INTERPRETING END-OF-COURSE EVALUATION RESULTS

Laura Winer, Lina Di Genova, Pierre-Andre Vungoc and Stephanie Talsma[1]

Teaching and Learning Services

McGill University

Please cite as follows: Winer, L., Di Genova, L., Vungoc, P.-A., & Talsma, S. (2012). Interpreting end-of-course evaluation results. Montreal: Teaching and Learning Services, McGill University.

[1] This document has benefitted from feedback and comments from the Course Evaluation Advisory Group: Evelina Balut, Robert Bracewell, James Brawer, Andre Costopoulos, Nancy Czemmel, Haley Dinel, Alfred Jaeger, Mairead Johnson, Bruce Lennox, Amber Saunders and Kevin Wade.

Contents

1. Executive Summary

End-of-course evaluations are designed to address three goals: 1) improve the delivery of courses in the future; 2) provide a forum for students to provide feedback to academic administrators on current performance; and 3) provide information to future students about a specific course and instructor. At McGill, these data are one component of the process of evaluating teaching for matters of merit, reappointment, tenure and promotion. There is a risk that numbers can be given more importance than warranted in the overall profile of an individual's teaching because of their relative ease of presentation and comparison. However, there are other equally valid factors, sometimes less-well documented or quantifiable, that must be given appropriate consideration in the overall assessment of an individual's teaching performance: participation in curriculum initiatives, innovative teaching strategies, personal development via workshops and undergraduate and graduate supervision. (See the Teaching Portfolio Guidelines for a more detailed discussion http://www.mcgill.ca/tls/teaching/portfolio/guidelines.)

There is a need for all involved to understand the uses of course evaluations:

1. Instructors need to understand that the main goal of evaluating teaching is to help improve teaching through constructive criticism.
2. Students need to view evaluations as a service to the instructor to improve teaching overall and the course in particular as well as a service to future students who will take the course. However, that will only happen when students are convinced that the results are considered seriously and do not simply fall into a void.
3. Both instructors and students should see evaluations as a significant benefit. They can be used as a starting point for conversations. Evaluations remain an important channel for improving one of the key pillars of our University and until they are seen in this positive light, they will not be appreciated by students or instructors.
4. Finally, administrators should use the evaluations to inform an ongoing dialogue about teaching, including positive reinforcement, and identification of areas in need of improvement, and to consider when assigning teaching duties within a unit. Results should not be used simply and primarily as a quick and easy way to assess performance.

2. Introduction

End-of-course evaluations are one of the ways that McGill works towards maintaining and improving the quality of courses and the students' learning experience. As such, their primary goal is to provide constructive feedback to instructors on course content and teaching delivery.[2] In addition, course evaluations form a valuable component of a comprehensive system for instructor evaluation. They should, however, be regarded as only one component of teaching evaluation, as they represent student reports concerning a particular course at a specific point in time, and cannot alone document the full range of teaching responsibilities and accomplishments of a faculty member.

Thus, the challenge for instructors and administrators is to interpret student ratings in ways that are meaningful and reasonable so that faculty can be assured that the decisions that are based on them are appropriate. The following suggestions are drawn from what is now a large body of research literature, representing over thirty years of investigation into the value and effective use of student course ratings.[3] They are supported by analyses of data from McGill end-of-course evaluations from 2008, 2009 and 2010.

3. Purpose

3.1 Target Audience

This document is intended to be used by instructors, teaching assistants and academic unit heads (Chairs, Directors and Deans)[4]. It provides guidelines and recommendations for interpreting both numerical data and written comments from course evaluations in reasonable and productive ways.

3.2 Purpose for Instructors

Course evaluation results are one input to an ongoing reflective process that instructors should engage in to improve their teaching and future offering of courses. The information that students provide, especially the comments, can be useful in identifying areas where changes and modifications have been effective, and those that still require attention. As well, the comments can provide documentation for academic administrators to help with the appropriate interpretation of results.

3.3 Purpose for Teaching Assistants

Teaching Assistants are at the beginning of their teaching careers and so stand to benefit significantly from feedback from students. These guidelines, especially on interpreting written comments, will help TAs benefit from the feedback; consultation with the instructor, Academic Unit Head or trusted peers is strongly encouraged.

---

[2] http://www.mcgill.ca/tls/teaching/course-evaluations/about/policy

[3] For an overview of research on course evaluations, both paper-based and online, see: Gravestock, P. & Gregor-Greenleaf, E. (2008) *Student course evaluations: Research, models and trends*. Toronto: Higher Education Quality Council of Ontario. This document provides a comprehensive review and serves as the basic reference for our interpretation guidelines.

[4] A companion document for students is available at: http://www.mcgill.ca/tls/teaching/course-evaluations/interpretation/students.

3.4 Purpose for Academic Unit Heads

As part of their responsibilities as academic administrators, academic unit heads[5] should discuss course evaluation results with instructors in the annual review process. The academic unit head should regularly review all numerical results and comments, especially for pre-tenure instructors. Course evaluation results, both numerical and written, can also be useful in the merit process. Feedback from courses within programs can also be useful to identify program strengths as well as curricular or other changes that may be warranted.

[5] An academic unit head is defined as a Chair for a Faculty with departments, a Director of a School, and a Dean for a Faculty without departments.

4. Guidelines for Instructors in Interpreting and Reporting Results

4.1 Interpretation for Improvement and Development

When using course evaluations to improve and enhance your teaching, numerical results are most useful for identifying strengths and weaknesses while comments provide insights for reflection. When looking at the results it is important to consider the following:

1. Ratings of **global items are the most useful as indicators of overall instructional effectiveness** (e.g., "Overall this instructor is an excellent teacher"; "I learned a great deal from this course"). Responses to these questions are found to correlate most consistently with measures of actual student achievement. Generally, results below 3.5 should be of concern, while 3.5 to 4 represent solid results, and mean scores over 4 are considered strong. As well, it is advisable to follow-up on any result that is more than .5 below or above the comparison mean (department, Faculty by level or class size).
2. **The mean is not sufficient to provide a picture of the distribution of responses.** When interpreting the numerical results, consider information such as the distribution of responses by item as well as the variation in responses. To understand the range of opinion, one should interpret the mean in conjunction with the shape and frequency of responses along the scale. Generally, differences that are less than .5 above or below the mean should be regarded as functionally equivalent.
3. **The standard deviation provides important additional information about the variability of student responses.** A standard deviation for a question greater than 1 indicates relatively high differences of opinion; in such cases, comments can be particularly useful to help understand the variation.
4. Mercury **results are reported to only 1 decimal place** to avoid overemphasis on differences that are not meaningful. If follow-up analyses are carried out on the data, do not look beyond 1 decimal place. As discussed in section 7.3, factors that have a statistically significant impact on course ratings do not usually result in meaningful differences.
5. **Written comments provide the most useful information for teaching improvement** because they can provide insight into why some students had difficulty learning or, conversely, why others succeeded. Written comments often help clarify and illuminate some of the observed numerical response patterns. (See the section below on Interpreting Students' Written Comments.)
6. **Course ratings are most useful in improving teaching effectiveness when they are coupled with appropriate consultation.** To help derive the most benefit from your results, we encourage you to discuss them with a trusted colleague, your academic unit head or someone from *Teaching and Learning Services* (TLS).

4.2 Reporting numerical results

Course evaluation results are reported in different contexts for different purposes. They are used in reappointment, tenure and promotion,[6] in the annual merit review and in nominations for teaching awards. They can also be used in teaching improvement consultations. Depending on the context, the focus may be on a specific course, a specific term or patterns over time. The following are examples that may be helpful models.

4.2.1 Reporting of Results for a Specific Course

Table 1 is an example of results for a specific course and semester presented in the context of a teaching portfolio. As illustrated, the results from a specific course can be compared to various subsets. The standard deviation[7] is included because it indicates the degree of agreement among the students. Note that the questions reported are listed below the table.

Question #	# of Respondents N=58; (%)	Course Mean (s.d.)*	Section (n=7) Average Mean (s.d.)	Average Mean for Faculty: Course Level** (s.d.)	Average Mean for Faculty: Class Size*** (s.d.)	Department Average Mean**** (s.d.)
1	26 (44.8%)	3.8 (1.0)	3.8 (0.8)	3.8 (0.5)	3.9 (0.5)	3.7 (0.4)
2	26 (44.8%)	3.8 (1.1)	3.8 (0.9)	4.0 (0.4)	4.0 (0.4)	3.9 (0.3)
3	26 (44.8%)	3.7 (1.2)	3.8 (0.9)	3.9 (0.6)	4.0 (0.5)	3.9 (0.5)
4	26 (44.8%)	3.8 (1.0)	3.7 (0.9)	3.8 (0.6)	3.9 (0.5)	3.9 (0.4)
5	26 (44.8%)	4.3 (0.7)	4.0 (0.8)	-	-	4.1 (0.4)
6	26 (44.8%)	4.4 (0.8)	4.3 (0.7)	-	-	4.3 (0.3)
7	26 (44.8%)	3.9 (0.9)	4.1 (0.9)	-	-	4.1 (0.4)
8	26 (44.8%)	4.0 (1.0)	4.0 (0.9)	-	-	3.9 (0.4)
9	26 (44.8%)	2.7 (1.2)	3.4 (1.1)	-	-	3.7 (0.4)

Table 1: Faculty of Management - 300-Level Course, Fall 2009

Rating scale: 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree

Note: As there are only four questions common to all questionnaires, the averages by level and class size are calculated only for them.

* Standard deviation

**300-Level courses for the Faculty of Management for that term. This can only be calculated for the four core questions. (N=64 courses, 74 instructors)

***Class size of 31 to 100 students. (N = 156 courses, 178 instructors)

****Based on the questionnaire – Faculty of Management (BCom-core). (N = 59 courses, 64 instructors)

[6] The full Teaching Portfolio guidelines are available at: http://www.mcgill.ca/tls/teaching/portfolio/guidelines#GUIDELINES

[7] A small standard deviation means that most students provided similar responses, while a larger standard deviation means that ratings were more varied.

List of questions: (Red: core course question, Green: core instructor question, Purple: instructor question, Blue: course question)

1. Overall, this is an excellent course.
2. Overall, I learned a great deal from this course.
3. Overall, this instructor is an excellent teacher.
4. Overall, I learned a great deal from this instructor.
5. Overall, the instructor responded to students' questions with clarity and expertise.
6. The instructor was organized and well prepared for each class.
7. The instructor was available for student consultation.
8. The instructor's teaching methods (the skills and effectiveness of the instructor, the style of the course, the kinds of assignments given, the encouragement of class participation, etc.) were effective and appropriate.
9. The evaluation methods used in this course were fair and appropriate.

Different forms of visual representation can aid in understanding the results; below is an example of course means compared to different groups using a column chart.

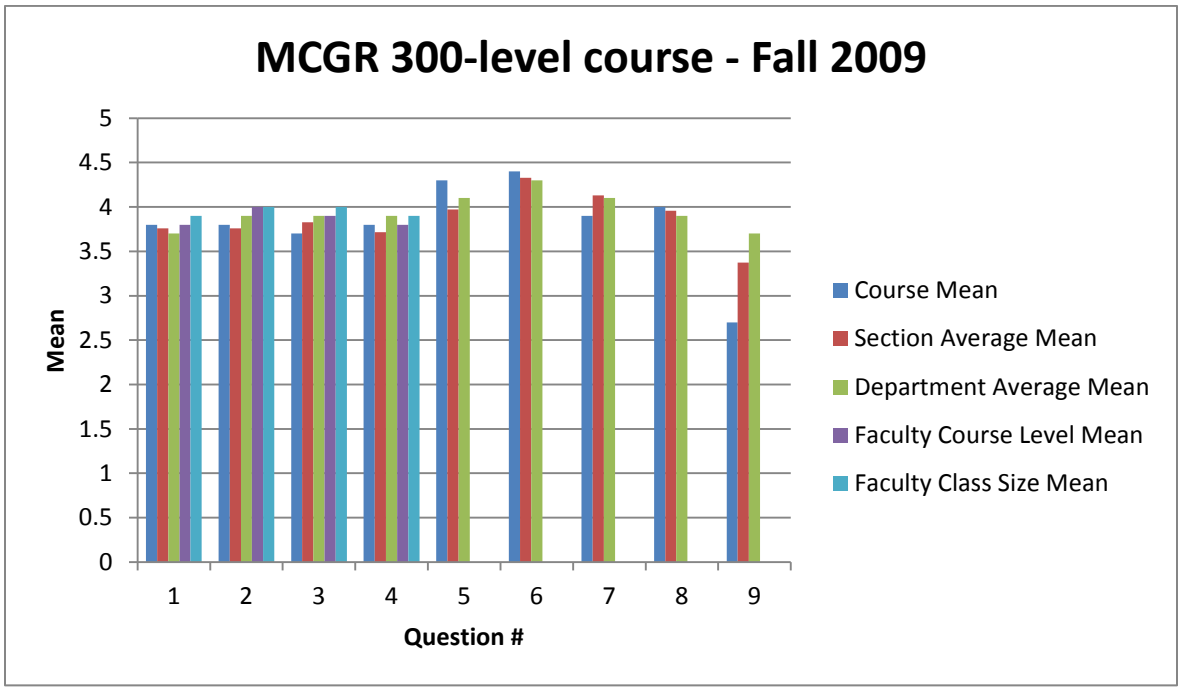


Figure 1: Faculty of Management - 300-Level Course

An Instructor should discuss any results that appear lower or higher than those of comparison groups by a meaningful amount. While there are small differences for the first eight questions between the means of the course and the comparison groups, they are minor and provide little useful information. As a guideline, differences of ±0.5 are generally not meaningful. However, Q9 shows a difference that should be examined for possible contributing factors. The question, "The evaluation methods used in this course were fair and appropriate" could be influenced, for example, by difficulties arising from a lack of coordination among the instructors teaching the different sections. Other possible reasons might be indicated in students' comments for this question.

4.2.2 Reporting of Results for a Specific Course over Time

In order to show course and teaching improvement over time, gathering data from previous course offerings is important. The table below shows results for a course over a span of 4 semesters:

Question #	Fall 2008 Mean (s.d.)*	Winter 2009 Mean (s.d.)	Fall 2009 Mean (s.d.)	Winter 2010 Mean (s.d.)
1	4.5 (0.9)	4.4 (0.7)	4.1 (1.0)	4.7 (0.5)
2	4.5 (1.0)	4.5 (0.6)	4.2 (1.0)	4.8 (0.5)
3	4.2 (0.8)	4.0 (0.9)	3.8 (1.0)	4.3 (1.0)
4	4.1 (0.9)	4.1 (0.8)	3.7 (1.0)	4.1 (1.1)

Table 2: Faculty of Engineering - 200-Level Course

Rating scale: 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree

***Standard deviation**

There are many different ways to represent the numerical results; choose the one that best illustrates the points to be emphasized. Here is a simple line graph showing the results of the four core questions over the span of 4 semesters:

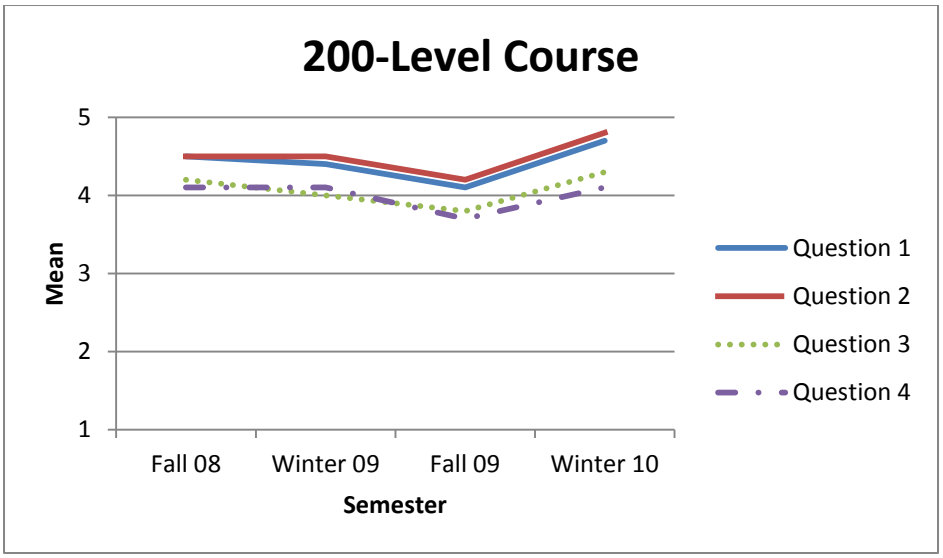


Figure 2: Faculty of Engineering –200-Level Course

Questions:

1. Overall, this is an excellent course.
2. Overall, I learned a great deal from this course.
3. This instructor is an excellent teacher.
4. Overall, I learned a great deal from this instructor.

5. Interpreting Students' Written Comments

Making sense of students' written comments can be a daunting task; however, comments provide the insights and richness that help in understanding and interpreting student feedback. Approaching the comments in a systematic way can make the process easier and more meaningful. Analyzing students' written comments helps to:

- avoid the frustration often caused by confusing or contradictory comments,
- avoid an overreaction to negative comments,
- gain insight to improve instruction, and
- distinguish areas that instructors can improve from those that should be referred to others.

In this section, we highlight key areas for consideration and outline categories that students frequently mention in their written comments[8]. To help organize and interpret students' written comments, the categories extracted from the literature have been mapped to the questions from the recommended pool of questions used at McGill.

Important Notes:

1) A common concern of instructors is students' ability to evaluate teaching reliably and competently. Research on this area dates to the 1970s[9] While students are not the best source for opinions on the instructor's knowledge of the discipline or the accuracy of the course content, there are many areas for which students are uniquely well-suited to provide an informed opinion. These include their own understanding and motivation, as well as the appropriateness and quality of the instructional and evaluation methods used in the course.

2) A common misconception is that students will only complete a course evaluation when they have had a very positive or very negative experience in the course. In fact, while students are more likely to add comments when they have strong opinions about the course, the same percentage of students complete the final evaluation regardless of overall student opinion of the course. (See section 7.2.2, Student comments.)

3) Respect the confidentiality of students who complete the course evaluations. Never assume the identity of the author of specific comments, and assume that those comments were written in good faith with the purpose of providing constructive feedback.

---

[8] The following references focus specifically on comments in student evaluations: Alhija, F. N. & Fresko, B. (2009). Student evaluation of instruction: what can be learned from students' written comments? *Studies in Educational Evaluation,* 35, 37-44; Lewis, K.G. (2001). Making sense of student written comments. *New Directions for Teaching and Learning,* 87, 25-32; Zimmaro, D.M., Gaede, C.S., Heikes, E.J., Shim, M.P. & Lewis, K.G. (2006). *A study of students' written course evaluation comments at a public university.* Austin, TX: Division of Instructional Innovation and Assessment, University of Texas at Austin.

[9] For a discussion of the conditions under which student evaluations are valid, see Scriven, M. (1995). Student ratings offer useful input to teacher evaluations: Practical Assessment. *Research & Evaluation*, 4(7), 4-5. The research is summarized in Gravestock, P. & Gregor-Greenleaf, E. (2008). *Student Course Evaluations: Research, Models and Trends.* Toronto: Higher Education Quality Council of Ontario.

Important considerations to keep in mind when interpreting student comments:

- When reading students' written comments, always balance them against the mean rating to keep them in context. Otherwise, negative comments may be given more weight than is appropriate.
- In general, student comments are strongly correlated with all quantitative measures on the course evaluation.
- Look for repeated patterns in the comments which can be useful to identify issues that are of importance to students as a whole. However, do not dismiss a comment out of hand if it comes from only one student.
- Positive comments tend to be more general in nature, whereas negative comments tend to focus on a particular aspect of a course.
- Comments on items such as scheduling, class length, timing and frequency or class composition tend to be critical. These items should be discussed with the academic unit head.
- After reading through your students' written comments, make an initial assessment. Overall, were the comments positive or negative regarding the course or your instruction?

Comments other than general ones usually fall into one of four main categories of topics:

1. Instructor influence
   Comments concerning clarity and difficulty, teaching strategies, course activities, assignments and assessments address areas that the instructor has the most opportunity to adjust and improve. For example, the comment: "He should present the material in a more structured and organized manner," refers to course design and organization, aspects that the instructor can address.

2. Shared instructor/students
   Comments that fall in this area often reflect a shared responsibility between the instructor and the students and include topics such as interest and communication. For example a student's course experience that elicits the comment: "Some students were just not interested in learning the material of this course" may have just as much to do with their motivation to learn in this area as it has to do with the instructor's ability to facilitate an interesting course. An open discussion with students can help identify how these areas may be improved.

3. Shared instructor/administration
   Comments related to organization and structure as well as the physical environment often require a concerted reaction from the instructor and the program administrator (i.e., Dean, Chair or Director). For example, the comment: "This course should be open to students in their first year" should be brought to the attention of the administration for consideration in program design.

4. Personal traits of the instructor
   Comments about the personal traits (for example, accent or apparent unfriendliness) of the instructor understandably often elicit strong emotions. These should be reflected upon and where appropriate and possible, potential strategies should be discussed with a trusted colleague or teaching support specialist.

Page intentionally blank.

5.1 Comments Analysis Worksheet

The Worksheet is intended for instructors and teaching assistants to use to make sense of student comments. Often multiple comments are related to the same category; for example, 10 students may all make comments about the assignments being unclear. This is not really 10 different comments but rather one comment 10 times. The multiple mentions give it weight, but it is only one area that needs to be addressed for improvement.

Tips for Analysis:

To facilitate organizing the comments, we have created a table which identifies the categories for the questions.

- The *Comments Analysis Worksheet*[10] helps organize student comments and make sense of the written data. The worksheet has been organized alphabetically in sections according to most frequently commented categories.
- Note any student comments that will help in interpretation.
- Indicate positive and negative comments.
- Record the frequency of comments surrounding each theme to help identify the areas where students felt most strongly.
- Add any personal notes that will help in the process of building on the feedback received.

Comments should be tracked according to the category(ies) they relate to and whether they are positive or negative. Note that one comment may contain multiple points related to different comment categories. Any comments that are particularly insightful or constructive should be noted.

[10] The Worksheet is available as a Word document for download at: http://www.mcgill.ca/tls/teaching/course-evaluations/interpretation.

Comments Analysis Worksheet					
Comment Category	**Sample Positive Student Comments**	**Total +**	**Sample Negative Student Comments**	**Total -**	**Personal notes**
Overall (Course or Instructor)					
Clarity & Difficulty					
Organization & Structure					
Interest					
Teaching Strategies					
Assessment & Feedback					
Outside of Classroom Communication					
Personal Traits					
Physical Environment					

6. Guidelines for Academic Unit Heads in Interpreting Results

Normally, academic unit heads should review the course evaluation results each term, especially for pre-tenure professors. Doing so on a regular basis will provide an ongoing picture of teaching strengths and areas needing attention. It is often tempting to look at the means and identify ratings below the mean as problematic. However, this can lead to oversimplified results that mask useful information. In addition to the guidelines for instructors, the following points should be kept in mind by unit heads when reviewing results:

1. **The mean is not sufficient to provide a picture of the distribution of responses.** When interpreting the numerical results, consider information such as the distribution of responses by item as well as the variation in responses. To understand the range of opinion, interpret the mean in conjunction with the shape and frequency of responses along the scale. Generally, differences that are less than .5 above or below the comparison mean (department, Faculty level or class size) should be regarded as functionally equivalent.

2. **The standard deviation provides important additional information about the variability of student responses.** A standard deviation for a question greater than 1 indicates relatively high differences of opinion; in such cases, comments can be particularly useful to help understand the variation.

3. Mercury **results are reported to only 1 decimal place** to avoid overemphasis on differences that are not meaningful. If follow-up analyses are carried out on the data, do not look beyond 1 decimal place. Small differences that are statistically significant are common with large sample sizes. As a result, it is important to ask whether the difference is large enough to have some practical implication. For example, if two instructors in a department receive average ratings of 4.7 and 4.8 on the question "*Overall this instructor is an excellent teacher*"; it would be difficult to argue that the difference of 0.1, although statistically significant, is large enough to claim that the instructor with a rating of 4.8 is a better teacher.

4. To form a comprehensive and meaningful understanding of an individual's teaching, it is important to **consider the entire pattern of results** from a number of different courses and classes of students over a period of time. Research suggests that data should be reviewed from courses from every term for at least two years, totaling at least five courses.

5. Ratings of **global items are the most useful as indicators of overall instructional effectiveness** (e.g., "Overall this instructor is an excellent teacher"; "I learned a great deal from this course"). Responses to these questions are found to correlate most consistently with measures of actual student achievement. How an individual instructor compares to the departmental mean is actually less important than the overall rating. It is reasonable to strive to have all instructors and courses obtaining solid or strong results, but not everyone can be above average. Generally, mean scores over 4 are considered strong, means from 3.5 to 4 represent solid results, and results below 3.5 should be of concern. As well, it is advisable to follow-up on any result that is more than .5 below or above the comparison mean.

6. **Avoid ranking instructors from "best" to "worst"** based on the course ratings, as even with a group of excellent instructors, rankings will present strong instructors as being "below average." Rankings by default diminish the accomplishments of some instructors, even if all are excellent teachers. An instructor who may be below the unit mean in a unit with a strong teaching culture may still be teaching well.

7. If comparisons are made, **the comparison group should be identified so as to be meaningful**. In many academic units, there are different forms used for different types of courses; e.g., labs, seminars, undergraduate vs. graduate. As the means are given by questionnaire, there is a subset automatically generated of those courses using the same questionnaire. When a course is a section of a multi-section offering, the other sections of the course are a good starting point. As well, courses of the same level within the Faculty can help contextualize results. Note that when comparisons are made, it is important to safeguard the confidentiality of results of the comparison group.

8. **Written comments provide extremely useful insights** for formative purposes. The *Comments Analysis Worksheet* (p. 13) is a tool to summarize comments to facilitate analysis.

    - Focus on specific, descriptive items and look for patterns. Specific teaching behaviors (e.g., clarity of objectives) are easier to change than personal characteristics (e.g., enthusiasm).

    - Consider ratings in relation to written comments to see if the latter provide indicators and suggestions for improvement.

    - If there is considerable variation in response on an item (e.g., some report assignments as appropriately challenging and others as too challenging.), it may represent important differences in the nature of the students, e.g., senior versus first year, or an uneven distribution of background preparation for the course.

7. Interpretation of Numerical Results

The quality of the data is key for any interpretation of numerical results from course evaluations to be meaningful. Data quality is influenced by several factors: the reliability of the instrument, the representativeness of the group of students providing the feedback, as well as a consideration of contextual factors. Therefore, before discussing how to interpret the numbers themselves, it is important to look at these aspects.

7.1 Reliability concerns

If the questions that are asked do not result in consistent results, their utility is severely limited. The reliability of the measurement instrument has two aspects: 1) test-retest consistency, i.e., whether the questions result in similar results over time; and 2) reliability of the scale, i.e., whether the questions meant to measure the same construct do so, often referred to as internal consistency.

Analyses were conducted at McGill of data from fall 2008 and fall 2009 on the four core questions[11] and the data showed consistent results in both terms across all Faculties. The reliability of the scale was measured with Cronbach's alpha, a statistic that ranges from 0 (unacceptable) to 1 (excellent) that is based on the correlations between different items of the same scale. Cronbach's alpha was consistently in the range of .9, which is considered excellent, when analysed by course level (100s, 200s, etc.), class size and Faculty.

7.2 Response Rates and Sample Representativeness

7.2.1 Response rates

A primary concern of instructors is the percentage of students in their classes who complete the course evaluations. It is common for response rates in an online system to be lower than in a paper-based system. Although generally more responses are obtained from in-class administration, high response rates alone do not ensure the validity of the results, nor do lower response rates necessarily mean that the responses are not representative of the class. Instructors and administrators often focus most of their concern on the response rates per se, whereas the concern is more appropriately focused on the representativeness of the respondents. Research at McGill and other institutions has shown no evidence of bias in the data resulting from smaller samples. In other words, those completing online questionnaires tend to be representative of the class as a whole.

[11] Q1: Overall, this is an excellent course. Q2: Overall, I learned a great deal from this course. Q3: Overall, this instructor is an excellent teacher. Q4: Overall, I learned a great deal from this instructor.

Desired response rates: Response rates vary according to the method of administration, with an average response rate of 30% for online surveys considered acceptable compared to approximately 50% for classroom administration.[12]

At McGill, there is a sliding scale of response rates required for results to be available to the community (see Table 3). This scale gives an indication of the reliability of the responses; for class sizes greater than 30, the desired response rates conform to or exceed recommendations in the literature.[13] For courses with enrolments of less than 30, a higher response rate would be required for similar reliability; therefore, responses from smaller courses with a response rate of less than 75% should be interpreted with particular care.

Class size	McGill Desired Response Rates (%)
5-11	minimum 5 responses
12-30	at least 40%
31-100	at least 35%
101-200	at least 30%
201-1000	at least 25%

Table 3. Recommended Response Rates [14]

In all cases, although especially for small classes, patterns of responses over time are more informative than results from a single assessment. Single results can provide indicators of areas of potential strengths or weaknesses, but patterns of results are needed to appropriately infer trends or come to reasonable assessments.

7.2.2 Sample representativeness

To confirm and extend results from the literature, analyses of McGill data from the 2008-09 and 2009-10 academic years were conducted to address the following questions:

1) Are there systematic differences on any demographic variables[15] between those who complete at least one evaluation and those who complete no evaluations?
2) Are stronger students (defined by CGPA) more or less likely to participate in the evaluation process than weaker students?

Demographic variables: Our analyses showed that no demographic group is disenfranchised by the online system as none of the examined characteristics distinguished between those students who participated in course evaluations and those who did not.

[12] http://www.utexas.edu/academic/ctl/assessment/iar/teaching/gather/method/survey-Response.php

[13] These response rates provide a confidence level of 80%; see 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.

[14] http://www.mcgill.ca/tls/teaching/course-evaluations/about/policy

[15] The categories were: Gender, Faculty, Mother Tongue, Curriculum Year, Program Load, Admit Region, Admit Category, Admit Type Group; Immigration Group and Citizenship Region.

Academic Performance: Instructors often express concern that students not doing well in courses (and therefore more likely to be disgruntled) are more likely to participate in the evaluation process. However, analysis of McGill data from fall 2009 revealed the inverse to be true; indeed, students with stronger academic performance were more likely to participate. The following academic units were chosen for analysis as they provide a cross-section of disciplines and teaching contexts: the departments of English, History and Classics, and Political Science (Faculty of Arts); the Faculty of Education; the Desautels School of Management; and Biology and Physics (Faculty of Science).

In all cases there was a statistically significant difference ($p<.001$) between those who participated and those who did not with the participants having higher grades. Participating students had a mean CGPA between 3.15 (sd, 0.45) and 3.46 (sd, 0.44); the mean CGPA for those who did not participate ranged from 2.94 (sd 0.59) to 3.14 (sd, 0.65). (See Appendix A.)

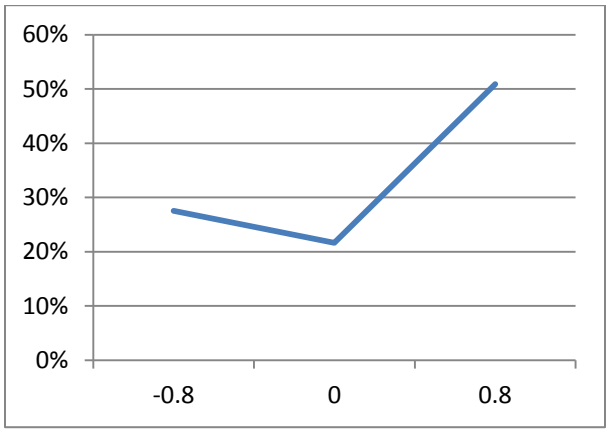
In summary then, there are no demographic characteristics distinguishing those who participate in evaluations from those who do not although students who complete course evaluations tend to have higher grades than those who do not.

Variations in response rates: Further analysis on response rates showed important variations within Faculties suggesting that students are making choices about which courses to evaluate. Analyses of fall 2009 data from the Faculties of Agricultural and Environmental Sciences, Education and Engineering indicate that neither instructor rank, class size nor course level explained the variation. Further research will be required to learn from students what influences their decision to complete a course evaluation questionnaire for a specific course (see Appendix B) .

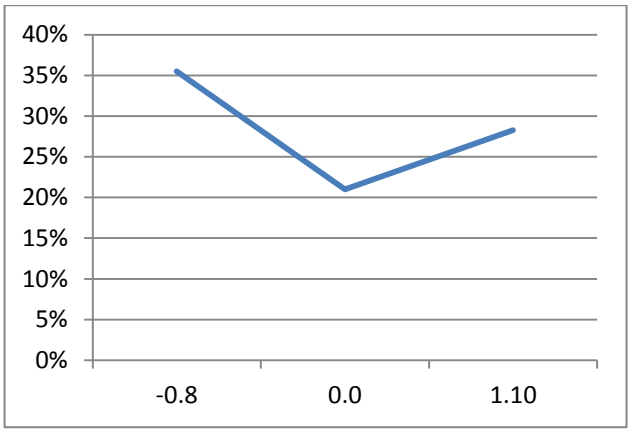
Response rates as indicators of ratings: A widespread perception is that students only complete course evaluations when they have extreme views. This would be seen in the distribution of response rates and ratings. If it were the case, one would expect to see bimodal distributions; in other words, courses with high and low ratings (as indicated by answers to the core questions) would receive higher response rates than those with average ratings. In the analysis of variation in response rates (see Appendix C), no correlation was found between the response rate and the rating.

Student comments: Students' comments on course evaluations may contribute to the persistent impression that students with extreme views fill out course evaluations. To see if the rating influenced the number of comments, courses in the departments of English, History and Classics, Political Science, and Physics for the 2010-2011 academic year were analysed (see Figure 3). The percentage of submissions with comments was identified and then plotted according to the mean rating on Q1: "Overall, this is an excellent course". The courses at the mean consistently produce the fewest comments with courses at either extreme receiving more comments. This means that if a student had an incredibly positive experience in Course A and an average experience in Course B, the student will be more likely to add comments for Course A and may not make the effort or feel the need to add written comments on the final evaluation for Course B. However, if the student completes the evaluation for Course A, he or she will be equally likely to complete the course evaluation for Course B.

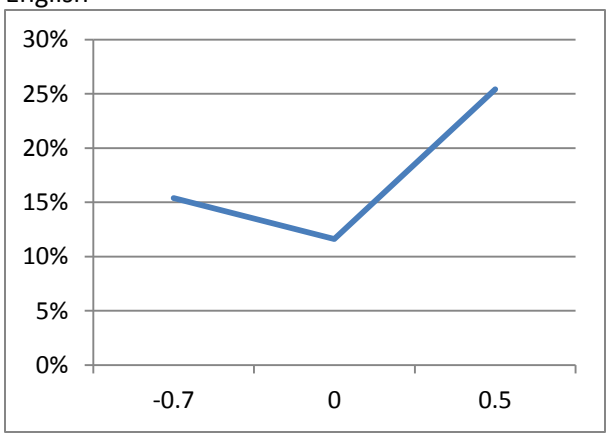
As reading the comments creates a strong impression of a course, the perception that students only complete evaluations when they have strong feelings is not surprising; extreme feelings are in fact more likely to provoke comments but they do not influence the probability of students' responding.



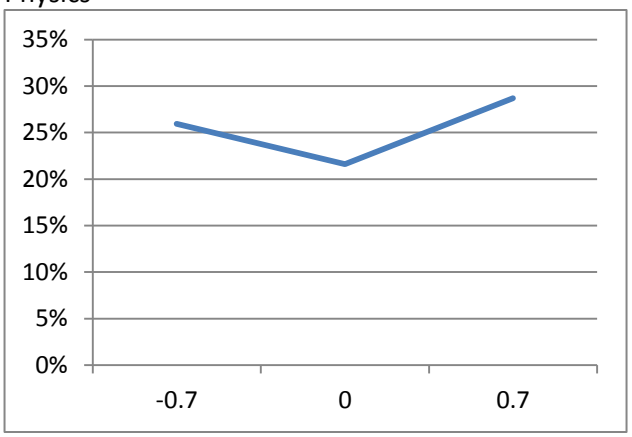
English



Physics



Political Science



History and Classics

Figure 3: % of Q1 responses with comments grouped by below the mean, at the mean and above the mean

It is important not to overinterpret response rates: a low response rate means that there was a low response rate and nothing else should be inferred.

7.3 Factors Influencing Course Evaluation Results

A number of factors have been mentioned in the literature as influencing course evaluation results; the most common are described below[16]. It is very important to note that even when statistically significant differences have been found, the resulting effects are very small, usually less than .01%.

7.3.1 Grading Leniency Concerns

Concerns are often expressed that instructors can "buy" good ratings with high grades. The relationship between grades and evaluations is a highly researched issue with a weak correlation having been demonstrated[17]. However, the correlation between good grades and high ratings comes from the fact that "good teaching…leads to better learning, and this in turn leads to both good grades and high course ratings."[18] An extensive literature review concluded that students who feel they have learned a lot expect high grades for their efforts and in turn rate their instructors high for good teaching.[19].

7.3.2 Class Size

The fall 2011 McGill data show that courses with enrolments of between 100 and 200 are the least favorably rated (see Figure 4). This finding differs from the literature whereby courses with enrollments between 35 and 100 are least favorably rated. The difference is usually on the order of 0.1.

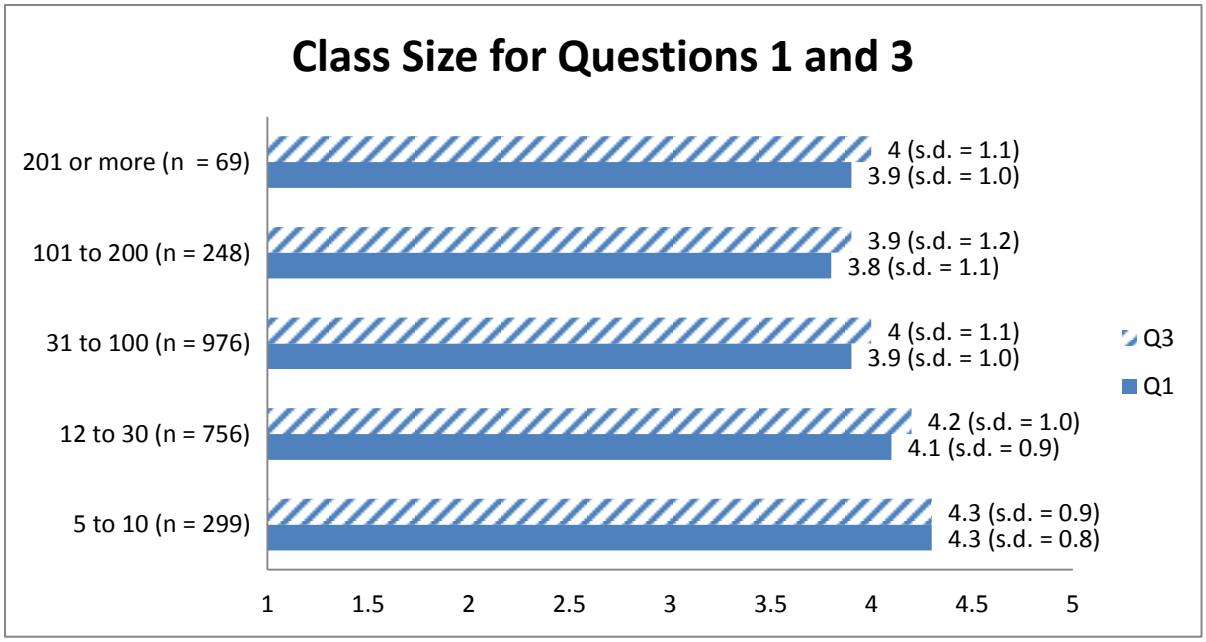


Figure 4: Mean ratings on core questions 1 and 3 by class size, fall 2009

[16] The factors were identified in the comprehensive literature review conducted by Gravestock & Gregor-Greenleaf in 2008. *Student Course Evaluations: Research, Models and Trends.* Toronto: HEQCO.

[17] Arreola, R.A. (1995) *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system.* Bolton, MA: Anker.

[18] Remedios, Richard and Lieberman, David A.(2008) 'I liked your course because you taught me well: the influence of grades, workload, expectations and goals on students' evaluations of teaching', British Educational Research Journal, 34: 1, 91 — 115, First published on: 22 September 2007 (iFirst)

[19] Gravestock, P. & Gregor-Greenleaf, E. (2008). *Student Course Evaluations: Research, Models and Trends.* Toronto: Higher Education Quality Council of Ontario.

7.3.3 Course Level

As the level of the course goes up, the ratings tend to be higher; for example, a 400-level class will usually have more positive results than a 200-level course (see Figure 5). In the same vein, graduate courses tend to be rated higher than undergraduate courses. The difference is usually on the order of 0.1.

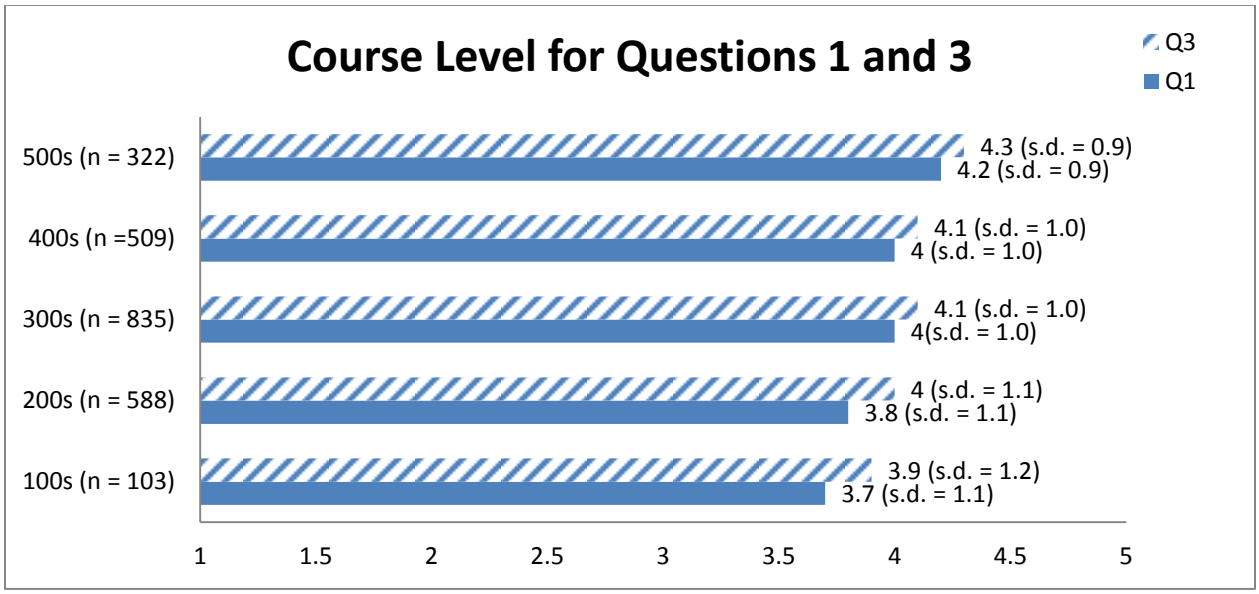


Figure 5: Mean ratings on core questions 1 and 3 by course level, fall 2009

7.3.4 General Discipline or Subject Area

Humanities, social sciences and education subject related classes usually have more positive results in course evaluations than engineering and natural science classes (see Figure 6[20]). The difference is usually on the order of 0.1. The differences within Faculties are often similar to those across the University.

7.3.5 Elective vs. Required Course

Students taking a course as an elective will frequently give it higher ratings than those taking it as a required course, although the impact is usually not significant. It is important to note that there are some courses that are required for some students but elective for others.

[20] The Faculty of Religious Studies and the McGill School of Environment were not included because of the small number of courses. The Faculty of Dentistry was combined with the Faculty of Medicine for the same reason.

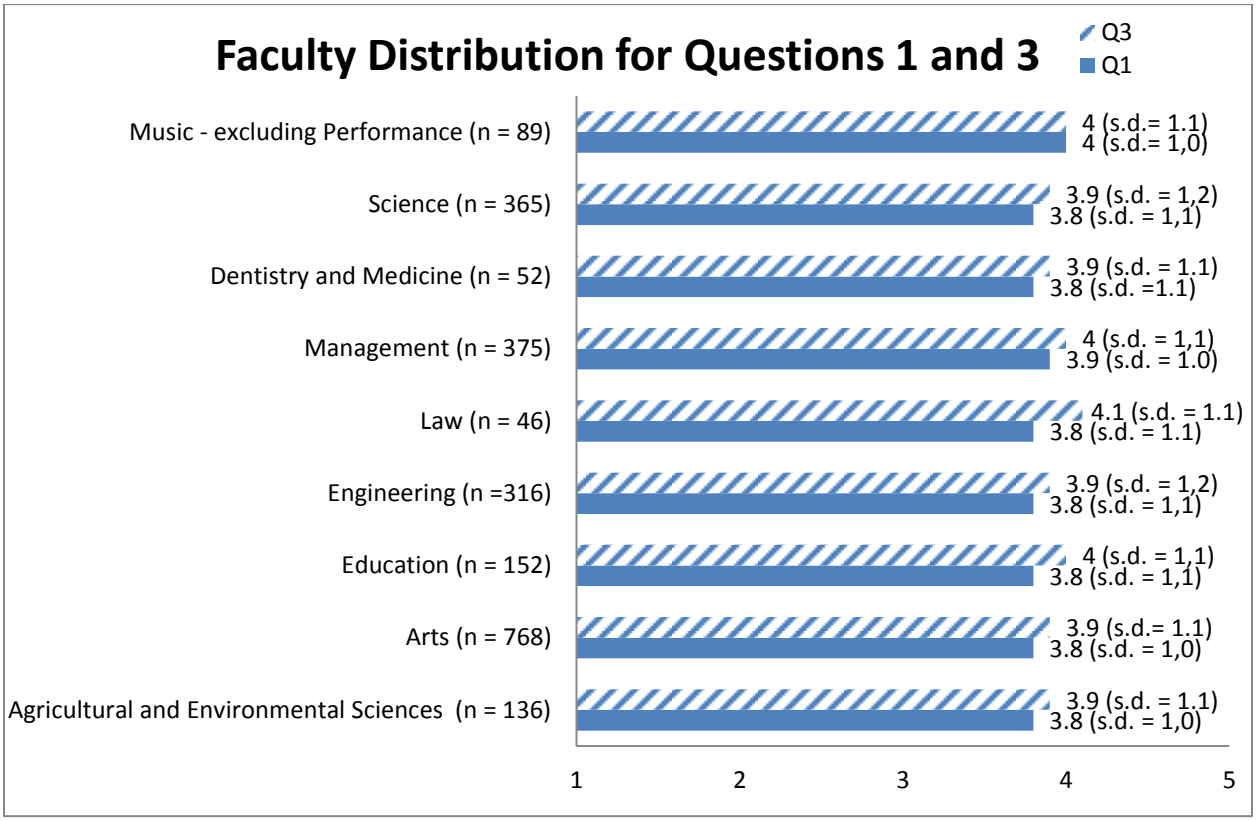


Figure 6: Mean ratings on core question 1 and core question 3 by Faculty, fall 2009

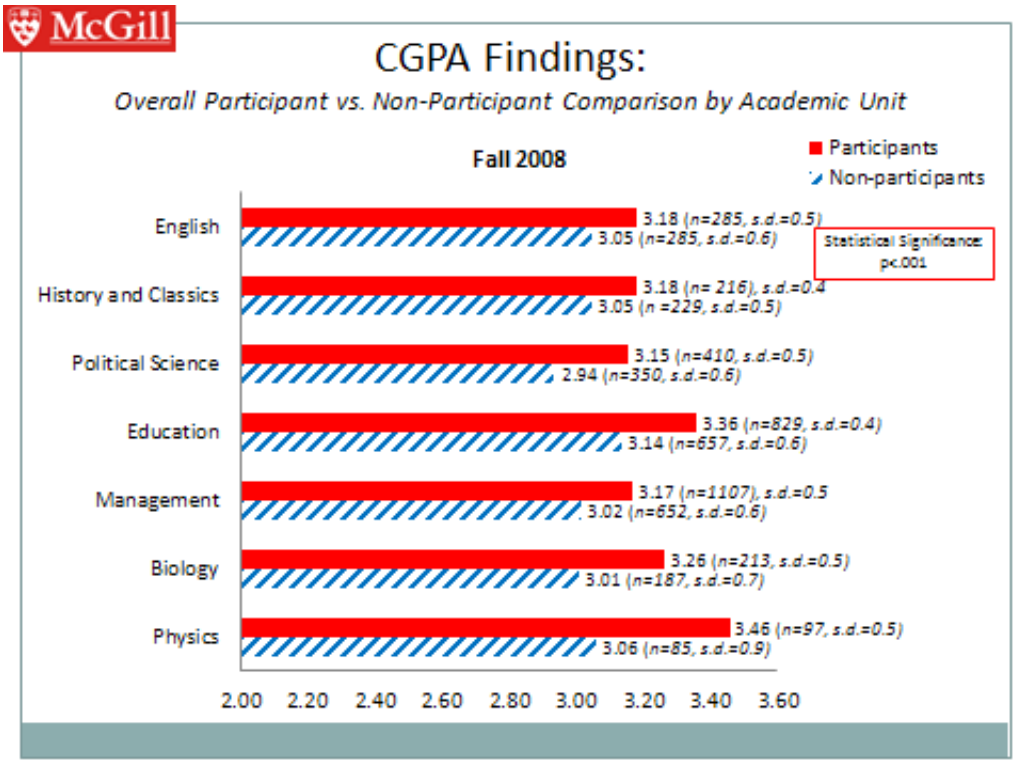
7.3.6 Timing of evaluation

Before the implementation of online evaluations, there was no real possibility of administering course evaluations after the end of classes. Previous research had shown stability in ratings even with different timing of their administration within the term. However, some instructors want to evaluate the entire course, including final examinations, so this option was examined. Some concerns were that students would not respond after the end of courses and that the ratings would be lower because of the exam experience. Results of a pilot at McGill in fall 2010 showed that there was no indication of systematically lower results for evaluations completed after the end of classes. Overall, response rates either stayed the same or increased slightly; however, many students used that time to complete the evaluations, with 40% of the responses coming after the end of classes. (See Appendix D.) The policy on end-of-course evaluations has now been revised to allow for academic units (departments, Schools or Faculties) to choose an extended dates option.

Appendix A: Comparison of course evaluation participants vs. non-participants on academic performance

Data from fall 2008 and fall 2009 were analyzed to determine if students' participation in the course evaluation process, defined as completing at least one course evaluation, was influenced by academic performance, as indicated by CGPA. Previous analyses showed that none of the demographic characteristics (Gender, Faculty, Year of Study, Admit Type, Admit Region, Admit Type Category, Immigration Group, Citizenship Region or Part-time/Full-time status) was a significant factor in influencing participation.

To answer the question about academic performance, data from seven academic units[21] were analyzed. The CGPA for the participant and non-participant groups were compared by academic unit for fall 2008 and fall 2009 and significant differences were found in all cases; students who participated tended to have higher CGPAs than students who did not. (See Figures A1 and A2.)



FigureA1: Fall 2008 CGPA comparisons for participants vs. non-participants

[21] In discussion with the Course Evaluation Advisory Group, the following academic units were chosen for follow-up analysis as they provide a cross-section of disciplines and teaching contexts: the departments of English, History & Classics and Political Science in the Faculty of Arts; the departments of Biology and Physics in the Faculty of Science; the Faculty of Education and the Desautels Faculty of Management.

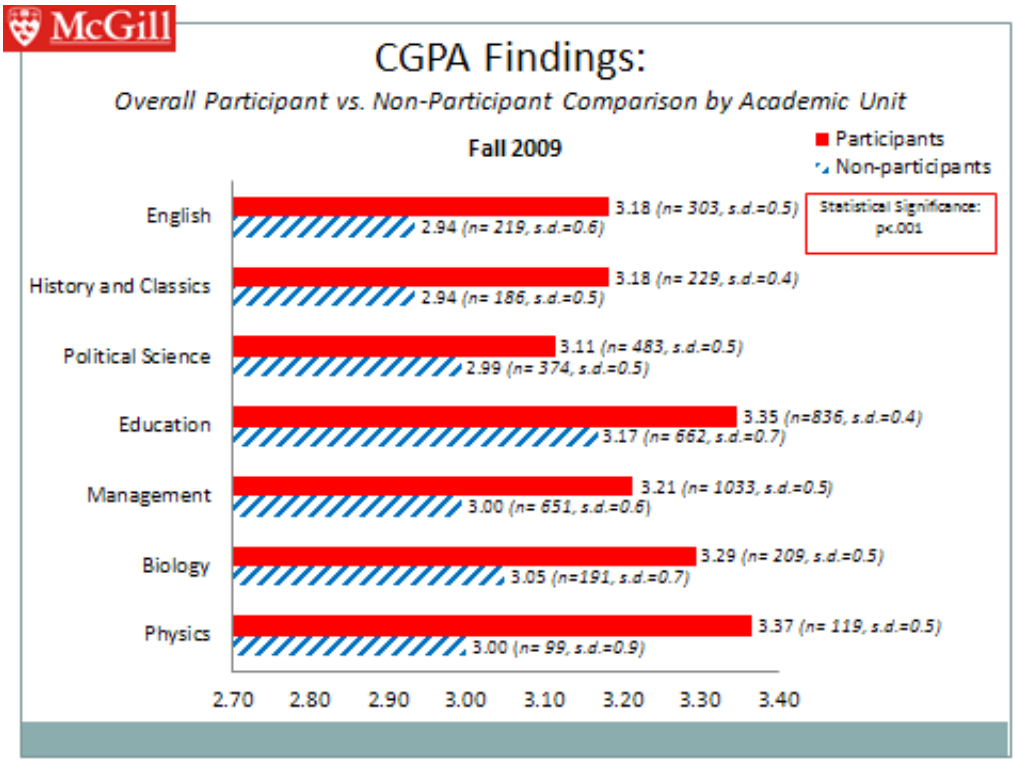


Figure A2: Fall 2009 CGPA comparisons for participants vs. non-participants

Appendix B: Response rate distributions within departments

Response rates are one measure of the success of a course evaluation system. Mean response rates by department and Faculty provide a basic measure; however, there are important variations *within* Faculties that are often larger than the differences *between* Faculties. Response rates were analyzed according to additional criteria for three Faculties, AES, Education and Engineering which had fall 2010 response rates of 53%, 50% and 43% respectively. To facilitate the analysis, response rates were grouped into four ranges: 0-24%; 25-49%; 50-74%; 75-100%.

Table B1 shows the distribution of response rates for each of the three Faculties; each had courses in the four response ranges. There were also obvious variations in distributions within Faculties according to Instructor Rank (Course Lecturer; Assistant Professor; Associate Professor; Professor) (see Tables B2, B3 and B4), Class Size (5-11; 12-30; 31-100; 101-200; 200+) (see Tables B5, B6 and B7), and Course Level (100, 200, 300, 400, 500, 600, 700) (see Tables B8, B9 and B10).

It would be premature to draw any conclusions from these preliminary analyses. However, as students within a Faculty do not complete course evaluations at a constant rate, the results suggest that response rates are subject to decisions by students that are made according to criteria that are not easy to discern. Going forward, detailed analyses will be conducted to learn more about the factors that influence the decision to complete a given evaluation.

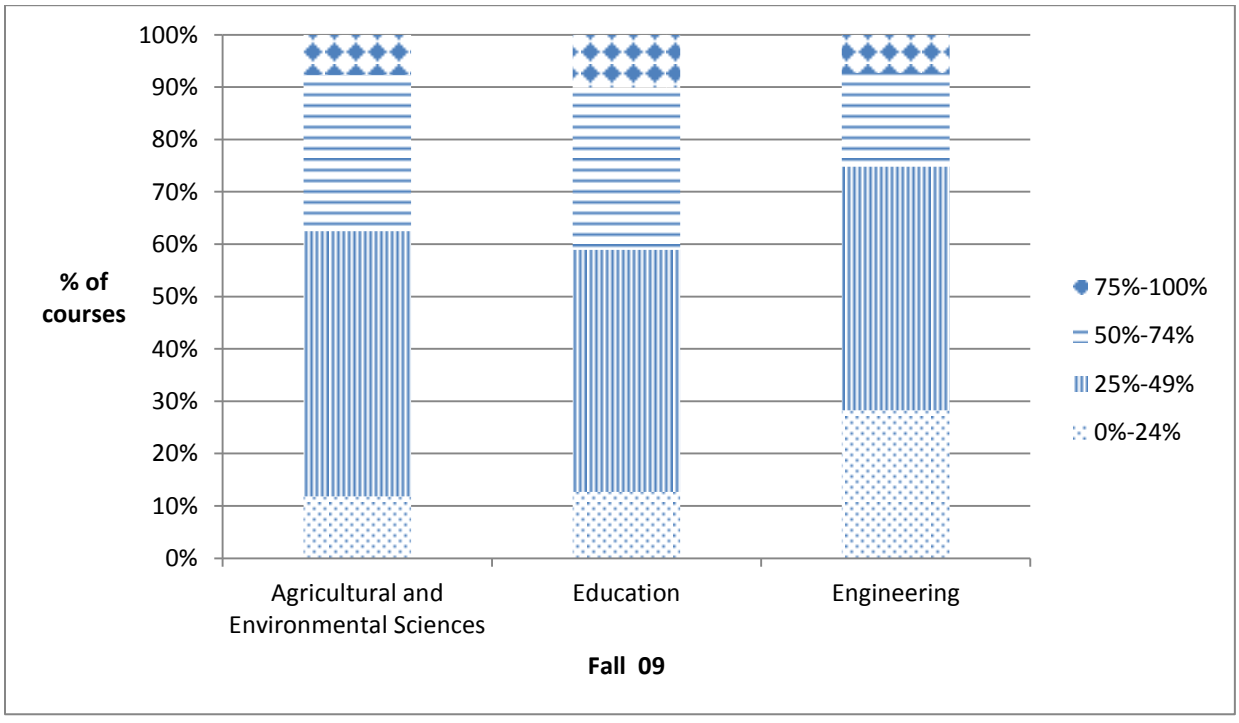


Table B1: Overall distribution of response rates

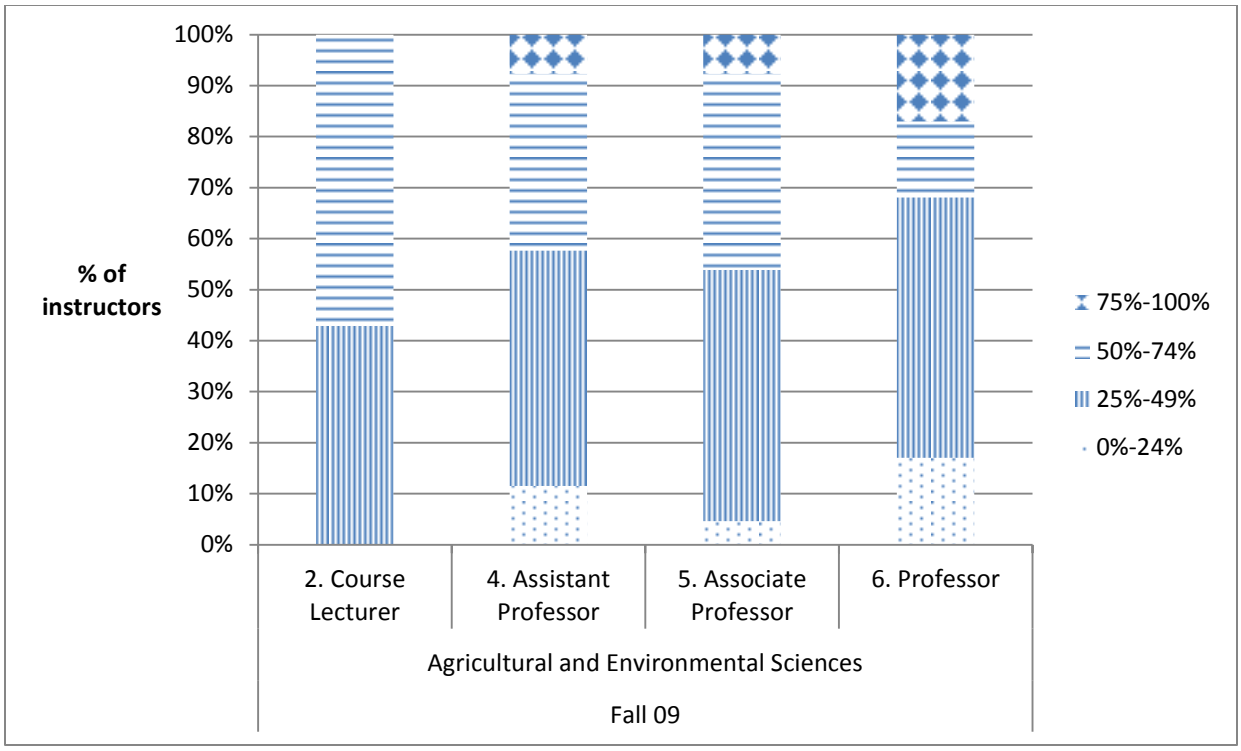


Table B2: FAES distribution of response rates by instructor rank

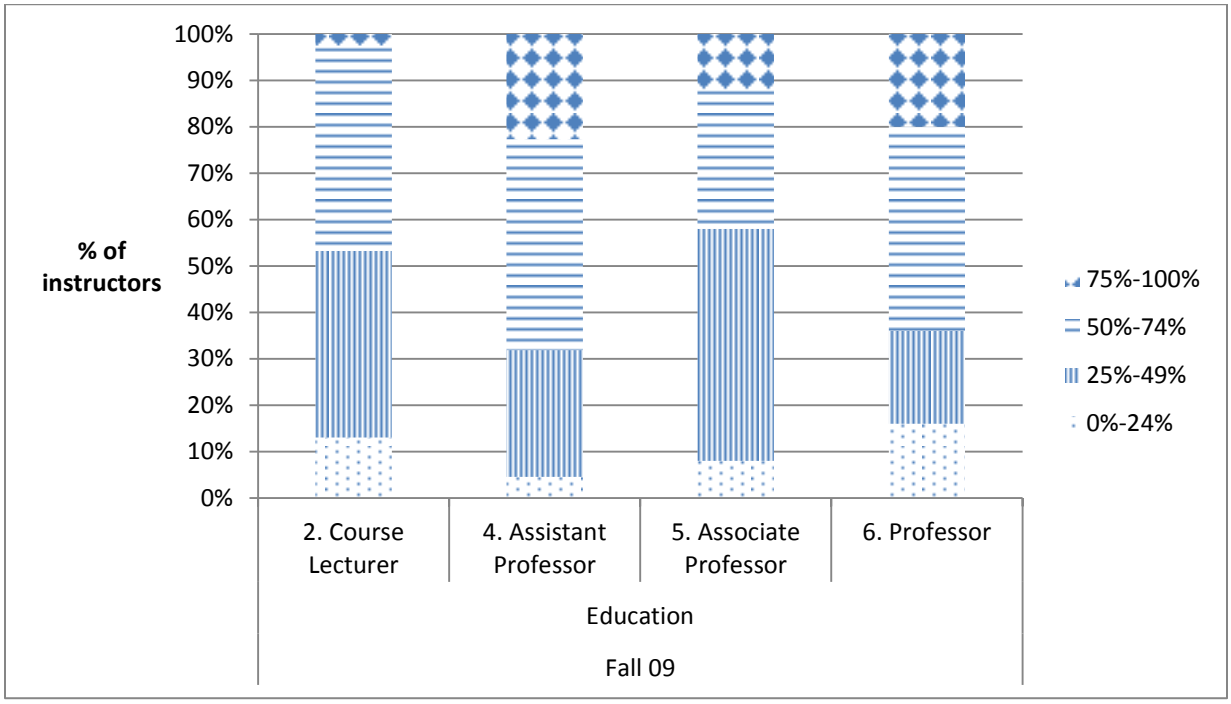


Table B3: Faculty of Education distribution of response rates by instructor rank

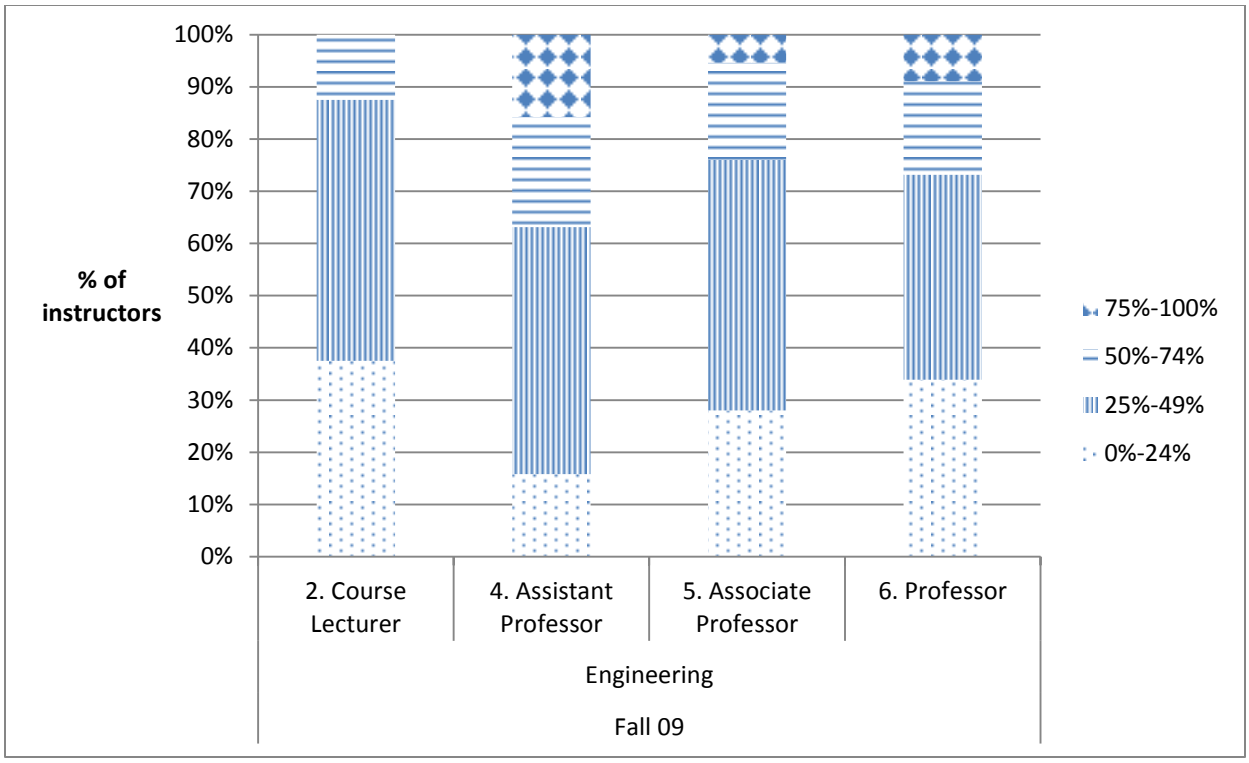


Table B4: Faculty of Engineering distribution of response rates by instructor rank

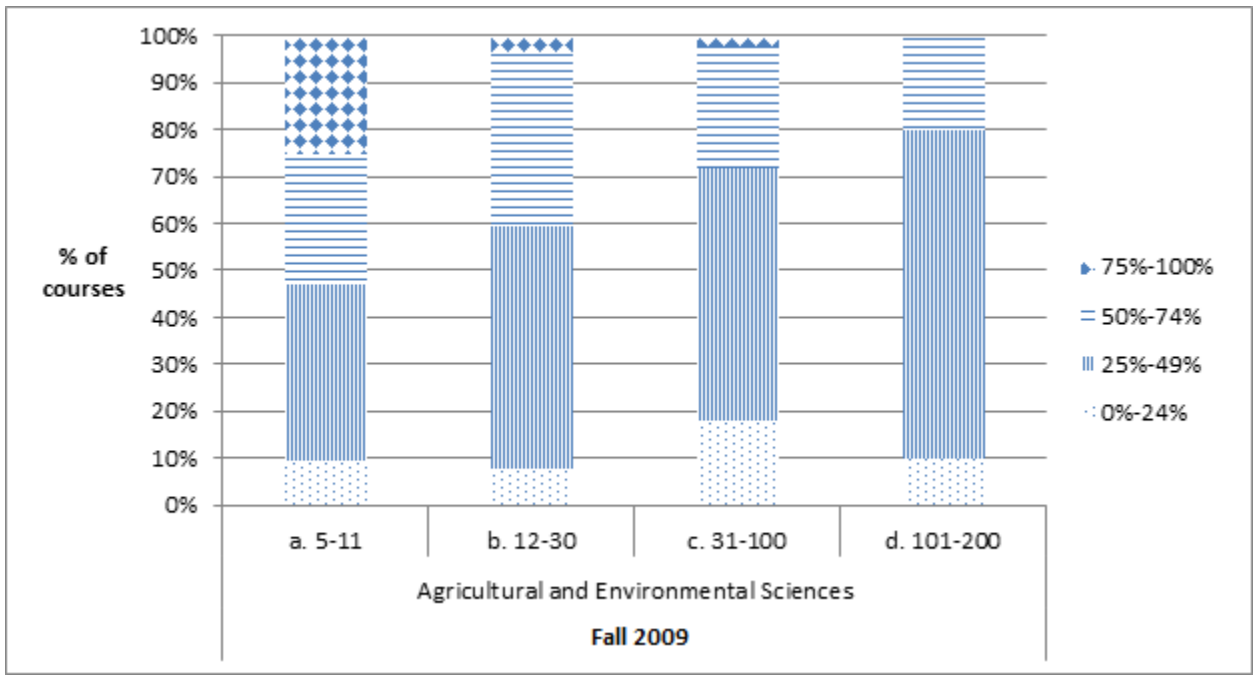


Table B5: FAES distribution of response rates by class size

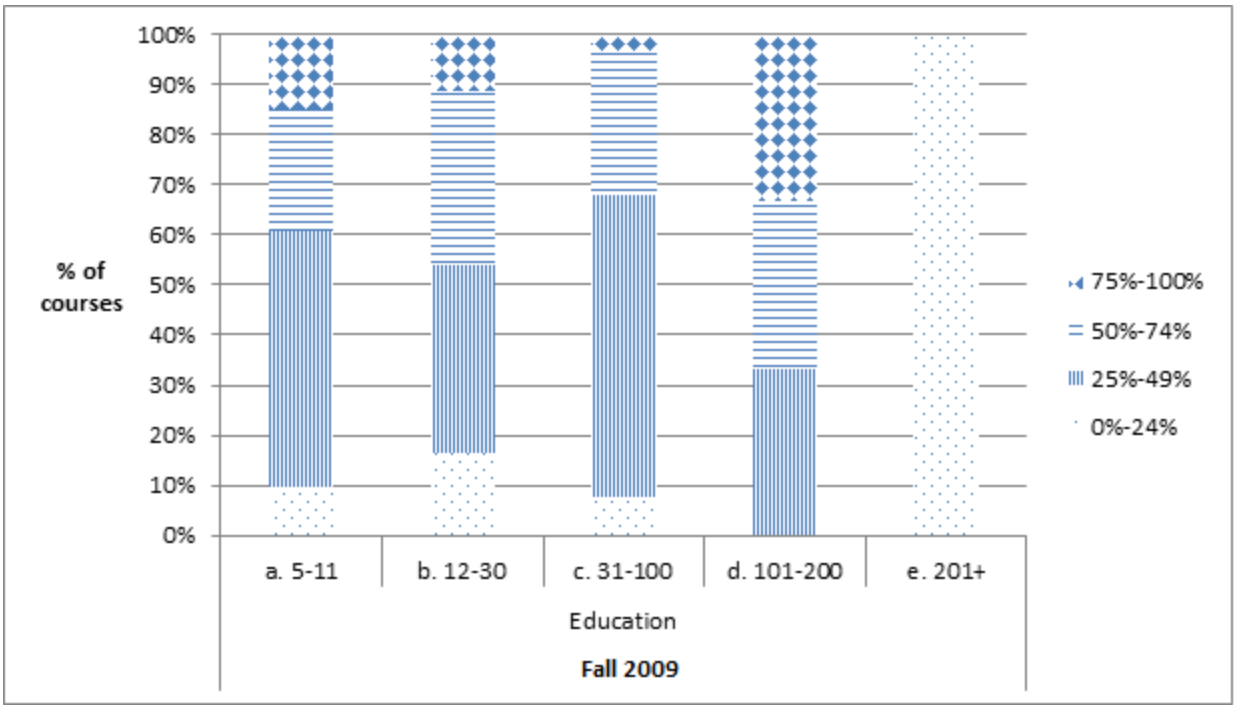


Table B6: Faculty of Education distribution of response rates by class size

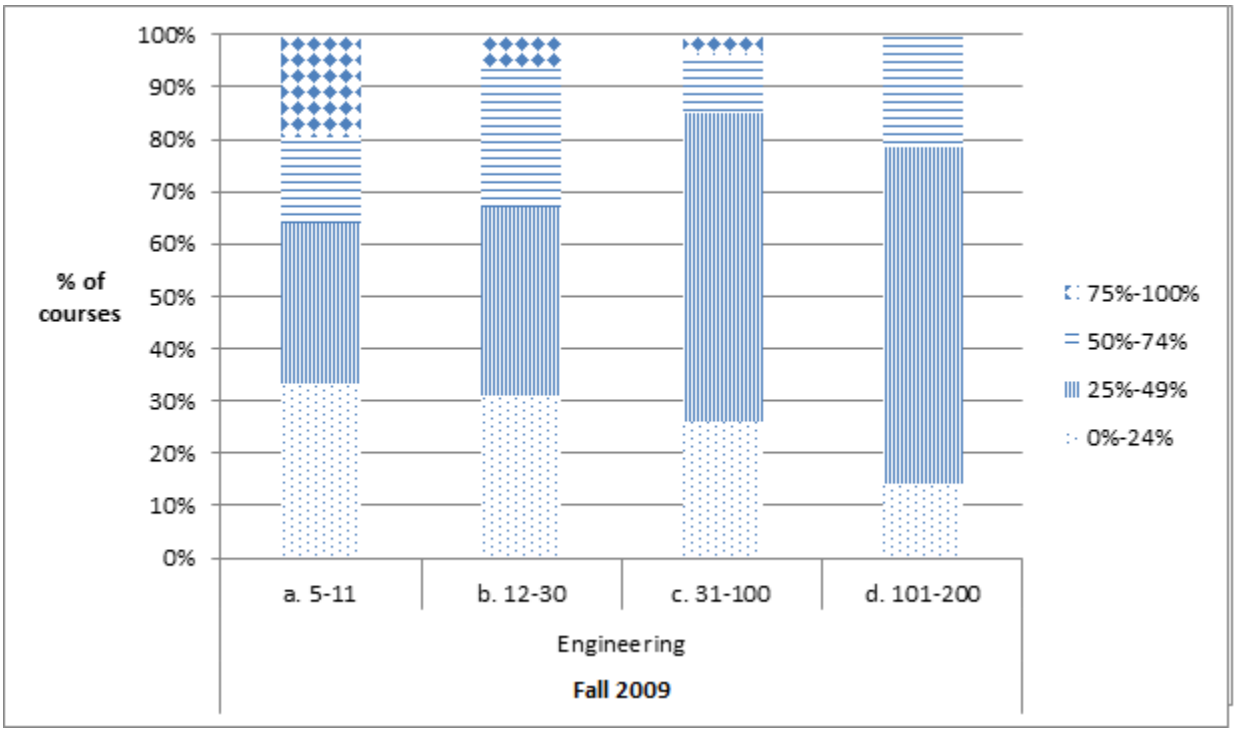


Table B7: Faculty of Engineering distribution of response rates by class size

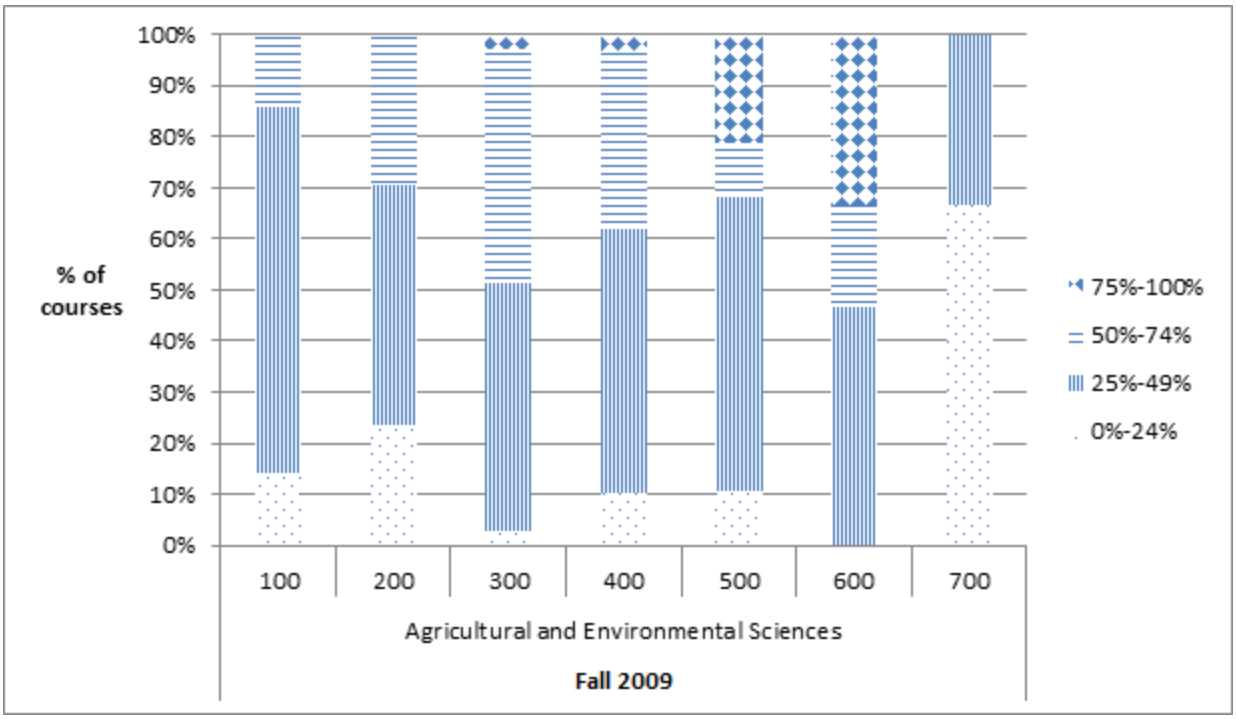


Table B8: FAES distribution of response rates by course level

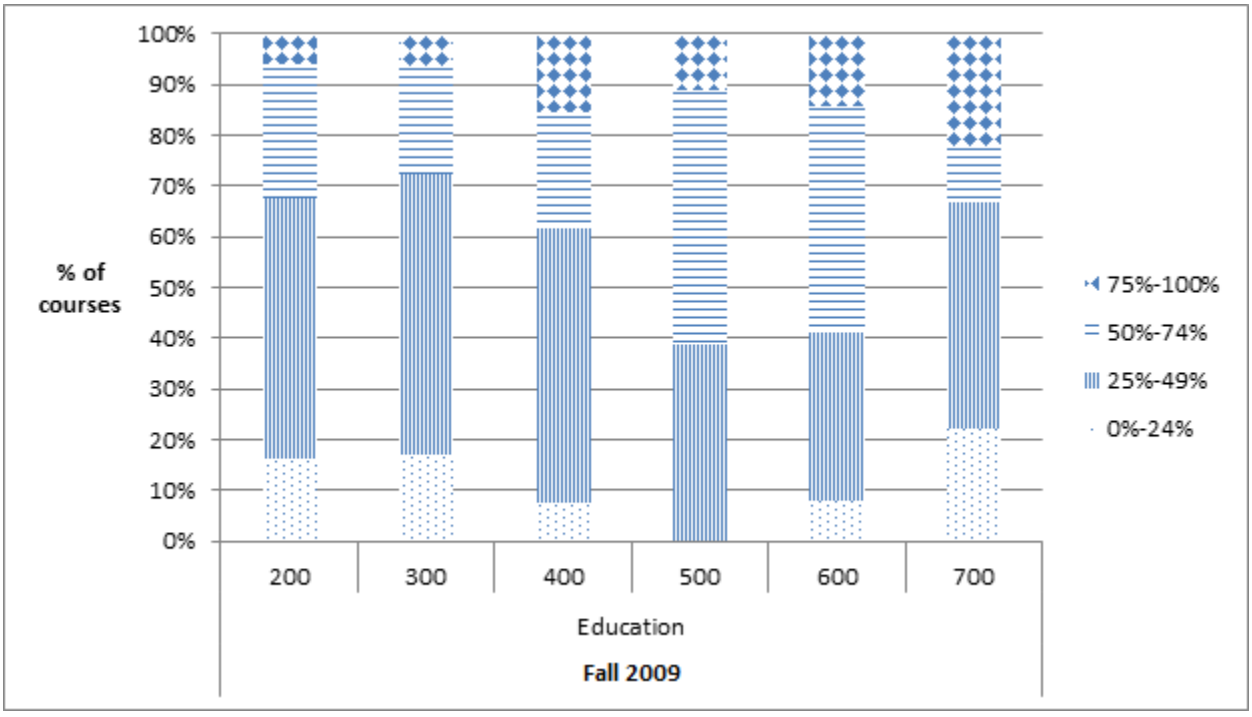


Table B9: Faculty of Education distribution of response rates by course level

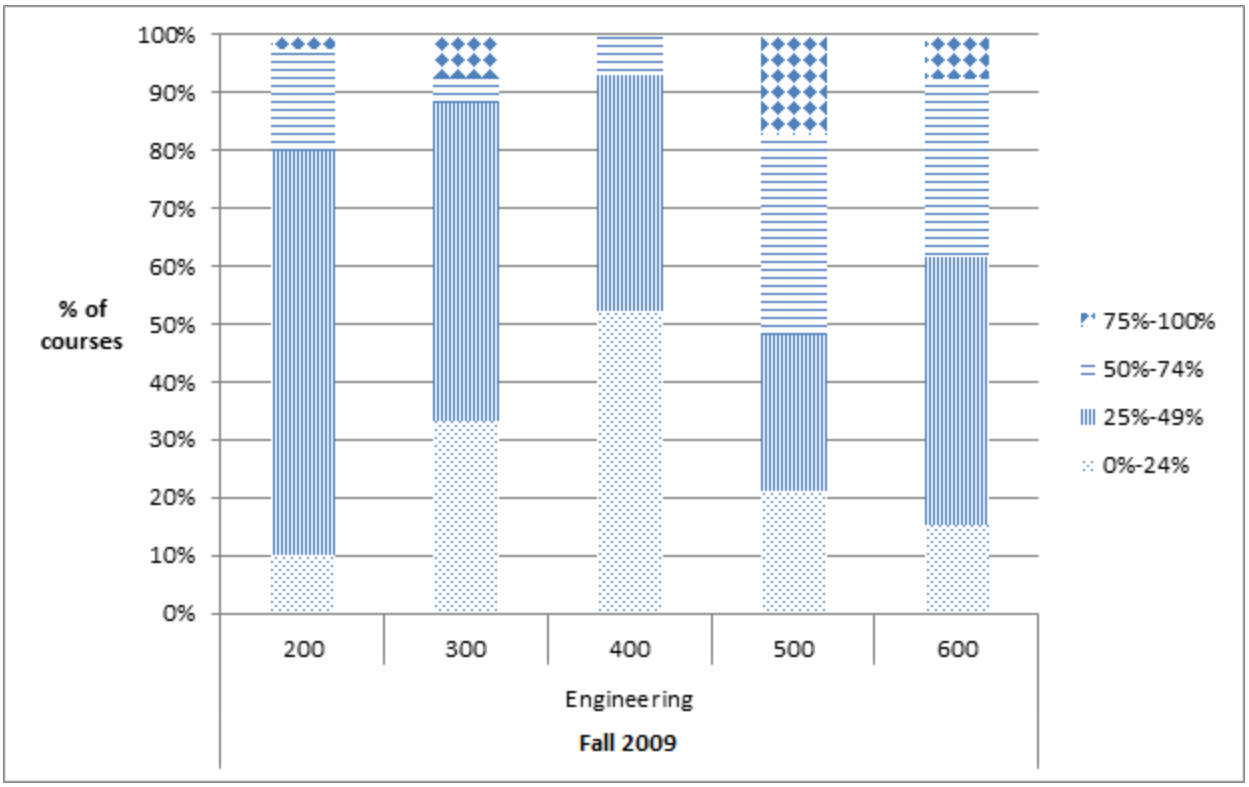


Table B10: Faculty of Engineering distribution of response rates by course level

Appendix C : Relationship between response rates and ratings

Both instructors and administrators often infer opinions from the response rates. In order to learn if a high response rate indicates a higher likelihood of a high rating, or the inverse, the correlation between response rates and ratings for two of the core questions was calculated for the Faculty of Arts. The departments within the Faculty were grouped into four categories: Humanities, Social Sciences, Foreign Languages and Area Studies.

Figures C1-C8 below show all response rates had a range of ratings, and all ratings had a range of response rates.

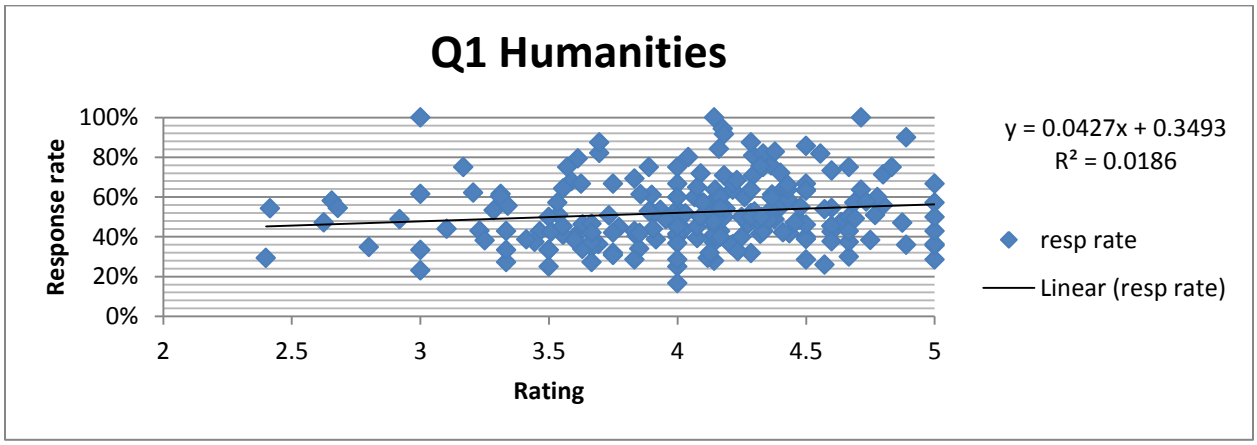


Figure C1: Correlation between response rate and rating for Q1 in Humanities

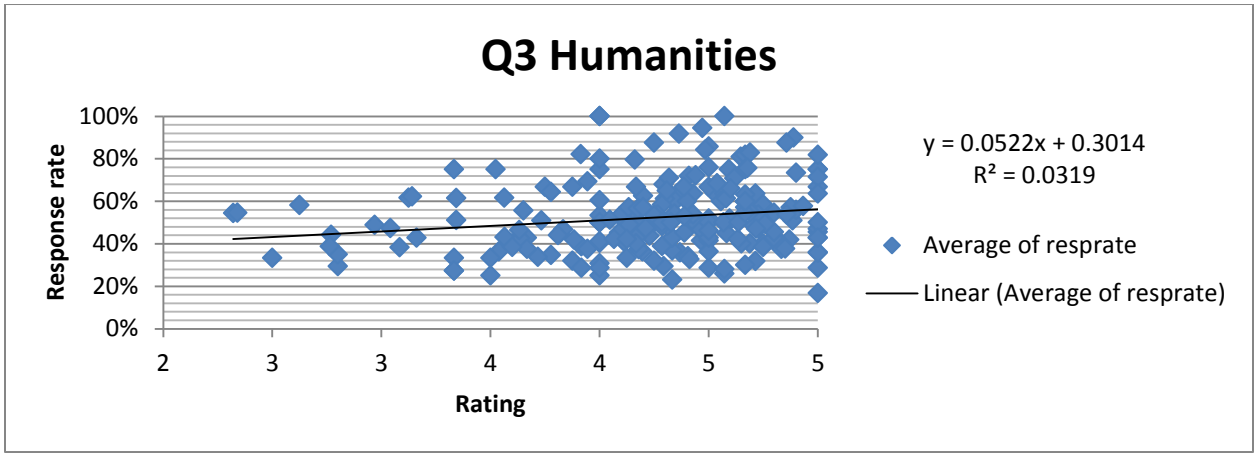


Figure C2: Correlation between response rate and rating for Q3 in Humanities

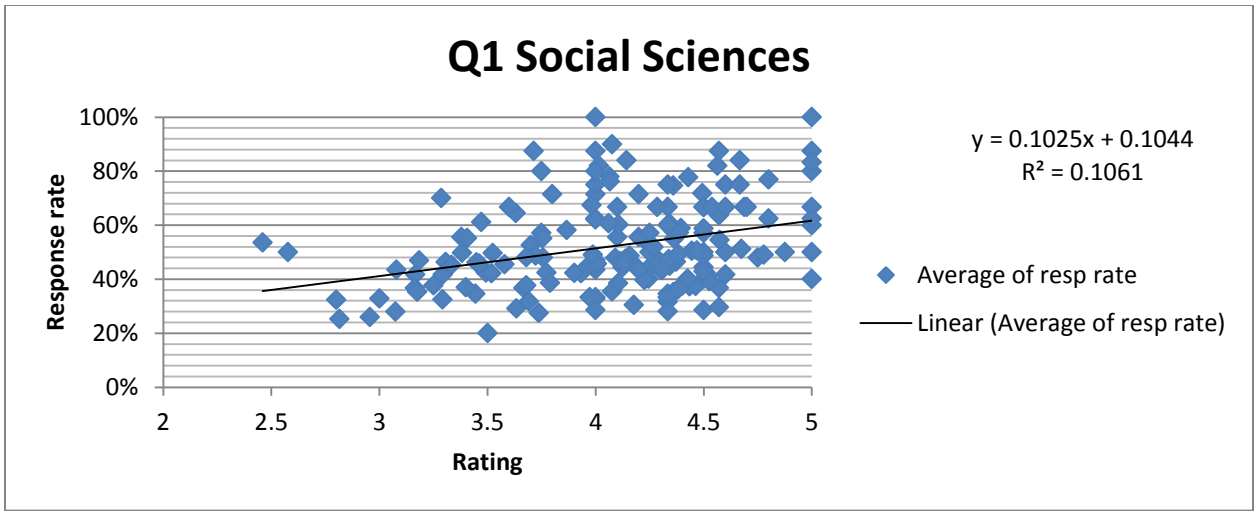


Figure C3: Correlation between response rate and rating for Q1 in Social Sciences

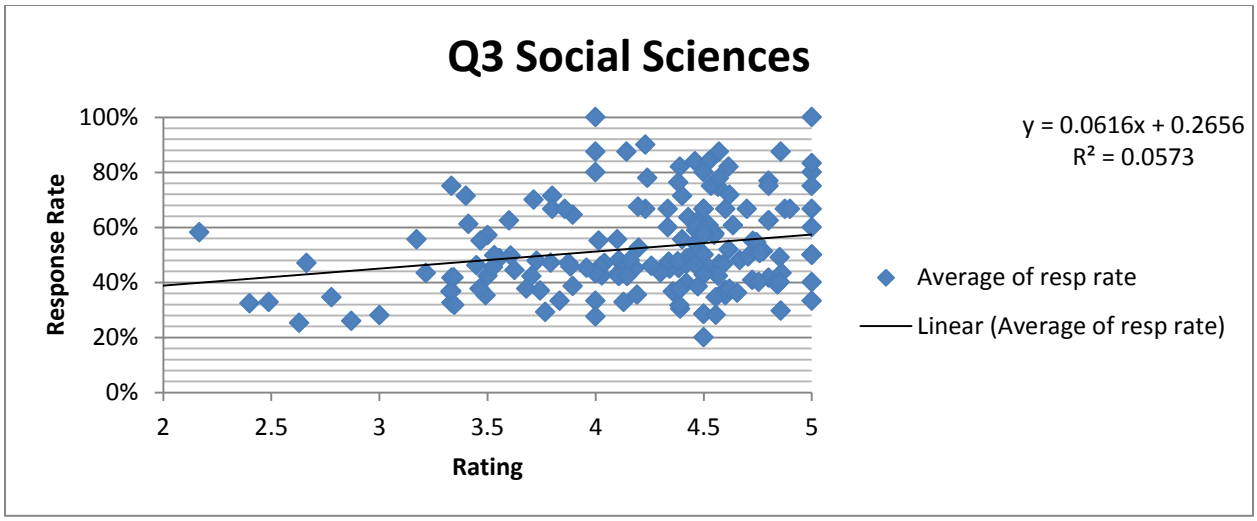


Figure C4: Correlation between response rate and rating for Q3 in Social Sciences

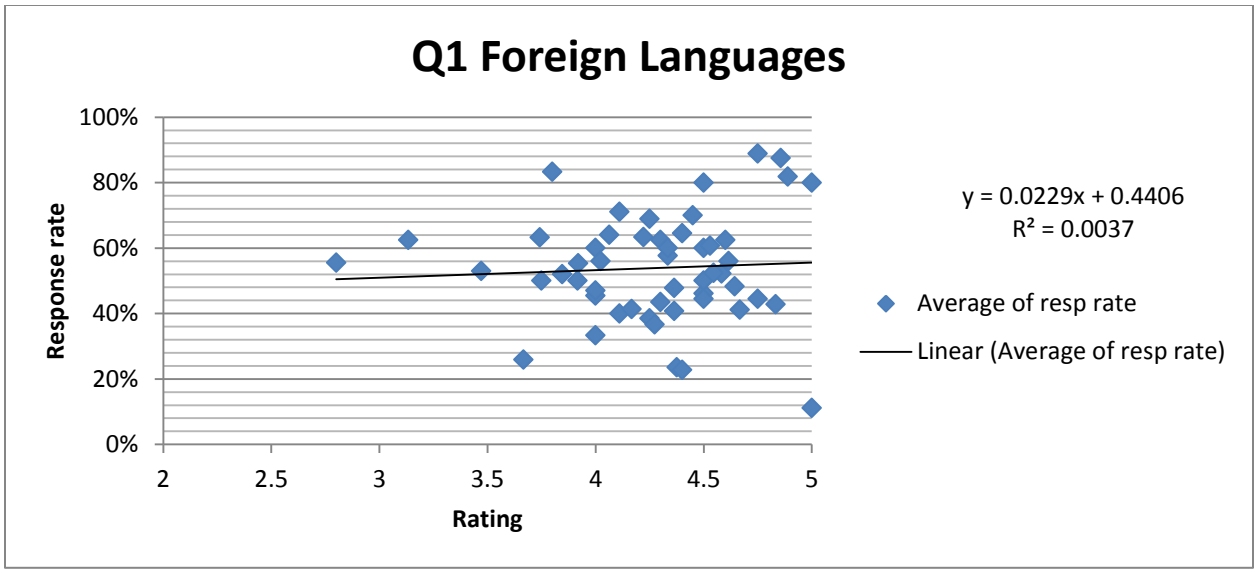


Figure C5: Correlation between response rate and rating for Q1 in Foreign Languages

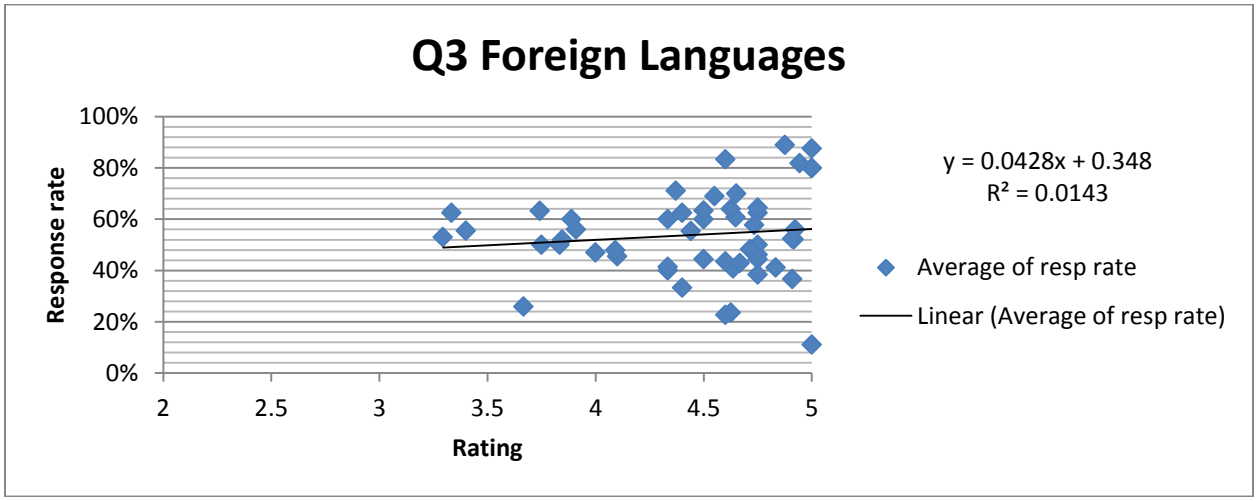


Figure C6: Correlation between response rate and rating for Q3 in Foreign Languages

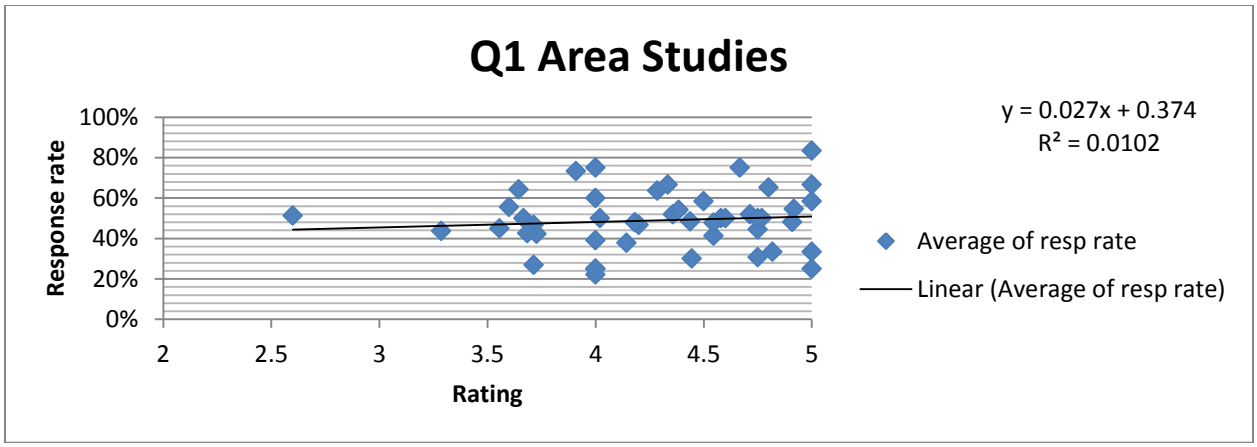


Figure C7: Correlation between response rate and rating for Q1 in Area Studies

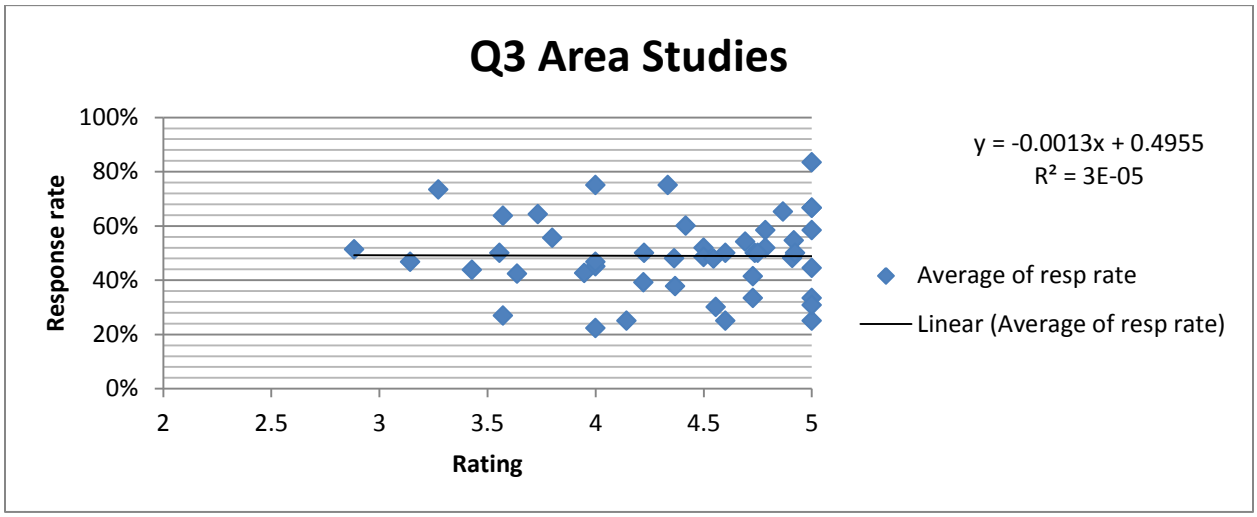


Figure C8: Correlation between response rate and rating for Q3 in Area Studies

Appendix D: Report on extended dates, fall 2010

In Fall 2010, seven units participated in a pilot to evaluate the impact of extending the dates for completing course evaluations. The impetus for the trial was the desire of several academic units to have their instructors receive feedback on all aspects of their courses, including a final exam if present. Numerous analyses were conducted to understand the impact, if any, of the extended dates. The main concern of instructors was that students might "punish" them for a hard exam or a low grade. There was also a concern that response rates might decrease due to students being less inclined to reply once classes stopped meeting.

1. Response rates by unit

Response rates by academic unit were compared for fall 2009 and fall 2010 (see Table D1). The rate was higher for 4 units (2 significant at p<.01), the same for 2, and lower for 1 (no significant difference).

	Fall 2009		Fall 2010 (pilot)			
Department	**n**	**%**	**n**	**%**	**diff**	**Significance**
Animal Science	14	51%	13	51%	0%	ns
Dietetics & Human Nutrition	24	37%	23	49%	12%	p<.01
Natural Resource Sciences	21	57%	23	60%	3%	ns
Parasitology	7	67%	6	68%	1%	ns
Information Studies	19	61%	21	61%	0%	ns
Integrated Studies in Ed	114	47%	103	55%	8%	p<.01
Religious Studies	28	50%	33	48%	-2%	ns
Overall	**227**	**49%**	**222**	**54%**	5%	p<.01

Table D1: Response rates 2009 vs. 2010 for units following extended dates

2. Response rate of extended dates vs. regular rates by Faculty

The response rates for the participating departments were then compared with the other departments in their Faculties. For both Faculties, the difference between the participating and non-participating departments was significant in fall 2010 (Education at $t(249) = 3.96$, $p<.001$; AES $t(149) = 2.94$, at $p<.01$). In 2009 in the Faculty of Education there was no significant difference between the departments who participated and those which did not, while for the Faculty of Agricultural and Environmental Sciences the difference that existed in 2009 was maintained $t(144) = 2.56$. (See Table D2.)

There was no reduction in response rate; if anything, the response rates trended upwards with the extended dates option.

Unit	**Fall 2009**		**Fall 2010**		
Education	**n**	**%**	**n**	**%**	**diff**
Extended dates	133	49%	124	56%	7%
Regular dates	96	47%	127	45%	-2%
Significance		ns		p<.001	
Overall	242	49%	251	49%	0%
AES					
Extended dates	66	50%	65	55%	5%
Regular dates	80	42%	86	47%	4%
Significance		p<.05		p<.01	
Overall	**174**	**51%**	**182**	**53%**	**2%**

Table D2: Response rates 2009-2010 within Faculty for extended dates vs. main evaluation period

3. Submission patterns for fall 2010

The pattern of responses by date was reviewed to assess when students completed their evaluations. Of the 3,736 submissions, 41% were completed during the exam period (see Figure D1 and Table D3). Clearly, a large number of students preferred to submit their evaluations after the end of classes.

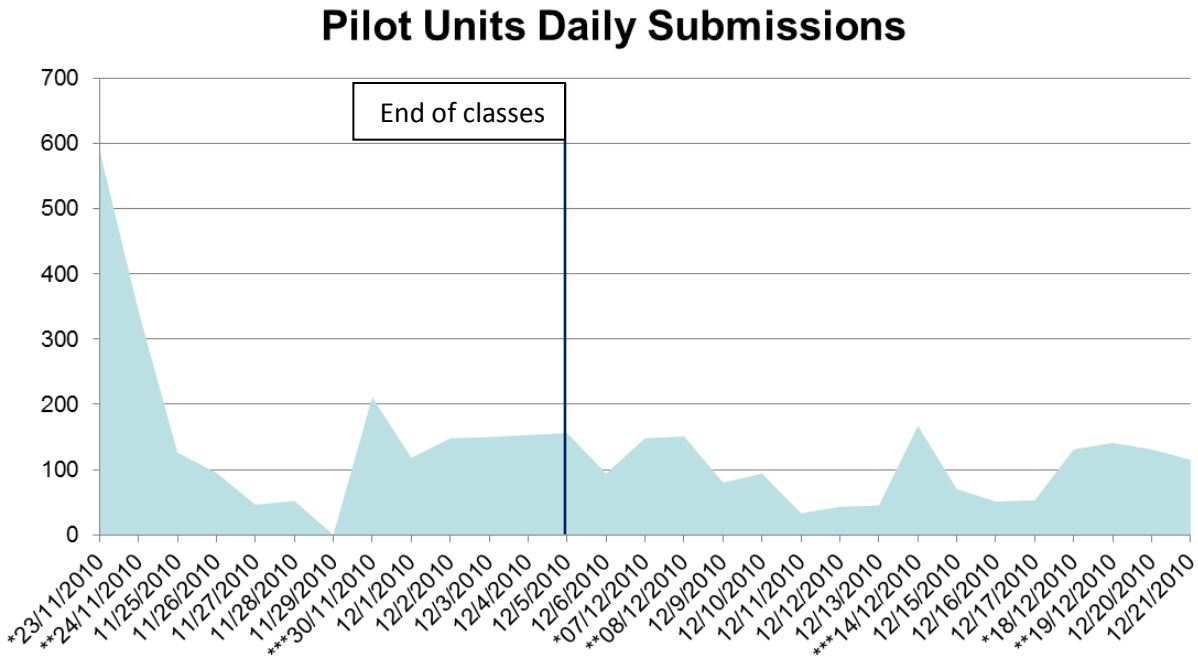


Figure D1: Pattern of responses during extended dates course evaluation

	Nov. 23- Dec. 5	Dec. 6 – Dec. 21	Total
Number of submissions	2,188 (59%)	1,547 (41%)	3,735

Table D3: Timing of responses for participating courses

Two questions were added to each questionnaire in participating units: 1) “Have you completed all of the work (including the final exam, if any) for this course?” and 2) “Have you received your final mark for this course?” (See Table D4 for the numbers of respondents who answered affirmatively.) Overall, 50% (range: 30%--64%) of the evaluations were completed after students had completed all of their work; only 4% (range: 0--7%) were completed after students received their final grade.

	Have you completed all of the work (including the final examination, if any) for this course?		Have you received your final mark for this course?	
Academic Unit	**n**	**% total**	**n**	**% total**
Animal Science	118	43%	12	4%
Dietetics & Human Nutrition	255	53%	25	5%
Natural Resource Sciences	168	43%	28	7%
Parasitology	74	31%	6	3%
Information Studies	230	64%	2	1%
Integrated Studies in Ed	1054	58%	100	6%
Religious Studies	179	30%	3	0%
Overall	**2078**	**50%**	**176**	**4%**

Table D4: Extra questions on extended dates questionnaires

Many students were able to comment on the course experience as a whole, including the evaluation instruments. Results on Q1 (excellent course) and Q3 (excellent teacher) were compared according to whether all work had been completed; the number of students knowing their grade was too small to permit a valid comparison to be performed according to whether they knew their grades. There was no significant difference on either question (see Table 5).

		Overall, this is an excellent course.			Overall, this instructor is an excellent teacher.		
		n	**Mean**	**Significance**	**n**	**Mean**	**Significance**
Have you completed all of the work (including the final examination, if any) for this course?	**No**	2102	4.0		2097	4.1	
	Yes	2076	4.0	ns	2074	4.2	ns

Table D5: Results for Q1 & Q3 by work completion

4. Comparison of repeated courses

There were 86 courses across the seven units that were taught in both fall 2009 and fall 2010 by the same instructor. To assess the potential impact of the response timeframe on course evaluations, the values for the first core question, "Overall, this is an excellent course" and the third, "Overall, this is an excellent instructor" were compared. There were no significant differences for any Faculty for either question.

The 86 repeated courses were then compared individually for the means on Q1 and Q3. From year to year, one would expect some variation in the means, and one hopes that there would be a trend towards improvement. For Q1 ("Overall, this is an excellent course") means, there were nine with significant differences: six increases and three decreases (see Table D6).

Q1 response	# of courses	% of total
Increased	6	(7%)
Same	77	(90%)
Decreased	3	(3%)
Total	86	

Table D6: Q1 response values for repeated course/instructor combinations

Of the 86 courses, there were 11 with significant differences in Q3 ("Overall, this is an excellent instructor") means: eight increases and three decreases (see Table D7).

Q3 response	# of courses	% of total
Increased	8	(9%)
Same	75	(87%)
Decreased	3	(4%)
Total	86	

Table D7: Q3 response values for repeated course/instructor combinations