Online Evaluation of Courses Initiative:

2012-2013 Pilot Completion Report

November 2013

Prepared for

Dr. Catherine Koshland

Vice Provost

Division of Teaching, Learning, Academic Planning, & Facilities

University of California Berkeley

Acknowledgements

Contributors to this Report included Dr. Mike Higgins, Postdoctoral Fellow at Woodrow Wilson School, Princeton University; Justin Lipp, Senior Instructional Designer, Educational Technology Services; Dr. Deborah Nolan, Professor of Statistics; and Dr. Laura Stoker, Associate Professor of Political Science.

Table of Contents:

I. Executive Summary of OEC Work for 2012-2013

II. Fall and Spring EvalSys Pilots

A) Observations of Response Rates Overall

B) Differences in Response Rate by Course Type

C) Challenges Encountered by OEC Initiative with Department Pilots

III. OEC Comparative Study Findings

A) Study Design and Procedures

B) Response Rates

C) Non-Response Bias

D) Response Completeness

E) Analysis of Rating Responses

F) Extrapolating from the Main OEC Experimental Study Results

IV. Instructional Format Questionnaire Item Analysis

A) Instructional Format Analysis Methodology

B) Proposal for Configurable Evaluation Templates

V. OEC Request for Proposal and Procurement

VI. Conclusion

I. EXECUTIVE SUMMARY

The Online Evaluation of Courses (OEC) Initiative has successfully conducted four full semesters of paper, online, and mixed-methods course evaluation pilots since its inception in Fall 2011. For the 2012-2013 academic year, nearly 11,000 evaluations from courses representing 10 separate academic units (including all of the evaluations for two entire departments) were distributed completely online, with 67.9% of the students returning evaluations over the course of the year. The OEC working group successfully implemented a business process for conducting course evaluations through the campus learning management system, bSpace, resulting in no recorded user complaints during the Spring 2013 term. However, we continued to face several administrative and technical challenges in using bSpace to implement online course evaluations for co-taught and cross-listed courses as well as for courses with complex results-reporting requirements. As a consequence, the OEC working group convened a diverse group of campus stakeholders from academic, technical, and administrative units at Berkeley to develop a Request For Proposal (RFP) for commercial course evaluation vendors and to select a system. This process resulted in the selection of a best-in-class solution from an experienced provider. A contract has been established with Explorance to license its evaluation product, Blue. Blue will be in place for a limited release with three academic departments for Fall 2013, with a wider campus rollout planned for a later date.

Finally, the OEC working group, in conjunction with faculty from the steering committee, the Office of the Registrar, and the Committee for the Protection of Human Subjects, successfully designed and conducted an experimental study of the online vs. in-class mode of course evaluation administration. The study included a sample of roughly 900 students from 10 upper- and lower-division courses in 3 departments (referred to as Departments A, B, & C in this

report to protect individual instructors). The study design allows us to compare the two evaluation modes in terms of the percentage of students who returned evaluation, the characteristics of those who did or did not respond, and the evaluation responses among those who did respond. The results, discussed in more detail below, showed several key findings.

- Response rates were comparable across the modes overall. However, they also suggest that response rates may decline when online evaluations are used in courses where class attendance is mandatory or otherwise high, and rise in courses where attendance rates are low.

- The characteristics of those who return evaluations were also comparable across the modes, with non-response more common in each mode among students who are struggling, academically.

- The only difference across modes in the pool of those who completed evaluations concerned the student's presence or absence in class on the day that evaluations were handed out. About a quarter of the online evaluations were received from absent students, compared to none of the in-class evaluations.

- With respect to the responses provided, we find that online evaluations tended to be both less complete—with more questions skipped and shorter responses provided to open-ended questions—with lower ratings provided as well. The least positive assessments tend to come from students in the online pool who were not in class when in-class evaluations were handed out.

II. OVERVIEW OF FALL 2012 AND SPRING 2013 EVALSYS PILOTS

The Fall 2012 and Spring 2013 OEC pilots were very successful in many respects. The open-source system, EvalSys, built for the bSpace (Sakai) learning management system successfully handled the deployment of nearly 11,000 total evaluations over both semesters, with periodic notification emails sent to both students and faculty as appropriate. Further, with workflow issues ironed, we received zero help tickets for the Spring 2013 evaluation cycle. The pilots included a total of roughly 400 primary and secondary course sections from a total of ten different academic units; though the bulk of these courses came from two primary departments (Department A and Department B). Of these evaluations, two-thirds were returned each semester (Fall: 4106/6070 = 67.6%; Spring: 3277/4802 = 68.2%). Evaluations utilized questionnaires tailored "instructional format" of the course (of 13 forms, only 7 were utilized).

Observations of Response Rates Overall

Looking inside the numbers, we see there appears to be a systematic difference in response rate between the two primary participating departments in the Fall and Spring pilots, Department A and Department B (see Table 1).[1] Overall, Department B exhibited lower response rates both semesters as compared to Department A.

Table 1. Fall and Spring OEC Pilot Response Rates by Department

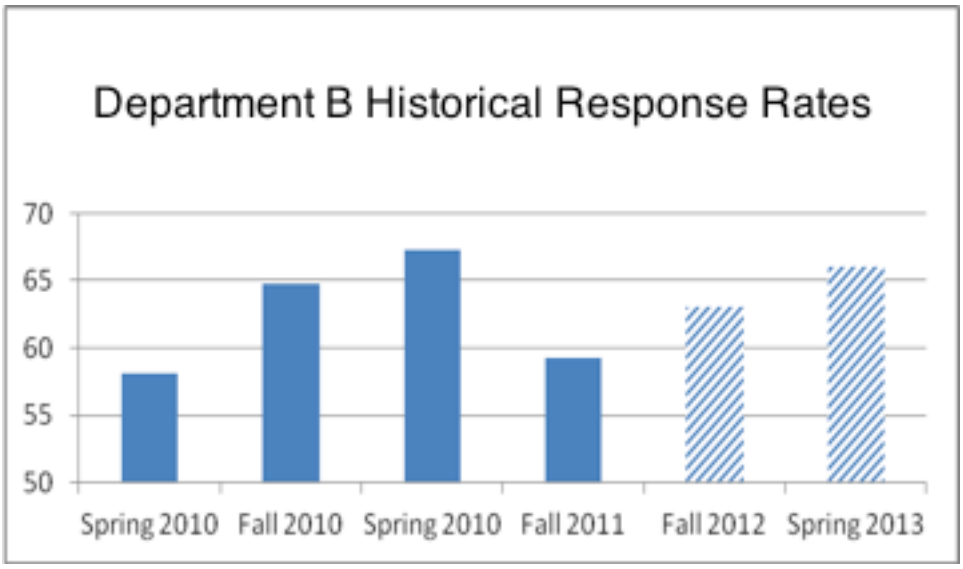
Department	Fall 2012			Spring 2013			Full Year
	# Completed	*# Enrolled*	*% Completed*	*# Completed*	*# Enrolled*	*% Completed*	*% Completed*
Department A	1119	1333	83.9	819	1078	76	**80.4%**
Department B	2987	4737	63.1	2458	3724	66	**64.4%**
Combined	**4106**	**6070**	**67.6%**	**3277**	**4802**	**68.2%**	**67.9%**

[1] The numbers of students are smaller for Spring 2013 than for Fall 2012 because in Spring 2013 we selected 8 courses from these departments (roughly 600 students) for inclusion in the experimental face-to-face/online study. The experimental study is discussed at length in Section III.

Several possible explanations for these differences emerged from formal interviews with instructors and informal discussions with staff. First, course size is a major differentiator between these two departments. Average course size in Department B is approximately double for what it is for Department A (34.5 vs. 17.3). Across all classes, we have observed a mild negative correlation between response rate and course size among these departments over both semesters ($r = -.13$, $p < -.01$). Furthermore, some Department B instructors reported that attendance was rather low in their courses (for both lectures and sections) and that attendance was not mandatory in large lecture courses. By contrast, the typical undergraduate Department A class meets 4-5 times per week and requires attendance. According to Department A's undergraduate advisor, "Our…classes are quite impacted so we are quite strict about attendance policy. From the third week on, the enrolled students are allowed 3 excused absences; that is, those absences won't affect their grade. Any more than those 3 absences and the instructor begins to subtract points." These differences between the departments may explain some of the disparity observed in response rates.

We asked Department B in Fall 2012 to provide the OEC working group with a list of historical instructor response rates; Department B has been independently maintaining these records in order to track their results over time (see Figure 1). The 2-year average response rate for paper-and-pencil evaluations in Department B was 62.4%, with only minor fluctuation over time. That rate is slightly lower than the department's average for the 2012-2013 online evaluation pilot of 64.5%, although the difference is within the margin of error (SE of difference = 3.72%). Overall, this comparison suggests that the online response rate for this department is in line with the response rate typically found for offline course evaluations.

Figure 1. Department of Department B Course Evaluation Response Rates: 2010-2011 Paper (Solid Bars) and 2012-2013 (Striped Bars)



Differences in Response Rate by Course Type

With regard to the thirteen instructional format questionnaires, no meaningful differences in response rate were observed in the pilot data. Generally, response rate did not vary within formats across semesters, or within departments for that matter. Department A did have somewhat higher response rates than Department B, but this difference mirrored the same overall trend between departments. See Table 2 for full reporting regarding instructional formats.

Table 2. Instructional Formats Response Rates for OEC by Semester 2012-2013

Instructional Formats	Fall 2012			Spring 2013		
	# Completed	*# Enrolled*	*% Completed*	*# Completed*	*# Enrolled*	*% Completed*
Clinic (Department B)	2	4	50	6	9	66.7
Discussion (Department B)	994	1556	63.9	1038	1591	65.2
Lab (Department B)	220	355	62	236	362	65.2
Seminar (Department B)	62	94	66	7	8	87.5
Lecture (Department B)	1709	2728	62.6	1171	1754	66.8
Language (Department A)	607	721	84.2	431	587	73.4
Lecture (Department A)	423	515	82.1	313	389	80.5
Seminar (Department A)	89	97	91.8	37	54	68.5
Writing (Only evaluated in Department A)	N/A	N/A	N/A	38	48	79.2

One differentiation in response rate occurred between primary (i.e., courses generally taught by professors and/or lecturers) and secondary sections (i.e., sections generally taught by GSIs and attached to a primary section). In general, primary sections elicited a somewhat greater rate of response relative to secondary sections (p-values for both Fall 2012 and Spring 2013 are near zero), suggesting some systematic preference among students to more likely evaluate course professors than GSIs. These data show how much larger departments like Department B use secondary GSI-led sections than smaller ones like Department A, with half of all completed evaluations in Department B coming from GSI-secondary sections and virtually none in Department A. See full results in Table 3.

Table 3. Response Rates by Department & Course Type 2012-2013

Department A	**Fall 2012**			**Spring 2013**			**Full Year Combined**
	# Completed	*# Enrolled*	***% Completed***	*# Completed*	*# Enrolled*	***% Completed***	***% Completed***
Primary	1119	1333	**83.9**	819	1078	**76**	**80.3**
Secondary	16	24	**66.7**	0	0	**0**	**66.7**
Undergrad	1062	1272	**83.5**	788	1078	**76**	**78.7**
Graduate	54	58	**93.1**	31	43	**72.1**	**84.2**
Department B	**Fall 2012**			**Spring 2013**			**Full Year Combines**
	# Completed	*# Enrolled*	***% Completed***	*# Completed*	*# Enrolled*	***% Completed***	***% Completed***
Primary	1773	2826	**62.7**	1184	1771	**66.9**	**64.3**
Secondary	1214	1911	**63.5**	1274	1953	**65.2**	**64.4**
Undergrad	2732	4405	**62**	2291	3448	**66.4**	**64**
Graduate	255	332	**76.8**	167	276	**60.5**	**69.4**

A final area of consideration in this dataset is the breakdown in graduate versus undergraduate courses. Overall, graduate students completed evaluations at slightly higher rates in both departments than undergraduate students. However, the broader trend of higher response rates in Department A as compared to Department B holds for both graduate and undergraduate classes. The largest change in response rate for any group was among Department B graduate

classes between Fall and Spring, with a one-third drop in participation, despite a similar number of total enrollments in both semesters. Any reason for this comparative drop in response rate is not apparent from other elements in the data.

Challenges Encountered by OEC Initiative with Department Pilots

The OEC working group encountered several important issues with regard to accurately deploying electronic course evaluations during the 2012-2013 pilots. These issues include customizing instructional format questionnaires by departments, handling multiple instructor (i.e., co-taught) courses, managing evaluations for cross-listed courses, and controlling permissions to access the results of evaluations.

Department A requested a customized set of evaluations for the courses in their department. Due to timing limitations, we were not able to accommodate those requests for the Fall 2012 term, but we did make the requested changes for Spring 2013. These changes largely involved the addition of five extra questions deemed useful to the department's language instruction. Interestingly, these questions were inserted into all of the department's courses, regardless of instructional format (including language, lecture, writing, and seminar) as a similar set of teaching issues is shared across all courses taught in Department A).

This in-depth customization for Department A informed a broader reorganization of the questionnaires for all courses. In the view of the OEC working group, the various instructional formats include three common themes for quantitative items, those being: 1) instructor, 2) course, and 3) self-evaluation question categories. By taking existing items and topically reorganizing them into these three categories, we feel there is a benefit to the overall organization of all 13 instructional format questionnaires by consistently priming students to

consider a similar set of items in sequence. This scheme was also used in a subsequent qualitative analysis of the instructional formats detailed at length in Section IV.

Department B presented a complex set of administrative challenges, mostly relating to data management regarding co-taught and cross-listed courses. During the Fall term, some students were given as many as 3 separate evaluations for their co-taught courses (one for each listed instructor, meaning that items pertaining to the "course" were repeated for each). Using the new question categorization scheme developed with Department A as a base, the process changed to repeat items specific to each instructor's individual performance thus only including items about the course once. This greatly shortened the survey completion process for students in co-taught courses by eliminating identical questions on different evaluations. In one specific set of co-taught Department B courses (98/198), average response rates moved up from 61.5% in Fall to 66.2% in Spring. The fact that students had fewer redundant questions to complete may have helped boost response rates in this instance.

Cross-listed courses continue to be a challenge in both the administration of and delivery of evaluation results. This stems from the state of campus course data. Specifically, department schedulers vary in methods used to identify courses that might be similar across departments. Issues of this nature have created significant problems around accurately assigning evaluations automatically by the online system, particularly in the case of cross-listed and co-taught classes. During the 2012-2013 pilots, OEC encountered several such issues including a cross-listed Department B course where the GSI for the cross-listed non-Department B discussion section was not entered as instructor, despite being entered for the Department B course code (a clerical error by the department). Frequently for cross-listed sections, adequate information simply does not exist. So, in this example, had we not caught and manually assigned an evaluation for this

GSI, roughly half of his class (those registered under the non-Department B course code) would not have received one.

Another relatively common case is "unofficially" cross-listed courses (a.k.a. 'room-shares'), where the department does not officially link two course control numbers for different departments together. This creates a very problematic situation for delivering course evaluations accurately. In order to determine "cross-listed" status, the evaluation system administrators are presently required to query the campus databases for a list of such "room-shares" where the same room location, date, and time for the courses in question all line up for two or more course codes. However, not all departments may wish to have such room-shares evaluated as cross-listed courses. From an administrative perspective creates a very difficult case for OEC as no common set of criteria can be used to describe the behavior of this type of data for all departments. Presently this requires us to work with departments to identify how courses should be evaluated. In implementing the new campus-wide system for 2013-2014, examples such as these present an important case where campus policy on data standards requires further input from the OEC steering committee and campus stakeholders regarding how to evaluate courses like these.

A final issue in need of further consideration from the 2012-2013 OEC pilots is courses with complex administrative reporting requirements. Specifically, three course codes for Departments B, C, and a third one were all cross-listed as a "Big Ideas" course for the College of Letters and Science. This particular course had 3 officially listed instructors, 3 unofficially listed instructors (who were also evaluated), and several GSIs whose sections also taught cross-listed sections across the 3 departments. There was some communication exchanged with "Big Ideas" sponsors as to the evaluation of this course; they had a set of questions to be included in the evaluation. However, the OEC team did not receive this communication until after the online

evaluations went live and thus could not be modified. Further, L&S needed to receive copies of the evaluation results as well. Again, for special programs not tied to a specific academic department such as Big Ideas, DeCal, Freshman Seminar and others, there does not exist a single data element to delineate these special programs. In practice, this means that for now, students in such courses evaluated by OEC may complete two evaluations: online for their academic department and in many cases for these programs a separate paper evaluation. This example is highly illustrative, demonstrating a need to establish: (1) greater clarity over evaluation administration responsibility, (2) system-level technical mechanisms to allow for feedback/input into the construction of course evaluation questionnaires from such programs, and (3) better tracking and automation of reporting back to these program administrators. With the state of campus data, such as it is, there does not appear to be a solution to the problem of automating this process for the time being.

III. OEC EXPERIMENTAL STUDY

To provide further evidence on how in-class and online evaluations differ, the OEC team carried out an experimental study in Spring 2013, partnering with faculty from Statistics (Deborah Nolan) and Political Science (Laura Stoker) and a Ph.D. student from Statistics (Michael Higgins). The study was approved by the Committee for the Protection of Human Subjects and supplemented with data from the Office of the Registrar. It involved approximately 900 students from 10 courses—both lower-division and upper-division—across 3 academic departments (Departments A, B, & C).

Study Design and Procedures

Faculty members were individually approached about having their students participate and consent was obtained from each one in advance. On a day arranged in advance with the

instructor, members of the OEC team came to individual courses to distribute envelopes containing the students' names on the outside. Each envelope contained either a paper evaluation form to be completed immediately during class or directions for accessing the evaluation online. Which version a student received was determined at random. In all courses, participants in both conditions (in-class and online) completed the exact same questionnaire, with eleven, 7-point rating scales evaluating instructor, course, and self-evaluation areas, as well as three open-ended questions asking about strengths and weaknesses in the course; this was the same as the standard OEC "lecture" instructional format questionnaire. Students in the online condition received an identical set of notifications and reminders as those participating in the regular OEC pilot. Faculty members received transcribed and compiled results for both in-class and online students, and in a format similar to what they otherwise would have received had their students completed the regular OEC pilot.

Using data supplied by the Registrar, we created a dataset containing prior-term GPA, course grade, major, gender, age, class standing (freshman, sophomore, junior, senior), and ethnicity on each enrolled student in the 10 participating courses. We also created a random identification number for each student that could be associated with the student's ID number via a secure file; the data was then de-identified for the purpose of analysis. The random identifier was printed on the paper questionnaires of those filling out their evaluations in class and associated with the evaluations of those submitting the evaluation online, enabling us to determine who did or did not fill out an evaluation. This procedure also enabled us to determine who was or was not in class when the course evaluations were handed out.

The strength of the study design is that it allows us to: (a) compare response rates across administrative modes, (b) examine how non-respondents differ from respondents within and

across modes, and (c) examine how evaluation responses differ across modes. The data can be analyzed separately by department and, for larger courses, by class. Analysis suggests that the randomization was successful; there were no sizeable (or statistically significant) differences by mode on any variable for which we had data.

The major weakness of the study design is that it, by necessity, utilized an administrative procedure that is unlike what students ordinarily experience. Students filling out in-class evaluations found the evaluation form inside an envelope that had their name on it. However, the evaluation itself was anonymous, as usual, and this point was stressed to the students. Students who were in class but who had been assigned to the online condition received an envelope with a notice that they were being asked to fill out an online evaluation and that emails containing instructions would be forthcoming, which is not a typical procedure during course evaluation. Finally, all of the students in attendance were aware that their course evaluations were part of a research project. These features may have had consequences for their responses (e.g., made the students more or less likely to respond, or more likely to answer questions in a given way), but whether that is so cannot be determined.

A further limitation of the study design is that it only involves ten courses in three departments. To be able to generalize to the Berkeley campus as a whole would require a more representative sampling of classes. In what follows, we primarily discuss the findings as they pertain to this sample, although we conclude with a limited set of extrapolations.

Response Rates

By definition, all students who returned in-class evaluations were present on the day that course evaluations were handed out (the "evaluation administration day"). However, not every student who was handed an in-class evaluation returned one. Of the students assigned to the in-

class condition, 63.3% were present, and 97.5% of those completed an evaluation, for a total response rate of 61.7%. Students did not, of course, need to be present on the evaluation administration day in order to return an online evaluation. However, we recorded whether or not they were present on that day. Of the students assigned to the online condition, 64.3% were present, and 70.1% of those completed an evaluation. Of the students who were absent (35.7%), 39.9% completed an evaluation. This brings the total response rate in the online condition to 59.3%, with about 76% of the responses coming from those who were present on the evaluation administration day and 24% coming from those who were absent.[2]

Thus, the overall response rates were comparable across the modes—61.7% in-class, 59.3% online—even though the composition of the students responding varied substantially by attendance. The difference of 2.4% is not statistically significant ($t = 0.763$, $p = .446$). Breaking the results down by course, however, yields a more nuanced conclusion.

Table 4 shows the response rates by mode for each course, with the exception of the small Department A courses (four primary sections in total), which are combined. A first point to make is that none of the mode differences are statistically significant at $p<.05$. However, there is only one course in which the online response rate is greater than the in-class response rate—Department C 1, a large, general education, lower-division lecture course that had low attendance on evaluation administration day.

Figure 2 shows how the gap in the response rate across modes relates to the percentage of students who were present on evaluation administration day. There are two ways to interpret the pattern we see in Figure 2. If we judge PS 1 to be an outlier, the remaining courses show a higher response rate for in-class evaluations than for online evaluations, regardless of attendance

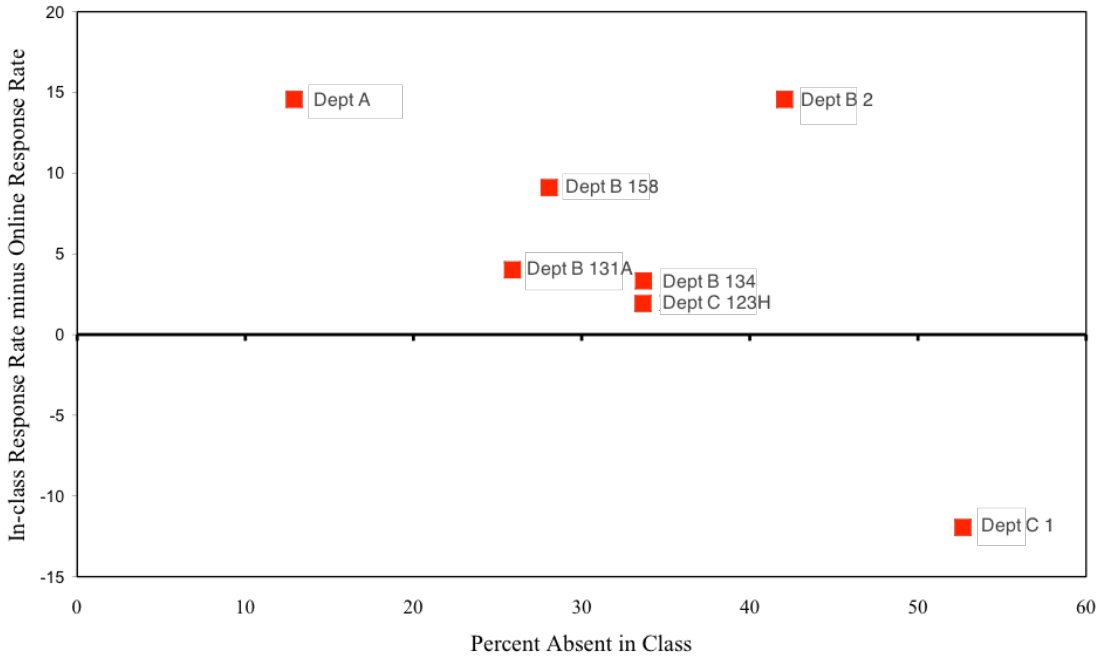
[2] Among students assigned to the online condition, 45.1% (.643*70.1=45.1) fell into the present & responded category, while 14.2% fell into the absent & responded category (.357*39.9=14.2%), for a total of 59.3% responding. Thus, 45.1/59.3 or 76% of the online respondents were present on the evaluation administration day.

Table 4. Response Rates by Mode and Course

Course	N	In-Class	Online	P-Value
Department C 1	243	43.0	54.9	.063
Department C 123H	104	65.4	63.5	.840
Department A (combined)	93	84.8	70.2	.095
Department B 2	145	63.9	49.3	.078
Department B 131A	170	72.6	68.6	.568
Department B 134	95	59.6	56.2	.746
Department B 158	57	64.3	55.2	.492
Total	*907*	*61.7*	*59.3*	*.446*

rates. Dropping the PS 1 cases and carrying out the mode comparison for the remaining courses yields a response rate difference of 7.8% (68.7% in-class vs. 60.9% online), which attains statistical significance (p = .036). On the other hand, Stat 2 could also be judged an outlier in the scatterplot. If Stat 2 is excluded, in-class evaluations response rate is strongly associated with how many students attend class. As the rate of attendance declines, so does in-class response rate, eliminating or even reversing the tendency for in-class response rates to exceed online ones.

Figure 2. Response Rate Difference across Modes, by Attendance



Thus, even though there are no overall significant differences in response rates across the modes, we believe that online course evaluations may possibly produce lower response rates than what is normally seen for in-class evaluations, especially, and perhaps only, for courses in which class attendance is typically high. However, for low attendance courses, one might expect higher total response rates using the online mode.

Non-Response Bias

The study design enables us to examine how non-responders and responders vary by a number of attributes, including age, sex, race/ethnicity, class standing at Berkeley, whether or not the student was a double-major, GPA as of the prior-term, and course grade. Table 5 shows these comparisons for each experimental group.

Table 5. Characteristics of Non-Respondents and Respondents, by Mode

	In-Class			*Online*		
	Non-Resp.	Resp.♦	P-Value	Non-Resp.	Resp.♦	P-Value
Age (mean, 16-42)	20.8	21.2	.208	21.1	21.0	.741
% Female	55.3%	63.2%	.099	55.6%	60.4%	.304
Class Standing (mean, 1-4)	3.2	3.4	.009	3.4	3.4	.960
% Double Major	6.4%	13.3%	.021	7.0%	12.9%	.043
Prior GPA (mean, 0-4)	3.15	3.35	.001	3.19	3.32	.009
Class Grade (mean, 0-4)	2.76	3.14	.001	2.75	3.09	.001
Race/Ethnicity						
% Caucasian	28%	32%	♣	26%	26%	♣
% Latino/Latina	13%	17%	—	16%	20%	—
% African American	5%	3%	—	3%	3%	—
% Chinese	20%	19%	—	23%	21%	—
% Other Asian	25%	22%	—	23%	23%	—
% Other/Missing	8%	8%	—	10%	8%	—

♣ The p-values for Race/Ethnicity are .637 and .894 for the in-class and online groups, respectively.

♦ The p-values on tests of differences in the characteristics of respondents in the in-class vs. online condition are: Age p=.381; % Female .513; Class Standing .722; % Double Major .891; Prior GPA .512; Class Grade .400; Race/Ethnicity .796.

Two findings stand out. First, the students who returned evaluations tended to be more high-performing, academically, than those who were non-responders regardless of administrative

mode. They entered the course with higher GPAs, earned higher grades in the course, and were more likely to be double-majors. These differences are present within both in-class and online samples. There are no other systematic differences between responders and non-responders. Although a significant difference by class standing arises in the in-class sample, no such difference is present in the online sample. Multivariate analyses predicting the probability of returning a course evaluation, which included all of the variables in Table 5 plus dummy variables for each course, yield a similar conclusion.[3]

Second, the sample of in-class responders is highly similar to the sample of online responders. On no variable shown in Table 5 is a difference statistically significant. A multivariate analysis examining whether the probability of returning an evaluation is influenced by different variables in the in-class sample than in the online sample yielded the same conclusion. In no case was there a significant difference across the experimental groups in the estimated effect of an independent variable on the probability of returning an evaluation.

Of course, as described earlier, the in-class and online responders do differ in one key respect—whether they were present or absent on the evaluation administration day. All of the in-class responders were present in class on the evaluation administration day while the online responders were a mix of those present (76%) and absent (24%). Because of this, we expanded the comparison to distinguish the two groups of online responders. Table 6 shows the characteristics of student who returned evaluations among (a) the in-class group, (b) the subset of the online group who were in class when evaluations were handed out, and (c) the subset of the online group who were absent when evaluations were handed out. On measures related to

[3] The estimated marginal effects on the probability of a response, from Logit analyses, are as follows. Full sample: Prior GPA, b=.12 (p=.012), Course Grade b=.07 (p=.003), Double-major b=.14 (p=.030). In-class sample: Prior GPA, b=.14 (p=.032), Course Grade b=.09 (p=.010), Double-major b=.12 (p=.223). Online sample: Prior GPA, b=.08 (p=.191), Course Grade b=.06 (p=.057), Double-major b=.16 (p=.078). Latinos/Latinas showed a modestly greater propensity to respond than did members of other race/ethnic groups. Course differences remained. No other variable showed significant effects.

academic performance, the absent group stands out. Compared to in-class and online respondents who were in class the day that evaluations were handed out, the absent group was less far along in their studies (3.2 vs. 3.4, 3.5 on class standing) though not younger, had a lower prior-term GPA (3.18 vs. 3.37, 3.35), and earned lower grades in the course (2.74 vs. 3.19, 3.14), on average. These variations within the online group of respondents may bear on how and why the evaluation responses differ across the modes, as discussed below.[4]

Table 6. Characteristics of Responders, by Mode and Class Attendance

	In-Class	Online & Present	Online & Absent	*P*-Value♦
Age (mean, 16-42)	21.2	20.9	21.2	.572
% Female	63.2%	61.3%	57.8%	.714
Class Standing (mean, 1-4)	3.4	3.5	3.2	.041
% Double Major	13.3%	14.6%	7.7%	.357
Prior GPA (mean, 0-4)	3.35	3.37	3.18	.020
Class Grade (mean, 0-4)	3.14	3.19	2.74	.001
Race/Ethnicity				
% Caucasian	32%	24%	32%	♣
% Latino/Latina	17%	17%	28%	—
% African American	3%	3%	0%	—
% Chinese	19%	22%	15%	—
% Other Asian	22%	26%	12%	—
% Other/Missing	8%	7%	12%	—

♦ The *p*-values shown for the first six variables come from F-tests. In the three cases where the F-test is significant, post-hoc T-tests indicate that the Online & Absent group is different from the other two groups.

♣ A chi-square test for Race/Ethnicity yields a *p*-value of .116.

Response Completeness

We evaluated the completeness of the questionnaires by considering the number of rating and open-ended questions answered as well as the word count of open-ended responses. The evaluation form asked eleven closed-ended, rating questions and three open-ended questions, as noted above. Students that completed an evaluation in class answered slightly more closed-

[4] These differences remain if we employ stricter definitions of respondents, e.g., those who returned an evaluation and answered all closed-ended questions, or those who returned an evaluation, answered all closed-ended questions, and answered at least one open-ended question.

ended questions per evaluation than did those filling out an evaluation online, on average (11.0 vs. 10.9, $p = .003$), with 99% and 93%, respectively, answering all eleven questions. Differences in responses to the open-ended questions were more substantial. Whereas 80% of the in-class evaluators answered all three open-ended questions (and only 5% answered none), just 55% of the online evaluators answered all three (and 27% answered none). On average, the in-class group answered 0.7 more questions (2.7 vs. 2.0, $p < .001$). These results are summarized in Table 7, which also breaks the overall results down by course. The tendency for online course evaluations to contain more missing data is evident within each course.

Table 7. Response Completeness by Mode and Course

Course	Rating Questions (Average Number Answered)			Open-Ended Questions (Average Number Answered)		
	In-Class	Online	P-Value	In-Class	Online	P-Value
Dept. C 1	11.0	10.9	.268	2.2	1.7	.034
Dept. C 123H	11.0	10.8	.247	2.9	2.1	.001
Dept. A (combined)	11.0	10.9	.144	2.7	2.3	.054
Dept. B 2	11.0	10.9	.047	2.7	2.1	.003
Dept. B 131A	11.0	10.9	.071	2.9	1.9	.001
Dept. B 134	11.0	11.0	.313	2.8	1.9	.004
Dept. B 158	11.0	10.9	.296	2.7	2.3	.241
Total	*11.0*	*10.9*	*.003*	*2.7*	*2.0*	*.001*

Course	Rating Questions (% Complete, 11/11)			Open-Ended Questions (% Complete, 3/3)		
	In-Class	Online	P-Value	In-Class	Online	P-Value
Dept. C 1	100%	96%	.124	62%	45%	.070
Dept. C 123H	97%	91%	.295	91%	58%	.002
Dept. A (combined)	100%	94%	.122	79%	70%	.346
Dept. B 2	100%	92%	.047	80%	64%	.096
Dept. B 131A	98%	92%	.087	89%	49%	.001
Dept. B 134	100%	96%	.313	82%	52%	.017
Dept. B 158	100%	94%	.296	83%	69%	.332
Total	*99%*	*93%*	*.001*	*80%*	*55%*	*.001*

Interestingly, while the online evaluators were less likely than the in-class evaluators to answer the open-ended questions, they gave longer responses when they did answer them. On

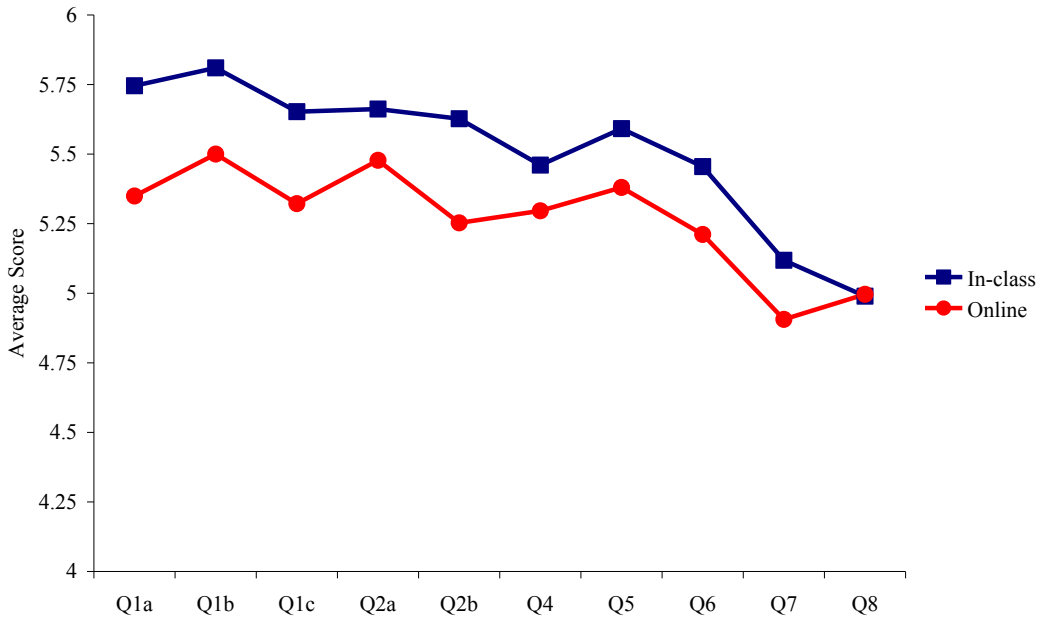
question 9, which concerns the strengths of the course, the online evaluators averaged 23 words per response while the in-class evaluators averaged 18 words. This gap of 5 words grows sizably to a gap of 20 words on question 10, which concerns areas where the course could be improved (42 vs. 22 words). On question 11, which allows students to provide feedback on the course to future students, the gap narrows to 7 words (24 vs. 17). These patterns suggest that online evaluators offered more critical or in-depth evaluations than the in-class evaluators, an inference that is supported by an analysis of the closed-ended responses, which we discuss next. Compositional differences across the two groups of evaluators may explain these differences. But students may also be more comfortable offering criticisms of the course when answering evaluation questions online.

Analysis of Rating Responses

In this section we report results on 10 of the 11 closed-ended questions, which asked students to provide ratings on a 1 (negative) to 7 (positive) scale: Q1a Explained concepts clearly; Q1b Presented organized lectures; Q1c Encouraged critical engagement; Q2a Effectively organized course; Q2b Assignments and lectures complementary; Q4 Instructor created conducive learning environment; Q5 Overall effectiveness of instructor; Q6 Overall effectiveness of course; Q7 How satisfied with course; Q8 How satisfied with own effort. The only question we do not analyze here, Q3, asked students to report on how many hours they worked per week. There were no differences across groups in responses to this question (means of 3.1 and 3.0 for the in-class and online groups, respectively). Because of the small number of cases within some courses, we aggregate across courses in our analysis. All of the results we present hold up if we factor out the differences across courses in mean evaluation ratings.

On average, students returning online evaluations evaluated the course and instructor lower than did the students returning in-class evaluations. Of the 10 closed-ended questions analyzed, only one—question 8—showed no differences at all, and that question asked how satisfied the student was with his or her own effort in the class. The differences on Q1a, Q1b, Q1c, Q2b, and Q6 are all statistically significant at $p<.05$. The p-values for the differences on Q2a, Q4, Q5, and Q7 are in the .08-.18 range. The p-value for Q8 is .952. Figure 3 illustrates these results.

Figure 3. Evaluation Responses by Mode

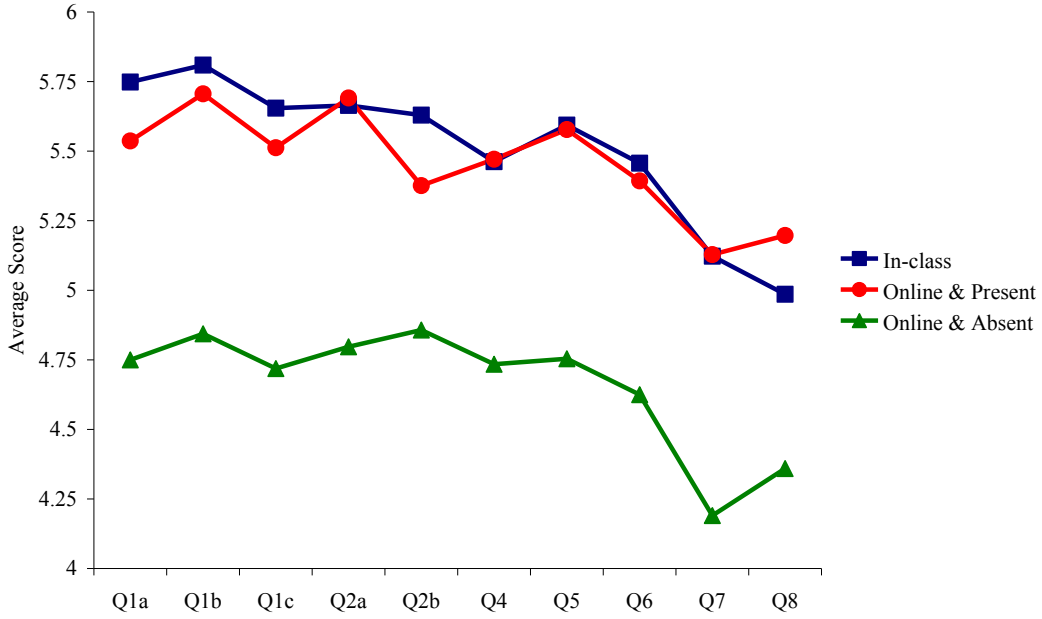


Note: Q1a Explained concepts clearly; Q1b Presented organized lectures; Q1c Encouraged critical engagement; Q2a Effectively organized course; Q2b Assignments and lectures complementary; Q4 Instructor created conducive learning environment; Q5 Overall effectiveness of instructor; Q6 Overall effectiveness of course; Q7 How satisfied with course; Q8 How satisfied with own effort

Because the online respondents include a mix of those present and absent from the class on evaluation administration day, whereas the in-class respondents were all present, we next considered whether this compositional difference was behind the lower ratings in online assessments. As Figure 4 shows, the in-class and online evaluation scores are quite similar

among the students where were present in class. Lower ratings in the online evaluations are primarily attributable to the assessments of those who were absent.

Figure 4. Evaluation Responses by Mode and Attendance



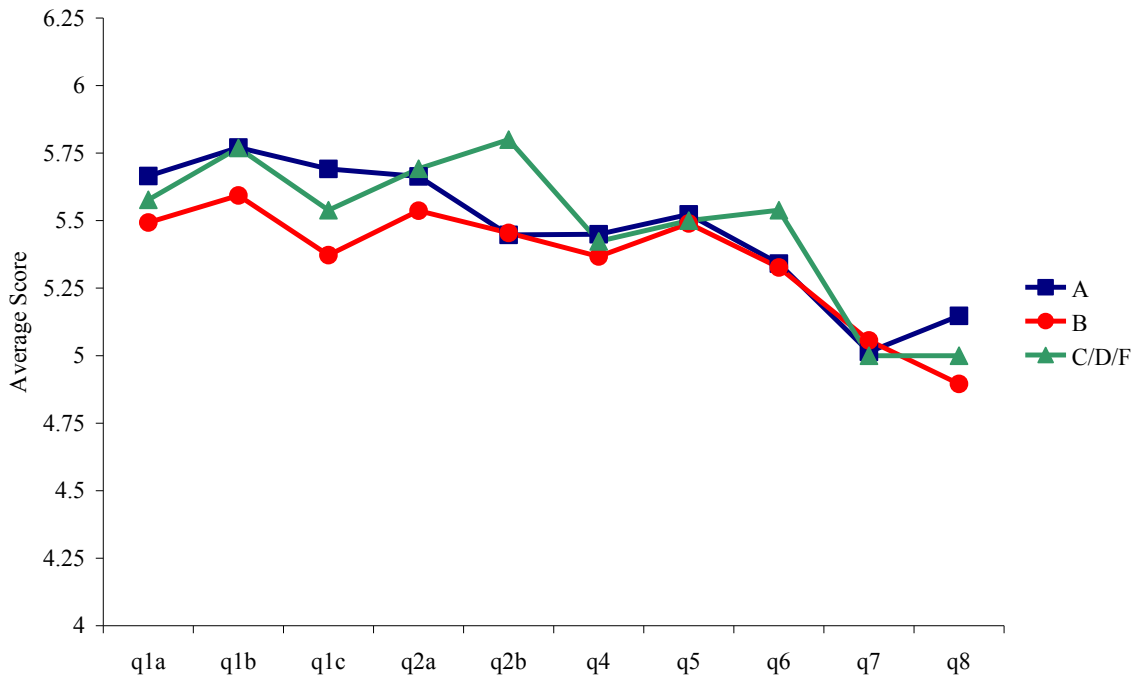
Note: Q1a Explained concepts clearly; Q1b Presented organized lectures; Q1c Encouraged critical engagement; Q2a Effectively organized course; Q2b Assignments and lectures complementary; Q4 Instructor created conducive learning environment; Q5 Overall effectiveness of instructor; Q6 Overall effectiveness of course; Q7 How satisfied with course; Q8 How satisfied with own effort.

The differences are large—on the order of one point—and evident in the responses to each evaluation question. T-tests comparing the "online & absent" group to either the "online & present" group or the "in-class" group yield *p*-values less than .001 in all but one instance (Q2b, comparing "online & absent" to "online & present," p=.028). By contrast, none of the differences between the "online & present" and "in-class" groups are large in size or statistically significant at p<.05. In other words, “online & absent” students disproportionately affected average question ratings in this sample.

As pointed out earlier, those who were absent from class on evaluation administration day and yet submitted an online course evaluation tended to have lower prior-term GPAs and,

especially, to have received worse grades in the class than did others returning evaluations (Table 6). As Figures 5 and 6 illustrate, evaluation scores are not associated with prior-term GPA, but are fairly strongly associated with course grades.[5] If we assume that these associations reflect a causal process whereby students doing poorly in the course are more likely to provide lower evaluations, then the poorer academic performance of the "online & absent" group could partly explain their more negative assessments. However, the difference between the absent and present groups persists in multivariate analyses that control for the course grade received.[6] It is, thus, just as plausible—or even more plausible—that the students' dissatisfaction or negative perception is responsible both for their absence and their more negative course evaluations.

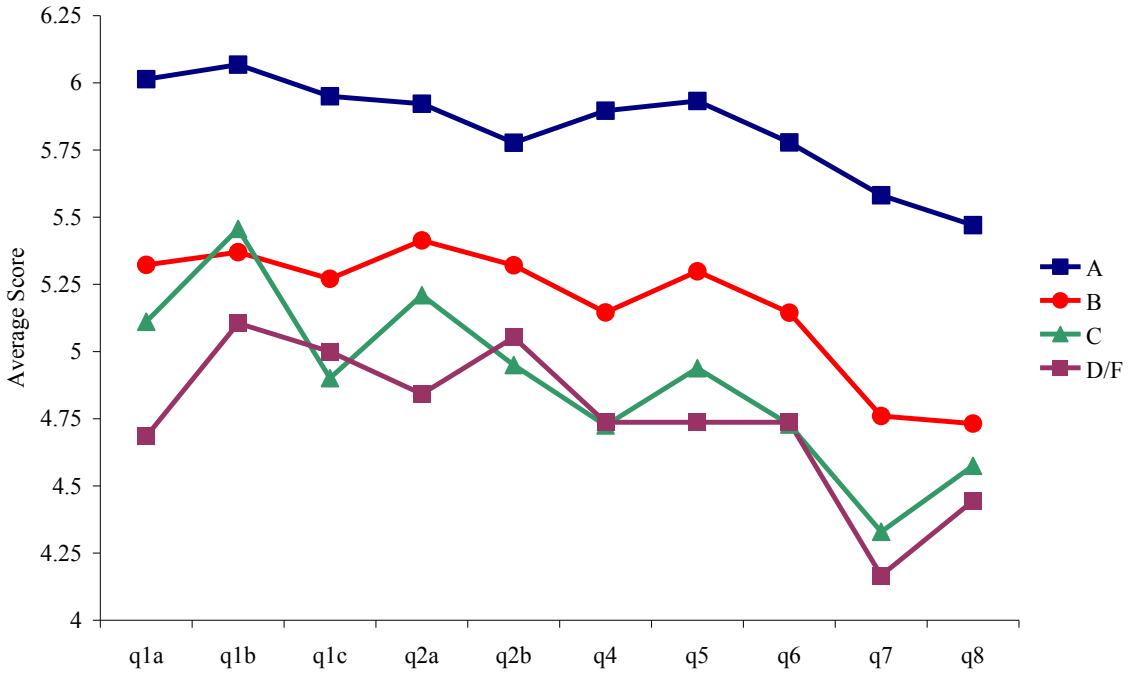
Figure 5. Evaluation Responses by Prior-Term GPA



[5] Figures 6 and 7 show the results with prior-term GPA and course grade grouped. Using the continuous measures, the Pearson's correlation between prior GPA and evaluation response averages .04 across the 10 measures, for the sample as a whole. For class grade, the average correlation is .25. Prior-term GPA and course grade are correlated at .51.

[6] We evaluated a number of specifications and functional forms, and in no case did the difference between the absent groups and the present groups disappear when class grade was controlled. Of course, these are naïve regressions since absence from class and course grade are presumably causally intertwined.

Figure 6. Evaluation Responses by Class Grade



Extrapolating from the Main OEC Experimental Study Results

Based on the experimental pilot study results, we believe that response rates for online evaluations are likely to be comparable to in-class response rates for many courses, but may go down for course with very high class attendance rates and go up for courses with very low attendance rates. The pool of online responders is likely to be similar in most respects to the pool of in-class responders, and to be just as biased toward those who have been doing relatively well in their prior courses and/or are performing relatively well in the class at hand. However, the online respondent pool includes some members of a group that in-class evaluations miss: those who are not present during course evaluation administration. That the online mode can pick up some responses from this group of students is quite useful, though the response rate for this group is likely to be quite low. These students tend to provide the most negative

evaluations, producing a tendency for online evaluations to be more negative than in-class evaluations on the whole. However, the inclusion of these students may represent a more ecologically valid method of course evaluation administration, as the opinions of this group of students are otherwise not captured via traditional in-class evaluations. Further, identifying how and why these students are disaffected may be aided by online evaluations, particularly with the use of midterm or formative assessments to make adjustments in the course before the end of the term.

IV. INSTRUCTIONAL FORMAT QUESTIONNAIRE ANALYSIS

After two-years of using the current thirteen forms, the exiting OEC model for conducting course evaluations has required adaptation to new information gleaned from our pilot experience, particularly in light of the level of customization sought by departments like Department A. The thirteen formats (activity, clinical with supervision, clinical without supervision, foreign language, lab (attached to lecture), lab (stand-alone), lecture, project, section (discussion), section (problem) seminar, studio performance, and writing) are organized around differences in the instructional content of a course, as defined by a large group of faculty focus groups during the initial phases of the OEC initiative. For instance, a physical education course evaluated using an "activity" form is instructionally very different from a chemistry lab. Additionally, given that the University has contracted to replace the bSpace EvalSys platform with Explorance's Blue evaluation system, the need for a rigidly defined set of course evaluation forms determined by instructional format may be somewhat unnecessary due to much greater technical flexibility in the new system to handle a wide range of department-level customizations.

Instructional Format Analysis Methodology

Having piloted these evaluation forms over four semesters with hundreds of courses in over a dozen academic departments, certain issues became apparent in both the consistency of items and organization of the forms themselves over the last two years. As per the aforementioned customization effort with Department A, items related to course, instructor, and student self-evaluation categories previously were interspersed together potentially creating confusion for students. Further, there is some redundancy across formats with subtle but systematic differences in question wording and inconsistent coverage in question themes. For instance, only 4 of the 13 questionnaires ask about instructor feedback to students, and among these, each item contains a different wording. From a data management perspective, automating the assignment of the current forms would be difficult. The 13 formats do not map to existing campus course data, and as such each format has had to be manually assigned thus far with the input of departmental staff and academic personnel. In some cases, the same class has been evaluated using two different forms between semesters depending on individual teaching personnel involved, causing potential downstream issues for data integrity, as items are inconsistent between questionnaires used to rate the same course.

Given these issues, the OEC working group conducted a systematic review of the current forms to explore ways to simplify the number of questionnaires and items given the current framework. Reviewing the instructional formats required a systematic procedure to evaluate all instruments as fairly as possible. As such, the OEC team chose to follow the Corbin and Strauss (2008) guidelines for qualitative 'axial coding' because of its rigor and well-regarded approach to qualitative text analysis. The Corbin and Strauss procedure involves evaluating a body of 'text' (i.e., instructional format question items) to identify emergent themes. Specifically,

analysis is concerned with "identifying, naming, categorizing, and describing phenomena found in the text" (p. 87). Analysis continues until the emergence of saturation across themes, which is the point where each theme contains a unique idea, discrete from other themes found in the overall body of text.

In brief, the axial coding procedure for analyzing the questionnaires involved "identifying, naming, categorizing, and describing phenomena found in the text" (Corbin & Strauss, 2008, p. 87). Analysis continues until the emergence of saturation across themes, or the point at which each theme is unique and discrete from other themes found in the corpus of text. The procedure here involved three steps; first all question items were plotted on one axis in a table against the instructional formats on the other axis to determine the extent of overlap among the forms for question items. This helped to determine that only rarely are questions repeated explicitly over all questionnaire types, but there is a degree of overlap in question content as suggested anecdotally earlier (e.g., four different wordings of an 'instructor feedback' item). This prompted a second step of grouping question items by evaluation subject (instructor, course, self-evaluation, and free response) and then further categorizing within each by question content (see Table 8 below). Questions were categorized by identifying key words and phrases that establish the basic premise of each question; similar items were then grouped together and ascribed an identifying label.

Table 8. Top-Level and Sub-Categories for Instructional Format Analysis of Rating Questions

Instructor Specific Items	Course Specific Items	Self-Evaluation Items
1) Presentation of Content	1) Course Content	1) Time Investment Estimates
2) Clarity of Expectations or Directions	2) Increased or Development of Skills/Knowledge	2) Personal Effort Satisfaction
3) Helpfulness & Availability	3) Useful/Clear Feedback on Performance	
4) Useful/Clear Feedback on Performance	4) Course Overall	
5) Encouraging of Participation & Discussion		
6) Overall Teaching Effectiveness		

In the third and final step, the new question content categories were plotted orthogonally against the existing instructional formats to determine the level of representation for each existing questionnaire within each question category. It is here where the current approach to structuring this existing set of questions by instructional format becomes very challenging. In attempting to streamline the number and complexity of the instructional formats, several intractable issues emerged. Firstly, the number of items and equivalence of question content categories varies between instructional formats. This manifests in practice with some questionnaires simply asking too many questions within certain content areas and not enough in others (See Appendix A: Question Category Frequency Counts). A good example is "Clarity of Expectations/Directions" items from the instructor-category. Here only 3 of the existing instructional formats are represented (Clinic with Supervision, Lecture, and Writing) and each asks a different version of a similar question. When looking by instructional format, there is a

range of between 7 and 10 of the 13 total question categories represented within the current questionnaires. Not a single instructional format asks questions within each category identified by this coding scheme.

Proposal for Configurable Evaluation Templates

After identifying these emergent categories, it was necessary to consider their validity. Within the existing constraints of the questionnaires, there is limited opportunity for further categorization, so the parsimony of the question categories seems to fit well within the current set of items. A final look at the face validity of the items themselves begs the question of whether assigning evaluations by instructional format is necessary given the current questions. With very few exceptions, the questions items articulated in this pool do not pertain to specific subject areas, disciplines, or approaches unique to one instructional format or another. Taking all this into consideration, the OEC working group recommends that in place of the current questionnaires, that OEC instead takes a different approach. Taking inspiration from both this analysis and the 2009 Taskforce on Teaching Evaluation's work in developing the question items, the OEC working group proposes to create a single generic evaluation instrument with items selectable within each question category and subcategory. This approach yields several important benefits.

Firstly, to deploy even a streamlined set of instructional format questionnaires would require the development and implementation of a complex new technical workflow that does not currently correspond to campus course data. Internally, the OEC team has discussed how instructional formats could be mapped to existing course metadata (TIE Codes), but any mapping with such sources would not be one-to-one. Even if a reasonable mapping could be made, there still remain other issues. As we have seen in the OEC pilots, individual departments will vary in

what they evaluate within an instructional format from term to term. For instance, Spanish R1A was evaluated as a lecture during Fall 2012, but R1B (the second semester in the sequence) was evaluated as a writing course during Spring 2013. Other instances where courses listed as a one thing in the schedule of (e.g., Department B 20 S 101 LAB) were evaluated using a different instructional format (e.g., discussion). Using available campus data sources, there is no reliable way to automate this kind of mapping. As many more departments and academic units come into the evaluation system, catching inaccurate mappings or manually reassigning large numbers of courses that were improperly assigned is potentially fraught with risk.

Comparatively, the benefits of a single, modifiable generic form are significant. Rather than presupposing that a particular course and instructional format mapping are correct, a more generic form can instead be amended or modified to reflect the particular needs of an academic department or school. And, if a department or school fails to act or declines to add any questions to a questionnaire, they can instead simply have a fully generic and transferable format that will allow for cross-comparison between and among courses in their department and school.

The proposed generic instrument (Appendix B) incorporates a single question from each of the identified question categories developed for the current instructional format analysis. One exception is that there were two questions drawn from the "Increased or Development of Skills/Knowledge" subcategory: one for both knowledge and skills separately as this was the second most numerous and most diverse question category in the analysis. This created a total of 13 Likert-type items, 2 interval questions (e.g., estimates of time-spent on course), and 3 free-response items for the questionnaire, netting 18 in total. The proposed generic instrument is somewhat lengthier than the typical OEC pilot questionnaires, which averaged closer to 15 questions. For ordinal time-estimate items from the student self-evaluation category, it was

decided to include both items. Originally, these items were geared toward the primary and secondary section time investment estimates. However, the question construction was sufficiently different that it was deemed useful to include slightly modified versions of both items. In some cases, question wordings were changed to better reflect the objectives of their question categories and to provide a more diagnostic response, working from the principles of survey questionnaire design. For instance, Lecture question #2 "The instructor: presented lectures in an organized manner" was changed to the slightly more generic "The instructor presented content in an organized manner". In this way the student can infer "content" to refer to both written and oral content.

The benefits of the proposed form are such that departments, should they choose to modify the items, can be done easily. One option to accommodate these could be to incorporate the present items from each category into a question bank. In this way, the items from the proposed form could be swapped out in favor of questions deemed better for a particular department. Given the flexibility of this proposed approach, a generic questionnaire with designated question slots for particular content areas seems both simpler, and requiring less significant adjustment to the original instructional formats approach developed by the Teaching Evaluation Taskforce.

Though slightly out of scope for this report, it is worth noting that this proposal has been presented to the COT as of September, with a follow-up discussion scheduled for early December. Several questions and recommendations have been exchanged as a result of this presentation, and the OEC working group is currently working with both the COT and our rollout partners in Department B, Department A, and Chemistry to conduct a test of this workflow for the Fall 2014 term.

V. OEC Request for Proposal and Procurement

Having evaluated the EvalSys bSpace (Sakai CLE) from pilots during 2011-2013 and deciding it could not provide a long-term solution for campus, we initiated a Request for Proposal (RFP) in March 2013. As stated in the executive summary, we recently completed the procurement process with the selection of Explorance's Blue (http://www.explorance.com/blue), a best-in-class technology solution that will meet the needs of the departments and administrative units across campus; it is a very flexible platform allowing for departmental and instructor-level evaluation customization, a feature we are very happy to be introducing to the campus this fall. Below, we will describe our process for leading a thorough and inclusive RFP process.

Having defined our requirements prior to piloting with the EvalSys (bSpace) system, the ETS team was able to confirm these requirements by way our recent department-level pilot efforts. Requirements gathering was supplemented with information provided by faculty and staff where necessary. We also aligned these requirements with Berkeley Extension goals, so they could ultimately use whatever system we selected. With assistance from the Office of Procurement, we authored the RFP, including functional and technical requirements as well as our selection criteria and review process. The RFP was posted April 1, 2013.

Additionally, a critical aspect of the review and selection process was the formation of and a University-wide review committee. The committee included staff from ETS, IST, Office of Planning and Analysis, the Office of Academic Personnel, the Center for Teaching and Learning, the Department of Department B as well as a faculty member from Chemistry. The review committee confirmed requirements, reviewed candidate systems, attended vendor product demonstrations, and voted on finalists.

Three vendors responded to our RFP. One was eliminated in the first review, which was conducted internal to ETS only, as it could not meet the must-have technical requirements stated in the RFP. We then worked with two finalists, both of whom provided live demonstrations as well as a sandbox for IST and ETS employees to conduct an accessibility review. Members of the review committee viewed the demos in real time or watched recordings subsequently and all had the chance to ask questions to ensure clarity on the systems' functionality as related to the RFP's requirements. We completed the demonstrations and reviews in early June.

The review committee opted for Explorance's Blue by a wide margin working within the predetermined selection criteria. Explorance was able to demonstrate that it met all of our "must-have" and many of our "nice-to-have" requirements, including:

- Evaluation administration and delivery that supports the university's organizational hierarchy
- Selectable question banks
- Advanced reporting, with aggregate and contextualized views
- Ability to build sub-accounts/projects for Extension and/or professional schools, if desired
- Ability to release and process evaluations across multiple schedules to accommodate mid-term evaluations, summer sessions, etc.
- Integration with campus Learning Management System and Enterprise Data Warehouse

Importantly, Blue is also the course evaluation vendor for many of our peers including University of Pennsylvania, the UCLA Anderson School of Business, the University of Southern California and the University of San Francisco.

We are currently underway with a limited rollout of Blue for Fall 2013. We will include our two existing partner departments, Department B and Department A, as one unit currently supported by IST's legacy system, Our Unit, which is currently slated for decommission at the end of 2014. The additional unit is the Department of Chemistry and potentially others in Spring. Integrating these new departments into our existing framework should present few problems

given Explorance's enterprise-class capacity and flexibility. Business analysis has been undertaken to understand Chemistry's exact evaluation needs, which has presented an interesting set of new course classification to evaluate (e.g., modular, co-taught classes). For Spring, we hope to continue to work with Chemistry and other prospective departments as we continue to learn about Blue's capabilities and how to adapt them to meet the needs of campus.

VI. CONCLUSION

With 2013 at a close, OEC continues to move forward taking the lessons learned from the past year into account. Among these, several things stand out. First, in moving from pilot project to functioning service, OEC will be known as "Course Evaluations" going forward, staffed by a group committed to ensuring high integrity and overall improvement in the course evaluation process. To further increase accuracy of evaluation delivery and reporting, certain challenges in campus data do need to be overcome. Further, the data we have collected from both the online evaluation pilots and experimental study suggest online evaluations present a complex reality requiring a nuanced understanding of their uses and effectiveness.

For instance, we may see greater variation in response rate by course and department as more units begin to participate in the Course Evaluations service, yet for large, low attendance classes, we may see increases in overall response rate. Ratings and completeness of evaluation responses may also decrease, due to the inclusion of a broader pool of student raters, but the depth of the open-ended responses that are provided seems to increase online. Further, no demographic factors other than academic achievement seem to impact the composition of those students completing evaluations online, eliminating some concern over equity of access or bias in the pool. Our data also suggest that among attending students, there was no difference

between online and offline ratings, but non-attending students received lower course grades in addition to providing lower ratings.

In whole, these new data are only telling a more complete picture of the state of course evaluations and by extension course instruction at Berkeley. Online evaluations present only a lens to reflect student perceptions, and in some respects, student academic performance. Responding to this newly identified pool of underperforming students represents both a challenge and an opportunity. Identifying the reasons for their disaffection is difficult and not entirely evident in our data. However, by learning of this gap, the opportunity identify ways to better serve underperforming students presents us with a new toolset for gauging their performance, perhaps with more frequent use of midterm and other formative evaluations.

In considering the questionnaires themselves, online evaluations present us with the opportunity to rethink how we deliver and structure evaluations. Recent work from Dr. Phillip Stark (http://teaching.berkeley.edu/blog/evaluating-evaluations-part-1) shows that the Berkeley campus community may need to broaden the types of evidence used in evaluating teaching overall. Further, departments and individual instructors going forward have much greater opportunities to inform their own teaching and programmatic efforts with the inclusion of customized questions designed to fit their own objectives. In providing access to this capacity, our hope is to deliver the campus more meaningful and relevant data about the students in their courses.

Appendix A: Question Category Frequency Counts

	Instructor Specific Items						Course Specific Items				Self-Evaluation Items		Open-Ended Items
	Presentation of Content	Clarity of Expectations/Directions	Helpfulness/Availability	Useful/Clear Feedback on Performance	Encouraging of Participation/Discussion	Overall Teaching Effectiveness	Course Content	Increased/Developed Skills/Knowledge	Useful/Clear Feedback on Performance	Course Overall	Time Estimates	Personal Effort Satisfaction	All Items
Activity	1		1			2	3			2	1	1	3
Clinic w/ Supervision	1	1		1		2		3		2	1	1	3
Clinic w/o Supervision						2		4	2	2	1	1	3
Foreign Lang				1	1	2	2	1		2	1	1	3
Lab (Stand-Alone)	1		1	1		2	2			2	1	1	3
Lab (w/ Lecture)			1	1		4	2			2	1	1	3
Lecture	1	1			1	2	2			2	1	1	3
Project			3	1		2		3		2	1	1	3
Section (Discussion)	1		1		1	3	1	2		2	1	1	3
Section (Problem)	1		1		1	3		3		2	1	1	3
Seminar					2	2		2		2	1	1	3
Studio	1		1	1		2		2		2	1	1	3
Writing		1	1	1	1	2		2		2	1	1	3
TOTALS	7	3	10	7	7	30	12	22	2	26	13	13	39

Appendix B: Revised OEC Instrument

Author's Note: This section includes bolding of questions that were selected for inclusion in the final instrument template for both their generic and representative qualities within a given category. In some cases the wording for these question items was modified slightly to accommodate a new question stem formatting. Note that bolded questions can be substituted for any other question within its respective category at the departmental level.

1. Instructor/GSI-Specific Question Themes:
    a. Presentation of Content
        i. **The instructor presented content in an organized manner**
        ii. The instructor clearly presented the skills to be learned
        iii. The instructor effectively presented the tools (e.g. materials, skills, and techniques) needed
        iv. The instructor effectively presented concepts and techniques

    b. Clarity of Expectations or Directions
        i. **The instructor explained concepts clearly**
        ii. The instructor made the elements of good writing clear
        iii. The instructor clearly articulated the standards of performance for the course
        iv. The instructor provided guidance for understanding course exercises
        v. The instructor increased my understanding of course material

    c. Helpfulness/Availability
        i. **The instructor was helpful when I had difficulties or questions**
        ii. The instructor helped me achieve my goals
        iii. The instructor helped me define the goals and scope of the project
        iv. The instructor helped me identify resources I needed to carry out the project
        v. The instructor was helpful when I had difficulty performing activities
        vi. The instructor was helpful to me individually (in conferences, email exchanges, etc.)
        vii. The instructor was readily available during the class
        viii. The instructor provided help when I had difficulties

    d. Useful/Clear Feedback on Performance
        i. **The instructor provided clear constructive feedback**
        ii. The instructor provided useful feedback on my writing
        iii. The instructor provided meaningful feedback on my work
        iv. The instructor provided meaningful guidance on my progress/work
        v. The instructor provided constructive feedback in response to difficulties with the language
        vi. The instructor gave me constructive feedback
        vii. The instructor gave me constructive feedback on assignments
        viii. The instructor clearly articulated the standards of performance

    e. Encouraging of Participation/Discussion
        i. **The instructor encouraged student questions and participation**
        ii. The instructor engaged the class in productive discussions
        iii. The instructor guided the discussion well
        iv. The instructor encouraged student contributions
        v. The instructor provided opportunities for class participation
        vi. The instructor encouraged critical engagement with the material

vii. The instructor encouraged participation

f. Overall Teaching Effectiveness:
   i. How successful was the instructor in creating an environment that was conducive to learning?
   ii. How would you rate the overall effectiveness of the instructor's teaching? [7]
   iii. MANDATORY: Considering both the limitations and possibilities of the subject matter and the course, how would you rate the overall effectiveness of this (graduate student) instructor?

2. Course-Specific Question Themes:
   a. Course Content (Organization, Clarity of Expectations/Directions, Balance/Appropriateness)
      i. **The course was effectively organized**
      ii. The course presented skills in a helpful sequence
      iii. The course provided an appropriate balance between instruction and practice
      iv. The course was appropriate for the stated level of the class
      v. The course was organized in a way that helped me learn
      vi. The course provided a mixture of explanation and practice
      vii. The course assignments and lectures usefully complemented each other
      viii. The course instructions (including, manuals, handouts, etc.) were clear
      ix. The course work helped me understand concepts more clearly
      x. Instructions for course materials (including manuals, handouts, etc.) were clear
      xi. The lab complemented my understanding of the lectures
      xii. The section helped to complement the lectures

   b. Increased or Developed Skills/Knowledge (Thinking, Writing, Application, Knowledge, Communication, Ethics): General/overall learning & content-specific learning are possible sub-categories within this set (see i and ii below).

      i. Application & Specific Skill Development:
         1. **The course developed my abilities and skills for the subject**
         2. The course developed my ability to interact with diverse groups of people
         3. The course provided guidance on how to become a competent professional
         4. The course developed my ability to read and think critically
         5. The course helped me improve my writing
         6. The course developed my ability to provide constructive critiques to others
         7. The course helped me make progress in my acquisition of the language
         8. The course helped me conceptualize and present my ideas in my artistic medium
         9. The course helped me understand ethical issues involved
         10. The course developed my communication/presentation skills

      ii. Theory/Content Knowledge:
         1. The course developed my ability to think critically about the subject
         2. The course developed my ability to apply theory to practice

[7] Items 1fi-ii represent a place where both items appear on most/all of the existing forms, making them de facto mandatory questions. Is it feasible to collapse these into a single item or drop them entirely, as they tend to get at a similar idea of evaluating teaching/learning overall? In this regard, they tend to replicate the mandatory campus question directly below them (1fiii).

3. The course provided the opportunity to practice the skills required in the course
4. The course allowed me to synthesize fundamental knowledge and skills
5. The course gave me a deeper insight into the topic
6. In this course, I learned a great deal
7. The course improved my problem-solving skills

c. Course Overall
    i. How satisfied were you with this course?
    ii. MANDATORY: Considering both the limitations and possibilities of the subject matter and the course, how would you rate the overall effectiveness of this course?

3. Self Evaluation:[8]
    a. Time Estimates
        i. How many class sessions did you attend?
        ii. On average, how many hours per week have you spent on this course , including attending classes, doing readings, reviewing notes, writing papers, and any other course-related work?
    b. Personal Effort
        i. How satisfied were you with your effort in this course?

4. Open-Ended
    a. Strengths of Course/Section
        i. Please identify what you consider to be the strengths of the course .
    b. Areas of Improvement for Course/Section
        i. Please identify area(s) where you think the course could be improved.
    c. Student-to-Student
        i. SEPARATE SPECIFIC DEPARMENT OPT-IN: Feedback for other students: What advice would you give to another student who is considering taking this course ? [9]

[8] The Self-Evaluation Questions appeared on all versions of the forms created by the faculty focus groups, so as such are de facto mandatory questions.
[9] This specific opt-in for student-to-student feedback is something that individual department's will have to opt-in to. Just as with the OEC initiative itself, which is opt-in at the departmental level, this question is a discrete, separate opt-in. So a department may participate in online evaluations, but is not additionally required to participate in the student-to-student question.