Preprint of: Francis Jones (2016): Impact assessment of a department-wide science education initiative using students' perceptions of teaching and learning experiences, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2016.1188057
Link to article: http://dx.doi.org/10.1080/02602938.2016.1188057

Impact Assessment of a Department-wide Science Education Initiative using Students' Perceptions of Teaching and Learning Experiences

Francis H.M. Jones*

Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia.

*Department of Earth, Ocean and Atmospheric Sciences,
University of British Columbia,
Earth Sciences Building (ESB),
2020 - 2207 Main Mall,
Vancouver, British Columbia V6T 1Z4,
Canada.
fjonese@eos.ubc.ca
604 822-2138

Acknowledgements

Thanks are extended to Sara Harris and Brett Gilley for helping develop and deploy the student learning experiences survey, Carl Wieman, Sarah Gilbert, Warren Code and Sara Harris for advice regarding data analysis and writing, Adrian Jones for assisting with data transcription and statistical advise, and the students and instructors who participated in these surveys. Feedback from anonymous reviewers has also been crucial in helping improve the article. This work was made possible by the Carl Wieman Science Education Initiative, The Faculty of Science and The Department of Earth, Ocean and Atmospheric Sciences at the University of British Columbia.

Abstract

Evaluating major post-secondary education improvement projects involves multiple perspectives, including students' perceptions of their experiences. In the final year of a seven-year department-wide science education initiative, we asked students in 48 courses to rate the extent to which each of 39 teaching or learning strategies helped them learn in the course. Results were related to the type of improvement model used to enhance courses, class size and course year level. Overall, students perceived unimproved courses as least helpful. Small courses that were improved with support from science education specialists were perceived overall as more helpful than similar courses improved by expert teaching-focused faculty without support, while the opposite was found for medium courses. Overall perceptions about large courses were similar to perceptions of medium courses. Perceived helpfulness of individual strategies was more nuanced and context dependent, and there was no consistent preference for either traditional or newer evidence-based instructional practices. Feedback and homework strategies were most helpful in smaller courses and independently improved courses. Results indicate that students are perceptive to benefits that arise when improvements are made either by expert educators or by research-focused faculty who received dedicated support from science education specialists.

Keywords

education improvement models, impact assessment, initiative evaluation, learning experiences, student perceptions

Introduction

Background

Education transformation initiatives have been in progress in many colleges and universities and several models for change are employed (Weaver et al. 2016), particularly in Science, Technology, Engineering and Mathematics or STEM disciplines (Henderson, Beach, and Finkelstein 2011). However, evaluating the impacts of such initiatives, especially large scale projects spanning many courses is challenging (Fairweather, Tarpani, and Paulson 2016) requiring at least three perspectives. A key perspective is whether the effectiveness and sophistication of student learning increases. The Valid Assessment of Learning in Undergraduate Education initiative (VALUE) of the Association of American Colleges & Universities is one example of how assessment of student learning represents the predominant aspect of evaluating education improvement efforts (Sullivan and Schneider 2015). Impacts on instructing faculty are also important, including changes in their efficiency, satisfaction, capabilities and habits as educators (McCrickerd 2012; Wieman, Deslauriers, and Gilley 2013). A third perspective concerns the perceptions of students regarding the learning experiences they encounter. Their motivation, and the way they think about learning and studying, is important because it determines how they tackle their learning tasks (Struyven, Dochy, and Janssens 2005; Ambrose et al. 2010, 66–69).

Students' perceptions are considered by various authors at many levels and for different purposes. At the departmental, institutional or national level they are used to assess student engagement (McCormick, Gonyea, and Kinzie 2013; Handelsman et al. 2005), or to evaluate student satisfaction with institutional practices or individual academic, social and personal factors that influence their own academic performance (Welsh 2010). The most commonly acquired form of student perceptions, and probably the most widely researched (Marsh 2007, 320), are Student Evaluations of Teaching (SET) (Wilson, Lizzio, and Ramsden 1997; Richardson 2005; Marsh 2007). In his comprehensive review of SET research and applications, Marsh, 2007, concludes that SETs can be reliable and useful and that few other indicators of teaching effectiveness are systematically supported by research findings. However, he also notes that two persistent critical issues are that teaching effectiveness is both multidimensional and context dependent, and that integrating SETs into programs to enhance teaching effectiveness is still challenging. We aimed to address these issues by asking students about multiple specific experiences and characterizing results in terms of the different models used to implement teaching enhancements.

At the individual course or activity levels, students' perceptions are often obtained by instructors to help assess and improve their own teaching strategies or as part of studies about specific learning or teaching activities. Several comprehensive guidelines for collecting and using such feedback exist (Procter et al. 2015; Boud and Molloy 2013). However, perceptions data obtained for such specific situations are not easily used for evaluation across different courses or settings because purposes and questions may differ. Instead, a broader assessment of students' perceptions across courses and contexts is desirable. The Classroom Survey of Student Engagement (CLASSE) (Smallwood and Oiumet 2009) is one such instrument that has been used to evaluate education improvements (Reid 2012). However we needed a set of questions more closely tailored to evaluating all courses across our department, therefore we developed an assessment of perceptions called the Student Learning Experiences Survey (SLES).

Context

This study was undertaken to help evaluate a seven-year department-wide education improvement initiative that changed our department's education culture by helping instructors permanently adopt evidence-based teaching practices (Wieman, Perkins, and Gilbert 2010; Chasteen et al. 2016). The initiative fostered change at the department level by injecting substantial short term funding and incentives, keeping costs and sustainability of delivering education essentially unchanged and focusing upon how students learn rather than what was taught (Carl Wieman Science Education Initiative, 2016). One key to success was engaging full time science education specialists (Chasteen et al. 2011) as primary agents of change. They have PhD or Masters level expertise in the department's subject area and have demonstrated a commitment to becoming experts in discipline-based post-secondary pedagogy. They collaborated with individuals or small groups of faculty to help increase faculty's knowledge of teaching and learning research, introduced new educational and assessment practices, collected quantitative and qualitative data about learning, and conducted scholarly research where appropriate. Measuring the success of this education initiative has entailed evaluations of many department, course and student indicators (Chasteen et al. 2016) but until the present study, none have assessed whether students perceive specific teaching and learning experiences differently in courses that were enhanced using different improvement models.

Goals of the study

This study focusses on assessing student perspectives as a means of evaluating the impact of a large education improvement initiative. Students' perceptions of 39 teaching or learning strategies were investigated in 48 courses taught by one department of a large publicly funded research university during the last year of the education improvement initiative. Our goal was to identify relationships between students' perceptions and improvement models. The intent was not to address why students perceive strategies as either helping them learn in the course or not, nor to somehow judge the quality of instruction. Instead we sought to gather a pattern of 'helpfulness' across a consistent set of specific strategies from as many students as possible who were learning in the department during the final year of the seven year initiative. With these data, two specific questions were addressed. First, were student perceptions about whether teaching and learning strategies helped them learn in the course (i.e. the strategy's 'helpfulness') related to the manner in which courses were improved, the class size and year level of courses? Second, which specific teaching or learning strategies were perceived as most or least helpful, and how did these preferences depend upon the same three factors? Answering these questions was anticipated to contribute towards evaluating whether, and how, this major education initiative positively impacted students' perceptions of their educational experiences.

Methods

Improvement models, Courses and Instructors

Forty eight courses provided data, with one offered in 4 sections, 4 offered in 2 sections, and 42 offered in one section, for a total of 54 course sections taught in either the fall term of 2013 (September through December) or the spring term of 2014 (January through April). Three models for effecting education improvement were employed sometime during the seven-year initiative. First, 21 course sections participated in major education transformation projects lasting four or more terms. Each project involved one science education specialist devoting roughly a third of their time to helping instructors implement evidence-based, active, student-centric practices and developing strategies to measure student capabilities and attitudes. Second, 10 course sections received less intensive support in the form of consulting with science education specialists to adjust specific laboratory, classroom, homework or assessment practices. Third, 8 course sections were improved by instructors experienced in proven discipline-specific pedagogy with little support from science education specialists. Finally, fifteen course sections remained largely unchanged during the seven year initiative, not because they were intentionally left out, but for a variety of reasons related to logistics and personal preferences of individual faculty. These intervention models are referred to as “transformed”, “consulted”, “independent” and “none”. Table 1 lists the number of course sections experiencing each improvement model with numbers of respondents and corresponding response rates.

Table 1: Number of courses (N), total respondents (TR), and response rates (RR), organized by improvement model (rows) and course year level.

	1st year courses			**2nd year courses**			**3rd year courses**			**4th year courses**		
	N	**TR**	**RR**	**N**	**TR**	**RR**	**N**	**TR**	**RR**	**N**	**TR**	**RR**
Transformed	6	496	29%	5	290	52%	9	369	50%	1	16	94%
Consulted	1	60	29%	2	54	73%	1	23	61%	6	123	55%
Independent	0	-	-	3	131	77%	3	242	72%	2	47	84%
None	1	64	47%	1	50	85%	4	136	55%	9	184	64%

Table 1 reflects the priorities of the education initiative, which were to help as many instructors as possible gain pedagogic expertise, and to target larger courses in favour of smaller more specialized courses. All but one of the large first year course sections and over half the second and third year course sections participated in multi-term transformation projects. In contrast, most fourth year courses received less official improvement support.

Regarding class size, we chose to define each course section as either small, medium or large, and then characterized dependence of perceptions on improvement model for each size group. Nine courses with over 200 enrolled students each were defined as large. As a priority of the improvement initiative, eight of these received transformation support and the ninth received consulting support. All other courses had fewer than 150 students. Choosing an enrolment value to distinguish between small and medium courses is challenging. We chose to define small courses as having fewer than 50 students because Freeman et al. (2014) found in their meta-analysis of 225 studies on active learning that active strategies had the highest impact on student success in courses with fewer than 50 students. Setting 50 students as a boundary value also ensured sufficient numbers of courses in this study for analyses within each group. Table 2 summarizes the number of courses and total number of respondents (not enrolments) in classes grouped by improvement model.

Table 2. Number of course (with total number of respondents in brackets) grouped by class size and improvement model.

	Large classes (> 150)	**Medium classes (50 - 150)**	**Small classes (< 50)**
Transformed	8 (574)	7 (431)	6 (164)
Consulted	1 (61)	2 (86)	7 (106)
Independent	0	4 (307)	4 (116)
None	0	5 (243)	10 (190)

A total of 39 instructors participated in this study. Twelve participated in more than one project. Seven taught in independently improved courses with four of those actually contributing to those independent improvements. Sixteen taught unimproved courses and ten of those were never involved in the seven year education initiative.

Survey questions

The Student Learning Experiences Survey (SLES) was designed to obtain focused student feedback about specific teaching or learning strategies. We wanted to avoid asking students to make judgements about their own general or specific capabilities or whether those capabilities improved because it is known that self-assessment is difficult for novices in any discipline (Kruger and Dunning 1999). Thirty nine questions were posed in a manner inspired by the Student Assessment of Learning Gains (SALG) (Seymour et al. 2000). First, nine questions were posed by asking "*How much did information provided help you learn in this course*" and listing nine information types. Next, thirteen questions asked "*To what extent did classroom strategies help you learn in this course*", and finally seventeen questions asked "*How much did homework or feedback help you learn in this course*". The five options for all questions were "*extremely helpful*", "*very helpful*", "*moderately helpful*", "*little or no help*" and '*NA*' (not applicable). For analysis these responses were assigned values of 4, 3, 2, 1 and 'NA' respectively. These, and various averages calculated on these values, are referred to in this paper as 'helpfulness' values. One compliance check question, included to help confirm students were reading questions carefully, instructed students to check the "little or no help" option.

In order to keep the survey anonymous no gender, ethnicity, age or other identifying information was obtained, although students were asked to identify their major (eg BSc), specialization (eg "geology") and their current program year (eg. freshman, senior, etc.). One question was asked about whether multiple instructors were advantageous. Additional questions asked students to compare their workloads and enthusiasm for the course to other courses they were taking concurrently. These and five final agree/disagree questions are not discussed in this paper although they were used for other purposes and results will be reported elsewhere. The complete questionnaire can be obtained via Jones, 2014.

Data collection and preparation

Students took roughly 20 minutes to respond to all questions on paper forms during a class towards the end of term. It is commonly challenging to obtain permission for 30 minutes of time in classes across a large department, however most instructors recognized this as a unique opportunity to hear feedback about their course and as an important contribution to the department-wide education initiative. Online surveying was considered but our own experience and that of others (Dommeyer et al. 2004) is that response rates are higher using in-class surveys unless a grade incentive is offered with online versions.

Sampling rates were 51% to 100% for thirty-eight course sections and 14% to 50% for sixteen. Fall term data were transcribed by hand. Classes with over 100 respondents were sub-sampled down to sampling rates chosen based on conclusions of Zumrawi, Bates, and Schroeder (2014) who recommended minimum response rates for institutional course evaluation surveys based on class size. These sub-samples were chosen randomly from respondents while maintaining the proportion of students in each year level and degree program, as expressed in complete class lists. For all other courses, the actual sampled demographics were checked for consistency with class demographics. The spring term data were gathered using automatically scored paper response forms so sub-sampling was not necessary.

Raw data consisted of responses to 39 questions from 2489 respondents representing 4871 students. Students who did not correctly answer the compliance check question were removed. Also, courses with fewer than 5 respondents and field-based courses were not included. The final count of respondents across all 54 course sections was 2278 students.

The data for every response from each student comprised a value of 4, 3, 2, 1 or NA for each answer to the 39 questions, plus seventeen more parameters characterizing details such as course code, class size, intervention model and student's degree. 'Not applicable' responses were never included in calculations. Averaged over all respondents, the mean perception values for each of the 39 questions ranged from 1.75 to 3.18. The total number of respondents for any one question ranged from 415 to 2244 students.

One further preparation step involved testing data validity (i.e. the assumption that students answered questions thoughtfully and truthfully) by comparing response patterns to pairs of questions. First, mean helpfulness values were found for each question in each of 54 course sections. Then the 54 values for one question were correlated against the 54 values for a second question. The resulting linear correlation coefficient indicates whether students' response patterns were similar for those two questions. When this was done for all questions, the pairs of questions that were most similar were most highly correlated (e.g. "clicker questions" and "discussions about clicker questions"); moderate correlations were found between less similar question pairs (e.g. "whole class discussions" and "Socratic lecturing"), and other pairings yielded correlation coefficients suggesting little or no similarity in response patterns.

Results

Consistency among students in different majors or different course sections

Did perception values depend upon the major in which students are enrolled? This question, relevant mainly in large courses taken by students both within and outside the department, was tested in one first year and one second year course. Third and fourth year courses did not have sufficiently diverse populations to make this a concern. In the second year course 37 respondents were engineering majors and 46 were science majors. Both a standard t-test and a two-sample Wilcoxon Rank Sum test were applied using the R language and environment for statistical computing (R Core Team 2015) to test the null hypothesis that the two subgroups do not have different perceptions about the helpfulness of each survey question. The viability of using these tests with Likert-scale data is affirmed by de Winter and Dodou (2010). Results from both tests were very similar with none of the 39 questions yielding mean response values that were significantly different after applying the Bonferroni correction; i.e. for $P<0.05\div39=0.0013$. The Bonferroni correction tends to be conservative (Abdi 2007) and so is perhaps not ideal if testing a null finding such as this, however, only two of 39 survey questions were individually significantly different at $P<0.01$, and only four more at $P<0.05$. The same procedure was applied in a large first year course with 91 respondents declared as BSc majors and 74 respondents declared in other disciplines. This test resulted in only one survey strategy being perceived as significantly different by these two groups at $P<0.01$. This strategy was "use of clickers" (personal response systems) and both BSc and others perceived this strategy as very helpful (means of 3.6 and 3.3 respectively). Students therefore appear to be perceiving teaching and learning strategies consistently enough regardless of their major, so that populations in each class are sufficiently uniform for our purposes.

Did perception values differ in different sections of the same course? In one first year course SLES data were obtained from four separate sections taught to over 280 students each. The similarity of responses was assessed using the same tests described above with results yielding no significant differences at $P<0.05$ after applying the Bonferroni correction. Similar results were obtained for a second year course with sections in fall and spring terms. For a third year course with fall and spring term sections there were significant differences for three strategies (online content, in-class help from teaching assistants and homework exercises). Differences were more significantly large for one other first year course. Mean helpfulness values for one homework strategy and nine classroom strategies were perceived as significantly less helpful in one section. This course was taught with similar content and ostensibly similar teaching strategies but by different instructors in fall and spring terms, implying that the implementation of strategies by one instructor was more effective (helpful).

Cumulative results for each question

Figure 1 presents combined Likert scale data from the 1410 respondents in 31 course sections that were improved with either the transformation or consulting model. This combination was chosen because overall means of Likert data for all respondents in these two groups of courses were similar, as discussed shortly. Each of the 39 teaching or learning strategies is represented with a stacked bar representing cumulative Likert Scale data from this subset of respondents. Strategies widely recognized as evidence-based best practices (e.g. Ambrose et al. 2010) are identified with a Carat '^' symbol. Some strategies were strongly endorsed, having relatively few counts of moderately, little or no help (e.g. instructor's notes, lecture presentations, clicker questions and solo studying). Some strategies were weakly endorsed with more than half the respondents saying the strategy was moderately, little or no help (e.g. text books, non-clicker questions and pre-readings). Several strategies in each category were encountered by relatively few respondents, especially five of the homework and feedback strategies.

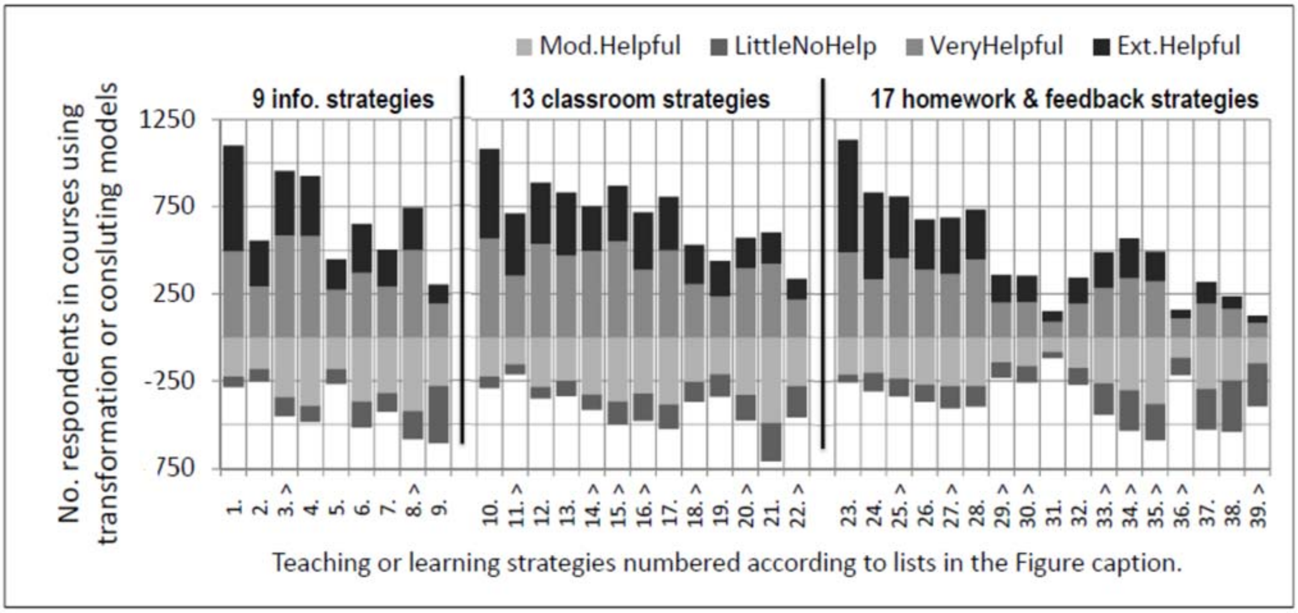

Figure 1. Likert scale data from respondents in courses improved using consulting or transformation models (N=1410). The x-axis is labelled with question (i.e. teaching or learning strategy) numbers based on the following list. The carat '^' symbol identifies strategies loosely defined as evidence-based best practices. Questions are sorted as explained in Figure 2.

How much did these types of information help you learn in this course?

1. Lecture notes provided by the instructor
2. Individual interactions with the instructor outside the class
3. > Learning goals describing knowledge and skills you are learning
4. The topic list or syllabus provided
5. Information accompanying lab instructions
6. Online content like notes or readings
7. Video material provided or pointed to by the instructor
8. > Learning goals related to attitudes, scientific thinking or professionalism
9. The text book

To what extent did these classroom strategies help you learn in this course?

10. Lecture presentations in class
11. > Clicker questions posed in class
12. Demonstrations, animations or simulations shown by the instructor
13. Help from the instructor during class
14. > Discussions before, during, and/or after those demonstrations
15. > Discussions about why material is useful, important or interesting
16. > In-class activities in groups using worksheets or other resources
17. "Socratic dialogues"; i.e. instructors teaching by constantly asking questions
18. > Discussions you had with other students about those clicker questions
19. Help from teaching assistants during class
20. > Whole class discussions moderated by the instructor
21. Questions asked of students in class NOT involving clickers of worksheets
22. > Opportunities you had to assess or comment on work of your peers

How much did these types of homework or feedback help you learn in this course?

23. Studying/reviewing on my own

24. Practice questions provided prior to midterm or final exams
25. > Studying/reviewing with a group of other students
26. Online quizzes and assignments
27. > Rubrics or grading schemes provided for assigned work
28. Homework exercises (problem sets, etc.)
29. > Instructor/TA feedback on preliminary work BEFORE final due date
30. > Projects you did with other students (written, oral, poster, etc.)
31. Homework explicitly related to lab work (reports, worksheets, etc)
32. Projects you did on your own (written, oral, poster, etc.)
33. > Feedback from instructors/teaching assistants on completed homework or projects
34. > Feedback provided on quizzes and exams
35. > Any requirements to complete some readings BEFORE classes
36. > Opportunities to comment on / reflect on learning and study habits
37. Readings of scientific or professional literature
38. Other "recommended" (not required) readings
39. > Online wiki or discussion board activity

Plotting raw data reveals which strategies were more or less common, but it does not clearly reveal which strategies were preferred. To better visualize the relative helpfulness of strategies, mean values of helpfulness were generated for each question by averaging the Likert-scale choice values from all respondents who were taking courses improved using either the transformation or consulting models. Figure 2a presents these mean values of helpfulness with strategies sorted within the three question categories to highlight those found to be most or least helpful. The most helpful evidence-based instructional practices included knowledge-based learning goals, clicker questions, group studying, and rubrics, while the most helpful traditional strategies included instructor's notes, lecture presentations, studying alone and practice tests. Least helpful strategies experienced by this subset of students included text books, non-clicker questions, assessment of peers, "other" readings and wikis or online discussions. Two other panels in Figure 2 present the difference between means of Figure 2a and corresponding means from respondents in unimproved courses (Figure 2b) or courses improved independently (Figure 2c). For each of the 39 strategies, an estimate of the significance of differences between the three mean values was obtained by applying a one way ANOVA test followed by Tukey Honest Significant Difference (TukeyHSD) post-hoc tests. Paired mean values that were significantly different are identified in Figures 2b and 2c as described in the caption.

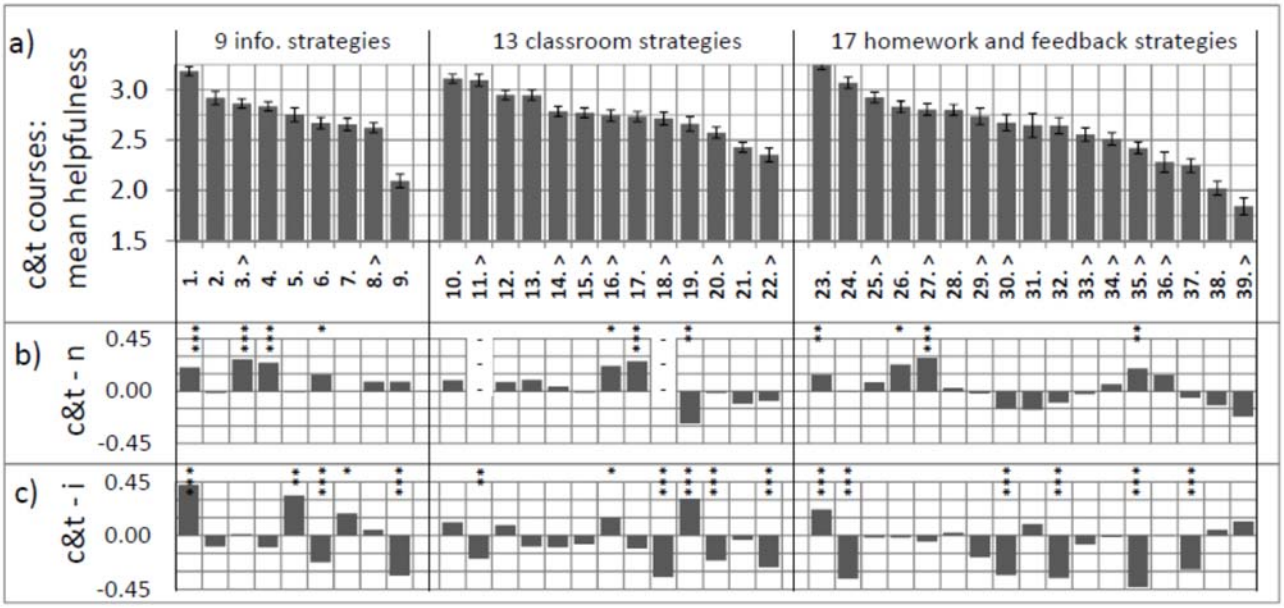

Figure 2. (a) Mean helpfulness values from respondents taking courses improved using consulting or transformation models, with 95% confidence intervals. The x-axis is labelled with question (i.e. teaching or learning strategy) numbers based on the list in Figure 1's caption, and is sorted to emphasize most and least helpful strategies within each category. (b) Differences between means from (a) and similar means from unimproved courses. (c) Differences between means from (a) and similar means from independently improved courses. In both (b) and (c), significance of differences between means: * for Padj < 0.05; ** for Padj < 0.01; *** for Padj < 0.001.

Results accumulated by intervention model, class size and year level

Figure 2 reveals that evidence-based instructional practices such as clicker questions, clicker question discussions, group work in classes, rubrics and online assignments were generally perceived as more helpful when encountered in improved courses. But what other factors may be contributing to overall patterns? Figure 3 summarizes perceptions from all students averaged over all strategies for courses experiencing each improvement model. Figures 3b and 3c summarize students' perceptions averaged over all 39 strategies in courses accumulated according to year level and class size respectively.

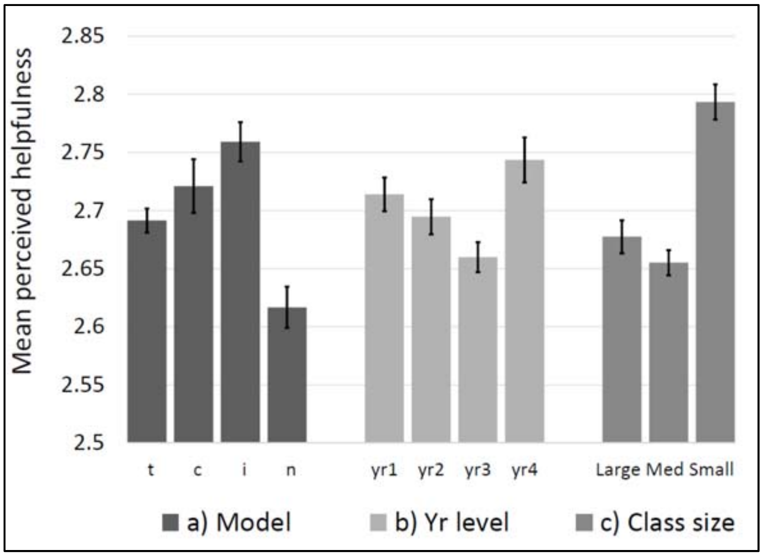

Figure 3. Aggregate mean perceived helpfulness averaged over all respondents, including 95% confidence intervals. Each aggregate mean value arises from averaging individual responses to all question answers from all relevant respondents. Three aggregations are: (a) Intervention model (t, c, i and n represent transformed, consulted, independent and none); (b) Course year level; (c) Class size.

Table 3. Significance between pairs of mean values in Figure 3 are summarized using the following codes: * for Padj < 0.05; ** for Padj < 0.01; *** for Padj < 0.001. Intervention codes represent course improvement models: t = transformed, c = consulted, i = independent and n = none. Class sizes are defined in Table 2.

Model	**Year level**	**Class size**
t-c	yr1-yr2	large-medium*
t-i***	yr1-yr3***	large-small***
t-n***	yr1-yr4	medium-small***
c-i*	yr2-yr3**	-
c-n***	yr2-yr4***	-
i-n***	yr3-yr4***	-

Significance of differences between pairs of means within each of these three groupings was examined using one way ANOVA followed by TukeyHSD post-hoc tests. Results are summarized in Table 3, which is organized to facilitate comparison with Figure 3.

Figure 3a and corresponding significance estimates in Table 3 can be summarized as follows. Overall average results are consistent with first impressions from Figure 2; that is, courses that received no intervention were perceived as significantly less helpful than courses improved using either of three improvement models. Courses improved by individual instructors with little official support were perceived overall as most helpful. The corresponding distinction between courses improved independently versus those receiving consulting support was marginally significant (Padj<0.05). The distinction between overall perceptions from courses improved with consulting versus transformation was insignificant.

Transformation projects concentrated initially on larger mostly first year courses, and Figure 3a shows that transformed courses were generally perceived to be relatively less helpful compared to other improved courses. This raises the question of whether class size and/or course year level influenced perceptions of helpfulness. From Figure 3(b), strategies were perceived overall as most helpful by students taking 1st or 4th year courses and least helpful by those in 3rd year courses. The distinctions between 1st year courses and 2nd or 4th year courses were insignificant. Regarding class size, Figure 3(c) shows strategies were generally perceived as much more helpful in small courses while the overall distinction between perceptions from large and medium courses was only marginally significant.

Averaged perceptions from courses grouped by improvement model are further broken down by class size in Figure 4 with corresponding significance estimates in Table 4. Small classes that were improved using transformation or consulting models were perceived overall as significantly more helpful than independently improved or unimproved courses. The opposite pattern occurred for the three types of improved medium sized courses. Also, overall experiences were perceived as less helpful in medium compared to small improved courses. Experiences in unimproved courses were perceived overall as more helpful in medium compared to small courses. In the eight large transformed courses, experiences were perceived overall to be similarly or more helpful compared to those in most medium-sized courses. However, the experiences in the large course that received only consulting support were perceived as significantly less helpful.

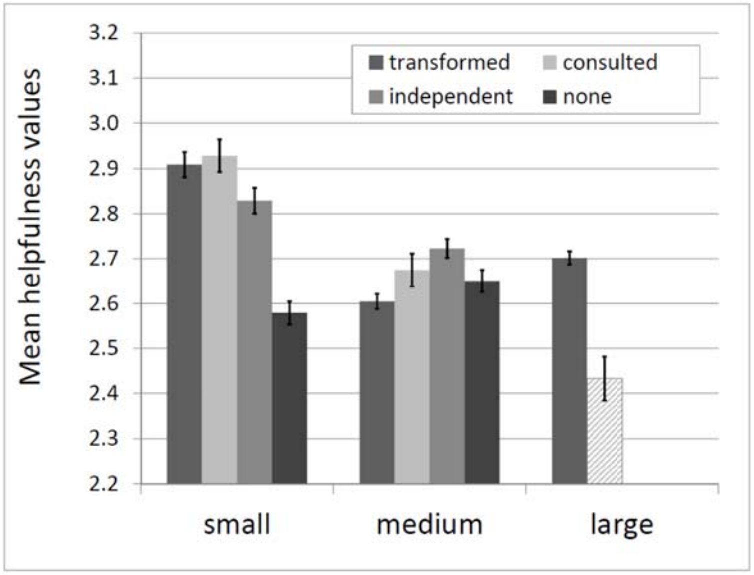

Figure 4. Aggregate mean helpfulness values from respondents in all courses improved by each intervention model, aggregated by class size, including 95% confidence intervals. Mean value shown with a patterned bar was derived from one course only.

Table 4. Significance between pairs of mean values in Figure 4 are summarized using the following codes: * for Padj < 0.05; ** for Padj < 0.01; *** for Padj < 0.001. Intervention codes represent course improvement models as defined for Table 3.

Small	Medium	Large
t-c	t-c*	t-c***
t-i***	t-i***	-
t-n***	t-n**	-
c-i***	c-i	-
c-n***	c-n	-
i-n***	i-n**	-

Perceptions for different improvement models were considered in terms of year level in a manner similar to class size. However, five of the resulting mean values involved respondents from only one course each, so considering variation of student perceptions in terms of course year level was considered less useful. However, three observations are relevant. First, strategies in the large transformed first year courses were perceived as significantly more helpful compared to those in transformed second or third year courses. Second, perceptions from independently improved courses varied less by year level than courses receiving other improvement styles. Third, strategies were generally perceived as most helpful in fourth year courses regardless of improvement model.

Figures 1 and 2 show which specific strategies were perceived as most or least helpful, while Figures 3 and 4 provide insight about the influence of class size and year level on averaged perceptions of helpfulness. Greater detail about the dependence of individual strategies on class size is provided in Figure 5 which compares mean helpfulness of specific strategies aggregated by class size using the same approach as Figure 2. Figure 5a displays mean helpfulness of each strategy (sorted in the same order as Figure 2) as perceived by all students taking medium sized courses. Figures 5b and 5c show differences in mean values between perceptions from medium versus small sized courses and between medium versus large courses. As in Figure 2, estimates of the significance of differences between three mean values for each of the 39 strategies were obtained by testing with a one way ANOVA followed by TukeyHSD post-hoc tests. From Figure 5b all but two strategies were perceived as more or equally helpful by students in small classes compared to medium classes. Comparing medium to large classes, Figure 4 suggested perceptions from these two were similar, however Figure 5c suggests this was not uniformly the case for specific strategies. Details are discussed below.

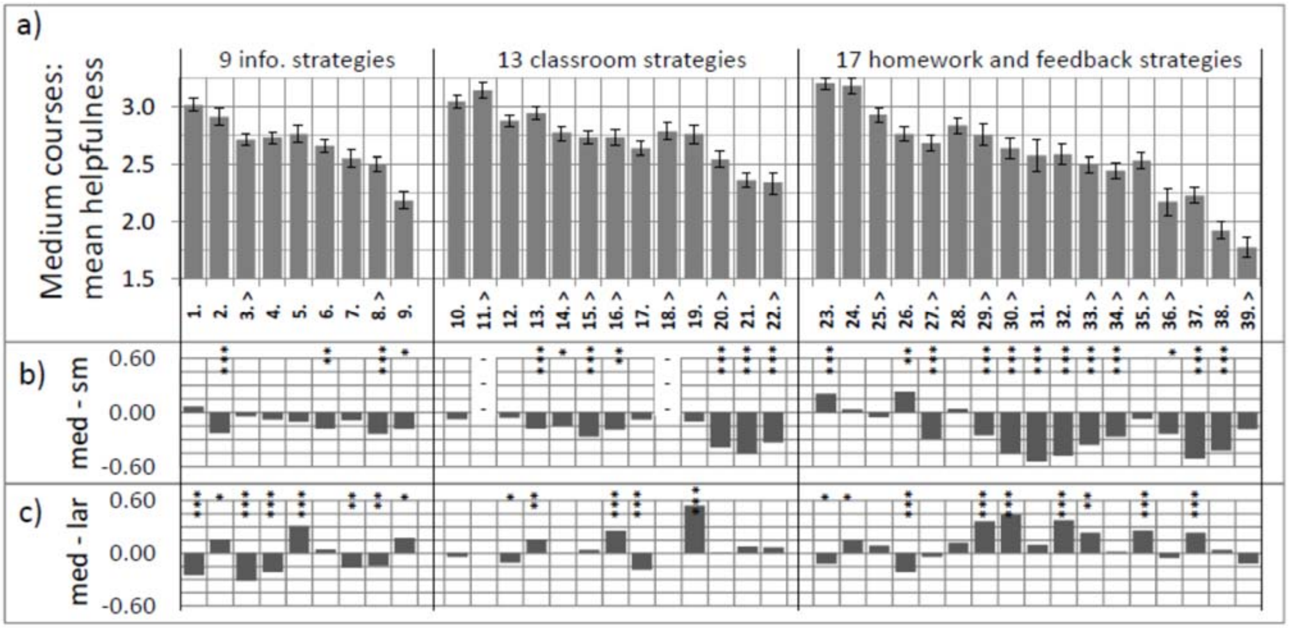

Figure 5: Mean helpfulness aggregated by class size, with x-axis labelled and sorted as per Figure 3. (a) Mean helpfulness from all students taking medium sized classes, with 95% confidence intervals. (b) Difference between mean values in (a) and corresponding mean values from respondents in small classes. (c) Difference between mean values in (a) and corresponding mean values from respondents in large classes. Significance of differences between means is indicated as per Figure 2.

Discussion

Regarding study question 1

Our first question is about evaluating the broader relationships between students' perceptions and intervention, class size or year level. Averaged over all 39 strategies, students' experiences were generally perceived as most helpful in courses improved by individual expert instructors, slightly less helpful in courses receiving official transformation or consulting support and much less helpful in courses that were not improved (Figure 4). This overall pattern is not unexpected because the eight independently improved courses were developed, enhanced and taught by particularly experienced instructors, all of whom have a recognized history of successful – sometimes award-winning – teaching innovation. In contrast, the 31 courses that received transformation or consulting support were taught by instructors with a wider range of experience but improvements were supported by science education specialists. It is encouraging to find that overall, students in those courses were nearly as enthusiastic about teaching and learning practices they encountered as students in the independently improved courses.

In terms of class size, experiences were perceived as less helpful in improved medium compared to improved small classes. This suggests that more effort is needed to make teaching and learning strategies as helpful in medium sized classes as they were in small classes. Regarding large classes, it is noteworthy that those students found their experiences to be similarly or more helpful compared to those in most medium classes. In other words, there seems little to distinguish overall helpfulness of experiences encountered in improved courses with enrolments ranging from 50 to 300. Specific factors influencing these overall results are discussed further below.

Regarding class year level, students in transformed first year courses (all large) perceived their experiences as significantly more helpful than students in transformed second or third year courses. This represents a positive outcome of the department's decision to focus improvements on the largest courses. Next, the lack of variation in terms of year level among independently improved courses suggests the benefits of expert pedagogues is perceived by students regardless of course year level. Finally, strategies were not surprisingly perceived generally as more helpful in fourth year courses regardless of improvement style, probably reflecting the facts that most classes are small and senior students are more committed, expert-like and experienced as learners.

These interpretations of broad, department-wide impacts of the initiative are based on averaged perceptions involving many courses. However each course has a unique history of improvement and unique context-specific characteristics. For example, in the course perceived overall as the least helpful of eight large transformed courses, students also responded to the SLES question about multiple instructors more negatively than students in 22 other courses taught by two or more teachers. Students are particularly sensitive to the consistency of experiences they encounter from multiple instructors teaching in a single course (Jones and Harris, 2012), therefore improvements in this course may have been compromised by difficulties coordinating multiple instructors. The point here is that results of assessing the overall impact of the initiative at the department level may not be directly applicable to individual courses.

Regarding study question 2

Our second question asks which specific strategies were perceived as most helpful and whether intervention model or class size affected these preferences. Raw Likert scale data were first presented in Figure 1. This form of presentation helps identify which specific strategies were encountered by relatively few students, indicating for example that several well-recognized best practices could be made more prominent in the department's curriculum. Examples are group projects, reflective practices and feedback on preliminary work. Also, strategies that were perceived as least effective can be recognized as those for which the majority of students chose the ‘moderately’, or ‘little or no help’ options (e.g. text books, non-clicker questions, feedback on tests, and three types of readings).

However, the calculated means and corresponding comparisons shown in Figures 2 and 5 were more useful for identifying strategies perceived as most helpful and for addressing the impact of intervention or class size on their helpfulness. One somewhat unanticipated outcome was that traditional strategies were still perceived as the most helpful, although importantly, three evidence-based instructional practices that were promoted during the initiative were not far behind, namely learning goals, clicker questions and group studying (Figure 2a). In terms of the impact of improvement model, 11 strategies were considered significantly more helpful in courses improved with science education specialist support. Four are evidence-based instructional practices while 7 are more traditional (Figure 2b), so active support for making improvements apparently helped make both new and traditional strategies more helpful. The only exception (i.e. the only strategy considered significantly more helpful in unimproved courses) was the use of teaching assistants, possibly because they often run labs and nine of the fifteen unimproved classes include laboratory components.

Similarly, comparing perceptions from classes improved either with or without science education specialists (Figure 2c) reveals that six strategies were perceived as significantly more helpful in courses improved with science education specialist support. Only one of these (in-class activities) may be considered an evidence-based instructional practice. Nine strategies were perceived to be very helpful but were not significantly dependent upon improvement style. Of these, 4 are evidence-based instructional practices (instructor help in class, discussions about usefulness, group studying and feedback on preliminary work) and 5 are more traditional strategies. Consequently we conclude that evidence-based instructional practices and traditional strategies were affected similarly by both styles of improvements, although with perhaps a few more evidence-based instructional practices being perceived as more helpful in independently improved courses.

Turning to the impact of class size on perceptions of specific strategies (Figure 5), students in small classes found 12 of 18 evidence-based instructional practices and 11 of 21 other strategies to be significantly more helpful compared to those in medium classes while only 2 were more helpful in medium classes (studying alone and online assignments). In particular, virtually all feedback strategies were perceived as significantly more helpful in small compared to medium classes, and likewise for medium compared to large classes. This result may be expected given the time it takes to provide meaningful homework and feedback using conventional methods. However these results do suggest that students perceive the benefits of well-managed homework and feedback. Therefore future improvement initiatives should perhaps concentrate on developing innovative and efficient ways to scale up these aspects of learning for classes of over 50 students.

Regarding perceptions in medium versus large classes (Figure 5c), strategies that are more easily scalable for large enrolments were considered more or equally helpful in large classes, including most information strategies, lecture presentations, clickers questions, clicker question discussions, demonstrations, Socratic lecturing, solo studying, online assignments, and rubrics. Apparently, our concerted efforts to scale these pedagogic practices for large classes has been at least somewhat successful. However, data indicate that other strategies that are more challenging to deliver to large classes continue to be more helpful in medium classes. Examples include in-class help from instructors and teaching assistants, classroom-based activities, projects and feedback. These results are consistent with Freeman et al. (2014) who found in their metastudy of 225 reports, that active learning had the strongest effects on student success in classes with fewer than 50 students and the weakest effects in classes of over 110 students. We conclude that students do recognize when strategies are less successfully scaled for larger enrolments and therefore student perceptions data can help identify strategies to prioritized for further research and development.

As mentioned above, perceptions from individual courses can differ from perceptions averaged over several courses. This is also true for specific strategies. Some that were generally perceived as ‘unhelpful’ were in fact perceived as particularly helpful in certain courses that featured these strategies. Therefore using student perceptions for evaluating improvement initiatives benefits from both overarching interpretations based on aggregate results and from simultaneous attention to the details in specific courses.

Recommendations

Can perceptions of students help identify how to make further improvements? In general, a strategy may be perceived as unhelpful because it was present but poorly implemented, because it was not present, or because students do not always recognize when a strategy is in fact helping them learn. Regardless of why students perceived a strategy as unhelpful, their opinions do suggest areas that deserve further targeted effort. Conveying such indicators to instructors and the department is an important follow-up action.

For example, perceptions from students in multi-section courses suggested that perceived helpfulness was usually independent of the section in which students were enrolled. However this was not the case for one course, indicating that student perceptions data can help identify when consistency of specific learning experiences needs improving in courses with multiple sections.

Also, aggregate results suggest that strategies involving homework and feedback were generally perceived as less helpful than both information and classroom strategies, especially in medium and large classes. Regarding specific strategies, several widely used evidence-based instructional practices were perceived favourably (eg. clickers and in-class activities), however others were rarely experienced or underappreciated, including assessment of peers, reflective practices, use of online discussions or wikis, and (surprisingly) text books, pre-readings, scientific or professional readings and ‘other’ readings. Evidently we should be working on making readings of all types more useful to students.

In order to make specific recommendations about individual courses available to interested instructors, we prepared course-specific summaries of results within a few weeks of gathering the data. Average perceptions of helpfulness were prepared and sorted for each course, then presented in a standard two page report with workloads, enthusiasm and multiple instructor results. These reports became useful catalysts for informal discussions about what students found most or least helpful. They also helped identify practical priorities for adjustments that suited individual instructor’s interests and abilities.

Further studies that are building upon this work include use of an adaptation of SLES to track longitudinal changes in student perceptions as improvements are implemented in courses offered in both face to face and distance learning settings. Also, subsets of SLES questions have been adapted for use as midterm checks on perceptions of small scale efforts to improve specific strategies in individual courses. Other studies involving these data are under way and will be reported in separate articles. In particular, student perceptions expressed by SLES data have been compared to observations of classroom practices using the Classroom Observation Protocol for Undergraduate STEM (Smith et al. 2013; Lund et al. 2015), to instructors’ teaching practices reported using the Teaching Practices Inventory (C. Wieman and Gilbert 2014), to the independently assessed level of experience in evidence-based instructional practice for each instructor, and to institution-mandated general instructor evaluations that are gathered for all courses at the institution.

Conclusions

This study helped evaluate our department’s education improvement initiative by characterizing students’ perceptions of specific learning strategies in terms of the improvement model, class size and year level of each course. The first conclusion is that students in unimproved courses perceived most strategies as less helpful compared to students in courses improved with either of three improvement models. The second conclusion is that most strategies were perceived as much more helpful in small courses than in medium or large courses. Third, most strategies in small courses improved using transformation or consulting models were more helpful than those in independently improved small courses, but the opposite was true in medium sized courses. In large first year courses improved with the transformation model, strategies overall were at least as helpful as those in medium sized second, third or fourth year courses. However, information-related strategies (including learning goals) were generally perceived to be more helpful in larger classes, while active in-class strategies and most forms of homework or feedback were more helpful in medium classes.

Regarding the 39 specific strategies, both evidence-based and more traditional instructional practices were perceived to be more helpful in improved courses compared to unimproved courses. Also, neither evidence-based nor traditional strategies dominated as ‘most helpful’. Students perceptions were found to depend upon the specific situations of individual courses, so evaluating the general impacts of a department-wide improvement initiative requires attention to both individual situations and aggregate results from as many courses as possible. Strategies identified as ‘unhelpful’ represent excellent indications of where further effort can be directed to improve the exposure to best practices, the effectiveness of those strategies and students’ abilities to recognize them as helpful.

There are three implications. First, science education specialists helped instructors who were less experienced in evidence-based instructional practices deliver courses in which students perceive experiences as similarly helpful compared to courses delivered by instructors who were already expert in evidence-based instructional practices. Second, it affirms the benefits of having expert pedagogues in a department who can independently elevate the teaching practices of their courses and set the bar for others who can then improve with help from science education specialists. Third, students’ self-reported perception data did provide valuable data that contributed to evaluation of the impacts of our department-scale education enhancement initiative.

References

Abdi, H. 2007. “The Bonferonni and Šidák Corrections for Multiple Comparisons.” In *Encyclopedia of Measurement and Statistics*, edited by N. Salkind, 3:103–107. Thousand Oaks (CA): Sage. http://wwwpub.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf.

Ambrose, S. A., M. W. Bridges, M. DiPietro, M. C. Lovett, and M. K. Norman. 2010. *How Learning Works: Seven Research-Based Principles for Smart Teaching*. 1sted. Jossey-Bass.

Boud, D., and E. Molloy, eds. 2013. *Feedback in Higher and Professional Education: Understanding It and Doing It Well*. Abingdon, Oxon ; New York, NY: Routledge.

Chasteen, S. V, K. K Perkins, P. D Beale, S. J Pollock, and C. E Wieman. 2011. “A Thoughtful Approach to Instruction: Course Transformation for the Rest of Us.” *Journal of College Science Teaching* 40 (4): 70–76.

Chasteen, S. V, K. K. Perkins, W. J. Code, and C. E. Wieman. 2016. “The Science Education Initiative: An Experiment in Scaling up Education Improvements at a Research University.” In *Transforming Institutions: Undergraduate STEM Education for the 21st Century*, edited by Gabriela C Weaver, Wilella D Burgess, Amy L Childress, and Linda Slakey. West Lafayette, Indiana: Purdue University Press. http://www.thepress.purdue.edu/titles/format/9781557537249.

Carl Wieman Science Education Initiative, 2016. “EOAS CWSEI Project Overview and Outcomes.” Accessed February 19, 2016. http://cwsei.ubc.ca/departments/earth-ocean.htm.

de Winter, J.C.F., and D. Dodou. 2010. “Five-Point Likert Items: T Test versus Mann-Whitney-Wilcoxon.” *Practical Assessment, Research & Evaluation* 15 (11): 1–12.

Dommeyer, C. J., P. Baum, R. W. Hanna, and K. S. Chapman. 2004. “Gathering Faculty Teaching Evaluations by in-Class and Online Surveys: Their Effects on Response Rates and Evaluations.” *Assessment & Evaluation in Higher Education* 29 (5): 611–23. doi:10.1080/02602930410001689171.

Fairweather, J., J. Tarpani, and K. Paulson. 2016. “The Role of Data in Promoting Institutional Commitement to Undergarduate STEM Reform: The AAU STEM Initiative Experience.” In *Transforming Institutions: Undergraduate STEM Education for the 21st Century*, 429–38. West Lafayette, Indiana: Purdue University Press.

Freeman, S., S. L. Eddy, M. McDonough, M.K. Smith, N. Okoroafor, H. Jordt, and M. Wenderoth. 2014. “Active Learning Increases Student Performance in Science, Engineering, and Mathematics.” *Proceedings of the National Academy of Sciences* 111 (23): 8410–15. doi:10.1073/pnas.1319030111.

Handelsman, M. M., W. L. Briggs, N. Sullivan, and A. Towler. 2005. “A Measure of College Student Course Engagement.” *The Journal of Educational Research* 98 (3): 184–92. doi:10.3200/JOER.98.3.184-192.

Henderson, C., A. Beach, and N. Finkelstein. 2011. “Facilitating Change in Undergraduate STEM Instructional Practices: An Analytic Review of the Literature.” *Journal of Research in Science Teaching* 48 (8): 952–84. doi:10.1002/tea.20439.

Jones, F., and S. Harris. 2012. “Benefits and Drawbacks of Using Multiple Instructors to Teach Single Courses.” *College Teaching* 60 (4): 132–39. doi:10.1080/87567555.2012.654832.

Jones, F., 2014. “Students’ Learning Experiences Survey.”, Retrieved May 4, 2016, from https://open.library.ubc.ca/cIRcle/collections/facultyresearchandpublications/37212/items/1.0300444 .

Kruger, J, and D Dunning. 1999. “Unskilled and Unaware of It: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self-Assessments.” *Journal of Personality and Social Psychology* 77 (6): 1134, 1121.

Lund, T. J., M. Pilarz, J. B. Velasco, D. Chakraverty, K. Rosploch, M. Undersander, and M. Stains. 2015. “The Best of Both Worlds: Building on the COPUS and RTOP Observation Protocols to Easily and Reliably Measure Various Levels of Reformed Instructional Practice.” *CBE Life Sciences Education* 14 (2). doi:10.1187/cbe.14-10-0168.

Marsh, H. W. 2007. “Students’ Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness.” In *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, 319. Springer New York.

McCormick, A. C., R. M. Gonyea, and J. Kinzie. 2013. “Refreshing Engagement: NSSE at 13.” *Change: The Magazine of Higher Learning* 45 (3): 6–15. doi:10.1080/00091383.2013.786985.

McCrickerd, J. 2012. “Understanding and Reducing Faculty Reluctance to Improve Teaching.” *College Teaching* 60 (2): 56–64.

Procter, S., J. Irwin, D. Mazhindu, J. Nayoan, and F. Smith. 2015. “Increasing Understanding of the Best Ways to Collect and Use Feedback from Students and Trainees in Order to Improve the Quality of Education and Training.” http://eprints.bucks.ac.uk/1669/.

R Core Team. 2015. “R: A Language and Environment for Statistical Computing.” R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Reid, L. 2012. “Redesigning a Large Lecture Course for Student Engagement: Process and Outcomes.” *Canadian Journal for the Scholarship of Teaching & Learning* 3 (2): 1–37.

Richardson, J. T. E. 2005. “Instruments for Obtaining Student Feedback: A Review of the Literature.” *Assessment & Evaluation in Higher Education* 30 (4): 387–415. doi:10.1080/02602930500099193.

Seymour, E., D. J. Wiese, A. Hunter, and S. M. Daffinrud. 2000. “Creating a Better Mousetrap: On-Line Student Assessment of Their Learning Gains.” In *National Meeting of the American Chemical Society*. https://aacu-secure.nisgroup.com/resources/sciencehealth/documents/Mousetrap.pdf.

Smallwood, R. A., and J. Oiumet. 2009. “Measuring Student Engagement at the Classroom Level.” In *Designing Effective Assessment: Principles and Profiles of Good Practice*, edited by T. W. Banta, E. A. Jones, and K. E. Black, 193–97. San Francisco: Jossey- Bass.

Smith, M. K., F. H. M. Jones, S. L. Gilbert, and C. E. Wieman. 2013. “The Classroom Observation Protocol for Undergraduate STEM (COPUS): A New Instrument to Characterize University STEM Classroom Practices.” *Cell Biology Education* 12 (4): 618–27. doi:10.1187/cbe.13-08-0154.

Struyven, K., F. Dochy, and S. Janssens. 2005. “Students’ Perceptions about Evaluation and Assessment in Higher Education: A review1.” *Assessment & Evaluation in Higher Education* 30 (4): 325–41. doi:10.1080/02602930500099102.

Sullivan, D. F., and C. G. Schneider. 2015. *The VALUE Breakthrough: Getting the Assessment of Student Learning in College Right*. Association of American Colleges & Universities.

Weaver, G. C, W. D Burgess, A. L. Childress, and L. Slakey. 2016. *Transforming Institutions: Undergraduate STEM Education for the 21st Century*. West Lafayette, Indiana: Purdue University Press.

Welsh, A. J. 2010. “Considering the Student Perspective: Factors That Undergraduates Perceive as Influential to Their Academic Performance in Science.” Master’s thesis, University of British Columbia, Vancouver. https://circle.ubc.ca/bitstream/id/98388/ubc_2010_fall_welsh_ashley.pdf.

Wieman, C., and S. Gilbert. 2014. “The Teaching Practices Inventory: A New Tool for Characterizing College and University Teaching in Mathematics and Science.” *Cell Biology Education* 13 (3): 552–69. doi:10.1187/cbe.14-02-0023.

Wieman, C., L. Deslauriers, and B. Gilley. 2013. “Use of Research-Based Instructional Strategies: How to Avoid Faculty Quitting.” *Physical Review Special Topics - Physics Education Research* 9 (2). doi:10.1103/PhysRevSTPER.9.023102.

Wieman, C., K. Perkins, and S. Gilbert. 2010. “Transforming Science Education at Large Research Universities: A Case Study in Progress.” *Change Magazine* March/April 2010 (March). http://www.changemag.org/Archives/Back%20Issues/March-April%202010/transforming-science-full.html.

Wilson, K. L., A. Lizzio, and P. Ramsden. 1997. “The Development, Validation and Application of the Course Experience Questionnaire.” *Studies in Higher Education* 22 (1). doi:10.1080/03075079712331381121.

Zumrawi, A.A., S. P. Bates, and M. Schroeder. 2014. “What Response Rates Are Needed to Make Reliable Inferences from Student Evaluations of Teaching?” *Educational Research and Evaluation* 20 (7–8): 557–63. doi:10.1080/13803611.2014.997915.