CSSHE
SCÉES

Canadian Journal of Higher Education
Revue canadienne d'enseignement supérieur
Volume 46, No. 4, 2016, pages 115 - 131

Addressing Common Concerns about Online Student Ratings of Instruction: A Research-Informed Approach

Laura Winer, Lina DiGenova, Andre Costopoulos
McGill University

Kristen Cardoso
Middlebury Institute of International Studies at Monterey

Abstract

Concerns over the usefulness and validity of student ratings of instruction (SRI) have continued to grow with online processes. This paper presents seven common and persistent concerns identified and tested during the development and implementation of a revised SRI policy at a Canadian research-intensive university. These concerns include bias due to insufficient sample size, student academic performance, polarized student responses, disciplinary differences, class size, punishment of rigorous instructor standards, and timing of final exams. We analyzed SRI responses from two mandatory Likert scale questions related to the course and instructor, both of which were consistent over time and across all academic units at our institution. The results show that overall participation in online SRIs is representative of the student body, with academically stronger students responding at a higher rate, and the SRIs, themselves, providing evidence that may moderate worries about the concerns.

Résumé

Avec les processus électroniques, les inquiétudes quant à l'utilité et à la validité des évaluations de l'enseignement par les étudiants (EEE) ne cessent de croître. Le présent document révèle sept problèmes communs et constants concernant l'utilité et la validité des évaluations électroniques de l'enseignement par les étudiants (EEE) en ligne qui ont été identifiés et testés dans une université canadienne centrée sur la recherche. Parmi ces problèmes,

on compte une déformation des résultats attribuable à un échantillon de taille insuffisante, une faible performance scolaire des étudiants, une polarisation des réponses des étudiants, des différences disciplinaires, des classes de taille inégale, une perception négative face aux attentes élevées de certains chargés de cours et l'horaire des examens finaux. Nous avons analysé les réponses à deux questions obligatoires, selon une échelle de Likert, et liées au cours et à son chargé de cours. Les deux questions ont conservé leur cohérence au fil du temps et au sein de l'ensemble des unités d'enseignement de notre institution. Les résultats démontrent que la participation à l'EEE en ligne est généralement représentative du corps étudiant, bien que le taux de participation des étudiants plus performants au niveau académique s'est révélé plus élevé. Cela nous fournit un argument important pour répondre aux inquiétudes souvent émises au sujet des problèmes liés à l'EEE.

Introduction

Reviews of student ratings of instruction (SRIs) have already addressed a number of common concerns regarding their validity and usefulness, but these concerns continue with the increasing use of online ratings. We have identified seven categories of persistent concerns often voiced by instructors. The identification process was informed by over a decade of professional consultations with instructors at McGill University and the research literature that has examined different facets of SRIs. Data analyses with online SRIs can help us address common concerns using an evidence-based approach (Adams & Umbach, 2012; Avery, Bryan, Mathios, Kang, & Bell, 2006; Gravestock & Gregor-Greenleaf, 2008). The concerns can be divided into three categories: respondents, context, and academic rigour (Carell & West, 2010; Cashin, 1990; Johnson, 2003). The concerns related to *respondents* are sample size, weak versus strong academic performance, and the "love it or hate it" response, which suggests that only students with extreme opinions complete SRI forms (Centra, 2003; Hakstian, Rawn, & Cutler, 2010). The concerns related to *context* are academic disciplinary differences and class size (Beran & Violato, 2010; Leung & Kember, 2011). Finally, *academic rigour* involves concerns about maintaining academic standards and the impact of final exams on SRIs (McNulty et al, 2010).

Approach

We approached this study as faculty, administrators, and educational developers working to establish and implement a comprehensive McGill University online SRI policy. Since 2008, Teaching and Learning Services (the unit responsible for overseeing SRIs) has worked with the Course Evaluation Advisory Group (CEAG), a committee of faculty, academic administrators, staff, and students mandated to develop SRI interpretation guidelines for use by our university. These guidelines are intended to help instructors improve the delivery of their courses; help administrators and faculty committees in their decision-making processes regarding reappointment, tenure, promotion, and merit; and educate students about how to provide constructive feedback to instructors. Our collaborative approach with CEAG and consultation with additional members of the community highlighted the need to demonstrate the outcome of each concern using evidence from

our institutional SRI data. Below we outline each common and persistent concern, as well as the analyses used to generate meaningful and engaging discussions about the validity and usefulness of SRIs across campus.

Common Concerns in the Literature

Concern 1: Sample Size

Online SRIs generally have a lower response rate than traditional paper ratings, creating a concern that online response rates may be too low to constitute a representative sample. The size of the sample is a pervasive concern and one of the most frequently studied SRI issues (Gravestock & Gregor-Greenleaf, 2008; Nulty, 2008; Sorenson & Reiner, 2003). Obtaining adequate response rates is particularly problematic with online ratings. Today's instructors are increasingly concerned with response rates because SRIs have a significant impact on administrative decision making (Marsh, 2007a). Although response rates were higher for traditional in-class paper SRIs than for online SRIs at most universities, there is no evidence that the quality of online SRIs data is inferior (Benton & Cashin, 2012). A study comparing paper and online SRIs found no significant differences in sex, class standing, or expected grade between online and paper respondents (Stowell, Addison, & Smith, 2012), supporting the idea that although online rating forms have lower response rates, the rates are high enough to be adequately representative of the class as a whole.

Negative response bias is often expressed as a concern related to sample size because instructors feel that dissatisfied students and those with lower grades are more likely to complete rating forms (Benton & Cashin, 2012; Johnson, 2003) or to write comments (Sorenson & Reiner, 2003). Johnson (2003) found no evidence that lower response rates for online forms result in lower instructor ratings.

Concern 2: Weak Student

Instructors have said they think that students who perform poorly in a course will punish them regardless of the quality of instruction and fairness of assessment. The weak student concern is similar to concerns over negative response bias. Instructors fear low ratings as retribution from students who perform poorly in a course, despite a number of studies that have demonstrated a positive bias. Students with higher cumulative grade point averages (CGPA) are more likely than peers with a lower CGPA to complete rating forms (Adams & Umbach, 2012; Avery et al., 2006; Hativa, 2014; Porter & Umbach, 2006; Sorenson & Reiner, 2003). Several studies have also found that students with higher grade point averages (GPA) are more likely to complete rating forms than those with lower GPAs (Porter & Umbach, 2006; Sorenson & Reiner, 2003), and that students who perform poorly in a course or anticipate a low grade are less likely to respond than better-performing students (Adams & Umbach, 2012; Avery et al., 2006). Adams and Umbach (2012) found in a four-year study that students with D and F grades were 0.77 times as likely to complete a rating form as students in the same course with better grades. In their reviews of past research, Benton and Cashin (2012) and Hativa (2013) point to a number of studies that found little or no relationship between student ratings and CGPA (e.g., Abrami, 2001; Marsh & Dunkin, 1997; Marsh & Roche, 2000).

Concern 3: Love It or Hate It

Instructors often voice concern that only students with extreme opinions respond, and that results do not form a fair representation of student opinion of their teaching. Of the seven concerns, the love it or hate it concern is the least researched. Gravestock and Gregor-Greenleaf (2008) acknowledge this concern by faculty and administrators and call for further investigation. A study of 3,067 written comments by students from Tel Aviv University found that students who wrote comments tended to have stronger views than those who did not add comments, and that the majority of written comments were positive rather than negative (Alhija & Fresko, 2009). Although further research in this area is needed, these findings suggest that while students with strong opinions may be more apt to leave comments, these students do not accurately reflect the opinions of all students who respond as represented in numerical ratings.

Concern 4: Discipline Specific

Instructors sometimes think student ratings fail to account for the inherent difficulty in teaching their particular discipline. Differences in teaching styles and goals among disciplines may account for some differences in ratings (Gravestock & Gregor-Greenleaf, 2008), and there is some evidence these disciplinary differences in style and goals affect SRI ratings. Studies have found that the humanities tend to receive the highest ratings, followed by the social sciences, and then natural sciences (Cashin, 1990; Hativa, 2013; Johnson, 2003; Neumann, 2001; Ory, 2001; Wachtel, 1998). Centra (2009) found that courses in the natural sciences, mathematics, engineering, and computer science had an average mean about a third of a standard deviation less than humanities courses. Because most ratings instruments are generic, Neumann (2001) calls for more research on the role of disciplines in shaping teaching and defining teaching effectiveness.

Recent studies suggest that the overall nature of effective teaching and learning are nevertheless consistent across disciplines (Hativa, 2013). Leung and Kember (2011) studied 3,305 students in four discipline areas—humanities, business, hard science (engineering and science), and health sciences and medicine—and concluded that data from each disciplinary group fit into a common model of good teaching. The differences that emerged seemed rooted in the epistemological nature of the disciplines, but socially constructed narratives may contribute as much to epistemological beliefs as true disciplinary differences.

Concern 5: Class-Size

Instructors teaching larger introductory courses are often concerned they will unfairly receive lower ratings than instructors teaching smaller, higher level courses. There is evidence that smaller, higher level courses receive slightly higher ratings than larger, lower level courses, especially if the higher level courses are graduate courses (Marsh, 2007b; Marsh & Dunkin, 1997). However, the correlation between class size and ratings is statistically insignificant and therefore does not impact validity (d'Apollonia & Abrami, 1997; Gravestock & Gregor-Greenleaf, 2008; Marsh, 1987; Marsh & Roche, 1997). Although smaller courses receive higher ratings, students also report learning more in them, suggesting that the effect of class size is a reflection of student learning rather than an indication of bias (Centra, 2009). Interestingly, very large classes have also been found to

receive higher ratings than medium-sized courses, suggesting that neither class size nor course level biases instructor ratings (Hativa, 2013).

In a study comparing instructors with the highest and lowest student ratings, negative correlations were found between the rating for overall teaching effectiveness and class size, with no correlation between rating and level of difficulty (Pan et al., 2009). Analyses of ratings for teaching award winners found no significant correlations between overall teaching effectiveness, expected grade, and level of difficulty. Teaching award winners were found to teach significantly larger classes on average than other instructors, and a negative correlation was found between effectiveness and class size. Another study, analyzing 294,692 student responses for 8,065 course sections, found no effect on ratings from class size or level (Pepe & Wang, 2012).

Concern 6: Rigorous Standards

Instructors sometimes think that students punish them for maintaining rigorous academic standards. They believe that lower grades will result in lower ratings scores and higher grades in higher scores, regardless of the quality of instruction. There is a persistent fear that instructors who give low grades will be unfairly punished, while those who give high grades will be rewarded. It is one of the most contentious issues around student ratings. Marsh (1987) refers to this as the leniency hypothesis, where leniency in assigning grades will result in more favourable ratings, as opposed to the validity hypothesis, which states that students who learn more in a course will likely receive higher grades and also rate their instructor more highly because of how much they learned.

A few studies have found significant correlations between expectations of high grades and positive ratings (Greenwald & Gillmore, 1997; Wachtel, 1998); however, Marsh and Roche (2000) found that Greenwald and Gillmore did not account for student learning. Studies continue to be published supporting the grading leniency hypothesis, although many of these studies are limited in scope or sample size. Crumbley, Flinn, & Reichelt (2010), for example, relied heavily on anecdotal evidence regarding the unethical behaviour of instructors giving higher grades due to the use of student ratings, while Carrell & West's (2010) methodology made their results non-generalizable to most universities.

Nonetheless, a considerable number of studies have found no significant correlations between expected grades and instructor ratings. Less demanding courses can get lower ratings than more challenging ones (Centra, 2003; Heckert et al., 2006; Marsh & Roche, 2000), and when grades are perceived as too high, instructors receive lower ratings (Abrami, 2001). Pepe and Wang (2012) found that communication is the most important consideration for students in giving an instructor an excellent score. This is supported by a similar study finding that students give high ratings to instructors who are clear in explaining and aiding understanding, while giving lower ratings to instructors who are unclear or ineffective lecturers (Pan et al., 2009).

A study at a major Canadian university found that students engaged in their own learning tend to receive higher grades and give higher ratings (Beran & Violato, 2010). Another study found that the average exam score increased as students' self-rated learning increased (Benton et al., 2013). Patrick (2011) found that, although expected grade and overall rating of the course were significantly correlated, the correlation between expected grade and overall teaching effectiveness was not, suggesting that expected grades

did not significantly affect students' opinions of the instructor. The majority of the research supports the idea that rigorous standards can engage students in their learning more, accounting for higher instructor ratings.

Concern 7: Final Exam

Prior to the introduction of online ratings, for logistical reasons all ratings were completed in class before the final examination. Online forms offer the possibility of keeping the ratings open until after exams; however, there is a widespread concern that students will punish instructors for a challenging final exam, resulting in the common practice of closing rating forms before the final is written. The concern that response rates or instructor ratings will be lower after the final exam in a course seems to be rooted in studies from the 1970s and 1980s (Ory, 2001; Wachtel, 1998) and has led to changes in the timing of administering SRIs in some programs and institutions. These studies only refer to paper-rating forms (Wachtel, 1998). A report on response rates at the University of British Columbia found no studies testing the effect of timing on ratings, and that all studies on response rates were conducted during the final weeks of the semester (Hakstian et al., 2010). Two recent studies experimented with the timing of ratings; however, both were interested in the effect of exam grades on ratings rather than response rates (Arnold, 2009; McNulty et al., 2010). No recent studies have focused on the effect of timing on response rates.

While the literature has examined each of the concerns presented above separately, our study provides an early perspective on SRIs from online data collection using a more longitudinal perspective for all of the concerns. In addition, our research provides new evidence for the love it or hate it and final exam concerns. Our study's objectives were to share our research observations for each of the concerns about SRIs and to contribute to advancing future research in online SRIs.

Hypotheses

We hypothesized that the preliminary evidence of online student ratings of instruction do not support the seven persistent concerns: sample size, weak student, love it or hate it, discipline specific, class size, rigorous standards, and final exam.

In particular, with respect to the *nature of the respondents*, we hypothesized that the sample is representative:

- There is no difference on student demographics (sex, geographic origin, year of study, academic load and discipline) when comparing course ratings of participants versus nonparticipants.
- There is no difference on student academic performance (weak vs. strong student).
- The responses are not bimodal, that is, not only the students who love the course or hate it complete the rating forms.

For the two concerns related to *context*, we hypothesized that there are no differences

- across academic disciplines, and
- class sizes.

Finally, for concerns related to *academic rigour*, we hypothesized that there are no differences

- related to academic standards, that is, students do not punish instructors who have high academic standards, and
- in student ratings based on the timing of student ratings of instruction into the exam period.

Materials and Methods

At McGill University, end-of-course SRIs have been conducted exclusively online with a university-wide student rating system since 2006. All courses with five or more registered students are rated. The typical evaluation period lasts for approximately three weeks and ends the day before final exams begin. Since 2011, individual academic units have had the option of extending the rating period to the last day of exams.[1]

Student participation in course and instructor rating is voluntary and anonymous; there is no way to link any individual student to a specific response. However, the university student information system retains information on whether students have completed one or more student rating forms in each academic term.[2]

We divided the student population in two groups: participants who completed at least one course rating in a given semester, and nonparticipants, who did not complete any ratings. The participation versus non-participation data was linked to student demographic information.

Questionnaire

The Student Rating of Instruction Questionnaire has three parts: (1) four mandatory questions, (2) up to 21 additional questions, and (3) a section for written comments. The data set of this analysis was the responses to the four mandatory questions that are consistent over time and across all academic units. More specifically, the analyses in this study focused on two of the four mandatory items:

> Course question
> (Q1) Overall, this is an excellent course
> Instructor question
> (Q3) Overall, this instructor is an excellent teacher.

The mandatory questions are answered on a scale from 1 to 5 (1 = strongly disagree to 5 = strongly agree).

Population

End-of-course ratings from students enrolled in undergraduate degree-seeking programs at the university were included in this study. The academic units included in this study were selected by the CEAG and considered representative. Data used to investigate the sample size, weak student, love it or hate it, and rigorous standards concerns were from academic years 2008 to 2009 and 2009 to 2010. The class size concern used fall 2009 data and the extended dates analysis are from fall 2009 and fall 2010, which coincides with policy changes to SRI timing at the university. Data were from courses taught by single instructors only. The specific time frame and academic disciplines included in the analyses per concern are outlined below.

Analyses

The methods of analyses consisted of means comparisons and correlations. Means comparisons using independent *t*-tests were conducted to compare participants versus nonparticipants, and academic disciplines over time. *T*-tests were also used to compare grades and class-size comparisons. Finally, a correlation was conducted to show the relationship between grades and academic discipline.

Results

Concern 1: Sample Size

Student demographic characteristics of participants versus nonparticipants in 2008 to 2009 and 2009 to 2010 were compared in order to investigate the sample size concern and weak student concern. The student demographic characteristics included in the analysis were sex, year of study, academic load, geographic origin, academic discipline, and CGPA. The total number of students for 2008 to 2009 was 18,699 with a distribution of 53% participants versus 47% nonparticipants. Similarly, for 2009 to 2010, the distribution of participants was 55% versus 45%, with a total of 19,383 students. A comparison of the participants versus nonparticipants showed no statistical differences across sex, year of study, academic load, geographic origin, or academic discipline. We find no evidence of bias in our data resulting from online samples (Table 1).

Concern 2: Weak Student

In order to answer the question about academic performance or the weak student effect, data from seven academic units were analyzed. The CGPA for the participant and nonparticipant groups was compared by academic unit using *t*-tests for fall 2008 and fall 2009. Statistically significant differences were found in all cases; students who participated tended to have higher CGPAs (ranged from 0.13 to 0.40; Table 2) than students who did not. Participants had a mean CGPA between 3.15 (SD = 0.45) and 3.46 (SD = 0.44); the mean CGPA for nonparticipants ranged from 2.94 (SD = 0.59) to 3.14 (SD = 0.65).[3] In summary, participants tended to be representative of the class as a whole for every characteristic examined except CGPA, where they tend to be stronger.

Concern 3: Love It or Hate It

Instructors often think that the students who respond are only those with extreme opinions, meaning that the results are not from a representative sample of students in a class (Alhija & Fresko, 2009; Sorenson & Reiner, 2003). If this were the case, the distributions of ratings should be bimodal; however, this is not shown in the data (Table 3). The mean ratings and standard deviations reported above indicate that the findings are not bimodal, and this is true across disciplines. Nevertheless, instructors may be left with the impression that strongly happy or unhappy students respond disproportionately because open-ended questions (comment fields) tend to be answered by students who express strong views.

Table 1.
Student Characteristics of Participants vs. Nonparticipants (2008–2009 and 2009–2010)

	2008–2009			2009–2010		
	Participants	Non participants	TOTAL	Participants	Non participants	TOTAL
N	**9,967**	**8,702**	**18,699**	**10,585**	**8,798**	**19,383**
	%	%	%	%	%	%
Overall	53	47	100	55	45	100
Sex						
Female	57	43	100	59	41	100
Male	48	52	100	49	51	100
Year of Study*						
0	54	46	100	58	42	100
1	59	41	100	59	41	100
2	54	46	100	55	45	100
3 &4	48	52	100	49	51	100
Academic Load						
Full-time	54	46	100	56	44	100
Part-time	43	57	100	39	61	100
Geographic Origin						
In-province	55	45	100	56	44	100
Out of province	52	48	100	53	47	100
Out of province International	53	47	100	54	46	100
Academic Discipline						
Health Sciences	52	48	100	54	46	100
Humanities & Social Sciences	56	44	100	57	43	100
Science & Engineering	52	48	100	51	49	100

* Similar to Freshman, Junior, Sophomore and Senior

Table 2.
CGPA Comparison of Participants vs. Nonparticipants (2008–2009 and 2009–2010)

Discipline	*Fall 2008*			*Fall 2009*		
	N	ΔMean CGPA+	Sig***	*N*	ΔMean CGPA++	Sig***
Humanities	570	0.13	$p<0.001$	522	0.24	$p<0.001$
Humanities	445	0.13	$p<0.001$	415	0.24	$p<0.001$
Sciences	400	0.25	$p<0.001$	400	0.24	$p<0.001$
Sciences	182	0.40	$p<0.001$	218	0.37	$p<0.001$
Social Sciences	1486	0.22	$p<0.001$	1498	0.18	$p<0.001$
Social Sciences	1759	0.15	$p<0.001$	1684	0.21	$p<0.001$
Social Sciences	760	0.21	$p<0.001$	857	0.12	$p<0.001$

++ Participants CGPA – Nonparticipants CGPA
*** 2-tailed

Concern 4: Discipline Specific Difficulty

The concern that some disciplines are inherently more difficult to teach and therefore result in instructors receiving lower ratings was assessed by comparing the mean ratings using an independent sample *t*-test for the two items examined across academic disciplines in fall 2009. The mean ratings on a scale of 1 to 5 are 3.8 (*SD* =1.0; range 3.8–4.0) and 3.9 (*SD* =1.1; range 3.9–4.1) for the course question and instructor question respectively. The difference between the disciplines with higher ratings (humanities and social science) and the others (engineering and sciences) is in the order of 0.1, signalling a detectable, but small, difference in ratings for the course or instructor (Table 3). Based on our preliminary analysis with the academic departments investigated, the differences were not considered large enough to be practically meaningful.

Table 3.
Mean Ratings by Academic Discipline by Course Question (Q1) and Instructor Question (Q3) (Fall 2009)

		Q1		Q3	
Discipline	*n*	Mean	*SD*	Mean	*SD*
Humanities / Social Sciences	768	3.8	1.0	3.9	1.1
Humanities / Social Sciences	152	3.8	1.1	4.0	1.1
Humanities / Social Sciences	46	3.8	1.1	4.1	1.1
Humanities / Social Sciences	375	3.9	1.0	4.0	1.1
Humanities / Social Sciences	89	4.0	1.0	4.0	1.1
Sciences / Engineering	136	3.8	1.0	3.9	1.1
Sciences / Engineering	316	3.8	1.1	3.9	1.2
Sciences / Engineering	52	3.8	1.1	3.9	1.1
Sciences / Engineering	365	3.8	1.1	3.9	1.2

Concern 5: Class Size

Larger courses are often considered more difficult to teach, and consequently produce lower ratings than smaller courses. Mean rating comparisons for the course and instructor questions (on a scale of 1 to 5) are presented in Table 4 for fall 2009. The ratings illustrate that smaller courses (up to 30 students) received higher ratings (between 4.1 and 4.3 for the course question and 4.2 and 4.3 for the instructor question). Mean ratings for classes ranging from 31–100, 101–200, and over 201 students were 3.9 for the course question and between 3.8 and 4.0 for the instructor question. However, the difference is usually on the order of 0.1 over a standard deviation of approximately 1, which means that the distributions overlap almost completely.

Table 4.
Class Size Comparisons for Course Question (Q1) and Instructor Question (Q3) (Fall 2009)

		Q1		Q3	
	N	Mean	SD	Mean	SD
5 to 10	299	4.3	0.8	4.3	0.9
12 to 30	756	4.1	0.9	4.2	1.0
31 to 100	976	3.9	1.0	4.0	1.1
101 to 200	248	3.9	1.2	3.8	1.1
201 or more	69	3.9	1.0	4.0	1.0

Concern 6: Rigorous Standards

A correlational analysis was conducted to address the rigorous-standards concern ("I have standards so because I'm a rigorous grader, I am unfairly punished"). We analyzed university data from fall 2008 and fall 2009 on the four mandatory questions. Although statistically significant ($p < 0.001$), the correlation between grades and ratings is weak across 5 disciplines ($r = 0.1$) thus accounting for little of the variance.[4]

Concern 7: Final Exams

To assess the concern that administering end-of-course ratings during the final exam period would adversely impact response rates and ratings, two analyses were conducted: response rate and mean rating comparisons prior to final exams versus the extended dates during final exam period by academic unit.

Response rates by academic unit were compared for fall 2009 and fall 2010 (Table 5) when a change in practice at the university began. A *t*-test mean comparison showed that the rate was statistically higher for two units identified in Table 3 as a humanities and social sciences unit, and a science unit ($p < 0.01$), and the same for the rest of the units. There was no reduction in response rate; if anything, the response rates trended upwards with the extended dates option.

Table 5.
Response Rates 2009 vs. 2010 for Units Following Extended Dates

Discipline	N	Fall 2009 (Regular dates) %	Fall 2010 (Extended dates) %	Sig***
Humanities / Social Sciences	28	50%	48%	*ns*
Humanities / Social Sciences	103	47%	55%	$p < 0.01$
Humanities / Social Sciences	13	51%	51%	*ns*
Sciences / Engineering	6	67%	68%	*ns*
Sciences / Engineering	23	37%	49%	$p < 0.01$
Sciences / Engineering	21	57%	60%	*ns*
Overall	194	52%	55%	$p < 0.01$

*** Two-tailed

The second analysis compared the mean ratings for courses taught by the same instructors in fall 2009 during the regular data collection period (prior to final exams) versus the extended dates in fall 2010. Table 5 shows that mean ratings during final exams were higher in three academic units (ranging from 0.03 to 0.14), and lower in four academic units (ranging from –0.11 to 0.23). The *t*-test mean rating differences were not statistically significant in six of the seven academic units, suggesting that extending end-of-course ratings during the exam period does not adversely impact student ratings.

Table 6.
Mean Ratings 2009 vs. 2010 Comparison for Units Following Extended Dates

Discipline	Fall 2009 Regular dates	Fall 2010 Extended dates (during final exams)	Δ Mean Ratings (2010-2009)	Sig
Humanities / Social Sciences	317	353	0.14	*ns*
Humanities / Social Sciences	461	460	0.03	*ns*
Sciences	226	216	-0.13	*ns*
Sciences	226	349	0.08	*ns*

Discussion and Conclusions

We did not find support for the most common concerns about the validity and usefulness of online SRIs. The few effects we found are either evidence against the common concerns or are too small to be considered meaningful. Overall, course-rating respondents are representative of the student body, except that academically stronger students are more likely to complete rating forms. The results related to the context showed that the differences in both academic discipline and class size were small, on the order of 0.1. In

particular, humanities and social sciences ratings were higher than those for science and engineering courses. Smaller courses received slightly more favourable ratings but the distributions of ratings overlap almost completely, and the difference is tiny compared to the standard deviation. Finally, two of the five academic disciplines we examined for timing of the final exam showed higher ratings when the rating period extends beyond the exam.

The literature does not provide any systematic validation of the concerns outlined, although sometimes the questions were not addressed specifically and require some inference. In contrast, we provide explicit tests of the most common concerns about online student ratings of teaching and show that none is supported; the literature was particularly vague regarding the love it or hate it concern. The analysis around the final exam concern has not been addressed in the context of online ratings of instruction with the important logistical advantages that they provide, and so these findings are of special note.

This study is, of course, limited since it draws on data from only one institution. Future studies will explore whether the findings are replicated elsewhere. A second limitation of this investigation is that the study is observational and based on mean comparisons and correlational analyses of select SRI data. A more systematic analytical approach using the entire SRI data would be ideal and allow for greater statistical understanding of the variable relationships and control of error terms. Future studies will explore online SRIs using different data-mining techniques and analyses, such as structural equation modelling, and will include additional variables beyond the four mandatory questions.

Our research shows promising results in support of online SRIs, which in turn also offer numerous advantages, notably ease of administration and completion, speed of analysis and reporting, complete confidentiality of comments, and the possibility of including the final exam as a part of the rating process.

Notes

1 Note that in 2014 the policy changed so that the default period extends to two days after the exam period, but academic units may choose a condensed evaluation period ending the day before final exams begin. See www.mcgill.ca/mercury for the policy and other relevant information.

2 The University Research Ethics Board granted approval for this use of the data.

3 For more information about CGPA, see http://www.mcgill.ca/mercury/files/mercury/course_evaluation_results_interpretation_guidelines.pdf

4 For more information, see http://www.mcgill.ca/mercury/files/mercury/course_evaluation_results_interpretation_guidelines.pdf

References

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. In M. Theall, P. C. Abrami, and L. A. Mets (Eds.), The student ratings debate: Are they valid? How can we best use them? [Special issue]. *New Directions for Institutional Research, 2001*(109), 59–87.

Adams, M. J. D., & Umbach, P.D. (2012). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education, 53*(5), 576–591.

Alhija, F. N. A. & Fresko, B. (2009). Student evaluation of instruction: What can be learned from students' written comments? *Studies in Educational Evaluation, 35*(1), 37–44.

Arnold, I. J. M. (2009). Do examinations influence student evaluations? *International Journal of Educational Research, 48*(4), 215–224.

Avery, R. J., Bryan, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *Journal of Economic Education, 37*(1), 21–37.

Benton, S. L., & Cashin, W. E. (2012) Student ratings of teaching: A summary of research and literature. *IDEA Paper No. 50*. Manhattan, KS: Kansas State University Center for Faculty Evaluation & Development.

Benton, S. L., Duchon, D., & Pallett, W. H. (2013). Validity of student self-reported ratings of instruction. *Assessment & Evaluation in Higher Education, 38*(4), 377–388.

Beran, T., & Violato, C. (2010). Student ratings of teaching effectiveness: Student engagement and course characteristics. *Canadian Journal of Higher Education, 39*(1), 1–13.

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, *118*(3), 409–432.

Cashin, W. E. (1990). Students do rate different academic fields differently. In M. Theall & J. Franklin (Eds.), Student ratings of instruction: Issues for improving practice [Special issue]. *New Directions for Teaching and Learning, 1990*(43), 113–121.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*(5), 495–518.

Centra, J. A. (2009). *Differences in responses to the Student Instructional Report: Is it bias?* Princeton, NJ: Educational Testing Service.

Crumbley, D. L., Flinn, R. E., & Reichelt, K. J. (2010). What is ethical about grade inflation and coursework deflation? *Journal of Academic Ethics, 8*(3), 187–197.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198–1208.

Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends*. Toronto, ON: Higher Education Quality Council of Ontario.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209–1217.

Hakstian, A. R., Rawn, C. D., & Cutler, C. (2010). Student evaluation of teaching: Response rates. Vancouver, BC: University of British Columbia.

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

Hativa, N. (2014). *Student ratings of instruction: A practical approach to designing, operating, and reporting*. *Second edition*. Charleston, SC: Oron Publications.

Heckert, T. M., Latier, A., Ringwald-Burton, A., & Drazen, C. (2006). Relations among student effort, perceived class difficulty appropriateness, and student evaluations of teaching: Is it possible to 'buy' better evaluations through lenient grading? *College Student Journal, 40*(3), 588–596.

Johnson, T. D. (2003). Online student ratings: Will students respond? In D. L. Sorenson & T. D. Johnson (Eds.), Online student ratings of instruction [Special issue]. *New Directions for Teaching and Learning, 2003*(96), 49–60.

Leung, D. Y. P., & Kember, D. (2011). Disciplinary differences in student ratings of teaching quality. *Research in Higher Education, 52*(3), 278–299.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253–388.

Marsh, H. W. (2007a). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht, Netherlands: Springer.

Marsh, H. W. (2007b). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99*(4), 775–790.

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241–313). New York, NY: Agathan Press.

Marsh, H. W., & Roche, L. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias and utility. *American Psychologist, 52*(11), 1187–1197.

Marsh, H. W., & Roche, L. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92*(1), 202–228.

McNulty, J. A., Gruener, G., Chandrasekhar, A., Espiritu, B., Hoyt, A., & Ensminger, D. (2010). Are online evaluations of faculty influenced by the timing of evaluations? *Advances in Physiology Education, 34*(4), 213–216.

Neumann, R. (2001). Disciplinary differences and university teaching. *Studies in Higher Education, 26*(2), 135–146.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education, 33*(3), 301–314.

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. In K. G. Lewis (Ed.), Techniques and strategies for interpreting student evaluations [Special issue]. *New Directions for Teaching and Learning, 2001*(87), 3–15.

Pan, D., Tan, G., Ragupathi, K., Booluck, K., Roop, R., & Ip, Y. (2009). Profiling teacher/teaching using descriptors derived from qualitative feedback: Formative and summative applications. *Research in Higher Education, 50*(1), 73–100.

Patrick, C. L. (2011). Student evaluations of teaching: Effects of the big five personality traits, grades, and the validity hypothesis. *Assessment in Higher Education, 36*(2), 239–249.

Pepe, J. W., & Wang, M. C. (2012). What instructor qualities do students reward? *College Student Journal, 46*(3), 603–614.

Porter, R. S., & Umbach, P. D. (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education, 47*(2), 229–247.

Sorenson, D. L. & Reiner, C. (2003). Charting the uncharted seas on online student ratings of instruction. In D. L. Sorenson & T. D. Johnson (Eds.), Online student ratings of instruction [Special issue]. *New Directions for Teaching and Learning, 2003*(96), 1–24.

Stowell, J. R., Addison, W. E., & Smith, J. L. (2012). Comparison of online and classroom based student evaluations of instruction. *Assessment & Evaluation in Higher Education, 37*(4), 465–473.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 29*(2), 191–212.

Contact Information

Laura Winer
Teaching and Learning Services
McGill University
laura.winer@mcgill.ca

Laura Winer (PhD, Educational Technology) is the director of Teaching and Learning Services at McGill University. She has been deeply involved in the development of online course evaluations since 2004. Other dossiers include graduate education initiatives, the appropriate and effective use of technology in teaching and learning, MOOCs, and policy development. She has extensive experience in faculty development and conducted research on educational and training applications of information and communications technology, most recently in the field of medical education. She has published and presented nationally and internationally on faculty development issues and integrating ICT into teaching and learning.

As an assessment professional in Student Services at McGill University, Lina Di Genova (PhD, Psychology) has over 10 years' experience in organizational performance metrics. Lina has led monitoring and evaluation of student affairs programs on topics ranging from orientation to academic advising, wellness, and alumni outcomes. Prior to joining Student Services, Lina worked in university institutional planning on national assessment initiatives, such as the National Survey of Student Engagement benchmarking program for the Canadian U15 Data Exchange and graduate education issues. Lina Di Genova is also a licensed organizational psychologist.

As McGill's dean of students, Andre Costopoulos (PhD, Archeology) facilitates the exchange, development, and dissemination of best practices in areas including academic

advising, academic integrity, the Code of Student Conduct, diversity, and the promotion of a community centred on learning. He studies human adaptation to environmental change in the North in the past 6,000 years and the application of evolutionary theory and computer simulation methods to the study of human evolution. He led a major international collaborative research project during the International Polar Year and has published articles and book chapters on northern prehistory, cultural evolution and environmental change, among other subjects.

Kristen Cardoso is the user experience librarian at the William Tell Coleman Library, Middlebury Institute of International Studies at Monterey (MIIS). She received her Master of Library and Information Studies and Master of Arts in English Literature from McGill University, and her Bachelor of Arts degree in English Literature from the University of Massachusetts-Dartmouth. She has experience in both public and academic libraries, and her areas of expertise include reference, information literacy, and instruction. She is an active committee member of the Monterey Bay Area Cooperative Library System (MOBAC).