This statement of response to the CEPT report comes from various members of the Department of Psychology. We are responding collectively because our discipline offers the technical expertise that is essential for understanding how student questionnaires for course evaluation should be designed and used.

Our statement takes the form of a summary of points that will be developed and documented further over the next few months when one of us serves on an external panel concerning student questionnaires for course evaluation. This panel was convened by the Ontario Council of University Faculty Associations (OCUFA) in recognition of the divergence of the increasing weight being placed on student questionnaire ratings at Ontario Universities from the increasing body of evidence indicating that student questionnaires cannot bear this weight because of inherent limitations.

The present statement addresses the following points.

1. Extraneous, "biasing" factors render student questionnaires invalid for summative evaluation.
2. Summative use of student questionnaires harms students' learning and instructors' integrity and academic freedom.
3. The 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 widespread use of student questionnaires at other 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. Decisions about student questionnaire redesign and use should take full account of the best available internal (University of Waterloo) and external expert analysis and opinion. Thus far, this has not happened.

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

The essential problem is that ratings on student questionnaires have been shown to be responsive to (i.e., "biased" as a result of) a host of factors, among them instructor gender, ethnicity, age, physical attractiveness, speaking style, etc.; whether a course is in or out of major (or otherwise "important"), its class size, time of day, unusualness (instructional innovation), etc. 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, including at our university—Sinclair & Kunda, 2000, as sketched in Point 3). The magnitudes of influence of these biasing factors can be quite sizable (Stark, 2015).

Another major scholar and critic on this topic (besides Stark and Frieshtat), Wieman (2015), summarizes 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)

This problem is cataclysmic for the way we use student questionnaires in summative evaluation (i.e., for performance evaluation linked directly with pay and promotion decisions). It means that the rank ordering of instructors in terms of student questionnaire ratings, as seen by a peer review committee, can be utterly scrambled in relation to whatever differences might exist for instructors' true instructional effectiveness. Thus the numbers are worthless for answering: Who are the more effective or less effective instructors in our department? Of course, instructors who are disorganized, mumbling, or unclear may generally obtain lower ratings, compared with crisp, clear instructors. But research has shown that an instructor with very poor content but good style can obtain high ratings, and there are many factors that can produce low ratings when content is excellent.

As explained later, there is no way for the peer review committee to unscramble the rank order to arrive at a valid rank ordering. The net effect is that pay and promotion are affected by the biasing factors, which is deeply unjust. Especially in light of research that has appeared in recent years, Stark (2015) joins Wieman (2015) in emphasizing that we should not expect sufficient residual of true instructional effectiveness to be dominant in the rank orders. Instead there is every reason to believe that a dominant influence is "liking" for the course and/or instructor in some global sense (e.g., Nilson, 2012)—which clearly is not necessarily tied to instructional effectiveness.

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

*2a1. 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 (as 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 that new equivalent findings have not emerged. Adding to doubts about the original justification for summative use of student questionnaires, recent re-analysis of the pre-1981 studies argues that the conclusions of the original review were fundamentally mistaken (Uttl et al, 2016). 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).

2a2. Students' learning and instructors' academic freedom suffers further from summative use of student questionnaires because this use can deter use of innovative teaching approaches that are recommended by experts in post-secondary instruction (e.g., at our CTE).

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). Stark corroborates actual experiences of detrimental effects.

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 or difficulty in countering it.

2b. CEPT's proposal to make student questionnaire ratings available to all at UW is disturbing in light of what has been presented in Points 2a1 and 2a2 of this statement.

Pressures can only increase toward pleasing students relative to educating them. We urge the university not to implement this public display for this reason. Another reason not to implement public display is that it strongly signals validity of the ratings, the very criterion that is severely lacking (Point 1). Because the ratings are biased, public display also magnifies discrimination for women, people of colour, and others.

As psychologists, we recognize that many people believe that they are immune to the forces discussed in Points 2a1 and 2a2, and to any further influence from making the ratings public. However, a sizable body of research shows that people often have no awareness of the reasons for their actions, and when they recognize the possibility of undesired influence, they mistakenly believe they are totally or mostly immune (see Wilson et al., 2002, for an entry point to this literature). We all seek integrity. Nevertheless, the trends described by Stroebe in 2a1, and the rest of Strobe's analysis in his full paper, indicate that instructors are not immune.

2c. 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. This belief follows "Theory X": workers need to be leveraged with carrots and sticks, which McGregor (1960) criticized long ago. McGregor went on to describe "Theory Y" as an alternative belief set which 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 primarily are the education and other development of students.

Academics and leadership coaches have been recommending adoption of a Theory Y mindset ever since—especially for knowledge workers (of which professors and other instructors are perhaps the exemplar)—because these experts see this mindset as the path toward truly superior performance. Those who cling to Theory X as a rationale for continuing summative use should be required to show that instructional performance is inferior by instructors in the many universities that prohibit direct summative use of student questionnaires. We call that the control group. We strongly doubt that their

instructional performance is inferior, because those instructors share with us the primary, "intrinsic" motivators: to do a good job as instructors and to produce good outcomes for the students.

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

It is difficult to properly critique the CEPT report's suggestion that bias and other sources of inaccuracy can be minimized, because the methods to be used were not described in any detail. The CEPT report asserts: "a properly designed and implemented training and orientation program can enhance the utility of course evaluations" (p. 7). This statement is without foundation. 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.

The CEPT report refers to how "training and orientation content must address … the importance/role of bias (especially concerning gender and race) when completing and interpreting evaluations, and ethical obligations generally" (p. 7). Evidently part of the thinking behind this statement is that with effort and good intentions, people can significantly reduce or eliminate biased response.

However, 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 most if not all judges did not wish to be biased and were unaware that they were. Neverthless, bias clearly had been operating.

On this basis, we from the Department of Psychology recommend that peer review committees, for annual performance evaluation, *not* routinely see *any* of the student questionnaire ratings. As Wilson et al. state, "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.

3.2 Numerical adjustments to compensate for bias are impossible.

The CEPT report recognized, "in a study done at the University of Waterloo, when students received low grades, they gave statistically lower overall ratings of quality (course and instructor quality ratings were combined) to female instructors than male instructors" (p. 6).

The upshot of this study runs deeper than the demonstration of gender-based bias at our own university. It illustrates how bias operates in a manner that is too complex to allow a numerical fix to adjust for bias that exists in the ratings.

In this Waterloo study, the ratings difference between male and female instructors was not consistent across the board; it was contingent on the instructors' actions. Thus, valid adjustment would

require somehow taking account of this contingency. The underlying contingency here probably involves gender role violation by female instructors who give lower grades. Other gender role violations probably impact students' ratings as well. But there is no way to track those violations and apply them in a scheme for numerical adjustment. Many other such interactions among the various other biasing factors probably exist. For example, a detrimental effect of time of day might be contingent on whether students are more conventionally versus professionally oriented; innovative instructional approaches might be fine for some audiences though not for most; and so forth.

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.

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.

The CEPT report states: "it is possible to reduce the potential for bias, in its many potential forms, through the design of the course evaluation instrument" (p. 7). The report also can be read as indicating that halo effects can be remedied through attention to "clarity of question intent." These statements reflect misconceptions. Halo 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?). Bias operates imperceptibly (as explained in Point 3.1) and can combine with halo to pervade survey responses.

3.5 Continuous improvement efforts will not yield an instrument that is valid for summative evaluation.

The CEPT report states: "The testing of course evaluation instrument results will determine the reliability and validity of the instrument, including the influence of variables that could bias results at Waterloo" (p. 7) and "Quality Assurance Office staff should also monitor the performance of the course instrument and platform on a term-by-term basis, and report findings."

We have no idea what CEPT has in mind here.

High reliability is likely but could be due largely to halo effects. Reliability without validity has little value.

Validation of the instrument would require study of the association between scores on the instrument and scores on a criterion variable, based on some other measure or indicator of instructional effectiveness. It is difficult to imagine how to arrive at a criterion variable at our university and how it would be used in validation research. Under Point 2a1, we described research which used amount learned as a criterion, in studies that have tracked students across sequenced courses. If someone at our university were able to replicate that kind of research in appropriate departments here, there is no

reason to believe that the results here would support the validity of the instrument. Most likely we would see a replication either of no association—or even of a negative association—with amount learned.

3.6 The lack of effective remedies for bias should be pivotal for deciding whether to continue to use student questionnaires in summative evaluation.

The importance of the information that we have provided in this Point 3 cannot be overstated. CEPT recognized the well-documented operation of bias in students' ratings, as in its statement:

> "There is no question that biases (e.g. sexism, racism, ageism) exist on any campus, and that these biases can be expressed in course evaluations" (p. 6).

The report claimed that various remedies existed for this serious problem, but that claim was unsubstantiated. Psychological research and theory strongly indicate that the proposed remedies cannot undo biases to an extent that will unscramble student ratings data to expose any true instructional effectiveness that the questionnaires might tap.

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

4a Issues of bias and validity are transformed when solely formative use occurs.

The CEPT report provides a very helpful framework, either for summative or formative evaluation, in its analysis (e.g., Table 1) of three dimensions of instructional effectiveness: course design, course delivery, and learning experience. Where CEPT took a wrong turn was in promoting students' answers on end-of-term questionnaires as a valid way of measuring all of these dimensions for purposes of summative evaluation.

As the report says elsewhere, students can tell us about their perceptions and experiences. They 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." This ignorance of one's ignorance may be greatest for the most ignorant, as documented as the Dunning-Kruger effect (Dunning et al., 2003; Wikipedia).

Students' questionnaire responses about their perceptions and experiences nevertheless can provide useful data for the instructor and for instructional support staff (as at CTE) 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 we in the Department of Psychology see. We 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. 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 our recommendation from Psychology is adopted—to stop allowing peer review committees to see any course ratings routinely—what is left is formative evaluation use, and then 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 our 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. The CEPT report is correct in the various places where it calls for careful design of survey items, but toward what end? Summative validity cannot be built into the items, as explained earlier. 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.

4b Simultaneous summative and formative use undermines the formative use that would be of greater benefit to students' learning.

Simultaneous summative and formative use creates a bind for instructors in terms of the non-compulsory survey items that they choose. 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 instead is to choose items that will document a lack of any shortcomings. Wieman (2015) adds:

> 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)

4c It is not at all clear that the university must maintain summative evaluation.

Some members of our university community may hear claims that Waterloo is constrained to summative use of student questionnaire ratings either by University policy or external entities. Those claims should be scrutinized.

CEPT states: "Waterloo's Policy 77 states that 'student evaluations are an important source of information' in the assessment of teaching" (p. 5). However, such policy statements are implemented

with more malleable procedures. For example, Waterloo's procedure in compliance with policy might make students' ratings routinely available to department chairs as a way of triggering further examination of whether instructional performance is deficient—without routinely passing those ratings to peer review committees. (This can be called the "canary in a coal mine" function of the ratings.)

CEPT most likely is aware of pressures from the provincial level for university "accountability," and some at our university may believe it would be helpful in this regard to, for example, publish students' course ratings. The OCUFA panel that was mentioned in our opening statement has begun to consider this angle. Initial indications are, again, that there is considerable flexibility in how each university may respond to such pressures or requirements. If anyone were to make claims of necessary summative use or publication of ratings based on the provincial context, the full basis of that claim should be spelled out.

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

Student questionnaires concerning perceptions of instruction were first devised many decades ago. At our university they were not used universally before 1980, and their significance in pay and promotion decisions has increased since that time. Roughly parallel developments have occurred at some other universities. As noted earlier, in the 1980s some evidence was said to have accumulated to justify summative use, but that evidence has been refuted. In particular, Uttl et al. (2016) concluded that "Re-analyses of previous meta-analyses… indicate that SET ratings explain at most 1% of variability in measures of student learning" and also that the newer research (as in Point 2a1) shows that students' ratings nowadays "are unrelated to student learning."

The CEPT report refers repeatedly to best practices at other universities, but is has become clear that there are no good practices for summative evaluation based on student questionnaires. Canada's most innovative university is following a misguided pack instead of leading the way toward instructional excellence, as it might through focusing student questionnaires on formative instead of summative evaluation.

As psychologists, several of us are familiar with another instance of 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 a while and very possibly harmful to some 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. 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 we believe, based on the evidence, that they are. Relatedly, we certainly are 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.

6a Student questionnaires are particularly vulnerable to bias because they are not "grounded" in any way.

Sample questions in the CEPT report 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 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 favorable 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 response can be complex and not-at-all what it appears to be about on the surface.

6b Proper peer evaluation is grounded in facts and explicit criteria and therefore less vulnerable to bias.

One salient alternative to student questionnaires for obtaining summative evaluation data is to use faculty peer evaluation. Although, as one member of CEPT stated, any measurement involving human judgment can be subject to bias, *properly* conducted peer evaluation is less obviously vulnerable. Such evaluation could draw on CEPTs framework of dimensions of instructional effectiveness, and possibly other sources, to establish criteria that are more specified and objective than many of the student questionnaire items about students' "perceptions and experiences." The array of measurable dimensions may expand with peer evaluation, because 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 followed 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. 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.

6c 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.

The present response from many UW Psychology professors is not the place to try to develop an alternative to the use of student questionnaires for summative evaluation. We recognize that developing such an alternative will require time and effort. However, because the current procedure has been shown to be glaringly invalid and therefore unjust, and because it undermines student learning

and academic freedom, there is no justifiable choice in the matter of whether to 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. Decisions about student questionnaire redesign and use should take full account of the best available internal (University of Waterloo) and external expert analysis and opinion, but so far this has not happened.

When the university considers changes to the aspect of compensation called "pensions," it draws heavily upon internal expertise (as from Actuarial Science) and external consultants. The same should have happened when CEPT considered changes to the student questionnaires that bear on pay, but it did not. Although external expertise was sought by CEPT to some degree by examining relevant literature, the now-prominent works by Stark, Freishstat, and Wieman (as reviews and analyses) appear not to have been given proper consideration, and other key works as by Stroebe and Nilson either have appeared very recently or were less prominent and understandably had to be brought to CEPT's attention. Although members of the Department of Psychology were consulted, their advice simply was not taken, with the explanation that opinions can differ on the matters under consideration. Opinions differ on climate change, too, but some opinions are much better informed than others.

The evidence 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 remedied. Members of the Department of Psychology stated these facts to CEPT but, admittedly, did not document them nearly as fully as they have now been documented in this statement. It is now time for CEPT to take full account of the facts and recommend that the university cease summative use entirely.

References

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

Boysen, G. A. (2015a). 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*, 150–162.

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.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3): 83–87

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

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

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

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.

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). Jossey-Bass.

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.

Uttl, B., White, C.A., & Gonzalez, D. W. (2016, in press). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*. 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.

Drafted or Endorsed by

Wendi Adair, Associate Professor

Britt Anderson, Associate Professor

Derek Besner, Professor

Kathleen Bloom, Associate Professor

Ramona Bobocel, Professor

James Danckert, Professor

John Holmes, Distinguished Professor Emeritus

Roxane Itier, Associate Professor

Colin MacLeod, Professor & Chair

Jay (John) Michela, Associate Professor

Ian McGregor, Associate Professor

David Moscovich, Associate Professor

Jonathan Oakman, Associate Professor

Christine Purdon, Professor

Abigail Scholer, Associate Professor

Paul Wehr, Lecturer

Joanne Wood, Professor

Erik Woody, Professor

Expertise of the Most-Cited Sources in the Statement

Richard Freishstat currently serves as Director of UC Berkeley's Center for Teaching and Learning (CTL). In this capacity, he creates, leads, and facilitates 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. Ha also has had campus-wide responsibilities regarding educational technology, including technology involved in teaching evaluations. have 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.

Expertise of the Psychology Faculty Members Who Have Provided Input Directly to CEPT

Ramona Bobocel is Professor of Industrial-Organizational Psychology in the Psychology Department. Her expertise lies in the study of fairness and trust in work organizations, and in survey research methodology and analysis. She has special expertise in the field of Psychometrics, having developed and taught our department's advanced undergraduate course in Psychological Measurement for over 20 years. Dr. Bobocel has published numerous book chapters and research articles in leading scientific journals in psychology and management and has presented her findings at conferences and universities around the world. Her research has been funded continuously by Social Sciences and Humanities Research Council of Canada since 1996, and she has received numerous awards for her contributions. These include the Ontario Premier's Research Excellence Award and an industry-research collaborative grant from Bell University Labs to examine fairness from employee and management perspectives using large scale field survey methodology and analysis. Most recently, she was awarded Fellow status in the Association for Psychological Science for her sustained scientific contributions. She has served as Editor and Consulting Editor for top journals in her field, is past President of the Canadian Society for Industrial and Organizational Psychology. In addition to her academic research, Dr. Bobocel has consulted with several Canadian organizations seeking to assess and promote employee fairness in work organizations.

Jay (John) Michela is an associate professor in the Industrial-Organizational area of the Department of Psychology. His earlier academic work, 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 analyzed 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 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 APA 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. 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 Psychology and in Accounting and Financial Management.

Erik Woody was awarded the UW Distinguished Teacher Award in 2006. He has taught courses in test construction and psychometric theory for 35 years, attracting graduate students and even faculty members from several departments at UW beyond Psychology, as well as students from the Business School and students and faculty from Psychology at WLU. In his courses, he uses real data from scholarship competitions to illustrate bias and how to detect it, and he has also had undergraduates in research seminars collect data from multiple real courses to study the determinants of course-evaluation ratings (which produce results that invariably surprise them). He has published articles regarding a wide variety of quantitative techniques in several peer-reviewed journals, including *Psychological Methods* and *Psychological Assessment*, and also served as an Associate Editor for submissions on psychometric analysis.