accordance with quality assurance processes mandated by the Ontario Universities Council on Quality Assurance.[34] To the extent they are considered, *en passant*, they are a single indicator overshadowed by more substantive methods of evaluation.

There are reasons to treat even these apparently innocuous uses and aggregations skeptically. First, the problem of having questionnaires do double duty is simply repeated. Far from respondent biases cancelling each other out, the prospect increases that they are ramified. The student survey instruments vary with respect to whether and how questions about instructors and courses are distinguished, but the most effective way of ensuring that bias does not spill over from one set of questions to another is to administer different questionnaires at different times.

---

30 Boatright-Horowitz and Soeung, 2009; Schueler, 1988; Williams and Ceci, 1997.

31 On contentious topics, or student claims of female, racialized or queer instructor bias, or "pushing an agenda": Abel and Metltzer, 2007; Anderson and Kanner, 2011; Anderson and Smith, 2005; Bilimoria and Stewart, 2009; Dua and Lawrence, 2000; Littleford et al, 2010; or more likely to receive personal attacks, Perry et al, 2015.

32 Serving cookies for example: Hessler et al, 2018.

33 Ewing, 2012; Hativa, 2013; Isely and Singh, 2005; Krautmann and Sander, 1999; McPherson, 2006; Stroebe, 2016; Vasta and Sarmiento, 1979.

34 Ontario Universities Council on Quality Assurance, 2016.

Second, and most critically, for any proposal to use aggregated SQCT scores as a basis for resource allocation between programs or as performance indicators to compare universities, the evidence should be based on appropriate populations of respondents. Simple aggregation of scores includes the responses of students who are not enrolled in the program; on the other hand, it double counts students who are in the program. This implies, of course, that there will be different questionnaires for courses and instructors, programs, and institutions.

As far as we were able to determine, in no other province are student questionnaire scores identified, let alone used, as an indicator for purposes of reporting to the provincial government. Instead, random sample surveys of graduates appear to be the norm. If that is not sufficient, creating robust instruments for the right populations and collecting the data is not a cheap proposition. If the data are to be relevant and meaningful, a commensurate additional investment on the part of the provincial government should be forthcoming.

Similar considerations suggest that the publication of SQCT scores for public consumption is of dubious value, even if the reading audience were limited to members of the university community – most notably students. Where scores for instructors or courses are published, respondent bias alone renders them unreliable as guides for selecting courses. Unless rigourous standards of confidentiality are met, publishing the scores for courses without identifying instructors is at best a fig leaf. Regardless of the degree of confidentiality, the use of SQCT scores as an exercise in consumer choice can compound the discrimination against an instructor. In as much as it undercuts the future of the courses themselves, it runs the risk of generating an intellectual conformity that is not consistent with the spirit of the academic enterprise.

Methodology – looking forward

As these considerations suggest, the working group is persuaded that student questionnaires should not be used summatively in the evaluation of teaching performance. We concur with Ontario Arbitrator William Kaplan's declaration that "a high standard of justice, fairness and due process is self-evidently required."[35] Student questionnaires fall short on that score, sufficiently so that we would go further than Kaplan to recommend student questionnaire results not play any role in reviews of faculty teaching, except at and in the manner of instructors' choosing.

Even student questionnaires designed and administered for formative purposes are not without challenges presented by respondent bias and the halo effect. Substituting an alphabetical rating system for a numerical one will not change that. Nor will reporting only the distribution of responses eliminate the prospect that respondent bias will not factor into reviewers' assessment of a faculty member's teaching. One might expect that graphs or tables illustrating the distribution of responses for each question or a substantial cross-section of them will be sufficiently varied that no simple picture

[35] Ryerson University v Ryerson Faculty Association, 2018, p. 4.

emerges.[36] The halo effect alone mitigates against that possibility. Add to that the already-mentioned propensity of administrators to rely on benchmarks – here it could be a notionally acceptable distribution – and the result is a different iteration of a familiar problem.

Student questionnaires cannot avoid reflecting degrees of student satisfaction. Important as it is that students are given the opportunity to communicate their impressions of a course to the instructor, the principal questions in teaching evaluation are about teaching effectiveness and student learning. Including the scores from student questionnaires redirects the focus to student satisfaction. If not directly, setting student satisfaction above student learning also comes into tension with academic freedom and intellectual diversity.

If, as suggested above, the value of formative questionnaires for individual instructors is in the responses to particular questions about specific courses over time, it is not clear to us what the value of knowing the distribution of the scores would be to reviewers in any case. As with other modes of representing the results, the numbers do not speak for themselves. They do not move us away from over-interpretation, or emphasizing positive responses and defending against others, which follow summative SQCT scores. Student questionnaire results require understanding of context, and the more pertinent information must surely be a faculty member's narrative reflecting on the scores and the implications for their own courses and teaching.

We are not proposing that student questionnaires be eliminated. Formative questionnaires should be valuable sources of feedback from students to instructors. Although the we are mindful that "better" questions solve none of the problems of respondent bias and the halo effect, and even though we are not in a position to propose questions for formative use, we do recognize that design matters.[37] The next two sections will address other considerations that must be taken into account whichever instruments are devised.

Student questionnaires and research ethics

The attention of the working group in the research ethics of student questionnaires was guided less by questions about whether any of the instruments in use in Ontario had been subjected to ethics review than by whether they are being conducted in a fashion that is ethically responsible to students, and to faculty members.[38] The practice of conducting student questionnaires, and some of the surveys in use, predate current sensibilities about research ethics. In light of the methodological limits of student

---

[36] It would go against the grain of Arbitrator Kaplan's award to choose one or two questions to be "representative" or to compile the results into one or two representations.

[37] Freishtat (2016b) provides illustrative critiques of questions similar to many in use however.

[38] The working group's survey of faculty associations asked if the student questionnaire at their institution has been expressly exempted from ethics review, to which there was one affirmative response.

questionnaires, changes in the way surveys are administered, potential other uses of the data collected, and the legal environment, there are multiple reasons to revisit consent and confidentiality.

It seems necessary also to return to the beginning, to the role of student questionnaires in facilitating communication between students about their experience to faculty members who can put that knowledge to use in making their course(s) a more effective teaching and learning experience. If the volume of research on re-engaging students to respond to online questionnaires and convincing faculty of the value of the questionnaires is anything to go by, dissatisfaction with the current state of affairs with summative evaluations, and alienation of both students and faculty from the process, has become commonplace. For us, this is about faculty and students as subjects, rather than as objects or bystanders with little or no agency of their own. We believe formative questionnaires are considerably more conducive to that type of active engagement than summative versions.

Most broadly, we are persuaded that student questionnaires must conform to high ethical standards. No research method or instrument should result in harm to anyone, and even then a prior test of methodological soundness is sensible lest the research tools magnify the risk. As things currently stand with student questionnaires, the application of a research ethics review in itself can neither validate the method nor repair methodological limitations. The ethical questions, in any case, are not simply about the nature of multiple-choice survey instruments; there are additional sensitivities with respect to written responses to open-ended questions. These apply whether a response alerts a faculty member to a student who is at risk, or if it amounts to harassment of a faculty member, or worse.

Research ethics and human participants

The 2014 *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (TCPS2) is the most applicable benchmark. It applies to any research conducted at any institution eligible to administer funding from one of the funding councils, including that undertaken by non-faculty researchers, such as institutional analysis and planning personnel.[39] It starts with the presumption that the research it covers can pass peer review muster. TCPS2 provides an exemption from research ethics review requirements for performance assessments, provided the data are not used for secondary research however.[40] "Student course evaluations" are named as an example of an acceptable exemption.

For research that falls outside the scope of research for which ethics review approval is required, the TCPS2 suggests recourse to other sources of ethical guidance or "best practices guidelines." As far as we have been able to determine, there are no established professional ethical guides for university

---

[39] The Tri-Council is comprised of the Canadian Institutes of Health Research (CIHR), Natural Sciences and Engineering Research Council of Canada (NSERC), and Social Sciences and Humanities Research Council of Canada (SSHRC). A review of research ethics policies indicates the norm is to refer to the TCPS2 and to state the scope extends also to non-academic staff.

[40] See Appendix C.

institutional analysis or for teaching and learning specialists. A (non-exhaustive) review of publications and conference presentations in the institutional analysis community,[41] and of some teaching and learning specialist web-sites, suggests that practitioners in these fields unproblematically take student questionnaires for granted, and discuss best practices more in terms of securing better response rates and providing proper training for interpretation of the results of student questionnaires.

Best practice with respect to the ethics of student questionnaires is essentially common practice. From the review of the policies, administrative guidelines, and the survey instruments, it seems clear that the primary concern has been to protect the anonymity and confidentiality of student respondents. Typically, these procedures include ensuring that: responses, including comments, are anonymous; the faculty member is not present while the questionnaire is being completed; responses are not released to the faculty member until after final grades have been submitted. Other than advice to students that the results may factor in personnel decisions, were this applicable, and that they should give considered responses, there is little that addresses the role of instructors as human subjects.

The position of faculty members and their status as researcher and human participant is idiosyncratic. For example, when they conduct their own student questionnaires for formative purposes, there is a reasonable expectation (reflected in some collective agreements) that they will protect the privacy of students. For all that university-mandated student questionnaires are supposed to yield results that are informative for faculty members, it is the institution that assumes the role of researcher, most forcefully when it imposes a summative character on the research instrument. And while faculty members are not human participants in the sense that they are respondents, the data generated surely are about them. Especially when student questionnaires are used for performance evaluation, instructors are the ones who bear the consequences, some more so than others. Neither of these unique circumstances of "researcher" and "participant" is contemplated by the TCPS2, but there is no reason to limit application of its core principles – respect for persons, concern for welfare, and justice – only to those circumstances it does consider.

Consent and confidentiality – students

Concerns about consent and confidentiality for students is more obvious and pressing with the increasing use of online surveys and third-party providers. Due to limitations of the documentation the working group was able to assemble, it is not possible to know what safeguards any institution has established to protect student confidentiality, much less how well they might work. We attended less to data security, which is virtually a universal hazard and about which there are overlapping concerns for students and faculty, and more on consent as a point of departure. In a sense, consent comes before confidentiality.

[41] Annual conferences of the Canadian Institutional Research and Planning Association, for example.

There are two tracks in our considerations, one of which pertains to the secondary use of data collected through the student questionnaires. Secondary use includes:

- what a third party commercial provider might make of the data;[42]
- research for academic purposes, and;
- linking individual students' responses to other student-specific data collected through other electronic platforms hosted by the institution – the development of predictive analytics, for example, which might be undertaken by institutional analysis researchers for student learning services.[43]

Whether or not the TCPS2 guidelines for secondary use and data linkage actually apply, and would require that any of these uses be subject to formal research ethics review, it seems to us the prospect of secondary use should require institutions to advise students of that possibility and seek their consent. As discussed below, there are other considerations that make active, continuing, and informed consent, as prescribed by the TCPS2, a prudent step. We believe the stakes are such that implied consent is not enough.

Without access to all the documents, instruments, and platforms where consent could be sought, we cannot make a definitive statement about whether any institution meets the ethical bar, as set by the TCPS2, for participants' consent. No purpose is served by outing any institution: we are more interested in identifying areas for improvement of which all institutions should be mindful. And if what is available is indicative, there is work to be done to catch up with the twenty-first century. It is common enough to advise students that the results will be used to improve teaching and the course, to evaluate programs, and in personnel processes. They are also advised of the steps to protect them from reprisal.

It is not obvious that active consent at the point of filling out the survey is the norm. If consent must also be continuing, it cannot be sufficient for consent to have been given when a student registers for the term or registers for electronic services, even if it includes an omnibus agreement to the university's privacy policy. It might be one matter if it were simply students agreeing to have their SQCT responses linked to other learning-relevant data they generate. Even then, there is an argument to seek active consent because of the time delay between registration and student questionnaires. These questionnaires are consequential for and specific to faculty members as well, though quite differently from the data footprint left in the course of a student's learning and general usage.

---

[42] Third party providers that offer pre-packaged student questionnaires are in a position to aggregate data of multiple institutions and proffer (invalid and spurious, in our view) analyses using external comparator groups and "benchmarks."

[43] TCPS2 also requires research ethics review where the proposed research entails linkage between data sets, although institutional analysis practice in this regard may not be subject to research ethics review, in as much as it is not "research" and perhaps exempted as "normal educational requirements."

The TCPS2 requires that consent must be not only active, but also informed. This is the second track we considered. In addition to being advised of the purpose of the research, how the data will be used (including the possibility of secondary use and data linkage), and how their identity will be protected, etc., participants are also entitled to know whether there are any circumstances in which their identity may be disclosed and to whom the disclosure may be made. Students need to know that anonymity sometimes must yield to confidentiality.

The TCPS2 recognizes professional and legal obligations for researchers to advise appropriate authorities when there is a risk of harm to participants or third parties. Tracing the identities of student respondents is more feasible with online questionnaires, but not impossible when in-class, paper questionnaires are used.[44] Instructors are not in a precisely analogous situation as researchers and professionals in this context, but it might be appropriate to initiate a trace for the identity of a student who has disclosed in their anonymous comments that they are the target of abuse, for example.

Instructors most certainly may be the target of harassment via anonymous comments on student questionnaires. And, whether they are participants or third parties, they do have legal rights to workplaces free of harassment under the *Occupational Health and Safety Act* (OHSA) and the *Ontario Human Rights Code* (the "Code"). The OHSA requires Ontario employers, universities included, to have policies on workplace harassment and violence and procedures for addressing them, including the conduct of investigations. To investigate an instructor's complaint of harassment, then, some means of following up is necessary. Anonymity of student respondents should not be an impediment.

We return to the issue of harassment in the section on student questionnaires and human rights. For present purposes, it is necessary to emphasize that students need to know, for example, that if harassing or threatening comments are made under the cloak of anonymity, and a complaint is made, the investigation will be confidential, but their anonymity is no longer protected. Students need to be given more than just advice enjoining them to treat the questionnaire seriously and responsibly.

Consent and confidentiality – faculty

Consent and confidentiality figure differently for faculty members. Those with tenure-stream and similar appointments participate in the collegial governance of their institution, and almost all faculty members and instructors are members of associations or unions that negotiate the terms and conditions of their professional working lives at the institution. There is also legislation which bears on the position of faculty members in ways that apply not at all, or differently, to students. As a practical matter, the precise constellation of these factors varies from institution to institution, faculty agreement to faculty agreement.

---

[44] Maintaining anonymity for students as far as the instructor is concerned, for example, while retaining student identifiers under the care of a designated custodian and according to appropriate protocols.

As a matter of principle, research ethics requires compliance with the relevant legal frameworks, but they also may demand higher standards than the law. An arbitrator might accept the argument of a university employer that it could implement a policy of mandatory student questionnaires because the policy had been passed by Senate,[45] for example, but it does not follow there is a general principle of implied consent. The mere fact of faculty association dissent contradicts such a notion. Nor would it make sense to claim implied consent for those who are not represented on Senate – typically contract faculty. Similarly, a court decision that SQCT scores are not considered private information for the purposes of the *Ontario Freedom of Information and Protection of Privacy Act* (FIPPA) is a ruling on what is permissible in a specific configuration of facts;[46] it does not suspend confidentiality as a matter of general ethical principle.

In personal terms, consent is less about granting consent to the undertaking of student questionnaires that are institutionally mandated than it is about motivation for simply acceding to student questionnaires or fully participating in them in the first instance. Consent then is more about the release of information gleaned from the responses. As we suggested at the outset, faculty members have an interest in student feedback. At one level, they have a reasonable expectation that the instrument used for that purpose be adequate to the task, i.e., that it is methodologically valid. On a different level, it makes little sense to suggest that anyone would consent to something that may have coercive consequences for them.

Design and use of student questionnaires for summative purposes essentially renders faculty members as objects, at best as subjects who must submit to the exercise. By contrast, and in much the same vein as occurs for faculty members who design and administer their own questionnaires, student questionnaires undertaken for formative purposes treat the faculty member as a subject who is participating, albeit differently, together in the exercise with their students. In either case – one in which student questionnaires are not valid for comparative purposes and another in which they are formative in nature – the responses should matter to no one other than the faculty member, and perhaps those competent to help interpret them and inform teaching strategies. Confidentiality is the real upshot of consent for instructors: disclosure should be at their discretion.

The working group acknowledges students' desire to know that their feedback is being heard and taken seriously.[47] How to signal student responses are being taken into account is not a simple matter. If the results are used for summative purposes, any consequences are a matter of personal employment

---

[45] Mount Allison University v Mount Allison Faculty Association, 2015 CanLII 94980 (ON LA), <http://canlii.ca/t/gp3c6>, retrieved on 2017-06-29.

[46] University of Windsor and University of Windsor Faculty Association (Policy grievance), (L.A., 2007-02-19), SOQUIJ AZ-51135637; University of Windsor Faculty Association v. University of Windsor, 2008 CanLII 23711 (ON SCDC), <http://canlii.ca/t/1wzvg>, retrieved on 2017-06-26.

[47] OUSA 2014, 2015a, 2015b, 2018. The 2017 edition of *Policy Paper: Accountability* is different from the 2014 version, and does not refer to student questionnaires.

and not for public consumption. And, in addition to the point made in the previous section (that SQCT scores are not a reliable guide for choosing courses), publicizing the scores offers no information about whether an instructor has made any changes in response to student responses, much less what kind of changes. The scores indicate even less about their plans for the next iteration of a course, least of all when using methodologically compromised questionnaires. The risk that published scores will influence later iterations and become self-fulfilling prophecies diminishes their value even further.[48]

Much of the answer, it seems to us, turns on conducive conditions. One is simply to have a student questionnaire that is methodologically valid (as nearly so as possible), formative in nature, and actually addresses factors that are under the control of the instructor and relevant to the course and teaching. Another is to have sufficient resources so that faculty members have the time and support to invest in adapting their teaching and courses as appropriate. Yet another is to eliminate disincentives against instructors taking student questionnaire results into greater account: treat scores as illustrative and informative, not comparative and evaluative. Voluntary disclosure makes questionnaire responses and faculty reports of them more compelling.

While the first two conditions are similarly advocated for by student groups, the case for voluntary disclosure may be no more satisfying for some students than "trust us." Still, it is in keeping with the principle of procedural fairness faculty agreements are intended to embody. Whether it is in the course of reviews at prescribed times or at other occasions when a review is prompted, how student voices are heard in the course of teaching evaluation is subject to the terms of faculty agreements. Where agreements include provision for reporting on student questionnaires, it is one of several elements of teaching evaluation: outcomes of the evaluations cannot be predicted by questionnaire results alone. In principle as well, outcomes are not guaranteed, but the process must be fair to the faculty member being assessed. This much is implied in the TCPS2, even in exempting student questionnaires from research ethics review.

If student questionnaires are intended to be formative, it makes little sense to us that their impact should be measured by whether an instructor is penalized or rewarded. That kind of logic implies a transaction between faculty and students in which notional learning outcomes are delivered by instructors to students, rather than a relationship built on teaching and learning as active behaviours on the part of both parties. It is the latter which evidently forms the basis for a consensus that mid-course formative feedback from students to faculty is good practice because it fosters goodwill, in addition to its pedagogical value. It is for those reasons that it makes sense to ensure that the necessary resources, opportunity, and support are available for instructors to engage in the practice.

---

[48] On the influence of Ratemyprofessor ratings on subsequent scores, see, e.g., Legg and Wilson, 2012.

Research ethics – looking forward

The working group believes that returning student questionnaires to a strictly formative purpose is consistent with the intent of research ethics. In one sense, doing so restores a two-way communication between students and instructors, without the intervention or oversight of an external body or third party imposing its own narratives on the exchange. The parallel might be similar to the relationship between a researcher and respondents, were it not for the role of the institution, for practical purposes, as researcher. In the context of student questionnaires, however, we think it is the institution which owes ethical obligations to the human participants, namely students and instructors.

Although we have not concluded that student questionnaires should be required to pass review by universities' research ethics boards, we do believe that doing so would assure that certain critical issues are addressed, and it should reassure students and faculty alike that it is undertaken by a disinterested third party. Research ethics review cannot repair the methodological deficiencies of student questionnaires, and the working group certainly would not suggest that questionnaires should fail such review on that basis: student feedback is too valuable to be dismissed for formal reasons. If not reviewed by an ethics panel, the questionnaires nevertheless should be administered according to the principles and standards articulated by the TCPS2.

Faculty and students should have confidence in the security measures undertaken to protect the data collected and their identity, and that any use that might be deemed third-party use is conducted according to the TCPS2.[49] Consent is a critical threshold in this regard. For students, this means being advised not only of the possible uses of their responses and the measures to protect their privacy, it requires that they be informed of the difference between anonymity and confidentiality and the circumstances in which confidentiality prevails. For instructors, as we have suggested, consent is more about the scope of confidentiality, which should be at their discretion.

For faculty members, it must be underlined, research ethics is not a replacement for protections secured in faculty agreements and the law. Research ethics does underscore the expectation that participants, including faculty members, are treated with human dignity, concern for their welfare, and equity and justice – with respect for human rights. It is to human rights we turn in the next section.

Student questionnaires and human rights

Methodology and research ethics converge in the working group's review of the use of student questionnaires with regard to human rights. To restate the obvious, the review of methodology shows that women, racialized, and LGBTQ2S+ faculty receive lower SQCT scores than their able-bodied, cis-

[49] Collective agreements offer a range of relevant protections for faculty members, but there are common issues for instructors and students. Among them are: confidentiality protocols covering identification of persons eligible for access; criteria for access, security measures and hacking prevention; data management, including location(s) of storage, numbers of copies stored, management of multiple entries, traffic between sites (including provisions to avoid the United States); length of time for data retention, and provision for data deletion.

gender, white male colleagues, and thus are at greater risk of negative career consequences whenever student questionnaires are used summatively for performance evaluation. They are also more likely to be the target of abusive comments.[50] Research ethics is based on the principle that every person is owed respect to their dignity and equitable treatment as an individual.

Human rights are not on a parallel plane to methodological and ethical principles: there are legal obligations under the *Ontario Human Rights Code* and the OHSA, and complementary provisions in faculty agreements.[51] The *Code* stipulates equal treatment and freedom from discrimination, including "constructive discrimination" in employment.[52] It bears repeating that special pleading about the statistical significance or effect size of bias is irrelevant in this light: any bias is unacceptable. The *Code* and the OHSA also establish rights to freedom from harassment in the workplace, and the OHSA requires employers to have policies and procedures in place to give substance to this right. It speaks volumes that both statutes also include specific and particular reference and provisions addressing sexual harassment.

There are three corresponding angles to our considerations with respect to human rights – freedom from discrimination, equity, and freedom from harassment. The first is concerned with harm arising from the use of student questionnaires compromised by bias in personnel decisions at an individual level. The second casts a wider net, with a view to the iterative and systemic effects of using student questionnaires for use in decisions on pay and employment; if they were used only for formative purposes, the prospect of inequitable results would hardly arise. The working group sees a parallel challenge to the diversity of academic content. The third angle is the conditions enabling harassment, and its implications for the learning and working environment and the general intellectual climate of a university.

It is clear to us that the summative use of student questionnaires does not meet the bar set by the *Code*, and that their use in deciding matters of pay, appointments, and career progress is a form of systemic discrimination. We also have questions about whether employers are meeting their obligations under the *Code* and the OHSA to minimize the risk to faculty members of harassment, to notify faculty of their rights in such circumstances, or to deal effectively with it when it does occur. We further believe that using student questionnaires for summative rather than for formative purposes poses risk to a diverse and healthy academic learning and working environment.

---

[50] For research on this in the context of student questionnaires, see National Tertiary Education Union, 2018.

[51] See Appendix D for excerpts from the Code, Appendix E for excerpts from the OHSA.

[52] Also known as "adverse effects discrimination," it is defined by the Ontario Human Rights Commission (2013) as "a rule or practice [which] unintentionally singles out a group of people and results in unequal treatment." http://www.ohrc.on.ca/en/part-ii-–-interpretation-and-application/constructive-discrimination

Student questionnaires and discrimination

Where the review of methodology draws on an abundance of scholarly research, and research ethics compares existing documentation with an independently set standard, evidence of injuries to human rights on which the working group could rely includes actual complaints by affected individuals. The survey of faculty associations provided responses about student questionnaire-related complaints and grievances that are also about discrimination and/or harassment. As was indicated in the section on the background and scope of this report, differences in collective agreement provisions and in case management make it difficult to make authoritative statements about the number and rate of occurrences. It is evident nonetheless that faculty associations discern a link between student questionnaires and discrimination and harassment, and are making use of the protections defined by the *Code* and their collective agreements to defend their members against personal injury arising from bias in student questionnaire results.

All cases involving individuals are confidential, of course, and published decisions about cases of this nature that make it to the arbitration or tribunal stage are few. There are reasons defenders of student questionnaires are no more justified in claiming that the findings to date – or lack thereof – are in their favour than critics of student questionnaires. These cases often deal with wider issues than just student questionnaires. They may be settled before reaching arbitration or, when these cases are heard, they may be decided in ways that do not necessarily turn on issues of bias in questionnaires. And, even in cases where discrimination is the heart of the matter, student questionnaires are often just one part of the evidence.

So far as we are aware, there are no rights cases in Canada in which a decision has been rendered on the question of whether student questionnaires are tainted by bias and therefore discriminatory. Whether simply because a mounting number of cases establishes a pervasive pattern or because faculty associations can draw on existing research to establish endemic respondent bias, as the Ryerson Faculty Association did in its interest arbitration,[53] it is impossible to avoid the prospect that the summative use of student questionnaires for performance evaluation results in systemic discrimination. The bar for determining whether it is discriminatory to use SQCT scores in decisions affecting the pay and career status of any single member is not whether the scores were the sole or determining factor; it is whether the scores played a role and, given the problem of respondent bias, were discriminatory in effect.

Student questionnaires and systemic discrimination

The Ontario Human Rights Commission defines "systemic or institutional discrimination" as "policies or practices that appear to be neutral on their surface but that may have discriminatory effects on

[53] Ryerson University v Ryerson Faculty Association, 2018. The association entered expert reports by Richard Freishtat and Philip Stark as evidence: Freishtat, 2016a; Stark, 2016.

individuals based on one or more *Code* grounds."[54] On the face of it, student questionnaires are neutral, in as much they ask the same set of questions and are administered in substantially the same way for all faculty, regardless of their gender, racialization, sexual identity, or any other characteristic identified in the *Code*. Equal treatment, if left at that, simply means being blind to differences without regard to their potentially consequential nature. Equitable treatment requires recognition of difference and the removal of barriers and practices that, however unintentionally, yield a disadvantage due to that difference. Women, racialized, and LGBTQ2S+ faculty will still face bias with formative student questionnaires, but they will not be subjected to the consequences that only summative use can produce.

As the research evidence about student questionnaires demonstrates, even such simple matters as the time of day a course is scheduled can yield lower scores for an instructor. If the scores factor into the instructor's salary increment and they receive a lower amount than their peers, it might represent a one-time setback. The effect on lifetime earnings is compounded, and the total effect depends on when in the career trajectory the impact of the scores is felt. It might be argued that, over time, other faculty members will probably experience the same course scheduling disadvantage and so, in the end, there may be no lasting consequence of this kind of survey response bias for the faculty member's earning relative to other members.

The same is not true for faculty members subjected to bias based on their gender, racialized status, or other grounds prohibited by the *Code*.[55] Particularly where scores are comparative, the effects on career status and career progress for faculty members are negative, iterative, and systemic. Salary patterns are one non-trivial example, as the number of pay equity cases and settlements should illustrate.[56] For those subject to intersecting biases, gender and race or ethnicity for example, the added effect can be even larger gaps in earnings.[57]

Moving through faculty ranks is challenging for women and racialized faculty in the first place. Women are also over-represented in the ranks of both contract faculty and full-time faculty who are neither on the tenure-stream nor tenured.[58] With a teaching evaluation tainted by scores from questionable

---

[54] Ontario Human Rights Commission, 2008: http://www.ohrc.on.ca/en/iii-principles-and-concepts/2-what-discrimination (retrieved September 20, 2018).

[55] Expert reports by Basow (2018) and Henry (2018) provided analyses, that were instructive for extending the working group's thinking in this regard, and references, many of which are cited in this section.

[56] OCUFA, 2016. See also CAUT, 2010, 2011.

[57] See CAUT, 2018b; Woodhams et al, 2015.

[58] Field and Jones, 2016; Foster and Birdsell Bauer, 2018; CAUT, 2017a. Foster and Birdsell Bauer ask self-identification questions on gender identity, aboriginal status, and race and ethnicity. Canadian data on faculty by race and ethnicity, gender identity, physical ability, etc. typically are scant to nonexistent.

SQCTs, career progress can be slowed when tenure and promotion are postponed, or sidetracked when contracts are not renewed or tenure denied and the career trail must be picked up elsewhere.[59]

To be sure, delays in career progress and diminished earnings for any one faculty member *might* be attributed to their research profile. Still, it would be appropriate to ask whether there is a bias against the member's research and whether matters have been confounded further by relying on scores for teaching which are obtained using instruments that are not reliably free from bias. Beyond the research or teaching metrics, however, there are questions about the other factors that shape faculty members' capacity to meet expectations. For example, what is the rate of "cultural taxation" or "identity taxation" on women, racialized, and LGBTQ2S+ faculty members?[60]

Cultural and identity taxation commonly take the form of higher service workloads – on university committees as a diversity representative and working with members of the communities of which they are a part. Similar investments of time and energy are not expected or demanded of white, cis-gender male faculty members. One could spend less time on other professional responsibilities with predictable academic career peril, but the pressure is to accept higher workload overall.

It is in this sense that student questionnaires present a kind of double jeopardy. They are not just methodologically compromised by respondent bias, they also are blind to the background conditions that set the stage for a faculty member's teaching. To return to the scenario with which this section opened, the effect of class assignments over time is far from neutral if a faculty member is systematically being assigned undesirable class times or larger classes. If the member is also at risk of respondent bias, the combined size effects of bias attributable to class and instructor characteristics cannot be dismissed as negligible. In a variation of identity taxation, even assigning a faculty member courses with the best of intentions – courses in an area presumed to be in their field of expertise because of their identity – can tilt the scales against them even further when unconscious bias combines with perceptions that instructors themselves are biased in their objectives for the course.[61]

For faculty members who are at a systemic disadvantage, there can also be a spillover into workload similar to the identity taxation entailed in community service. On the one hand, instructors can find it necessary to devote more time and energy to teaching to raise their scores to the levels received by their white male counterparts simply to offset perceptions of relative competence and the effects of bias – especially for those with teaching-intensive appointments. While competence is assumed for white males, women and racialized instructors tend to be viewed more critically and have to work

---

[59] The effect on career paths, and ultimately faculty diversity, starts early: Huston, 2005; Mengel et al, 2018. See also Henry et al, 2017; Luther et al, 2001; Samuel and Wane, 2005.

[60] Amado Padilla (1994) is commonly cited as the originator of the term "cultural taxation."

[61] Abel and Metltzer, 2007; Anderson and Kanner, 2011; Anderson and Smith, 2005; Bilimoria and Stewart, 2009; Dua and Lawrence, 2000; Littleford et al, 2010; Perry et al, 2015.

harder to be seen as equally competent.[62] The phenomenon is not limited to perceptions of teaching ability: perceptions of leadership competence in other contexts are gendered and racialized.[63]

There is a parallel in language which tends to characterize (white) men in terms of their intellectual qualities and women on the basis of interpersonal skills. On the other hand, the expectations for women, for example, tend to be more labour intensive and entail more emotional labour.[64] As to mentoring or other kinds of psychological support, it is not just women but racialized and LGBTQ2S+ faculty members whose support students seek, precisely because of their membership in the respective community.

To repeat a point made in the section on methodology and student questionnaires, the absence of evidence of bias in SQCT scores is not evidence there is no bias or discriminatory effect. In these cases it is instead that comparatively higher investments of time and energy are not recognized or validated. For contract faculty teaching on a per-course basis, well over half of whom are women,[65] it is yet another example of unrecognized and unremunerated teaching activity.

Hazards to the intellectual and learning environment are less or more calculable, but easily overlooked by advocates of quantitative measures for the educational enterprise. It is already difficult enough for women to enter and succeed in traditionally male-dominated fields only then to face the prospect that a predominantly male enrolment will skew student questionnaire scores against them.[66] One can argue about whether a lack of diversity amongst faculty members also limits intellectual diversity, but if using student questionnaires for summative purposes limits or reduces faculty diversity, the questionnaires clearly are not an innocent tool. And where SQCT scores are published, to the extent students select courses based on the scores, and thereby tilt the choice of courses against the avenues of academic inquiry offered by the faculty members at the receiving end of the bias, the implications for intellectual diversity are only negative.

---

[62] Basow et al, 2013; Biernat et al, 2010; Boring, 2015; Foschi, 2000; Ho et al, 2009; Sinclair and Kunda, 2000. More generally, see *Presumed Incompetent: The Intersections of Race and Class for Women in Academia* (Muh et al 2012). One contributor, Mary-Antoinette Smith (pp. 418-419) writes: "I began to become increasingly more concerned about consciously performing to maintain my acceptably high student evaluations and what the cost might be to my self-esteem and fulfilling my responsibilities as a teacher. Aside from the time and effort involved, I found myself concentrating more on earning high evaluations and less on the pedagogical goals of my courses."

[63] On gender, Johnson et al, 2008; Scott and Brown, 2006; on gender and race, Rosette et al, 2016.

[64] On language and descriptions conforming to gender expectations, Basow, 2000; Basow et al, 2006; Schmidt, 2015; Storage et al, 2016. On expectations for women being more labour-intensive, and emotional labour, El-Alayli et al, 2018; Sprague and Massoni, 2005.

[65] Field and Jones, 2016; Foster and Birdsell Bauer, 2018.

[66] Male students are more biased towards female faculty, and business and engineering students are more prone to such bias; intersections of student and faculty gender and race may also feature. Basow, 1995; Basow and Martin, 2012; Boring et al, 2016; Mengel et al, 2018; Meyer et al, 2017; Pittman, 2010.

The simplification and standardization of university education that follows from those consequences may be an effective low-cost means of increasing the number of students with credentials and learning outcomes corresponding to labour market demands and employers' needs. But there is good reason to doubt that such an approach and the way student questionnaires are used as registers of student satisfaction, rather than student learning, are compatible with the academic mission of universities, or that they contribute to the critical learning and engagement faculty members hope to foster.

Student questionnaires and harassment

It is one thing for unconscious bias to skew responses against a faculty member; it is quite a different matter to use the questionnaire to express hostility towards them with lower scores and, as is sometimes the case, direct abuse through anonymous comments. Harassment and the harm and suffering it causes predate online student questionnaires, but their increased use and the cultural changes accompanying the rise of social media combine in ways that are particularly vexing for faculty members. Put simply: online surveys + anonymous comments = trolling heaven.

Comments solicited in student questionnaires are mostly anonymous. And faculty agreement provisions typically prohibit inclusion of anonymous materials in faculty members' personnel files. However, this does not mean that members are protected from harassing or threatening comments. As the responses to the survey of faculty associations suggest, the incidence of harassing comments is clearly on the rise. Similar findings have been reported by faculty organizations in Australia, the United Kingdom, and the United States.[67]

Anonymous online comments pose two types of problems for faculty members and their associations. The first set of issues include: that incidents of harassment occur at all; who is targeted by offensive remarks, and; the conditions that enable harassment. The *Code,* the OHSA mandate, and collective agreements similarly stipulate a safe and harassment-free workplace. With the passage of "Bill 132," which obliges universities and colleges to address sexual violence involving students,[68] it should be safe to say that the accepted legal and cultural norm is that no harassment or threat of violence is acceptable at Ontario universities.

Women, racialized, and LGBTQ2S+ faculty are not exclusively the targets of "contrapower" harassment by students, but they are more likely to be on the receiving end.[69] It should be understood by now that

---

[67] National Tertiary Education Union, 2018; University and Colleges Union – Queen's University Branch, 2013; Vasey and Carroll, 2016.

[68] Passed in March 2016, Bill 132 amended the OHSA to include provisions regarding sexual harassment and sexual violence specifically, and the Ministry of Training, Colleges and Universities Act to require postsecondary institutions to establish policies and procedures to deal with sexual violence, as well provisions for reporting.

[69] Blizard, 2016; Cassidy et al, 2014; Lampman et al, 2009; Lampman et al, 2016; National Tertiary Education Union, 2018. Contrapower harassment refers to harassment of someone who is in a structurally more powerful position by another person who is in a subordinate or less structurally advantageous position.

women and LGBTQ2S+ faculty are also at greater risk of sexual violence. Evidence about women's experience indicates they also suffer greater distress and are considerably more likely to consider quitting their position.[70] Any and all of this alone should give universities reason to revisit what they expect to gain from student questionnaires and how they administer them.

How likely harassment is to occur depends on the level of anonymity on one hand and the degree to which social interaction is mediated on the other – online versus face-to-face or in physical proximity, for example. Students may still be harder on female than male faculty, but when their responses are not anonymous the scores they award to the female instructors are higher and the rate of negative comments are lower.[71] Anecdotally, working group members found that even anonymous paper questionnaires were less likely to contain negative commentary than online questionnaires, perhaps because students thought faculty might be able to identify them by their handwriting, but also perhaps because the classroom setting imparts a sense of social responsibility.

Although being online does not confer anonymity by itself, it has become common wisdom that online behaviour is characterized by greater levels of incivility than other social settings. Online contexts combined with anonymity seem to strip participants of fundamental social inhibitions, and harassment and bullying are more likely to occur. Even if it does not always rise to the level of harassment, the incidence of aggressive and hurtful student comments is higher.[72] The social setting matters, and it seems that being present in a classroom inclines students to be more constructive in their (anonymous) responses, even if not entirely free of unconscious bias.[73]

Research examining response rates to online instruments indicates that having students complete them in class, just as is expected for paper versions, is one way to increase student participation.[74] If that were to reduce the incidence of abusive and threatening comments, it would be a good thing. If it were also to increase scores, let us reiterate the point that if the questionnaires are formative and for the faculty member alone, it matters little whether the effect is to raise their scores relative to other faculty: what matters for each faculty member is how the scores on specific questions for each course change from one iteration to another.

---

70 Lampman et al, 2016; National Tertiary Education Union, 2018.

71 Fries and McNinch, 2003.

72 Lindahl and Unger, 2010; more generally, see, for example, Lapidot-Lefler and Barak, 2012.

73 Strictly speaking, the research by Rhea et al (2007) does not show that the presence of others and the possibility of face-to-face interaction at the time of filling out student questionnaires yields higher scores and more constructive comments – online questionnaires were completed for online versions of the courses in the sample; in-class questionnaires for the face-to-face, in-class versions of the same courses – but the differences in behaviour they note with respect to non-anonymous online activity are suggestive nonetheless.

74 See, for example, Morrison, 2011; Nowell et al, 2010; Risquez et al, 2015; Stowell et al, 2012.

The second set of problems has to do with what is done when comments do cross the line of what is acceptable. It should be emphasized that there is no necessary relation between the use of student questionnaires for summative purposes and harassment, although there are problems peculiar to summative use when harassment occurs. Even if the comments do not figure in the evaluation of a faculty member, a double bind is created about how to deal with scores from the person making offensive comments. If the comments are known only by the instructor, but not the identity of the source, how are associated scores to be removed from the sample? If comments are vetted for offensive comments by the institution or a third-party provider, is the member's right to know being respected, who defines harassment, what is the definition, and are the scores removed from the sample?

In the first instance, the initiative is left to the instructor, obviously, but it must be incumbent on the university to permit instructors the option of removing the discredited scores, to advise faculty members to that effect, and to have a procedure for doing so that is reliable and credible to faculty members. Students too should be able to trust their identity will not be disclosed, except as may occur in accordance with an investigation into workplace harassment or similar proceeding arising from a faculty member's exercise of their rights to make a complaint in this regard. A similar conundrum may arise even in the case of formative questionnaires, which also depend on instructive comments and valid and reliable responses.

More urgently, whatever the design and use of the questionnaires, instructors must have confidence that if they do have cause to pursue a complaint arising from student questionnaires, there are policies and procedures in place to accord them due process. It seems to us that student questionnaires cannot be exempt from any university human rights, workplace harassment, or sexual harassment policy. No purpose is served in identifying any institution as having failed in this regard. The language of these policies is typically inclusive enough that it should be understood that comments made in student questionnaires could meet the definition of unacceptable behaviour. Nothing is lost by making it explicit in policy and making it plain to faculty and students alike. As the discussion of research ethics suggested, the main barrier should not be anonymity.

As for comments that are vetted, the survey of faculty associations revealed that comments are reviewed, and offensive ones evidently expurgated, before they are transmitted to the faculty member. If the institution uses a third-party provider, it may be the third party reviewing and editing the comments section. In addition to concerns about whether or not the corresponding scores are also removed from the pool, it is troubling that parties other than the faculty member may be choosing what the faculty member may know of the comments. It is difficult to navigate what and how much to tell an employee of course, but advising them of the existence of offensive comments and giving them a choice about reading them is a simple enough step and the bare minimum that should be expected.

We are not in a position to state conclusively that faculty members are being denied knowledge about offensive comments or the opportunity to assess them, but we do see a need to underscore two reasons why the knowledge and the opportunity to initiate a complaint must be available to affected members. First, simple fairness dictates that faculty members should know about any circumstance that is prejudicial to their prospects. Second, the OHSA obliges employers, including universities, to have policies and programs in place that address workplace violence and workplace harassment from whatever source. Failure to disclose to the target the fact of offensive comments abridges the faculty member's rights under the OHSA.

The adage that an ounce of prevention is worth a pound of cure is apt in light of a recent arbitration decision which held that, where the exposure arises because of employer-mandated use, employers have an obligation to protect employees from harassment through social media.[75] As mentioned in the section on student questionnaires and research ethics, university administrations seem to have foregone the opportunity of using an active consent process to advise students of the institutional policies and procedures regarding harassment, including disclosures necessary for investigations. Even if such measures are taken, we cannot be sure that discriminatory attitudes will not be reflected in the scores students assign to instructors.

Human rights – looking forward

When examined with a view to human rights, the weight of argument comes down against the use of student questionnaires for summative evaluations of performance, just as is the case when the lens is methodology and research ethics. The evidence is compelling that the summative use of student questionnaires is discriminatory in its effects on faculty compensation and career progress, however unintentional that may be.

Assuming equality and being indifferent to gender/racialized/LGBTQ2S+ characteristics is not an option. Only an equity lens makes sense in this context. Proposals to try and offset the lower scores assigned to faculty members at the receiving end of bias by adding some fraction to their scores is naive at best.[76] It would be perverse to try to resolve a social problem by using what is essentially an arbitrary adjustment to an already methodologically compromised instrument. Far too many data are necessary for it to seem, let alone be, fair to all. It makes far more sense to remove the barrier or the source of the differential treatment.

There remain very good reasons to continue using student questionnaires, albeit for formative purposes only. The working group is under no illusion that limiting questionnaires to formative use will eliminate harassment and threats of harm to faculty members, but we do believe that universities can

---

[75] Amalgamated Transit Union, Local 113 v. Toronto Transit Commission (Use of Social Media Grievance) [2016] O.L.A.A. No. 267.

[76] See, for example, McPherson and Jewell, 2007; McPherson et al, 2009.

do better at fulfilling their duties as employers by making it clear to students and faculty that universities cannot bracket the *Ontario Human Rights Code* and the *Occupational Health and Safety Act* or parallel provisions in faculty agreements when it comes to student questionnaires.

Equity and safety are not simply incidental to the academic mission of universities and the quality of student learning. They are critical conditions for cultivating academic freedom and open inquiry, and an atmosphere that thrives on demographic and intellectual diversity. None of this is without challenges of course. We do think, however, that formative student questionnaires are a far more constructive contribution to the qualitative dimensions of the teaching and learning experience of faculty and students than summative versions.

Moving forward on student questionnaires

The working group finds it mystifying that student questionnaires are still used for summative purposes despite the evidence on bias in student questionnaires and their tenuous connection to teaching effectiveness, the discriminatory impact on faculty careers, greater sensitivity to harassment, and in light of legal standards which enjoin against abetting discrimination or harassment. Add to that the subversive effect the same use has on student learning and the academic objectives of teaching, and the tenacity with which some advocates cling to summative end-of-course student questionnaires becomes utterly perplexing.

Since we set out on this journey in December 2016, others have anticipated where we would arrive. The Canadian Association of University Teachers changed its model clause on teaching evaluation to exclude the use of student questionnaires; the University of Southern California Provost is reported to have said "I'm done. I can't continue to allow a substantial portion of the faculty to be subject to this kind of bias"; the former head of Rice University's instructional support service has recommended "a moratorium on using student ratings results to rank and compare individual faculty to one another"; and, of course, Arbitrator Kaplan has declared that student questionnaires cannot be used to measure teaching effectiveness.[77] They did not tread all the ground over which we have ranged, but our paths have converged on important points.

Before discussing our own suggestions about the future for student questionnaires, it is necessary to return to principles about teaching articulated in faculty agreements. First and foremost amongst these are the role of teachers as scholars and the goal of fostering student learning in an environment of academic freedom and inquiry. In other words, teaching is recognized as one responsibility, complemented by others. The relative mix of responsibilities varies from faculty agreement to faculty agreement, faculty member to faculty member, and time to time, but teaching and learning as an

[77] CAUT, 2017b; for updated CAUT policies on student questionnaires and teaching evaluation, see CAUT, 2016a, 2016b; for USC Provost remark, see Flaherty, 2018; for former Rice head of instructional services, see Barre, 2018; Ryerson University v Ryerson Faculty Association, 2018.

academic activity is a constant that is not reducible to simplistic quantitative measures, despite what the widespread use of SQCT scores seems to convey.

We are hesitant to call any of the principles and guidelines we propose "best practices." For one thing, best practice frequently is little more than common practice. For another, the superlative form implies there is no room or need for improvement, when clearly there is much that can be done better if students and faculty, and the academic mission of universities, are to be served.

In the first instance, best practice should be a matter of consistency with faculty agreement provisions negotiated between faculty associations and universities. Faculty agreements already embody and articulate critical principles that should frame the use of student questionnaires. These include prohibitions against discrimination, adherence to natural justice and procedural fairness, standards that are reasonable and fair, and protection of academic freedom.

With those in mind, we propose seven general principles and guidelines for the use of student questionnaires. Each of them is encoded in provisions of one or more faculty association agreements. Not all of them are found in all agreements, and how any of them may be incorporated in or conform to existing agreements is something that only faculty associations themselves can determine.

In the second instance, changes to the use and administration of student questionnaires must be negotiated or done in consultation with faculty associations. Some matters, confidentiality protocols perhaps, may not require changes to existing agreements, but the standard must be that nothing is implemented unilaterally by the institutions or, for that matter, the provincial government. Top-down measures are counterproductive to what must be the starting point: teaching and learning as an active relationship between faculty and students.

Use of student questionnaires should be limited to formative purposes

The first principle is that no more use should be made of student questionnaires than they can sustain methodologically. The working group was asked to comment on the appropriate scope, methods and valid use of student questionnaires. We concur fully with Professor Michela's assessment of the methodological issues and conclusion that student questionnaires should be used for formative purposes only: they should not be used for purposes of summative evaluations or in career progress decision-making. The review from the perspective of research ethics and human rights supports the same conclusion.

The implications are not trivial for faculty members. The use to which student questionnaires are put affects their livelihood and their career trajectory. The methodological, ethical, and human rights reasons to take this position apply to all instructors, regardless of their appointment status, but contract faculty are particularly vulnerable. Student questionnaires should not play a role in determining pay

increments, in tenure or and promotion for tenure stream faculty, for permanent or continuing status and promotion for other full-time faculty, or for appointment and renewal for contract faculty.

Formative questionnaires, we firmly believe, are more consistent with the premise we started with – that student learning must be at the centre of why student questionnaires are used in the first place, and that the purpose of student questionnaires should be to enable faculty members to develop their course and their teaching. Extraneous purposes, like performance evaluation, bend the logic and lend themselves to distraction and misguided motives.

Student questionnaires should provide useful feedback for instructors

This still leaves questions about what the characteristics of a good student questionnaire are for formative purposes. Answering those questions is beyond the mandate or capacity of the working group, but it does follow that, if the questionnaire is for formative purposes, it needs to be designed to provide responses that will be useful for the instructor. Formative questionnaires are already used by many instructors to make mid-course adjustments and there is no reason a survey at the end of course cannot be used similarly – for preparation of a subsequent iteration of the course. Doubtless they will look quite different from many in use at Ontario institutions today.

On that score, the surveys would have to be designed to provide the instructor with meaningful information about the impact of the course content and pedagogical model from students' perspectives. Summative questions are beside the point. And if, as we have suggested, student questionnaires should give instructors instructive feedback on specific dimensions in different iterations of a course, it will not be a one-size-fits-all model. Common questions might be a feature, but this should follow from, rather than guide, the process of questionnaire development.

Ease of administration, (false) transparency, or (invidious) comparisons are not appropriate points of departure. To the contrary, getting the most value out of formative questionnaires requires resources and commitment, starting with universities.

At one level, there is the staff time simply to develop and supply questionnaires appropriate to instructors' varied needs. Whether it is to help an instructor understand the responses to their questionnaire, or to help select appropriate questions for the task at hand, the logical role of TLCs is to provide the type of consultation services instructors actually need to make the best use of the questionnaires.

At a different level, some investment may be necessary to enable faculty members to make best use of the new instruments and to take advantage of professional development opportunities. Accommodating demand for these types of assistance by full-time faculty might initially stretch existing methods for recognizing and developing pedagogical development, but contract faculty teaching on a per-course basis should not be expected to undertake this work on their own dime. Either increasing

remuneration to recognize pedagogical and professional development on instructors' own time or providing direct payment for teaching development would serve as a good indicator of universities' serious intentions to recommit to putting student learning before earning favour.

Student questionnaire results should be confidential

It follows from the formative nature of the questionnaires that the results should be confidential. There is no rationale for making the results available to anyone other than those with whom the instructor chooses. Each of the considerations on methodology, research ethics, and human rights lead us to conclude that the scores should not be made public. The size of the audience is irrelevant. Nor is there any need for the results of student questionnaires to be included in instructors' personnel files. Any departures from this default position should be subject to faculty agreements to ensure no one is penalized for not sharing scores or comments, and that any results that are shared are not used contrary to the faculty member's intentions or their rights under their agreement.

Although actual scores do not need to be shared, nothing prevents a faculty member from doing so. Aside from the dubious value of reporting averages and distributions, the more relevant data are how the scores change. More to the point, it is the instructor's narrative about what the results tell them, their reflections and self-evaluation, that speaks to how they have used them in the fashion they were intended – to inform their teaching. But there is no "need to know" that would warrant anyone who is not the intended audience – the designated academic administrators, for example, and review committees – having access to them. Similarly, an instructor may choose to share the scores on a confidential basis with experts in teaching and learning for the purpose of enhancing their teaching effectiveness.

In point of fact, just as formative questionnaires may look very different from the status quo versions, we should expect the use to which the scores and comments could be made will be quite different as well. If student questionnaires are summative, the incentive for instructors reporting results is to pick the most positive comments and score summaries, and to defend their teaching against negative ones. If the questionnaires are for formative purposes instead, faculty members have every reason to demonstrate how they monitored aspects of their pedagogical approach and course content and modified their teaching or course in light of students' responses.

Student questionnaires must seek informed and active consent from students

The working group was also asked to consider safeguards against discrimination and harassment. Because statutory obligations and clear ethical guidelines already exist, this is one area in which the working group should have no need to propose standards for best practice. Yet, as the section on research ethics indicates, this is the one area in which it most clearly must.

We are aware that use of student questionnaires for performance evaluation remains prevalent. Until strictly formative end-of-course student questionnaires are the norm, in any instance that scores and/or comments may be included in a faculty member's file for review by others, it should be made clear to students that this may occur. Responses and comments may be anonymous, but the principle of informed and active consent requires this much. By the same principle, students should be advised of any secondary use of the data, by whom and for what purpose(s). This is little different and no more difficult than is expected for agreement to a privacy policy.

Informed and active consent is imperative if institutions are to deal with harassment and threats of violence in a forthright and good faith manner. A consent statement must advise students of the institution's policy on harassment, and the scope of confidentiality in the event of an investigation of alleged harassment or threat of violence. We recognize that it would be impractical to have students read a lengthy disclaimer about privacy and the entirety of a university's policies concerning harassment and violence as part of the consent process. However, students should be advised that, in circumstances in which someone's health or safety is at risk or in circumstances where a faculty member has deemed comments to be harassment, their identity will be shared confidentially, on a "need to know" basis, to investigate the matter.

While the technology underlying online questionnaires allows the author of a harassing or threatening comment to be identified, doing so where paper questionnaires are in use is a function of the way in which they are administered. We do not know which institutions have procedures that make it possible to identify students writing abusive comments on paper returns. We do believe that, if we are to challenge harassment wherever it appears, student comments on questionnaires should not be an exception, and any institution that does not have the capacity to investigate these comments should implement the necessary changes to ensure they do.

Surveys for other reviews should be separately administered

There are a range of reasons why institutions, and the provincial government, seek feedback from students. The working group was not asked to comment on other surveys of students, except in so far as student questionnaires are used to answer questions beyond their primary mandate. As already indicated in the section on methodology, it is contrary to good policy sense simply to tally up skewed data from problematic surveys in the hopes of insights which are, at best, of questionable value. We have concluded, moreover, that student questionnaires should not do double duty, especially if they are formative in nature, and that questions meant for other purposes should be administered separately. Doing so does a disservice to faculty and students alike.

If there is reason to administer survey items to an instructor's students to ask questions about a course which is not intended to solicit information for or about the instructor, the bare minimum that should be expected is that questions about the course and instructor are clearly and strictly differentiated.

To avoid tainting the results by incipient bias, they should be administered separately from student questionnaires on courses and teaching. Similarly, instruments to solicit student feedback on courses for the purposes of program and/or curriculum review should be administered separately. Further, they should be aggregated and reported in such a way as not to risk revealing the identity of any instructor.

Other surveys of students are carried out for other reasons, including reporting to the provincial government. As long as these are conducted according to accepted methodologies and protocols for random sample surveys, they should pose no specific problem for individual faculty members. We are skeptical, however, of preoccupations with "accountability" and metrics that have redirected the logic of student questionnaires in the first place. For reasons similar to why we argue student questionnaires should not be used comparatively, we identify two further safeguards.

First, any instrument used to assess teaching of individuals, in-class observation rubrics for example, should not be reduced to or represent the results solely as summative scores, whether as a supplement to or substitute for SQCT scores. Any reduction to a decontextualized and uniform measure runs counter to the principles and the practical reality of academic freedom and intellectual diversity. Second, we also believe that surveys for program review or public reporting should not be used to compare, evaluate, or allocate resources between programs or institutions. They are differentiated, and the only instructive metrics are those used to assess how responses on relevant and specific questions have shifted, or not, in light of changes in the program or institution. As with program reviews, they should be only one of multiple methods.

Teaching evaluation requires a suite of tools

The working group was also asked to identify alternative and complementary methods of assessment. There is an extensive literature on the evaluation of teaching. It is not without its share of advocates for student questionnaires. Positions on the relative weight to be given them vary, but the consensus even for defenders of summative use seems to be that the questionnaires are one tool in a bigger kit.[78] One need only refer to the *Guidelines for the OCUFA Teaching Award* and the CAUT model clause on Evaluation of Teaching Performance for the elements and factors that define teaching.[79] They are too varied and complex to be represented in survey results. The two principal methods of teaching evaluation – what Arbitrator Kaplan called the "gold standard" in his arbitration at Ryerson University – are careful examination of teaching dossiers and in-class observation by peers.[80]

The breadth of material recommended for inclusion in teaching dossiers is reflective of the range of activities which comprise teaching at universities. The CAUT guide to teaching dossiers identifies seven different categories of documentation, and a review of Ontario faculty agreements finds dozens

[78] For example: Gravestock and Gregor-Greenleaf, 2008; Wright et al, 2014.

[79] OCUFA, 2018; CAUT, 2017b.

[80] Ryerson University v Ryerson Faculty Association, 2018, p.8.

of examples of documentary evidence of teaching performance.[81] Teaching dossiers are not mere catalogues. The list of possible materials includes the reports from in-class peer observers, course curriculum evaluations, and faculty self-evaluation, each of which addresses aspects of teaching not captured by student questionnaires. Course curriculum evaluation and in-class observation by peers attend to the match between the methods and materials used and the objectives for the course. Self-evaluation is the point of student questionnaires in the first place.

It is as a part of self-evaluation and reflection that the working group is persuaded that student questionnaire results can still be a feature of teaching evaluation, if they are used for formative purposes. There is value in an instructor reporting on the responses to formative versions of student questionnaires, and how the instructor has interpreted and used the responses to enhance their teaching effectiveness. This is the approach suggested in the *Guidelines for the OCUFA Teaching Award*, for example.[82] The focus is not on the results but on the role responses play in faculty members' assessment and development of their courses and their teaching.

We are not in a position to comment on the preferred models and practices for use of teaching dossiers, course curriculum evaluation, in-class peer observation, and self-evaluation. We will say, however, that they should be subject to faculty agreements and none should be used to the exclusion of the others. Teaching evaluation is multi-modal in its methods – a triangulation, as it were. We also repeat that rubrics and similar tools for in-class observation should not be used merely to generate scores. Simple scores, SQCT or otherwise, are not informative on their own and essentially invite readers to skip over the matter at hand.

We also think it is necessary to emphasize that these methods of evaluating teaching, including reading instructor self-evaluation of student questionnaires, require time to process with the seriousness they deserve. One would expect no less for any review required for probationary renewal, tenure or promotion. Annual reports or periodic reviews might require fewer of the tools, and briefer versions of them, but the principle is the same. Any outcome of any consequence from evaluation processes should be based on substantive information that reflects the value of teaching. The current widespread practice of using SQCT scores as a substitute for real evaluation of teaching is more a matter of convenience than commitment. Commitment we think is embodied best in peer evaluation.

Peer evaluation should be the rule

Peer review is foundational for the evaluation of faculty. This principle is firmly established for research and service activities – our fellow faculty members are best qualified to evaluate our work. It should

---

[81] CAUT, 2018a. A full list of documentation for teaching dossiers identified in Ontario faculty collective agreements and related sources is included in the working group's report on faculty association agreements and student questionnaires, available to OCUFA member associations only.

[82] The working group does not, however, endorse the use of comparative scores.

be just as fundamental a principle for the evaluation of teaching. University education is a collective responsibility; evaluating teaching is a collegial responsibility which should not be contracted out. This does not mean there is no role for teaching specialists, as in-class observers for example, although it seems to the working group that this should be limited to formative purposes. There is no substitute for peer knowledge of the content, the nature and value of activities occurring outside the classroom that form part of teaching, and the differences between courses and modes of delivery – all of which is not captured solely by classroom observation, much less by student questionnaires.

A fully functioning peer review process for teaching evaluation requires time and resources. It would be naive not to recognize the additional demands on faculty when simple reporting of responses is not enough. For faculty whose teaching is being reviewed, there is not only the time to build and maintain their teaching dossier and to produce an evaluation of their formative questionnaire results, there is an investment in gaining and maintaining the knowledge and capacity to put their best foot forward. Full-time faculty with already burdensome workloads will find it difficult to accommodate the demands without some kind of relief; contract faculty who currently are expected to meet the expectations on their own time and dime should expect remuneration and support.

We do not expect that peer reviewers, including in-class observers, will "know it when they see it" simply because of their own experience teaching and seeing others' teaching. In-class observation follows protocols intended to establish common points of reference. And, as Arbitrator Kaplan and the granting councils acknowledge and provide for, reviewers also need training on how to recognize bias and the ways it is manifested in the review process. As such, faculty members reviewing the teaching of their peers should be provided with the time and resources to ensure their effectiveness as peer evaluators. If the contract faculty teaching contributions are to be accorded respect equivalent to full-time tenure-stream faculty, affording them the opportunity to participate as reviewers of their contract faculty peers, and the time and remuneration that goes along with it, is the least that should be expected.

The cost of a peer evaluation system for teaching should be neither overlooked nor overstated. To be sure, getting the system up to speed entails more money for resources and faculty and TLC staff time than is currently provided. Once the practice is established, the principal cost difference is the time for in-class peer reviewers.[83] The margin of difference is less if in-class peer reviews are not conducted annually, but principally at career milestones, and perhaps at additional multi-year intervals or other occasions where this is provided in faculty agreements. For equity reasons, the working group is already recommending against the use of student questionnaire results in determining pay increments; in this case, full peer review of teaching is a disproportionate investment for the stakes. The real value of this type of evaluation is that it can be designed with greater emphasis on formative development and teaching improvement than student questionnaires alone can provide.

---

[83] Stark, 2015.

Conclusions – putting student questionnaires into perspective

Nobody is against effective teaching, faculty members least of all. Teaching is acknowledged as a professional responsibility, part of an integral academic mission. From their original purpose as tools for development of courses and teaching, however, student questionnaires have become a cheap, and wholly inadequate, substitute for substantive teaching evaluation.

As instruments of end-of-course summative evaluation, student questionnaires do not follow two basic rules of performance evaluation. First, no one should be evaluated by inappropriate mechanisms. Even the scores from a seemingly well-designed instrument can be misleading. And because scores are confounded by bias and the halo effect, student questionnaires have discriminatory effects if they are used for any purpose that compares scores of one instructor against those of some set of other faculty members.

Second, everyone should be given an opportunity, and the means, to improve if the accepted forms of evaluation reveal a need. Student questionnaires are summative only from the perspective of students completing the course; it is over. For faculty members, however, there will be future iterations of the course, modified as changes in content and form may require. Whether mid-course or end-of-course, formative student questionnaires are tools to assess how well students are doing and how to direct teaching efforts, i.e., to improve the learning experience for them.

No one doubts the value of formative student questionnaires. Not least, their focal point is teaching and learning as an active relationship between faculty and students. As we have suggested, the usefulness of student questionnaires as guides to effective teaching and student learning has been severely compromised by expectations placed on them by institutional administrators and provincial governments, with different objects in mind.

Refocusing attention on teaching and learning, and putting into practice the seven principles we have outlined requires resources, not a doubling down on metrics. Recommitting to teaching excellence and academic achievement starts with more funding from the provincial government. It requires also the willingness of university administrations to allocate resources to support faculty, students, and teaching, in keeping with universities' academic mission, and consistent with the terms they negotiate with faculty associations. Implementing the principles we have recommended is no more a one-way street than is teaching and learning.

References

Abel, Millicent H.; Meltzer, Andrea L. 2007. Student ratings of a male and female professors' lecture on sex discrimination in the workforce. *Sex Roles*, Vol. 57, No. 3-4, pp. 173-180. DOI: 10.1007/s11199-007-9245-x.

Algozzine, Bob; Beattie, John; Bray, Marty; Flowere, Claudia; Gretes, John; Howley, Lisa;Mohanty, Ganesh; Spooner, Fred. 2004. Student Evaluation of College Teaching: A Practice in Search of Principles. *College Teaching*. Vol. 52, No. 4, pp. 134-141. DOI: 10.3200/CTCH.52.4.134-141.

Amalgamated Transit Union, Local 113 v. Toronto Transit Commission (Use of Social Media Grievance) [2016] O.L.A.A. No. 267.

Ambady, Nalini; Rosenthal, Robert. 1993. Half a Minute: Predicting Teacher Evaluations from Thin Slices of Nonverbal Behavior and Physical Attractiveness, *Journal of Personality and Social Psychology*, 64, 431-441. DOI: 10.1037/0022-3514.64.3.431.

Anderson, K. J.; Kanner, M. 2011. Inventing a gay agenda: Students' perceptions of lesbian and gay professors. *Journal of Applied Social Psychology*. Vol. 41, pp. 1538-1564. DOI: 10.1111/j.1559-1816.2011.00757.x

Anderson, K. L.; Smith, G. 2005. Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences*, Vol. 27, No. 2, pp. 184-201. DOI: 10.1177/0739986304273707.

d'Apollonia, Sylvia; Abrami, Philip C. (1997). Navigating student ratings of instruction. *American Psychologist*, 52(11), 1198-1208.

Arbuckle, J. and B.D. Williams, 2003. Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations, *Sex Roles*, 49, 507-516. DOI 10.1023/A:1025832707002.

Barre, Elizabeth. 2018. Research on Student Ratings Continues to Evolve. We Should, Too. Rice University Centre for Teaching Excellence, February 22, 2018. http://cte.rice.edu/blogarchive/2018/2/20/studentratingsupdate

Basow, Susan. A. 1995. Student evaluations of college professors: When gender matters. *Journal of Educational Psychology*, Vol. 87, pp. 656-665.

Basow, Susan A. 2000 Best and worst professors: Gender patterns in students' choices. *Sex Roles*, Vol. 43, pp. 139-149. 10.1023/A:1026655528055

Basow, Susan A. 2018. *Expert Opinion (unpublished proceeding)*. Commissioned with the support of the Ontario Confederation of University Faculty Associations.

Basow, Sandra A.; Martin, Julie L. 2012. Bias in student ratings. In M.E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 40-49). Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/evals2012/index.php.

Basow, S.A.; Codos, S.;Martin, J. 2013. The Effects of Professors' Race and Gender on Student Evaluations and Performance, *College Student Journal*, Vol. 47, No. 2, pp. 352-363.

Basow, S. A.; Montgomery, S. 2005. Student evaluations of professors and professor self-ratings: Gender and divisional patterns. *Journal of Personnel Evaluation in Education*, Vol. 18, pp. 91-106.

Basow, S. A.; Phelan, J.; Capotosto, L. 2006. Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly*, Vol. 30, no. 1, pp. 25-35. 10.1111/j.1471-6402.2006.00259.x

Bedard, K., and P. Kuhn, 2005. Where Class Size Really Matters: Class Size and Student Ratings of Instructor Effectiveness, Department of Economics, University of California, Santa Barbara. http://econ.ucsb.edu/~kelly/ucsb4.pdf.

Bianchini, S., F. Lissoni, and M. Pezzoni, 2013. Instructor Characteristics and Students' Evaluation of Teaching Effectiveness: Evidence from an Italian Engineering School. *European Journal of Engineering Education*, Vol. 38, pp. 38-57. DOI: 0.1080/03043797.2012.742868.

Biernat, M.; Fuegen, K.; Kobrynowicz, D. 2010. Shifting standards and the inference of incompetence: Effects of formal and informal evaluation tools. *Personality and Social Psychology Bulletin*, Vol. 36, pp. 855-868. 10.1177/0146167210369483.

Bilimoria, D.; Stewart, A. J. 2009. "Don't ask, don't tell": The academic climate for lesbian, gay, bisexual, and transgender faculty in science and engineering. *NWSA Journal*, Vol. 21, pp. 85-103.

Bjork, Robert A.; Dunlosky, John; Kornell, Nate. 2013. Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology*. Vol. 64:417-444.

Blizard, Lida Marie. 2016. Faculty Members' Experiences of Cyberbullying by Students at One Canadian University: Impact and Recommendations, *International Research in Higher Education*. Vol. 1, no. 1, pp. 107-124. DOI: 10.5430/irhe.v1n1p107.

Boatright-Horowitz, S. L.; Soeung, S. 2009. Teaching White privilege to White students can mean saying good-bye to positive student evaluations. *American Psychologist*, Vol. 64, No. 6, pp. 574-575. DOI: 10.1037/a0016593.

Boring, Anne. 2015. Gender Biases in Student Evaluations of Teachers and their Impact on Teacher Incentives. Working Paper 2015-13. OFCE-PRESAGE-SCIENCES PO and LEDa-DIAL.

Boring, Anne; Ottoboni, Kellie; Stark, Philip. 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

Boysen, Guy A. 2015. Significant Interpretation of Small Mean Differences in Student Evaluations of Teaching despite Explicit Warning to Avoid Overinterpretation. *Scholarship of Teaching and Learning in Psychology* 1: 150-62. DOI: 10.1037/STL0000017.

Braga, Michel; Paccagnella, Marco; Pellizzaric, Michele. 2014. Evaluating students' evaluations of professors. *Economics of Education Review*. Vol. 41, pp. 71-88. DOI: 10.1016/j.econedurev.2014.04.002.

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

Canadian Association of University Teachers (CAUT). 2010. The Changing Academy? *CAUT Education Review*. Vol. 12, No. 1.

Canadian Association of University Teachers (CAUT). 2011. The Persistent Gap: Understanding male-female salary differentials amongst Canadian academic staff. *CAUT Equity Review*. No. 5.

Canadian Association of University Teachers (CAUT). 2016a. Evaluation of Teaching: CAUT Policy Statement. Ottawa: CAUT.

Canadian Association of University Teachers (CAUT). 2016b. Use of Student Opinion Surveys: CAUT Policy Statement. Ottawa: CAUT.

Canadian Association of University Teachers (CAUT). 2017a. *CAUT Almanac of Post-Secondary Education*. Ottawa: CAUT.

Canadian Association of University Teachers (CAUT). 2017b. *Model Clause on the Evaluation of Teaching Performance*. Ottawa: CAUT.

Canadian Association of University Teachers (CAUT). 2018a. *Teaching Dossier*. Ottawa. CAUT.

Canadian Association of University Teachers (CAUT). 2018b. *Underrepresented & Underpaid: Diversity & Equity Among Canada's Post-Secondary Education Teachers*. Ottawa: CAUT.

Canadian Federation of Students (CFS). 2014. Instructor Evaluations. In *Policy Manual: Issues Policy – As amended at the June 2015 National General Meeting*. Ottawa: CFS.

Canadian Federation of Students – Ontario (CFS-O). 2014. Quality in Higher Education. In *Issues Policy as amended at the 2017 Semi-Annual General Meeting*. Toronto: CFS-O.

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada. 2014. *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*. Ottawa: Secretariat on Responsible Conduct of Research.

Carrell, Scott E.; West, James E. 2010. Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy*, Vol. 118, No. 3, pp. 409-432. DOI: 10.1086/653808.

Cassidy, Wanda; Chantal Faucher, Margaret Jackson. 2014. The Dark Side Of The Ivory Tower: Cyberbullying Of University Faculty And Teaching Personnel. *Alberta Journal of Education Research*. 60(2): 279-299.

Clayson, Dennis 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*. Vol. 31, No. 1, pp. 16-30.

Commission on the Reform of Ontario's Public Services. 2012. *Public Services for Ontarians: A Path to Sustainability and Excellence*. Toronto: Commission on the Reform of Ontario's Public Services.

Cranton, P. A.; & Smith, R. A. 1986. A new look at the effect of course characteristics on student ratings of instruction. *American Educational Research Journal*, Vol. 23, No. 1, pp. 117-128.

Dua, E., & Lawrence, B. 2000. Challenging white hegemony in university classrooms: Whose Canada is it? *Atlantis*, Vol. 24, no. 2, pp. 105-122.

El-Alayli, A., Hansen-Brown, A.A. & Ceynar, M, 2018. Dancing Backwards in High Heels: Female Professors Experience More Work Demands and Special Favor Requests, Particularly from Academically Entitled Students. *Sex Roles*. Vol. 79, No. 3-4, pp 136-150. DOI: 10.1007/s11199-017-0872-6.

Ellis, Donna E. 2013. *Students' Responses to Innovative Instructional Methods: Exploring Learning-Centred Methods and Barriers to Change*. PhD Dissertation. Management Science: University of Waterloo.

Ellis, Donna E. 2014. What Discourages Students from Engaging with Innovative Instructional Methods: Creating a Barrier Framework. *Innovation in Higher Education*. DOI 10.1007/s10755-014-9304-5

Ewing, Andrew M. 2012. Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review*. Vol. 31, No. 1, pp. 141-154. DOI: 10.1016/j.econedurev.2011.10.002.

Ewing, V. L.; Stukas, A. A.; & Sheehan, E. P. 2003. Student prejudice against gay male and lesbian lecturers. *The Journal of Social Psychology*. Vol. 143, pp. 569-579. DOI: 10.1080/00224540309598464

Feeley, Thomas Hugh. 2002. Evidence of Halo Effects in Student Evaluations of Communication Instruction. *Communication Education*. 51:3, 225-236. DOI: 10.1080/03634520216519.

Feldman, Kenneth. A. 1978. Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education*. Vol. 9, No. 3, pp. 199-242.

Feldman, K. A. 1984. Class size and college students' evaluations of teachers and courses: A closer look. R*esearch in Higher Education*. Vol. 21, No. 1, pp. 45-116.

Field, Cynthia C.; Jones, Glen. A. 2016. *A Survey of Sessional Faculty in Ontario Publicly-Funded Universities*. Toronto: Centre for the Study of Canadian and International Higher Education, OISE-University of Toronto.

Flaherty, Colleen. 2018. Teaching Eval Shake-Up. Inside Higher Ed. May 22, 2018. https://www.insidehighered.com/news/2018/05/22/most-institutions-say-they-value-teaching-how-they-assess-it-tells-different-story

Foschi, M. 2000. Double standards for competence: Theory and research. *Annual Review of Sociology*. Vol. 26, pp. 21-42. DOI: 10.1146/annurev.soc.26.1.21.

Foster, Karen; Birdsell Bauer, Louise. 2018. *Out of the Shadows: Experiences of Contract Academic Staff*. Ottawa: Canadian Association of University Teachers.

Freishtat, Richard L. 2016a. *Expert Report on Student Evaluations of Teaching (SET)*. Prepared for the Ryerson Faculty Association and the Ontario Confederation of University Faculty Associations. Re. Ryerson University v Ryerson Faculty Association, 2018 (William Kaplan, arbitrator).

Freishtat, Richard L. 2016b. *Expert Supplemental Report on Student Evaluations of Teaching (SET)*. Prepared for the Ryerson Faculty Association. Re. Ryerson University v Ryerson Faculty Association, 2018 (William Kaplan, arbitrator).

Fries, Christopher J., and R. James McNinch. 2003. "Signed versus Unsigned Student Evaluations of Teaching: A Comparison." *Teaching Sociology*, vol. 31, no. 3, pp. 333–344. DOI: 10.2307/3211331

Gravestock, Pamela; Gregor-Greenleaf, Emily. 2008. *Student Course Evaluations: Research, Models and Trends*. Toronto: Higher Education Quality Council of Ontario.

Gurung, R.; Vespia, K. 2007. Looking good, teaching well? Linking liking, looks, and learning. *Teaching of Psychology*. Vol. 34, pp. 5-10.

Hamermesh, D. S.; Parker, A. M. 2005. Beauty in the classroom: Professors' pulchritude and putative pedagogical productivity. *Economics of Education Review*. Vol. 24, pp. 369-376. DOI: 10.1016/j.econedurev.2004.07.013

Hativa, Nira. 2013. *Student Ratings of Instruction: Recognizing effective teaching*. Oron Publications.

Hativa, Nira. 2015. Almost everything you ever wanted to know about Student Ratings of Instruction (SRI)… Presented at *Weighed in the Balance: Evaluating Teaching in Higher Education*. University of Windsor, Windsor (Nov. 30). http://ctl.uwindsor.ca/ctl/system/files/Hativa-PPT-Windsor.pdf

Henry, Frances. 2018. *Expert report (unpublished proceeding)*. Commissioned with the support of the Ontario Confederation of University Faculty Associations.

Henry, Frances; Dua, Enakshi; James, Carl E.; Kobayashi, Audrey; Li, Peter; Ramos, Howard; Smith, Malinda S.; eds. 2017. *The Equity Myth: Racialization and Indigeneity at Canadian Universities*, Vancouver: UBC Press.

Hessler, Michael; Pöpping, Daniel M.; Hollstein, Hanna; Ohlenburg, Hendrik; Arnemann, Philip H.; Massoth, Christina; Seidel, Laura M.; Zarbock, Alexander; Wenk, Manuel. 2018. Availability of cookies during an academic course session affects evaluation of teaching. *Medical Education*. Vol. 52, No. 10, pp. 1064-1072. DOI: 10.1111/medu.13627.

Hill, M.C.; K.K. Epps, 2010. The Impact of Physical Classroom Environment on Student Satisfaction and Student Evaluation of Teaching in the University Environment, *Academy of Educational Leadership Journal*, Vol. 14, pp. 65-79.

Ho, A. K., Thomsen, L.; Sidanius, J. 2009. Perceived academic competence and overall job evaluations: Students' evaluations of Black and European American professors. *Journal of Applied Social Psychology,* Vol. 39, pp. 389–406. 10.1111/j.1559-1816.2008.00443.x

Huston, Therese A. 2005. "Race and Gender Bias in Higher Education: Could Faculty Course Evaluations Impede Further Progress Toward Parity?" *Seattle Journal for Social Justice*: Vol. 4: No. 2, Article 34.

Isely, Paul; Singh, Harinder. 2005. Do Higher Grades Lead to Favorable Student Evaluations? *The Journal of Economic Education*. Vol. 36, No. 1, pp. 29-42. DOI: 10.3200/JECE.36.1.29-42.

Johnson, Stephanie K.; Murphy, Susan Elaine; Zewdie, Selama; Reichard, Rebecca J. 2008. The strong, sensitive type: Effects of gender stereotypes and leadership prototypes on the evaluation of male and female leaders. *Organizational Behavior and Human Decision Processes*. Vol. 106, No. 1, pp. 39-60. DOI: 10.1016/j.obhdp.2007.12.002.

Keeley, Jared W.; English, Taylor; Irons, Jessica; Henslee, Amber M. 2013. Investigating Halo and Ceiling Effects in Student Evaluations of Instruction. *Educational and Psychological Measurement*. Volume: 73 issue: 3, page(s): 440-457. DOI: 10.1177/0013164412475300.

Krautmann, A. C.; Sander, W. 1999. Grades and student evaluations of teachers. *Economics of Education Review*. Vol. 18, pp. 59-63.

Lake, D.A., 2001. Student Performance and Perceptions of a Lecture-based Course Compared with the Same Course Utilizing Group Discussion. *Physical Therapy*, Vol. 81, pp. 896-902.

Lampman, Claudia; Earl C. Crew, Shea D. Lowery & Kelley Tompkins (2016) Women Faculty Distressed: Descriptions and Consequences of Academic Contrapower Harassment, *NASPA Journal About Women in Higher Education*, 9:2, 169-189, DOI: 10.1080/19407882.2016.1199385.

Lapidot-Lefler, Noam; Barak, Azy. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*. 28: 434-443. DOI: 10.1016/j.chb.2011.10.014.

Legg, A. M.; & Wilson, J. H. 2012. RateMyProfessors. com offers biased evaluations. *Assessment & Evaluation in Higher Education*. Vol. 37, no. 1, pp. 89-97. DOI: 10.1080/02602938.2010.507299.

Levin, W. C. 1998. Age stereotyping: College student evaluations. *Research on Aging*. Vol. 10, no. 1, pp. 134-148.

Lindahl, Mary W; Unger, Michael L. 2010. Cruelty in Student Teaching Evaluations. *College Teaching*. Vol. 58, No. 3 pp. 71-76. DOI: 10.1080/87567550903253643.

Littleford, L. N.; Ong, K. S.; Tseng, A.; Milliken, J. C.; Humy, S. L. 2010. Perceptions of European American and Black instructors teaching race-focused courses. *Journal of Diversity in Higher Education*, Vol. 3, pp. 230-244. DOI: 10.1037/a0020950

Luther, Rashmi; Whitmore, Elizabeth; Moreau, Bernice. 2001. Making Visible the Invisible: The Experience of Faculty of Colour and Aboriginal Faculty in Canadian Universities. In Rashmi Luther, Elizabeth Whitmore, Bernice Moreau (eds.), *Seen but Not Heard: Aboriginal Women and Women of Colour in the Academy*. Ottawa: Canadian Research Institute for the Advancement of Women.

MacNell, Lillian; Driscoll, Adam; Hunt, Andrea N. 2015. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education*, 40: 291. DOI: 10.1007/s10755-014-9313-4.

Marsh, H.W.; Roche, L.A. 1997. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, Vol. 52, No. 11, pp. 1187-1197.

McKeachie, W.J., 1997. Student ratings: The validity of use. *American Psychologist*, Vol. 52, No. 11, pp. 1218–1225. DOI: 10.1037/0003-066X.52.11.1218

McPherson, M. A. 2006. Determinants of How Students Evaluate Teachers. *Journal of Economic Education*. Vol. 37, no. 1, pp. 3-20.

McPherson, M. A.; Jewell, R. T. 2007. Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly*, Vol. 88, No. 3, pp. 868-881. DOI: 10.1111/j.1540-6237.2007.00487.x.

McPherson, Michael A; Jewell, Todd R; Kim, Myungsup. 2009. What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes. *Eastern Economic Journal*. Vol. 35, No. 1, pp. 37-51. DOI: 10.1057/palgrave.eej.9050042.

Mengel, Friederike; Sauermann, Jan ;Zölitz, Ulf. 2018. Gender Bias in Teaching Evaluations, *Journal of the European Economic Association*, jvx057, DOI: 10.1093/jeea/jvx057

Meyer, J. P.; Doromal, J. B.; Wei, X.; Zhu, S. 2017. A criterion-referenced approach to student ratings of instruction. *Research in Higher Education*, Vol. 58, pp. 545-567. doi:10.1007/s11162-016-9437-8

Ministry of Advanced Education and Skills Development, Ontario (MAESD). 2017. Strategic Mandate Agreement, 'X' University, 2017-20, Draft Submission Template between the Ministry of Advanced Education and Skills Development and 'X University'. Toronto: MAESD.

Ministry of Advanced Education and Skills Development, Ontario; Carleton University. 2018. 2017-20 Strategic Mandate Agreement: Carleton University.

Ministry of Advanced Education and Skills Development, Ontario; McMaster University. 2018. 2017-20 Strategic Mandate Agreement: McMaster University.

Ministry of Advanced Education and Skills Development, Ontario; Western University. 2018. 2017-20 Strategic Mandate Agreement: Western University.

Ministry of Training, Colleges and Universities, Ontario (MTCU). 2015. *Focus on Outcomes – Centre on Students: Perspectives on Evolving Ontario's University Funding Model*. Toronto: MTCU.

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

Morrison, Rodger. 2011 A comparison of online versus traditional student end-of-course critiques in resident courses, *Assessment & Evaluation in Higher Education*, Vol. 36, No. 6, pp. 627-641, DOI: 10.1080/02602931003632399

Mount Allison University v Mount Allison Faculty Association, 2015 CanLII 94980 (ON LA), <http://canlii.ca/t/gp3c6>, retrieved on 2017-06-29.

Muhs, Gabriella Gutiérrez y; Yolanda Flores Niemann, Carmen G. González, Angela P. Harris (eds.). 2012. *Presumed Incompetent: The Intersections of Race and Class for Women in Academia*. Boulder: University Press of Colorado.

National Tertiary Education Union (NTEU). 2018. *Staff Experience of Student Evaluation of Teaching and Subjects/Units*. Melbourne: NTEU.

Nisbett, Richard E.; Wilson, Timothy D. 1977. The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, Vol. 35, no. 4, pp. 250-256. DOI: 10.1037/0022-3514.35.4.250.

Nowell, Clifford; Lewis R. Gale and Bruce Handley. 2010. Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment & Evaluation in Higher Education*. Vol. 35, No. 4, pp. 463–475.

Office of the Auditor General of Ontario (OAG). 2012. *Annual Report*. Toronto: Queen's Printer for Ontario.

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

Ogier, J. 2014. *Annual Report*. Toronto: Queen's Printer for Ontario.

Ontario. *Human Rights Code*, R.S.O. 1990, CHAPTER H.19. Accessed: June 14, 2017.

Ontario. *Occupational Health and Safety Act*, R.S.O. 1990, CHAPTER O.1. Accessed: June 14, 2017.

Ontario Confederation of University Faculty Associations (OCUFA). 2016. Pay equity among faculty at Ontario's universities: OCUFA's Submission to the Ontario Gender Wage Gap Steering Committee. Toronto: OCUFA.

Ontario Confederation of University Faculty Associations (OCUFA). 2018. *Guidelines for Teaching Awards Nominations*. Toronto: OCUFA.

Ontario Human Rights Commission. 2008. *Human Rights at Work*. Toronto: The Commission: Carswell.

Ontario Human Rights Commission. 2013. *Guide to your rights and responsibilities under the Human Rights Code*. Toronto: Government of Ontario.

Ontario Undergraduate Student Alliance (OUSA). 2014. *Policy Paper: Accountability*. Toronto: OUSA.

Ontario Undergraduate Student Alliance (OUSA). 2015a. *Those Who Can, Teach: Evolving Teaching and Learning Strategies in Ontario's Universities*. Toronto: OUSA.

Ontario Undergraduate Student Alliance (OUSA). 2015b. *Policy Paper: Teaching and Assessment*. Toronto: OUSA.

Ontario Undergraduate Student Alliance (OUSA). 2017. *Policy Paper: Accountability*. Toronto: OUSA.

Ontario Undergraduate Student Alliance (OUSA). 2018. *Policy Paper: Teaching and Assessment*. Toronto: OUSA.

Ontario Universities Council on Quality Assurance (OUCQA). 2016. *Quality Assurance Framework and Guide*. Toronto: OUCQA.

Padilla, Amado M. 1994. "Ethnic Minority Scholars, Research, and Mentoring: Current and Future Issues." *Educational Researcher*, Vol. 23, no. 4, pp. 24-27.

Perry, A. R.; Wallace, S. L.; Moore, S. E.; Perry-Burney, G. D. 2015. Understanding student evaluations: A black faculty perspective. *Reflections*, Vol. 20, no. 1, pp. 29-35.

Pittman, Chavellla. T. 2010. Race and gender oppression in the classroom: The experiences of women faculty of color with White male students. *Teaching Sociology*, Vol. 38, no. 183-196. DOI: 10.1177/0092055X10370120.

Reid, Landon. 2010. The Role of Perceived Race and Gender in the Evaluation of College Teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education*, Vol. 3, No. 3, DOI: 10.1037/a0019865

Rhea, Nancy; Rovai, Alfred; Ponton, Michael; Derrick, Gail; Davis, John. 2007. The Effect of Computer-Mediated Communication on Anonymous End-of-Course Teaching Evaluations. *International Journal on ELearning*. 6(4): 581-592.

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

Risquez, Anglica; Elaine Vaughan and Maura Murphy. 2015. Online student evaluations of teaching: what are we sacrificing for the affordances of technology? *Assessment & Evaluation in Higher Education*, Vol. 40, No. 1, pp. 120–134, DOI: 10.1080/02602938.2014.890695.

Rosette, Ashleigh Shelby; Zhou Koval, Christy; Ma, Anyi; Livingston, Robert. 2016. Race matters for women leaders: Intersectional effects on agentic deficiencies and penalties. *The Leadership Quarterly*. Vol. 27, No. 3, pp. 429-445. DOI: 10.1016/j.leaqua.2016.01.008.

Ryerson University v Ryerson Faculty Association. 2018. CanLII 58446 (ON LA), <http://canlii.ca/t/hsqkz>, retrieved on 2018-06-28.

Rytmeister, Cathy. 2013. "Go away and just let me teach!" The escalating use of student evaluation to measure teaching. Presented at the *NTEU National Teaching Conference*. Sydney, Australia (4-5 April). http://www.nteu.org.au/library/download/id/3739

Samuel, E.; Wane, N. 2005. "Unsettling relations": Racism and sexism experienced by faculty of color in a predominantly white Canadian university. *Journal of Negro Education*, Vol. 74, no. 1, pp. 76-87.

Schmidt, B. 2015. Gendered Language in Teacher Reviews, http://benschmidt.org/profGender (last retrieved 30 September 2016)

Schueler, G. F. 1988. The evaluation of teaching in philosophy. *Teaching Philosophy*. Vol. 11, no. 4, pp. 345-348.

Scott, Krystyn A.; Brown. Douglas J. 2006. Female first, leader second? Gender bias in the encoding of leadership behavior. *Organizational Behavior and Human Decision Processes*. Vol. 101, No. 2, pp. 230-242. DOI: 10.1016/j.obhdp.2006.06.002.

Short, H.; Boyle, R.; Braithwaite, R.; Brookes, M.; Mustard, J.; Saundage, D. 2008. A comparison of student evaluation of teaching with student performance. In *OZCOTS 2008: Proceedings of the 6th Australian Conference on Teaching Statistics*, pp. 1-10. OZCOTS.

Sinclair, Lisa; Kunda, Ziva. 2000. Motivated Stereotyping of Women: She's Fine if She Praised Me but Incompetent if She Criticized Me. *Personality and Social Psychology Bulletin*. Vol. 26, No. 11, pp. 1329-1342.

Smith, Bettye. P. 2007. Student ratings of teacher effectiveness: An analysis of end-of-course faculty evaluations. *College Student Journal*, Vol. 41, no. 4, pp. 788-800.

Smith, G., & Anderson, K. J. 2005. Students' ratings of professors: The teaching style contingency for Latino/a professors. *Journal of Latinos and Education*, Vol. 4, pp. 115–136.

Smith, Mary-Atoinette. 2012. Free at Last! No More Performance Anxieties in the Academy 'cause Stepin Fetchit Has Left the Building. In Gabriella Gutiérrez y Muhs, Yolanda Flores Niemann, Carmen G. Gonzalez and Angela P. Harris (eds.), *Presumed Incompetent: the intersections of race and class for women in academia*. (pp. 408-420) Boulder: University Press of Colorado.

Sproule, Robert. 2000. Student Evaluation of Teaching: A Methodological Critique of Conventional Practices. *Education Policy Analysis Archives*. Volume 8 Number 50.

Stark, Philip B. 2015. Teaching evaluations: Truthful or truthy? Presented at the *Third Lisbon Research Workshop on Economics, Statistics, and Econometrics of Education*. Lisbon, Portugal (23-24 January). http://www.stat.berkeley.edu/~stark/Seminars/setLisbon15.htm

Stark, Philip B. 2016. *Expert Report on Student Evaluations of Teaching (Faculty Course Surveys)*. Prepared for the Ryerson Faculty Association and the Ontario Confederation of University Faculty Associations. Re. Ryerson University v Ryerson Faculty Association, 2018 (William Kaplan, arbitrator).

Stark, Philip B.; Freishtat, Richard. 2014. An evaluation of course evaluations. *ScienceOpen Research*. DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1.

Stowell, Jeffrey R.; William E. Addison and Jennifer L. Smith. 2012. Comparison of online and classroom-based student evaluations of instruction. *Assessment & Evaluation in Higher Education*. Vol. 37, no. 4, pp. 465-473. DOI: 10.1080/02602938.2010.545869

Stroebe, Wolfgang. 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, No. 6, pp. 800-816. DOI: 10.1177/1745691616650284.

Storage, D.; Horne, Z.; Cimpian, A.; Leslie, S. J. 2016. The Frequency of "Brilliant" and "Genius" in Teaching Evaluations Predicts the Representation of Women and African Americans across Fields. *PloS one*, 11(3), e0150194.

Subtirelu, N.C., 2015. "She does have an accent but…": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society*. Vol. 44, pp. 35-62. DOI 10.1017/S0047404514000736

University and Colleges Union-Queen's University Branch (UCU-QUB). 2013. *Students' Evaluation of Teaching (SETS) Report*. Belfast: UCU-QUB. https://www.ucuqub.org/reports/

University of Windsor and University of Windsor Faculty Association (Policy grievance), (L.A., 2007-02-19), SOQUIJ AZ-51135637.

University of Windsor Faculty Association v. University of Windsor, 2008 CanLII 23711 (ON SCDC), <http://canlii.ca/t/1wzvg>, retrieved on 2017-06-26.

Uttl Bob; Smibert, Dylan. 2017. Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ* 5:e3299. DOI: 10.7717/peerj.3299.

Uttl, Bob; White, Carmela A.; Wong Gonzalez, Daniela. 2017. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, Volume 54, Pages 22-42, DOI: 10.1016/j.stueduc.2016.08.007.

Vaillancourt, Tracy. 2013. Students Aggress Against Professors in Reaction to Receiving Poor Grades: An Effect Moderated by Student Narcissism and Self-Esteem. *Aggressive Behavior*. Volume39, Issue1, Pages 71-84.

Vasey, Craig; Carroll, Linda. 2016. How Do We Evaluate Teaching? Findings from a survey of faculty members. *Academe* (Association of American University Professors), Volume 102, Number 3, May-June, 2016.

Vasta, Ross; Sarmiento, Robert F. 1979. Liberal Grading Improves Evaluations but not Performance, *Journal of Educational Psychology*, 71(2), 207-211.

Weinberg, Bruce A.; Hashimoto, Masanori; Fleisher, Belton M. 2009. Evaluating Teaching in Higher Education. *The Journal of Economic Education*. Volume 40, Issue 3, Pages 227-261. DOI: 10.3200/JECE.40.3.227-261.

Wendorf, C. A., & Alexander, S. (2005). The influence of individual- and class-level fairness-related perceptions on student satisfaction. *Contemporary Educational Psychology*, 30:2, 190–206.

Wilson, J. P., Remedios, J. D., & Rule, N. O. 2017. Interactive effects of obvious and ambiguous social categories on perceptions of leadership: When double-minority status may be beneficial. *Personality and Social Psychology Bulletin*. Vol. 43, pp. 888-900. doi:10.1177/0146167217702373

Wolbring, T., and P. Riordan, 2016. How Beauty Works. Theoretical Mechanisms and Two Empirical Applications on Students' Evaluations of Teaching, *Social Science Research*, Vol. 57, pp. 253-272. DOI: 10.1016/j.ssresearch.2015.12.009.

Woodhams, Carol; Lupton, Ben; Cowling, Mark. 2015. The Snowballing Penalty Effect: Multiple Disadvantage and Pay. *British Journal of Management*. Vol. 26, No. 1, pp. 63-77. DOI: 10.1111/1467-8551.12032.

Worthington, Andrew.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. DOI: 10.1080/02602930120105054.

Wright, Alan W.; Hamilton, Beverley; Mighty, Joy; Scott, Jill; Muirhead, Bill. 2014. *The Ontario Universities' Teaching Evaluation Toolkit: Feasibility Study*. Windsor: Centre for Teaching and Learning Reports.

Appendix A: Institutional documentation collected

Institutional documents in the collection are mostly those available publicly on university websites during early to mid-2017. Some were made available by OCUFA member faculty associations. The following table does not include institutions for which research ethics but no other documentation was available. The categories are explained below.

	Senate policy	Admin. guidelines	Survey instrument	Other related	Research ethics
Algoma			✓	✓	✓
Brock		✓		✓	✓
Carleton		✓	✓	✓	✓
Guelph		✓		✓	✓
Lakehead	✓		✓	✓	✓
Laurentian	✓		✓		✓
McMaster	✓				✓
Nipissing			✓	✓	✓
NOSM				✓	✓
OCAD University			✓	✓	✓
Ottawa		✓	✓	✓	✓
Queen's			✓	✓	✓
Royal Military College				✓	✓
Ryerson			✓	✓	✓
Saint Paul				✓	✓
Toronto	✓	✓	✓	✓	✓
Trent	✓			✓	✓
UOIT	✓	✓		✓	✓
Waterloo				✓	✓
Western	✓			✓	✓
Wilfrid Laurier			✓	✓	✓
Windsor	✓	✓	✓		✓
York	✓	✓		✓	✓

Senate policy: policies on student questionnaires specifically, and on teaching evaluation.
Administrative guidelines: instructions for administering questionnaires and results.
Survey instruments: samples of the questionnaires in use; some are department or faculty specific; others are university-wide instruments.
Other related: other related policies; informal directives for administration; frequently asked questions.
Research ethics: policies, terms of reference, guidelines.

Appendix B: Methodological issues in use of student questionnaires to assess teaching effectiveness

J. L. Michela, Ph.D., Department of Psychology, University of Waterloo, Ontario, Canada - February, 2019

"When the facts change, I change my mind. What do you do?" – John Maynard Keynes (maybe)[1]

This document summarizes and interprets empirical research findings concerning issues that arise when students' ratings on post-course questionnaires are used for evaluation of quality or effectiveness of instruction in university courses. The document has been revised for a working group that was convened by the Ontario Confederation of University Faculty Associations (OCUFA) to study these issues.[2] This working group was convened in recognition of the divergence of (a) the increasing weight being placed on student questionnaire ratings at Ontario universities and (b) the increasing body of evidence indicating that student questionnaires cannot bear this weight because of inherent limitations.

This unbearable weight has been accumulating since the 1920s. Algozzine et al. (2004) state:

> Originally intended to represent private matters between instructors and students regarding strengths and weaknesses, [student ratings-based] course evaluation information often has been put to another, more controversial use: to provide input in the annual evaluations, as well as for salary, promotion, and tenure decisions. (p. 135)

Drawing on language used in the field of evaluation research, these latter uses are "summative" uses. In contrast, "formative" use of students' ratings disconnects the ratings from personnel decisions, focusing instead on obtaining diagnostic information usable by an instructor for improving instructional performance. Primarily formative use, as from the 1920s, continued until the 1981s (Hornstein, 2017). This statement will vigorously endorse formative use while strenuously urging an end to summative use.

This document is not a literature review as such. It cites several authoritative literature reviews and, as author, I have read many of what I consider to be the more important and more recent primary sources as well. The related literature is, however, vast.

The right way to assess the associated issues is not to tabulate, in some quasi-meta-analytic manner, how many writers have taken one position and how many have taken another. Unwisely, proponents of summative use of students' ratings often approach the literature essentially in this manner, insisting that, with many voices on both sides, the question of student ratings' validity for summative use is merely a matter of opinion. However, from my extensive reading, my discussions locally, and my

---

[1] https://quoteinvestigator.com/2011/07/22/keynes-change-mind/

[2] The earlier version of this statement was produced with support from several members of the Department of Psychology at the University of Waterloo, who submitted their statement of objection to a project team for the redesign of student questionnaires.

correspondence with academics elsewhere, it has become clear that proponents of summative use of student questionnaires rely on claims in the literature that are outdated or transparently weakly supported—claims that, unfortunately, are numerous. Either by cherry-picking their sources or by choosing not to read sources that are manifestly pertinent, proponents pretend something other than the truth of the matter: The bulk of empirical studies and thorough reviews published in approximately the last 8 years are utterly damning about summative use.

One way to read the present statement is as a *prima facie* case against summative use. A *prima facie* case is "a case in which the evidence produced is sufficient to enable a decision or verdict to be made unless the evidence is rebutted."[3] When similar research findings and lines of analysis as in this document were presented in a recent arbitration decision against a university administration in my Canadian province of Ontario, what happened was telling. The evidence, "which came in the form of expert testimony and peer reviewed publications," "was largely uncontested."[4] That is, in that arbitration, the *prima facie* case stood up. I believe that the reason for that administration's non-response was that a counterbalancing body of analysis, which takes account of contemporary evidence, simply does not exist. In the nearly forty years since publication of the Cohen (1981) paper that has been so central to proponents' defense of summative validity (e.g., Benton & Cashin, 2012), the emperor has been not only defrocked, but totally denuded (Nilson, 2012; Uttl et al., 2017).

Before presenting the case against summative use, it may be helpful to highlight some further points that arose in the arbitration decision just referenced. First it should be noted that arbitrators are chosen by disputing parties as someone with no axe to grind and with the sophistication to receive and evaluate relevant facts and analysis dispassionately, if not fully expertly. After receiving input from those whom I consider to be the top experts in the field (see Stark, 2015; Stark & Freishstat, 2014), the arbitrator declared definitively and accurately that student surveys of instruction and instructors are not valid as measures of teaching effectiveness. The key reason for this invalidity is the large impact of many factors extraneous to instruction, pertaining to instructor characteristics (e.g., gender) and other factors (e.g., course characteristics such as elective versus required; class size; traditional teaching versus innovative). More will be said about these factors later here, labeled as biasing factors because of their extraneous nature.

The arbitrator's report addressed the ensuing question: What is being measured by rating scales which, on their face, misleadingly appear to be measures of teaching effectiveness? In essence: "student satisfaction." I hold that although this is a helpful characterization, it is imprecise, and it gives too much credit to the notion that student's ratings on course evaluation scales "measure" student satisfaction. Satisfaction is neither self-defining nor simple (Seth, Deshmukh, & Vrat, 2005). We can properly say that the ratings "reflect" satisfaction to some degree. But it is not meaningful to say that many of the

---

[3] https://www.dictionary.com/browse/prima-facie-case

[4] https://www.canlii.org/en/on/onla/doc/2018/2018canlii58446/2018canlii58446.html

bias-based components of variation in students' ratings (i.e., the extent of variation as a function of instructor gender or attractiveness; of many non-instructional course characteristics such as elective versus required course, etc.) are aspects of satisfaction as ordinarily conceived. The term "affect," as will be explained more later, is better. For example, there is a tendency for students to like or prefer to be taught by younger, more attractive male instructors. This tendency in the ratings provides no useful information either for summative or formative evaluation and teaching improvement.

The basis for the case against summative use of students' ratings is as follows.

1. Extraneous, "biasing" factors render student questionnaires invalid for summative evaluation.

2. Summative use of student questionnaires harms students' learning and instructors' integrity.

3. Commonly proposed remedies for bias and other sources of inaccuracy (e.g., "halo") will not be effective and bias will remain.

4. Student questionnaires nevertheless may be useful for formative evaluation and other purposes.

5. The fact that student questionnaires are used so widely across universities for "summative" evaluation gives no assurance of their appropriateness for that purpose. So-called "best practices" are ineffective.

6. The alternatives to student questionnaires that have been proposed in the literature can be expected to carry *less* bias and to do *more* to promote effective instruction.

7. When the facts change, beliefs and behaviour should change, too.

1. Extraneous, "biasing" factors render student questionnaires invalid for summative evaluation.

Validity of students' ratings for summative use requires that instructors who *truly* are higher performers receive relatively higher ratings by students. Formative use, to which we will return later, does not share this requirement, because a given instructor's trajectory over time is what matters primarily. That is, formative use operates one instructor at a time.

The essential problem for summative validity of students' ratings is that these ratings have been shown to be responsive to (i.e., "biased" as a result of) a host of factors, including instructor gender, ethnicity, age, physical attractiveness, speaking or personal style, whether a course is in or out of major (or otherwise "important"), its class size, time of day, and so forth. They are called biasing factors because, in themselves, these factors have nothing to do with instructional effectiveness. The relevant research has been conducted in various countries over recent decades (e.g., Stark & Freishstat, 2014). The magnitudes of influence of these biasing factors can be quite sizable (Stark, 2015). Collectively the biasing factors have more impact on instructors' rank orders of ratings than whatever impact is exerted by true instructional effectiveness (Stark, 2015; Boring, Ottoboni, & Stark, 2016).

Wieman (2015) summarizes and illustrates the problem as follows:

> Many researchers who argue for the value of student evaluations do so by showing that, within a limited context, the evaluations correlate with desirable outcomes. But that is not a sufficient condition to be suitable for evaluating an instructor's teaching as a guide for improvement or as part of the incentive system. The correlation with desirable outcomes must hold over a broad range of contexts and courses and be much larger than the correlations with other factors not under the instructor's control for that range of contexts and courses. Student evaluations fall far short of meeting that condition.
>
> To put this in more concrete terms, the data indicate that it would be nearly impossible for a physically unattractive female instructor teaching a large required introductory physics course to receive as high an evaluation as that of an attractive male instructor teaching a small fourth-year elective course for physics majors, regardless of how well either teaches. (p. 9)

In Weiman's analysis there is a breakdown in the necessary alignment between the rank orders, highest to lowest, for how instructors are rated in relation to how well they teach. Similar effects of breaking this necessary alignment can arise from the many other biasing factors besides these of instructor gender, attractiveness, and nature of course.

Such misalignment is cataclysmic for fairness in personnel decisions. For example, even if there is a "small" unfavourable effect upon ratings of being a female instructor, the effect on pay accumulates over the years. The same can be expected for being an innovative teacher (an effect that arises because

some students do not view instruction as of high quality when they must "teach themselves" in active learning) or for speaking with a foreign accent (in some instances), for example. And the effects of the biases are additive not only across time but also at any given point in time. A particular instructor could get lucky and have the biasing factors balance out, but another could provide innovative instruction in a poorly suited room scheduled at an undesired time—and be out of luck.

Some defenders of summative use insist that instructors who are disorganized, mumbling, or otherwise unclear are obviously likely to obtain lower ratings, compared with crisp, clear instructors. However, the issue is *not* whether there is *any* association between students' ratings (see Stark, 2015) and instructional effectiveness. The issue is that, with the large number of factors that can inappropriately depress or elevate ratings, any expectation of *sufficient* alignment of ratings and actual effectiveness is beyond unrealistic. The net effect is that pay and promotion are affected mightily by the biasing factors, which is deeply unjust.

Another line of defense for summative use is that people who use students' ratings in personnel decisions, such as within a peer review committee for annual performance evaluation, can adjust for biasing factors and thus recover summative validity. This is another fantasy, for reasons given later (point 3).

Instructors and administrators with an open mind will then ask: What *is* measured by course ratings questionnaires if not primarily teaching effectiveness? As suggested earlier in this document, it is some variation on the themes of liking of the instructor or the course itself (e.g., Nuhfer, 2010). One estimate holds that 50 to 75 per cent of the variance in student ratings is explained by personal style, involving a combination of congeniality, confidence, optimism, and enthusiasm (Clayson, 2011, cited in Nilson, 2012). Satisfaction with, or enjoyment of a course has a similarly huge association with students' ratings of effectiveness per se. Stark (2015) reports a correlation of 0.75 between rated enjoyment and "instructor effectiveness" and 0.80 between rated enjoyment and "course effectiveness." Clearly such ratings of "effectiveness" must not be taken at face value.

Unfortunately instructors or courses can be liked for the wrong reasons. Nilson (2012, p. 220) tells us what we already know about many students, if we are willing to admit it:

> Today's students want an instructor they can relate to, who is expressive and energetic, and who cares about and empathizes with them (Chonko, 2004; Clayson, 2011; Kelley, Conant, & Smart, 1991; La Lopa, 2011; Walsh & Maffei, 1994). If an instructor's ability to project such a persona motivated students to learn more, then ratings and learning would be positively related, but . . . they are not. Students also want the good grades that they were accustomed to getting before college (Pryor et al., 2011) and to preserve their positive self-concept. Thus, they accordingly reward faculty who give them high grades with high ratings.

2. Summative use of student questionnaires harms students' learning and instructors' integrity.

2.1 It has become clear from recent research that student questionnaire ratings can be inversely related to instructional effectiveness.

Some decades ago student questionnaire ratings were more defensible for summative evaluation in light of research that was interpreted as showing positive associations with learning (Cohen, 1981). Since that time, the role and weight of student questionnaire ratings have changed, as has the university context (toward a more customer-responsive organization, at least in part) along with the student body (which demands customer responsiveness, among other things). Nilson (2012) argues that these factors alone render the pre-1981 findings of little relevance, and she notes that more recent, equivalent findings have not emerged. Adding to doubts about Cohen's research as a major justification for summative use, recent re-analysis of the pre-1981 studies argues that the conclusions of Cohen's review were fundamentally mistaken (Uttl et al, 2017). This would, of course, explain why new, equivalent findings have not emerged.

In the best contemporary research, several studies have tracked students across sequenced courses, examining whether students who gave higher ratings in the earlier course performed better in the later course. A study involving introductory and later economics (Weinberg et al., 2009) encompassed approximately 45,000 enrollments in almost 400 offerings over 10 years. Strobe's (2016, p. 808) summary states:

> As in all previous research, course ratings were positively associated with the grades in the concurrent course. However, when course evaluation was used as a predictor of student performance in subsequent courses (controlling for current grades) no association was found.
>
> Two further studies involved random assignment of students to particular sections of courses.

Strobe summarizes the first, involving more than 10,000 students, as follows:

> Student evaluations of a concurrent course were significantly negatively correlated with [later] grades. Carrell and West (2010) concluded that their "results show that student evaluations reward professors who increase achievement in the contemporaneous course being taught, not those who increase deep learning."

The second was conducted in a business school in Italy (Braga et al., 2014). Stroebe states:

> When performance in future courses was used as criterion of learning, teacher evaluations showed a negative association. As Braga et al. (2014) concluded, "Teachers who are more effective in promoting future performance receive worse evaluation from their students. This

> relationship was statistically significant for all items of the rating instrument, (except for ratings of course logistics), and was of sizeable magnitude."

Strobe (2016) cites these findings within a larger analysis of the likely connection between summative use of student questionnaires and grade inflation that has been widespread across academia during the period in which this summative use has been given ever-greater weight. Greater student learning over this period is not a likely explanation for the rise in grades, because students' hours of effort on courses have declined considerably. Also, Scholastic Aptitude Test (SAT) scores have not increased since 1963.

Stroebe concludes that grade inflation and potential reduction in student learning stem largely from the unfortunate incentives from summative use of student questionnaires:

> Because many instructors believe that the average student prefers courses that are entertaining, require little work, and result in high grades, they feel under pressure to conform to those expectations.

Even some of the defenders of student questionnaires acknowledge that summative use of student questionnaires creates an incentive for instructors "to grade higher and to lower the level of difficulty/workload," or to manipulate other non-instructional factors, "in order to receive higher ratings from students" (Hativa, 2013).

Tellingly, more lenient grading was shown in an experiment to yield higher ratings on student questionnaires (Vasta & Sarmiento, 1979, cited in Stroebe).

Stark (2015) cites additional sources on the specific point that emphasis on student ratings has led to grade inflation (slide 64).

2.2 Students' learning suffers further from summative use of student questionnaires because this use can deter use of innovative teaching approaches that are recommended by experts in postsecondary instruction.

The literature review section of a related dissertation (Ellis, 2013) included the following:

> Felder and Brent [1996] indicate that "when confronted with the need to take more responsibility for their own learning, students may grouse that they are paying tuition—to be taught, not to teach themselves… course-end ratings may initially drop. It is tempting for professors to give up in the face of all that, and many unfortunately do" (p.43). Hockings (2005) corroborates this finding...." (p. 10).

Wieman (2015) also provides corroboration that instructors "fear that adopting more effective research-based teaching methods will lower student evaluation scores" (p. 10).

More difficult to substantiate are suggestions that instructors are deterred from addressing emotionally challenging topics in their courses, and that courses that do address such questions are inherently disadvantaged in the student questionnaire ratings. Ironically, some instructors of courses on gender bias itself (and related topics) will insist that this disadvantage does operate, yet they feel they would be hypocrites to soften the emotional challenges that they pose to students in an attempt to obtain more favourable ratings on student questionnaires. These instructors' perceptions are entirely consistent with the observation in point 1 of this statement—that ratings on student questionnaires predominantly reflect how students *feel* about a course or instructor. It does not feel good to be challenged about one's biases, including about one's lack of awareness of bias toward social groups in society, or being told about the high difficulty of countering it.

2.3 Summative use of students' ratings is not beneficial "on balance."

A final point to address in this section concerns the belief that the risks just described (e.g., watering down one's course) are worth taking to avoid a countervailing risk, namely that without summative use of student questionnaires, instructors will put less effort into their teaching, and student learning will suffer for that reason. This belief reflects a very incomplete psychological analysis. It epitomizes McGregor's (1960) "Theory X": the notion that workers need to be leveraged with carrots and sticks. McGregor went on to describe "Theory Y" as an alternative belief set that should guide organizational leaders. Under this mindset, leaders can best promote worker motivation by creating conditions in which workers can pursue work outcomes that they value intrinsically. For professors or other instructors, these outcomes include education and other development of students.

This idea expressed so well almost 60 years ago is recurrent in management (e.g., Alderfer, 1972; Hackman & Oldham, 1980; Likert, 1967), in education (Ramirez, 2001), and in society broadly (Kohn, 2018). Ramirez (2001) addressed merit pay policies across the educational sector. He reached the same conclusion as McGregor based on scholarship by Herzberg and Glasser (cited in Ramirez). He also quotes the preeminent writer and practitioner for quality improvement, W. Edwards Deming (1993), as having decried extrinsic reward systems . . .

> that squeeze out from an individual, over his lifetime, his innate intrinsic motivation, self-esteem, dignity. They build into him fear, self-defense, extrinsic motivation. We have been destroying our people, from toddlers on through university, and on the job. (p. 124)

Suggesting applicability specifically to the university sector, Wieman (2015) has commented:

> Faculty almost universally express great cynicism about student evaluations and about the institutional commitment to teaching quality when student evaluations are the dominant measure of quality. At every institution I visit, this sentiment is voiced. (p. 10)

Bringing this matter up to the present, Kohn (2018) explains:

> The problem is the outdated theory of motivation underlying the whole idea of treating people like pets—that is, saying: Do this, and you'll get that.
>
> Indeed, various researchers over the last half-century have admitted to being surprised by the ineffectiveness or destructiveness of rewards when money was offered to adults for succeeding at a tricky task, when movie tickets or praise was offered to children for tasting an unfamiliar beverage (the kids liked the beverage less than those who received neither a tangible nor a verbal reward), when merit pay failed to improve teachers' performance, and when incentives didn't increase seatbelt use or help people lose weight and keep it off.
>
> The best that carrots – or sticks – can do is change people's behavior temporarily. They can never create a lasting commitment to an action or a value, and often they have exactly the opposite effect ... contrary to hypothesis.

In my statement here, I may perhaps be coming across as overly argumentative on this point. But actually, when I have discussed summative use with colleagues at my university and elsewhere, this topic of motivation consistently comes up as a justification for continued summative use, even in light of, and recognition of the operation of bias and other issues. As one leader at my university said publicly, it is human nature to respond to incentives. Well, Kohn and the other sources cited here agree in a sense, but they emphasize that this response by instructors is merely some degree of compliance, without commitment, to whatever is incentivized. Granted, summative use might incentivize teaching that is better in some respects (e.g., doing more preparation for a lecture). But there are other ways to accomplish this without simultaneously incentivizing worse teaching, as Stroebe (2016) and other previously cited sources point to (e.g., due to avoiding workload challenge; avoiding instructional innovation; and marking too easily).

Wise practitioners of quality *improvement* do not focus on rewards and punishments. Deming (1993), cited earlier, made clear that the *systems* surrounding work and the *capabilities* of workers (not their motivations) are among the first places to look and intervene. Along these same lines, Kohn states:

> Working with people to help them do a job better, learn more effectively, or acquire good values takes time, thought, effort and courage. Doing things to people, such as offering them a reward, is relatively undemanding for the rewarder, which may help to explain why carrots and sticks remain stubbornly popular despite decades of research demonstrating their failure.

3. Commonly proposed remedies for bias and other sources of inaccuracy (e.g., "halo") will not be effective and bias will remain.

A common misconception was illustrated at my own university, where a study panel's report asserted that a properly designed and implemented training and orientation program can enhance the utility of course evaluations. This statement is without foundation with regard to summative use. To the contrary, such programs of training and orientation—either for students providing the ratings, or peer review committees using the ratings data—more likely will do little or nothing to reduce the impact of the biasing factors, given the findings of research sketched next.

3.1 People are very poor at recognizing the operation of bias or adjusting for it.

Evidently part of the thinking behind claims about the utility of bias training and orientation is that, with effort and good intentions, people can significantly reduce or eliminate biased response. Unfortunately, as expressed by a member of the OCUFA working group for which this statement of issues has been produced: Bias isn't like frosting that sits on the surface of an evaluation or belief; it is baked into the cake.

Psychologists have studied not only people's vulnerability to bias, but also their awareness of this vulnerability and their capacity either to avoid it or to compensate for it. Reviews of this literature include Wilson, Centerbar, & Brekke (2002). Studies show that people are generally terrible at perceiving, avoiding, and remedying bias.

Wilson et al. (2002) conclude: "People often get it wrong, either failing to detect bias or failing to correct for it.… People appreciate that other people are not very good at avoiding biased influences on their beliefs, but have a misplaced faith in their own ability to control their beliefs and avoid unwanted influences" (p. 194). Consider what happened to female membership in orchestras when those auditioning played behind screen so that they could not be seen by judges: the numbers of women hired increased strikingly (Golding & Rouse, 2000). It is likely that many judges did not wish to be biased and were unaware that they were. Nevertheless, bias clearly had been operating.

Similarly, we can expect that today's students very predominantly do not want to show bias, yet they do. My speculation is that when asked whether a course was of high quality (in whatever terms are operationalized on the questionnaire, and taking "halo" bias into account), they do *not* really know whether it was of high quality. The attributes of high-quality instruction aren't even obvious to the typical instructor; at our university we have had two panels engage a researcher or instructional developer to try to identify the elements of instructional quality, as input to the design of assessments. So how would students know? But students do know whether their instructor was male or female, and they have been acculturated to hold gender-specific stereotypes or schema about who is more competent. Students also have a schema for "good teaching" which probably includes energetic, outgoing, entertaining

"stand and deliver" presentation. Given students' lack of perspective and vacuum of knowledge about instructional quality, these schemas receive some degree of application even when, or perhaps especially when, students are trying to do a good job on the ratings. This speculation is consistent with many studies involving gender and perception, including a study done at my university (Scott & Brown, 2006).

3.2 Numerical adjustments to compensate for bias are impossible.

Ratings differences between male and female instructors vary considerably in magnitude from study to study. The same is true for estimates of the other biasing factors (required/elective, time of day, etc.). Probably this variation is not all just "normal" variation that we should expect from study to study based on, for example, sampling error. Instead, most likely there are moderating factors in the various contexts of assessment with which biasing factors, such as instructor gender and time of day, interact.

Many studies have directly demonstrated interaction of this kind. For example, at my university it was shown that female instructors were evaluated less favourably than male instructors by students with low marks, but this effect for instructor gender disappeared among students with high marks (Sinclair, & Kunda, 2000). Thus, valid adjustment would require taking account of this contingency, which is impossible in practical terms. The underlying contingency here probably involves gender role violation by female instructors who give lower grades (e.g., violation of the female gender schema for being nurturing). Other gender role violations probably impact students' ratings as well. But there is no way to track those violations and apply them *contingently* in a truly fair scheme for numerical adjustment.

As further illustration: Are morning classes okay for students in later years but not for those in early years in program? Are classes taken mostly by out-of-major students rated low if required by a different department, but not if their topic is of general interest? And so forth. Given this variability of effects, statistical adjustments will merely add additional scrambling to the already scrambled positions of instructors on whatever dimension is being measured in the first place.

This crucial problem of interactions was further documented in research by Boring, Ottoboni, and Stark (2016), which (because of its strong design) provides one of the strongest demonstrations to date that gender bias is a serious, insoluble problem for summative use of students' ratings. Referring to student questionnaires by the acronym "SET" (student evaluations of teaching), they state:

> SET are biased against female instructors by an amount that is large and statistically significant. The bias affects how students rate even putatively objective aspects of teaching, such as how promptly assignments are graded. The bias varies by discipline and by student gender, among other things. It is not possible to adjust for the bias, because it depends on so many factors. SET are more sensitive to students' gender bias and grade expectations than they are to teaching effectiveness. Gender biases can be large enough to cause more effective instructors to get lower

> SET than less effective instructors. These findings are based on nonparametric statistical tests applied to two datasets: 23,001 SET of 379 instructors by 4,423 students in six mandatory first-year courses in a five year natural experiment at a French university, and 43 SET for four sections of an online course in a randomized, controlled, blind experiment at a US university. (p. 1)

3.3 Subjectively derived adjustments will not solve the problems with numerical adjustment.

Any counter-suggestion that a peer review committee could make useful numerical adjustments on a more subjective basis should be dispelled.

First, the committee will not have very much (if any) of the kind of information that the preceding point shows to be required (e.g., about extent of gender role normativity) to make corrective adjustments. Further, the literature on numerical models for assessment (e.g., Dawes, 1979; Meehl, 1954) suggests that if we cannot arrive at an arithmetic numerical adjustment, substituting a subjective judgment will not help.

The basic notion that peer review committee members can adjust appropriately for biasing factors is contrary to everyday experience that once a number is put on an attribute (i.e., a number for teaching effectiveness, purportedly, even if only to some degree), the number sticks. Bias in students' ratings gets transferred to committee members by this route. On this basis many members of the Department of Psychology recommended to my university that peer review committees should not routinely see any of the student questionnaire ratings. As Wilson et al. (2002) stated in their research on the psychology of bias, "The most effective defense [against bias] is preventing exposure to contaminating information, and the least effective defense is trying to undo or ignore contamination once it has occurred" (p. 194). In the present context, student questionnaire ratings are "contaminating information" with respect to valid performance evaluation.

To make matters worse, many professors are quite poor in the first place at using numerical arrays of the kind provided by student ratings. The title of Boysen's (2015) article on this topic is: "Significant interpretation of small mean differences in student evaluations of teaching despite explicit warning to avoid over-interpretation." There are other articles in this vein.

3.4 Careful design of the items on the survey questionnaire will not solve the problems of bias and halo.

A prevalent misconception about survey questionnaires for course evaluation (e.g., Usher, 2018; cf. Michela, 2018) is that problems of halo bias and other biases can be solved or prevented at the survey design stage with proper development and selection of survey items.[5]

---

[5] The arbitrator's report, to which I referred earlier in this document, reflected this misconception in recommending more careful design of survey items. The arbitrator appeared to be drawing on the experts consulted for the arbitration. However,

Careful design can indeed improve usability of the survey data for formative evaluation to the extent that survey items' meanings are made less ambiguous, more focused, and so forth. But halo bias is so potent that items pertaining to objective facts (e.g., To what degree did the instructor reliably begin class on time?) tend to be answered more or less favourably in line with the rest of the indications of favourability on the survey questionnaire (e.g., How clearly did the instructor present the course material in lectures?). Other bias also operates imperceptibly (as explained in point 3.1) and can combine with halo to pervade survey responses.

The "halo" effect in this context is the tendency for survey respondents to bring all survey item ratings into some degree of conformance with an overall reaction to the course, positively or negatively. Thus, for example, there have been empirical demonstrations that even when an instructor goes to great lengths to return all marked assignments completely on time, even to the point of documenting students' ongoing (real time) acknowledgments of timely return of assignments, ratings of this matter on a reasonably-well-worded item do not square with this reality (see sources in Nilson, 2012, p. 218).

Finally, with regard to survey item design, it should be acknowledged that some students will deliberately give ratings that are inaccurate for reasons such as retaliation for a poor grade. Nilson (2012, p. 212) addresses instances in which ratings do not square with evident facts (such as whether assignments were returned on time):

> Were these misrepresentations of the truth due to students' forgetting, misunderstanding, or lying? Clayson and Haley (2011) surveyed students about their honesty in their ratings and written comments, and the disturbing results confirmed Stanfel's and Spoule's worst suspicions: about one-third of the students confessed to "stretching the truth," 56 percent said they knew peers who had, and 20 percent admitted to lying in their comments. Moreover, half the students did not think that what they did constituted a kind of cheating.

These are not the only empirical reports of this kind.

---

the experts may have had in mind improvement for formative, not summative evaluation, given the thrust of all their other input to the arbitration. The arbitrator also recommended changing the level of survey measurement from ordinal to interval, and to present survey results in terms of a score distribution instead of summary statistics of the distribution (such as the mean and standard deviation). These changes may indeed promote usefulness of the data for formative purposes, and this may be what the arbitrator's expert sources had in mind when these changes were mentioned in the experts' reports. However, there is no evidence supporting improvement for summative use from these procedures. The ways in which students generate their responses and the ways in which performance evaluators use those responses argue against the likelihood of ever seeing such evidence.

4. Student questionnaires nevertheless may be useful for formative evaluation and other purposes.

4.1 Issues of bias and validity are transformed when solely formative use occurs.

Defenders of summative use commonly assert that students can tell us about their perceptions and experiences of courses and instruction. However, students cannot tell us very accurately whether the course content was appropriate for the subject matter (because they are taking the course to learn it in the first place), whether the instructional approaches were well-suited (because they don't know very much about instructional approaches, and they prefer the familiar), nor even whether they learned a lot (relative to what they could have learned, given the most appropriate course design and course delivery). As Wieman (2015) puts it, "People are poor at evaluating their own learning because it is difficult to know what you do not know."

Students' questionnaire responses about their perceptions and experiences nevertheless can provide useful data for the instructor and for instructional support staff or peers, as they look for ways to improve the effectiveness of instruction. Within an orientation of "formative" evaluation, unfavourable answers could prompt a closer look at matters such as course organization and framing, delivery elements, and learning outcomes. However, even within a formative evaluation orientation, student ratings data should not be taken at face value. For example, Ellis (2013) questions the value of these data for innovative courses.

The issues with bias are vastly transformed in the switch from a summative to a formative evaluation orientation. An instructor's disadvantage by gender, attractiveness, time of day, topic (e.g., statistics for psychology) often can be "held constant" from one term to the next. Thus, a "within-instructor/within-course" comparison of student ratings across terms can be meaningful. Again, this is not to say that instructors should water down their courses if that's what it takes to get higher ratings. It is to say that higher ratings within-instructor/within-course could signal improved actual instructional performance, depending on the changes to instruction that were accomplished.

As an aside, instances of effective use of students' ratings in formative evaluation may be one reason why many in the university do not see the fundamental inappropriateness for summative evaluation that is argued here. I agree that for a given instructor, appropriate revisions to course design and delivery, triggered by student ratings data, may yield improved ratings. The problem for summative use of the ratings is that comparisons are no longer within-instructor/within-course, when summative. The many extraneous factors, across-instructor and across-course, scramble instructors' positions relative to one another, and there is no way to fix this.

Many issues of student questionnaire design, administration, and use are connected with whether the purpose is summative or formative evaluation. If the recommendation to stop allowing peer review

committees to see any course ratings routinely is adopted, then what is left is formative evaluation use. The various issues of design and so forth can be addressed to optimize formative use. For example, in survey administration, the issue of whether to provide incentives for students to respond (such as offering a prize in a lottery) takes a distinct cast. Under summative evaluation, some university administrators may fear that too-low response rates make the data invalid, and that incentives are an answer. But from the present point of view, obtaining survey responses in this way from unmotivated respondents is not a path to improved validity either for summative or formative purposes and easily could make matters even worse.

With a switch in orientation to formative evaluation, the concept of validity itself is transformed. For example, formative "validity" can be conceived not only as the accuracy of students' ratings of their perceptions and experiences (which depends partly on item clarity) but also on interpretability of those ratings. Interpretability might be enhanced by obtaining information from student questionnaires or other sources on contextual factors such as course enrollees' characteristics (e.g., in/out of major) or various course characteristics.

4.2 Simultaneous summative and formative use undermines the formative use that would be of greater benefit to students' learning through instructional improvement.

Simultaneous summative and formative use creates a bind for instructors in terms of the non-compulsory survey items that they choose when discretion is allowed. With exclusively formative use, the proposed bank of survey questions could be developed with emphasis on areas for improvement, and instructors would have an incentive to select items that might warrant improvement. With summative evaluation, the incentive is to choose items that will document a lack of any shortcomings.

5. The widespread use of student questionnaires at many universities for summative evaluation gives no assurance of their appropriateness for that purpose. So-called "best practices" are ineffective.

Many sources document widespread adoption internationally of student questionnaires for summative evaluation of university teaching performance. As noted earlier, by the 1980s, evidence was said to have accumulated to justify summative use, but that evidence has been refuted. In particular, Uttl et al. (2017) concluded that "Re-analyses of previous meta-analyses… indicate that SET ratings explain at most 1 per cent of variability in measures of student learning" and also that the newer research (as in point 2.1) shows that students' ratings nowadays "are unrelated to student learning." Nilson's (2012) critique as sketched earlier also is fundamental here.

Psychologists are familiar with another instance of the use of a very consequential yet invalid measurement method, namely the use of polygraphs as lie detectors. This use has been discontinued in court proceedings on the whole, but it was widespread for some time and very possibly harmful to some people in the 20th century. An American Psychological Association web page (APA, 2004) describes denunciations of polygraphs as lie detectors, as offered by a US National Research Council "blue-ribbon" panel and others. Of course this has not put the American Polygraph Association out of business. Similarly, the remaining defenders of summative validity prominently include people who sell SETs or otherwise have a vested interest.

The point is: Yes, just as many legal jurisdictions were mistaken in the weight they gave to polygraph results, all those universities making summative use of student questionnaires could be mistaken—and I believe, based on the evidence, that they are. Relatedly, I certainly am not reassured about the prospects for reducing bias through orientation and training merely because some other universities have the good intention to produce this effect this way.

6. The alternatives to student questionnaires that have been proposed in the literature can be expected to carry *less* bias and to do *more* to promote effective instruction.

6.1 Student questionnaires are particularly vulnerable to bias because they are not "grounded" in any way.

Typical questions in the student questionnaires include:

- The instructor was a clear communicator
- The instructor created a supportive environment that helped me learn
- Overall, the quality of my learning experience in this course was excellent

As items of this kind go, these are mostly fine provided that students' perceptions in those terms are what you want to assess. However, at the same time that evaluators seek to assess particular perceptions, students want to express their experience in overall terms, and they will use whatever items they receive to reflect this. This is an aspect of halo bias and explains how it can turn out that an instructor who returns assignments on time, and *gets students to acknowledge receipt on time for research purposes*, can still get lukewarm responses concerning on-time marking from end-of-term student questionnaires (e.g., see Wieman, 2015). In key respects, students are not really, or merely, answering the questions posed. They are using the available survey items to express broader attitudes (including the primitive "liking" described in Point 1)—expressions which are particularly vulnerable to all kinds of bias.

Students are no different from other survey respondents in this respect of using whatever survey items they are given to express broader attitudes. During Barack Obama's presidency, Republicans' responses to a seemingly factual matter, the unemployment rate, were less favourable than those of Democrats. (The reason to believe that Republicans were not merely less well informed is based on experimental manipulations that changed their responses.) During Donald Trump's pursuit of the U.S. presidency, while Trump was praising Putin, Republicans' attitudes toward Putin became more favourable over time. Given that Putin is little more than a cartoon character to nearly everyone in the U.S., what these respondents were reflecting in their answers seems more to have concerned their attitudes toward Donald Trump, not Putin. The point is that survey responses can be complex and not at all what they appear to be about on the surface.

6.2 Proper peer evaluation is grounded in direct observation and explicit criteria and is therefore less vulnerable to bias.

One salient alternative to student questionnaires for obtaining summative evaluation data is to use faculty peer evaluation. Although any measurement involving human judgment can be subject to bias,

properly conducted peer evaluation is much less vulnerable. Qualified peers can judge matters such as appropriateness of content, of assignments, of presentations, and so forth. In this context, evaluator orientation and training can be expected to have real benefit and to be applied reasonably well by the evaluator in most instances. The evaluator will have accountability for his or her evaluations, and an appeals mechanism could be established. Crucially, the evaluator, most likely, will not be put off by high workload or other challenges posed to students unless there is evidence of actual detriment to students. A properly prepared evaluator will not care about the course's time of day, dinginess of the classroom, extent of in- or out-of-major students, and many other factors that bias *students'* ratings.

6.3 Peer evaluation need not be as onerous as many of its critics contend, and other sources of information or factors can be considered in summative evaluation.

This document is not the place to try to elaborate alternatives to the use of student questionnaires for summative evaluation. Developing such alternatives require significant time and effort. However, because the current procedure has been shown to be glaringly invalid and therefore unjust, and because it undermines student learning, there is no justifiable choice in the matter of whether academics should exert that time and effort.

As to whether peer evaluation is impractical and even onerous, Stark (2015) describes how his department's experience with peer evaluation has been the opposite. He states that in periodic instances in which instructors receive peer evaluation, classroom observation took the reviewer about four hours, including the observation time itself. The process included conversations between the candidate and the observer, and included an opportunity for the candidate to respond to the written comments, along with a provision for a "no-fault do-over." Candidates and the reviewer generally reported that the process was valuable and interesting. Stark states that if this were done "for every milestone review," it would require approximately "16 hours over a 40-year career: *de minimis*."

Wieman (2015) prefers use of a teaching practices inventory, which ties to use of innovative teaching practices that have been shown to be more effective than traditional approaches (Wieman, 2016). Various writings by Stark, Frieshstat, and others provide additional alternatives to consider.

7. When the facts change, beliefs and behaviour should change, too.

The evidence that has accumulated in recent years against the validity and utility of summative use of student questionnaires is overwhelming, and there is no evidence that the biases that foul summative use can be sufficiently remedied. Denial of the harms to instructional practice (e.g., by incentivizing traditional and non-challenging teaching) and to fairness (due to gender bias and many other biasing factors) resembles denial of climate change, in the steadfast refusal by many summative use proponents to rebut or otherwise genuinely address existing facts.

Nevertheless, notable movement in the right direction has occurred in some quarters. Early in this document a favourable arbitration decision within the past year was described. Also within the past year, the provost at the University of Southern California declared summative use to be inadmissible (Flaherty, 2018):

> [The Provost] just said, "I'm done. I can't continue to allow a substantial portion of the faculty to be subject to this kind of bias," said Ginger Clark, assistant vice provost for academic and faculty affairs and director of USC's Center for Excellence in Teaching. "We'd already been in the process of developing a peer-review model of evaluation, but we hadn't expected to pull the Band-Aid off this fast."

The former head of Rice University's instructional support service, E. Barre, who had been widely cited by summative use proponents until the past year, described her 180 degree turn on the appropriateness of summative use in a web posting entitled "Research on student ratings continues to evolve. We should, too:"

> The most important recommendation I would now make is the following: we should put a moratorium on using student ratings results to rank and compare individual faculty to one another.

Barre (2018) went on to recommend a focus on formative use, with a within-person and within-course frame, much as has recommended in the present document.

Obviously, in this point of my case against summative use, I have returned to the quotation at the beginning, "When the facts change, I change my mind." A closing quotation suggests a further dimension to remaining resistance to change:

> *"It is difficult to get a man to understand something, when his salary depends upon his not understanding it." – Upton Sinclair*[6]

---

[6] https://quoteinvestigator.com/2017/11/30/salary/

References

Alderfer, C. P. (1972). *Existence, relatedness, and growth: Human needs in organizational settings*. New York: Free Press.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., Mohanty, G., & Spooner, F. (2004). Student evaluation of college teaching: A practice in search of principles. *College teaching, 52*(4), 134-141.

American Psychological Association (APA). (2004). The polygraph in doubt. http://www.apa.org/monitor/julaug04/polygraph.aspx.

Barre, E. (2018). Research on student ratings continues to evolve. We should, too. Rice University Center for Teaching Excellence. http://cte.rice.edu/blogarchive/2018/2/20/studentratingsupdate

Benton, S. L., & Cashin, W. E. (2012). Student ratings of teaching: A summary of research and literature. IDEA Paper 50. Manhattan, KS: The IDEA Center.

Boysen, G. A. (2015a). Significant interpretation of small mean differences in student evaluations of teaching despite explicit warning to avoid over interpretation. *Scholarship of Teaching and Learning in Psychology, 1*, 150–162.

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. (DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1)

Canadian Association of University Teachers (CAUT). (2018, November). The end of student questionnaires? *CAUT Bulletin*. https://www.caut.ca/bulletin/2018/11/end-student-questionnaires

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, 51, 281-309.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571-582.

Demming, W. E. (1994). *The new economics for industry, government, education*. Cambridge, MA: Massachusetts Institute of Technology Centre for Advanced Engineering Study.

Flaherty, C. (2016). Bias against female instructors. *Inside Higher Ed*, January 11.

Flaherty, C. (2018). Teaching eval shake up. *Inside Higher Ed*, May 22.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of 'blind' auditions on female musicians. *American Economic Review, 90*(4), 715-741.

Hackman, J. R., & Oldham, G. R. *Work redesign*. Reading, Mass.: Addison-Wesley, 1980.

Hativa, N. (2013). *Student ratings of instruction: Recognizing effective teaching*. Oron Publications.

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4*(1), 1304016.

Kohn, A. (2018, October 27). Science confirms it: People are not pets. *New York Times*, p. SR10.

Likert, R. (1967). *The human organization: Its management and values*. New York: McGraw-Hill.

McGregor, D. (1960). *The human side of enterprise*, New York, McGraw-Hill.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Michela, J. L. (2018, September). Feedback on feedback questionnaires' use and misuse. University of Waterloo Faculty Association Blog. https://fauw.blog/2018/09/13/feedback-on-feedback/

Nuhfer, E. B. (2010). A fractal thinker looks at student ratings. Retrieved from http://profcamp.tripod.com/fractalevals10.pdf

Nilson, L. B. (2012). Time to raise questions about student ratings. In J. E. Groccia and L. Cruz (Eds.), *To improve the academy: Resources for faculty, instructional, and organizational development* (Vol. 31).

Ramirez, A. (2001). How merit pay undermines education. *Educational Leadership, 58*(5), 16-20.

Scott, K. A., & Brown, D. J. (2006). Female first, leader second? Gender bias in the encoding of leadership behavior. *Organizational behavior and human decision processes, 101*(2), 230-242.

Seth, N., Deshmukh, S. G., & Vrat, P. (2005). Service quality models: a review. *International Journal of Quality & Reliability Management, 22*(9), 913-949.

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin, 26*(11), 1329-1342.

Stark, P. B. (2015). Teaching evaluations: Truthful or truthy? Presented at the *Third Lisbon Research Workshop on Economics, Statistics, and Econometrics of Education*. Lisbon, Portugal (23-24 January). http://www.stat.berkeley.edu/~stark/Seminars/setLisbon15.htm

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

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, 11*(6) 800-816.

Usher, A. (2018, September). Time to talk teaching assessments. Toronto, Ontario: Higher Education Strategy Associates. http://higheredstrategy.com/time-to-talk-teaching-assessments/

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22-42. doi: http://www.sciencedirect.com/science/article/pii/S0191491X16300323

Wieman, C. (2015) A better way to evaluate undergraduate teaching. Change: *The Magazine of Higher Learning, 47*(1), 6-15.

Wieman, C. (2016). Taking a scientific approach to science education.

https://circle.wustl.edu/wp-content/uploads/2016/08/Carl-Wiemans-Presentation-Slides.pdf

Wilson, T. D., Centerbar, D. B., & Brekke, N. (2002). Mental contamination and the debiasing problem. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 185-200). Cambridge University Press.

Biographical sketch of J. L. Michela, Ph.D.

John (Jay) Michela is an associate professor in the Industrial-Organizational area of the Department of Psychology. This subfield of Psychology encompasses many aspects of the issues at hand, including measurement of work performance, work motivation as tied to contingent pay versus intrinsic motivators, and customer satisfaction processes and surveys. Some of his earlier academic work, including at the graduate school of education and applied psychology at Columbia University, concerned social perception of people's traits and behaviours. In applied work at Columbia he teamed with others to analyze employee survey data from worldwide AT&T operations. He also designed and implemented a survey-based performance-alignment tool as consultant to the management consulting group of the worldwide Hay organization. At the University of Waterloo, he founded the Waterloo Organizational Research and Consulting Group (WORC Group), where he has led survey design and analysis projects including for Babcock & Wilcox, U.S.-based Ascension Health hospitals, Saville Software Systems, and Rogers Communications. Dr. Michela has taught psychological measurement, research methods, and multivariate statistics periodically at the graduate level since 1980. His publications include works that provide early or first-time demonstrations of statistical methods including multidimensional scaling, structural equation modeling, and hierarchical linear or non-linear models. He has written and reviewed for *Organizational Research Methods*, published in the *Journal of Applied Psychology* and elsewhere, served on two American Psychological Association journals' editorial boards, and served as a journal associate editor. Signalling his commitment to educational innovation, his use of blended learning in instruction has been featured along with others' on a website of the Centre for Teaching Excellence at his university. His original, on-line modules for instruction of students in teamwork, which he developed with the WORC Group, have been used in his and others' courses in his Department of Psychology and in the School of Accounting and Finance at his university.

Expertise of the most cited sources in the statement

Richard Freishstat served as Director of University of California at Berkeley's Center for Teaching and Learning (CTL). In this capacity, he created, led, and facilitated a variety of faculty development programs, including the Teaching Excellence Colloquium for new faculty, and the Presidential Chair Fellows Curriculum Enrichment Grant program. He has been an invited speaker and leader of international programs on faculty development, teaching and learning, and the evaluation of teaching—having delivered talks or programs at the Kuwait Foundation for the Advancement of Society, the UC Berkeley Center for Studies in Higher Education, and the University of Toronto, among others.

Philip B. Stark holds the titles of Professor of Statistics, Associate Dean of Mathematical and Physical Sciences, and Director of the Statistical Computing Facility at the University of California, Berkeley, where he is also a faculty member in the Graduate Program in Computational Data Science and Engineering; a co-investigator at the Berkeley Institute for Data Science; principal investigator of the Consortium for Data Analytics in Risk; director of Berkeley Open Source Food; and affiliated faculty of the Simons Institute for the Theory of Computing, the Theoretical Astrophysics Center, and the

Berkeley Food Institute. Previously, he was Chair of the Department of Statistics. He also has had campus-wide responsibilities regarding educational technology, including technology involved in teaching evaluations. He published more than one hundred and fifty articles and books. He has served on the editorial boards of archival journals in physical science, Applied Mathematics, Computer Science, and Statistics. He currently serves on four editorial boards. He has lectured at universities, professional societies, and government agencies in twenty-five countries. He was a Presidential Young Investigator, a Miller Research Professor, and a Velux/Villum Foundation Visiting Professor of Theoretical Computer Science. He received the U.C. Berkeley Chancellor's Award for Research in the Public Interest and the Leamer-Rosenthal Prize for Open Social Science. He is a member of the Institute for Mathematical Statistics and the Bernoulli Society; and he is a Fellow of the American Statistical Association, the Institute of Physics, and the Royal Astronomical Society. He is professionally accredited as a statistician by the American Statistical Association and as a physicist by the Institute of Physics.

Wolfgang Stroebe is Emeritus Professor of social psychology at Utrecht University and now at the University of Groningen. He is a past president of the European Association of Experimental Social Psychology and founding director of the Dutch Research Institute for Psychology and Health. He received the research award for outstanding scientific achievements concerning death and loss of the American Association of Death Counseling and Education in 2002, the Tajfel Award for outstanding scientific achievements and contribution to the development of social psychology of the European Association of Experimental Social Psychology in 2005, the lifetime achievement award of the German Psychological Association and an honorary doctorate from the University of Louvain (Belgium) in 2002. He is a member of the German National Academy of Science, Fellow of APS (resigned), SPSP, BPS and SPSSI. He has authored numerous books, chapters and articles, and for 25 years was co-editor of the European Review of Social Psychology.

Carl Wieman a professor of physics and a professor in the Graduate School of Education at Stanford University. He is the founder of the Carl Wieman Science Education Initiative (CWSEI) at the University of British Columbia and the Science Education Initiative at the University of Colorado. He received the Nobel Prize in Physics (2001) and the Carnegie Foundation's U.S. University Professor of the Year Award (2004). He served as the Associate Director for Science in the White House Office of Science and Technology Policy.

Appendix C: Excerpt – Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans[1]

...

Activities Not Requiring REB Review

The following distinguishes research requiring REB review from non-research activities that have traditionally employed methods and techniques similar to those employed in research. Such activities are not considered "research" as defined in this Policy, and do not require REB review. Activities outside the scope of research subject to REB review (see Articles 2.5 and 2.6), as defined in this Policy, may still raise ethical issues that would benefit from careful consideration by an individual or a body capable of providing some independent guidance, other than an REB. These ethics resources may be based in professional or disciplinary associations, particularly where those associations have established best practices guidelines for such activities in their discipline.

Article 2.5 Quality assurance and quality improvement studies, program evaluation activities, and performance reviews, or testing within normal educational requirements when used exclusively for assessment, management or improvement purposes, do not constitute research for the purposes of this Policy, and do not fall within the scope of REB review.

Application Article 2.5 refers to assessments of the performance of an organization or its employees or students, within the mandate of the organization, or according to the terms and conditions of employment or training. Those activities are normally administered in the ordinary course of the operation of an organization where participation is required, for example, as a condition of employment in the case of staff performance reviews, or an evaluation in the course of academic or professional training. Other examples include student course evaluations, or data collection for internal or external organizational reports. Such activities do not normally follow the consent procedures outlined in this Policy.

If data are collected for the purposes of such activities but later proposed for research purposes, it would be considered secondary use of information not originally intended for research, and at that time may require REB review in accordance with this Policy. Refer to Section D of Chapter 5 for guidance concerning secondary use of identifiable information for research purposes.

...

[1] http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/Default; accessed June 26, 2017.

Appendix D: Excerpt – Ontario Human Rights Code[2]

Human Rights Code

R.S.O. 1990, CHAPTER H.19

Consolidation Period: From December 5, 2016 to the e-Laws currency date.

...

Preamble

Whereas recognition of the inherent dignity and the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world and is in accord with the Universal Declaration of Human Rights as proclaimed by the United Nations;

And Whereas it is public policy in Ontario to recognize the dignity and worth of every person and to provide for equal rights and opportunities without discrimination that is contrary to law, and having as its aim the creation of a climate of understanding and mutual respect for the dignity and worth of each person so that each person feels a part of the community and able to contribute fully to the development and well-being of the community and the Province;

And Whereas these principles have been confirmed in Ontario by a number of enactments of the Legislature and it is desirable to revise and extend the protection of human rights in Ontario;

Therefore, Her Majesty, by and with the advice and consent of the Legislative Assembly of the Province of Ontario, enacts as follows:

...

PART I
FREEDOM FROM DISCRIMINATION

...

Employment

5. (1) Every person has a right to equal treatment with respect to employment without discrimination because of race, ancestry, place of origin, colour, ethnic origin, citizenship, creed, sex, sexual orientation, gender identity, gender expression, age, record of offences, marital status, family status or disability. R.S.O. 1990, c. H.19, s. 5 (1); 1999, c. 6, s. 28 (5); 2001, c. 32, s. 27 (1); 2005, c. 5, s. 32 (5); 2012, c. 7, s. 4 (1).

Harassment in employment

(2) Every person who is an employee has a right to freedom from harassment in the workplace by the employer or agent of the employer or by another employee because of race, ancestry, place of origin, co-

[2] https://www.ontario.ca/laws/statute/90h19; accessed June 14, 2017.

lour, ethnic origin, citizenship, creed, sexual orientation, gender identity, gender expression, age, record of offences, marital status, family status or disability. R.S.O. 1990, c. H.19, s. 5 (2); 1999, c. 6, s. 28 (6); 2001, c. 32, s. 27 (1); 2005, c. 5, s. 32 (6); 2012, c. 7, s. 4 (2).

Section Amendments with date in force (d/m/y)

…

Sexual harassment

…

Harassment because of sex in workplaces

(2) Every person who is an employee has a right to freedom from harassment in the workplace because of sex, sexual orientation, gender identity or gender expression by his or her employer or agent of the employer or by another employee. R.S.O. 1990, c. H.19, s. 7 (2); 2012, c. 7, s. 6 (2).

Sexual solicitation by a person in position to confer benefit, etc.

(3) Every person has a right to be free from,

(a) a sexual solicitation or advance made by a person in a position to confer, grant or deny a benefit or advancement to the person where the person making the solicitation or advance knows or ought reasonably to know that it is unwelcome; or

(b) a reprisal or a threat of reprisal for the rejection of a sexual solicitation or advance where the reprisal is made or threatened by a person in a position to confer, grant or deny a benefit or advancement to the person. R.S.O. 1990, c. H.19, s. 7 (3).

…

Reprisals

8. Every person has a right to claim and enforce his or her rights under this Act, to institute and participate in proceedings under this Act and to refuse to infringe a right of another person under this Act, without reprisal or threat of reprisal for so doing. R.S.O. 1990, c. H.19, s. 8.

Infringement prohibited

9. No person shall infringe or do, directly or indirectly, anything that infringes a right under this Part. R.S.O. 1990, c. H.19, s. 9.

PART II
INTERPRETATION AND APPLICATION

Definitions re: Parts I and II

10. (1) In Part I and in this Part,

...

"equal" means subject to all requirements, qualifications and considerations that are not a prohibited ground of discrimination; ("égal")

...

"harassment" means engaging in a course of vexatious comment or conduct that is known or ought reasonably to be known to be unwelcome; ("harcèlement")

"marital status" means the status of being married, single, widowed, divorced or separated and includes the status of living with a person in a conjugal relationship outside marriage; ("état matrimonial")

...

Constructive discrimination

11. (1) A right of a person under Part I is infringed where a requirement, qualification or factor exists that is not discrimination on a prohibited ground but that results in the exclusion, restriction or preference of a group of persons who are identified by a prohibited ground of discrimination and of whom the person is a member, except where,

(a) the requirement, qualification or factor is reasonable and *bona fide* in the circumstances; or

(b) it is declared in this Act, other than in section 17, that to discriminate because of such ground is not an infringement of a right. R.S.O. 1990, c. H.19, s. 11 (1).

Idem

(2) The Tribunal or a court shall not find that a requirement, qualification or factor is reasonable and *bona fide* in the circumstances unless it is satisfied that the needs of the group of which the person is a member cannot be accommodated without undue hardship on the person responsible for accommodating those needs, considering the cost, outside sources of funding, if any, and health and safety requirements, if any. R.S.O. 1990, c. H.19, s. 11 (2); 1994, c. 27, s. 65 (1); 2002, c. 18, Sched. C, s. 2 (1); 2009, c. 33, Sched. 2, s. 35 (1).

Idem

(3) The Tribunal or a court shall consider any standards prescribed by the regulations for assessing what is undue hardship. R.S.O. 1990, c. H.19, s. 11 (3); 1994, c. 27, s. 65 (2); 2002, c. 18, Sched. C, s. 2 (2); 2009, c. 33, Sched. 2, s. 35 (2).

...

Appendix E: Excerpt – Ontario Occupational Health and Safety Act[3]

Occupational Health and Safety Act

R.S.O. 1990, CHAPTER O.1

Consolidation Period: From December 8, 2016 to the e-Laws currency date.

...

Definitions

...

"workplace harassment" means,

(a) engaging in a course of vexatious comment or conduct against a worker in a workplace that is known or ought reasonably to be known to be unwelcome, or

(b) workplace sexual harassment; ("harcèlement au travail")

"workplace sexual harassment" means,

(a) engaging in a course of vexatious comment or conduct against a worker in a workplace because of sex, sexual orientation, gender identity or gender expression, where the course of comment or conduct is known or ought reasonably to be known to be unwelcome, or

(b) making a sexual solicitation or advance where the person making the solicitation or advance is in a position to confer, grant or deny a benefit or advancement to the worker and the person knows or ought reasonably to know that the solicitation or advance is unwelcome; ("harcèlement sexuel au travail")

"workplace violence" means,

(a) the exercise of physical force by a person against a worker, in a workplace, that causes or could cause physical injury to the worker,

(b) an attempt to exercise physical force against a worker, in a workplace, that could cause physical injury to the worker,

(c) a statement or behaviour that it is reasonable for a worker to interpret as a threat to exercise physical force against the worker, in a workplace, that could cause physical injury to the worker. ("violence au travail")

...

[3] https://www.ontario.ca/laws/statute/90o01; accessed June 14, 2017.

PART III.0.1
VIOLENCE AND HARASSMENT

Policies, violence and harassment

32.0.1 (1) An employer shall,

(a) prepare a policy with respect to workplace violence;

(b) prepare a policy with respect to workplace harassment; and

(c) review the policies as often as is necessary, but at least annually. 2009, c. 23, s. 3.

Written form, posting

(2) The policies shall be in written form and shall be posted at a conspicuous place in the workplace. 2009, c. 23, s. 3.

...

Program, violence

32.0.2 (1) An employer shall develop and maintain a program to implement the policy with respect to workplace violence required under clause 32.0.1 (1) (a). 2009, c. 23, s. 3.

Contents

(2) Without limiting the generality of subsection (1), the program shall,

(a) include measures and procedures to control the risks identified in the assessment required under subsection 32.0.3 (1) as likely to expose a worker to physical injury;

(b) include measures and procedures for summoning immediate assistance when workplace violence occurs or is likely to occur;

(c) include measures and procedures for workers to report incidents of workplace violence to the employer or supervisor;

(d) set out how the employer will investigate and deal with incidents or complaints of workplace violence; and

(e) include any prescribed elements. 2009, c. 23, s. 3.

...

Assessment of risks of violence

32.0.3 (1) An employer shall assess the risks of workplace violence that may arise from the nature of the workplace, the type of work or the conditions of work. 2009, c. 23, s. 3.

Considerations

(2) The assessment shall take into account,

(a) circumstances that would be common to similar workplaces;

(b) circumstances specific to the workplace; and

(c) any other prescribed elements. 2009, c. 23, s. 3.

Results

(3) An employer shall,

(a) advise the committee or a health and safety representative, if any, of the results of the assessment, and provide a copy if the assessment is in writing; and

(b) if there is no committee or health and safety representative, advise the workers of the results of the assessment and, if the assessment is in writing, provide copies on request or advise the workers how to obtain copies. 2009, c. 23, s. 3.

Reassessment

(4) An employer shall reassess the risks of workplace violence as often as is necessary to ensure that the related policy under clause 32.0.1 (1) (a) and the related program under subsection 32.0.2 (1) continue to protect workers from workplace violence. 2009, c. 23, s. 3.

Same

(5) Subsection (3) also applies with respect to the results of the reassessment. 2009, c. 23, s. 3.

...

Domestic violence

32.0.4 If an employer becomes aware, or ought reasonably to be aware, that domestic violence that would likely expose a worker to physical injury may occur in the workplace, the employer shall take every precaution reasonable in the circumstances for the protection of the worker. 2009, c. 23, s. 3.

...

Duties re violence

32.0.5 (1) For greater certainty, the employer duties set out in section 25, the supervisor duties set out in section 27, and the worker duties set out in section 28 apply, as appropriate, with respect to workplace violence. 2009, c. 23, s. 3.

Information

(2) An employer shall provide a worker with,

(a) information and instruction that is appropriate for the worker on the contents of the policy and program with respect to workplace violence; and

(b) any other prescribed information or instruction. 2009, c. 23, s. 3.

Provision of information

(3) An employer's duty to provide information to a worker under clause 25 (2) (a) and a supervisor's

duty to advise a worker under clause 27 (2) (a) include the duty to provide information, including personal information, related to a risk of workplace violence from a person with a history of violent behaviour if,

(a) the worker can be expected to encounter that person in the course of his or her work; and

(b) the risk of workplace violence is likely to expose the worker to physical injury. 2009, c. 23, s. 3.

Limit on disclosure

(4) No employer or supervisor shall disclose more personal information in the circumstances described in subsection (3) than is reasonably necessary to protect the worker from physical injury. 2009, c. 23, s. 3.

...

Program, harassment

32.0.6 (1) An employer shall, in consultation with the committee or a health and safety representative, if any, develop and maintain a written program to implement the policy with respect to workplace harassment required under clause 32.0.1 (1) (b). 2016, c. 2, Sched. 4, s. 2 (1).

Contents

(2) Without limiting the generality of subsection (1), the program shall,

(a) include measures and procedures for workers to report incidents of workplace harassment to the employer or supervisor;

(b) include measures and procedures for workers to report incidents of workplace harassment to a person other than the employer or supervisor, if the employer or supervisor is the alleged harasser;

(c) set out how incidents or complaints of workplace harassment will be investigated and dealt with;

(d) set out how information obtained about an incident or complaint of workplace harassment, including identifying information about any individuals involved, will not be disclosed unless the disclosure is necessary for the purposes of investigating or taking corrective action with respect to the incident or complaint, or is otherwise required by law;

(e) set out how a worker who has allegedly experienced workplace harassment and the alleged harasser, if he or she is a worker of the employer, will be informed of the results of the investigation and of any corrective action that has been taken or that will be taken as a result of the investigation; and

(f) include any prescribed elements. 2009, c. 23, s. 3; 2016, c. 2, Sched. 4, s. 2 (2).

...

Duties re harassment

32.0.7 (1) To protect a worker from workplace harassment, an employer shall ensure that,

(a) an investigation is conducted into incidents and complaints of workplace harassment that is appropriate in the circumstances;

(b) the worker who has allegedly experienced workplace harassment and the alleged harasser, if he or she is a worker of the employer, are informed in writing of the results of the investigation and of any corrective action that has been taken or that will be taken as a result of the investigation;

(c) the program developed under section 32.0.6 is reviewed as often as necessary, but at least annually, to ensure that it adequately implements the policy with respect to workplace harassment required under clause 32.0.1 (1) (b); and

(d) such other duties as may be prescribed are carried out. 2016, c. 2, Sched. 4, s. 3.

Results of investigation not a report

(2) The results of an investigation under clause (1) (a), and any report created in the course of or for the purposes of the investigation, are not a report respecting occupational health and safety for the purposes of subsection 25 (2). 2016, c. 2, Sched. 4, s. 3.

...

Information and instruction, harassment

32.0.8 An employer shall provide a worker with,

(a) information and instruction that is appropriate for the worker on the contents of the policy and program with respect to workplace harassment; and

(b) any other prescribed information. 2016, c. 2, Sched. 4, s. 3.

...