OFFICE OF
EDUCATIONAL
ASSESSMENT
Improving Learning Through Assessment

OEA Report 16-05

Online Versus Paper Student Ratings of Instruction: A Follow-up Study

Debbie McGhee
November 2016

INTRODUCTION

Since online student ratings of instruction were first introduced at the University of Washington (UW), faculty have expressed concern regarding the equivalence of ratings collected online and those obtained from traditional paper-based evaluation forms. In an earlier study,[1] we found that classes evaluated using paper forms did, in fact, receive slightly higher ratings than those evaluated online. To verify this difference and determine implications for use of course evaluations, we replicated some of the original analyses using a larger dataset. The results are described below.

METHOD

The UW course evaluation system (*IASystem*™) utilizes several evaluation forms containing a common subset for four "summative" items. These items and their combined average (the "global median") provide an overall evaluation of each course and allow cross-course comparisons. For the purpose of the present study, we collected median ratings of all five indicators for all classes taught in Autumn 2014 – Summer 2016. We limited our sample to surveys using standard *IASystem*™ evaluation forms (i.e., excluding custom forms and forms I, J, L, M, S, and W) and that were based on responses from five or more students (*M* = 21). In total there were *N* = 30,776 surveys with medians for all four summative items. Nearly three-quarters (73.6%) of the surveys were conducted online.

Ratings were given on a six-point scale from 0 (poor) to 5 (excellent).

RESULTS

Evaluation Mode (Online vs. Paper)

Initial analyses confirmed a small difference (range .05 - .07) in average ratings collected online compared to those collected on paper. Paper ratings were slightly higher across all five measures, as shown in Table 1.

Table 1. Average *IASystem*™ summative medians by evaluation mode

	Paper			Online			Diff	
Measure	Mean	SD	n	Mean	SD	n	Mean	95% CI
Course as a whole	4.11	0.61	8,119	4.06	0.66	22,657	0.06	[0.04, 0.08]
Course content	4.11	0.57	8,119	4.06	0.62	22,657	0.05	[0.03, 0.06]
Instructor overall	4.32	0.61	8,119	4.26	0.67	22,657	0.07	[0.05, 0.09]
Instructor's contribution	4.20	0.68	8,119	4.14	0.73	22,657	0.06	[0.04, 0.08]
Global Median	4.20	0.60	8,119	4.14	0.64	22,657	0.06	[0.04, 0.07]

The minimal size of the differences in ratings is apparent in Figure 1. Because *IASystem*™ reports course ratings to a single decimal, the obtained differences are visible for only one of the four summative items and the global median.

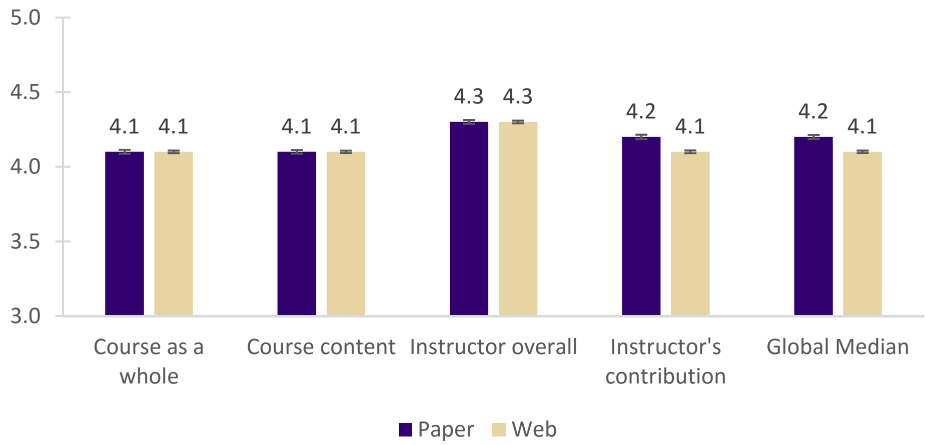

Figure 1. Average *IASystem*™ summative medians by evaluation mode

Correction for Bias

Although the observed differences in ratings were very small, we thought it might still be worthwhile to modify *IASystem*™ to adjust ratings of summative items for evaluation mode. The system currently computes adjusted medians for the four summative ratings and the global median by correcting ratings for (a) student's reason for taking the course, (b) class size, and (c) student's relative expected course grade. After carrying out the requisite regression analyses, we found that the magnitude of the evaluation mode effect was significantly smaller than the three existing variables. Figure 2 shows the standardized coefficients and R^2 from a multiple regression predicting the global median from evaluation mode and the three variables. Evaluation mode was a weak predictor which did not substantially improve the model; there would be no utility in adding this variable to the adjustment variables.

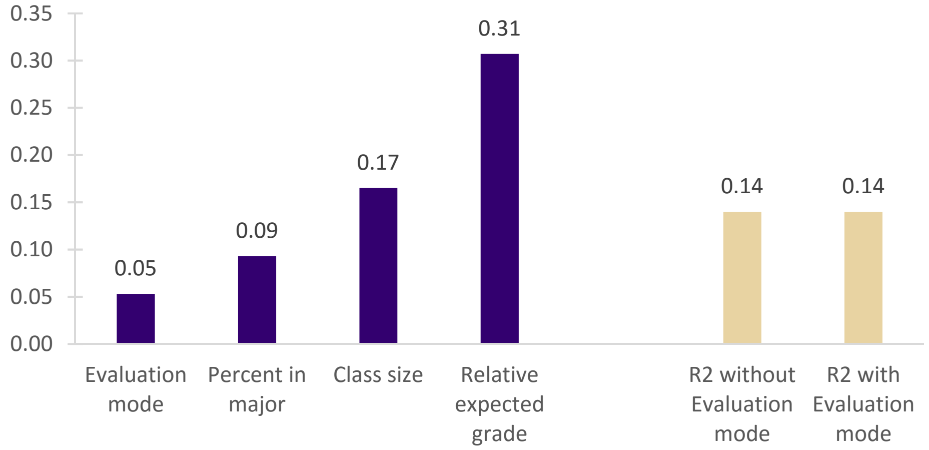

Figure 2. Standardized coefficients and R2 from a multiple regression predicting the global median from standard adjustment factors and evaluation mode

Course Mode (Online vs. Face-to-Face)

The results of our earlier study indicated that students tend to rate face-to-face courses more highly than online or hybrid courses. To verify this finding and extend it, we carried out two-way ANOVAs with evaluation mode (paper vs. online) as one factor and course mode (face-to-face vs. online or hybrid) as the other. This analysis allowed us to determine whether there might be an effect of congruence of evaluation and course modes., i.e., would students rate online courses differently if they were also evaluated online, etc. The results of these analyses are shown in Table 2.

Table 2. Two-way analyses of variance of course ratings by evaluation mode and course mode

		Pap			Online			
Measure	Course mode	Mean	SD	n	Mean	SD	n	*F*-tests
Median01	FTF	4.13	.60	8046	4.07	.66	21386	Evaluation mode: 9.98, *p* = .002
	Other	3.60	1.04	45	3.96	.66	1219	Course mode: 41.48, *p* < .001
								Eval. x Course: 18.81, *p* < .001
Median02	FTF	4.11	.57	8046	4.07	.62	21386	Evaluation mode: 8.17, *p* = .004
	Other	3.69	.84	45	4.00	.62	1219	Course mode: 28.32, *p* < .001
								Eval. x Course: 15.28, *p* < .001
Median03	FTF	4.34	.61	8046	4.27	.67	21386	Evaluation mode: 9.83, *p* = .002
	Other	3.75	1.23	45	4.13	.73	1219	Course mode: 52.92, *p* < .001
								Eval. x Course: 19.79, *p* < .001
Median04	FTF	4.20	.68	8046	4.15	.73	21386	Evaluation mode: 10.23, *p* = .001
	Other	3.61	1.27	45	4.02	.76	1219	Course mode: 44.48, *p* < .001
								Eval. x Course: 17.50, *p* < .001
Global	FTF	4.20	.60	8046	4.15	.64	21386	Evaluation mode: 10.83, *p* = .001
	Other	3.66	1.09	45	4.04	.66	1219	Course mode: 45.68, *p* < .001
								Eval. x Course: 19.36, *p* < .001

Note. FTF refers to face-to-face classes; Other refers to hybrid and all-online classes.

Consistent with earlier findings, both main effects were significant. Courses rated on paper received somewhat higher ratings than did courses rated online, and higher ratings were given to face-to-face courses than to online or hybrid courses. However, the interpretation of these main effects is not straight forward because there was also a significant interaction. As shown in Figure 3 for the global median, face-to-face classes rated online (*N* = 21,386) tended to receive slightly lower marks than face-to-face classes rated on paper (*N* = 8,046), while the opposite was true for online and hybrid classes (N_{online} = 1,219; N_{paper} = 45). It is uncertain how much of the difference for online/hybrid classes was due to the fact that all paper-rated courses were hybrid and hybrid courses in general received lower ratings than online courses.

The high variability of ratings given to online/hybrid courses evaluated on paper may be due to the extremely small number of cases. It may also reflect the diversity of instructional approach used in hybrid classes.

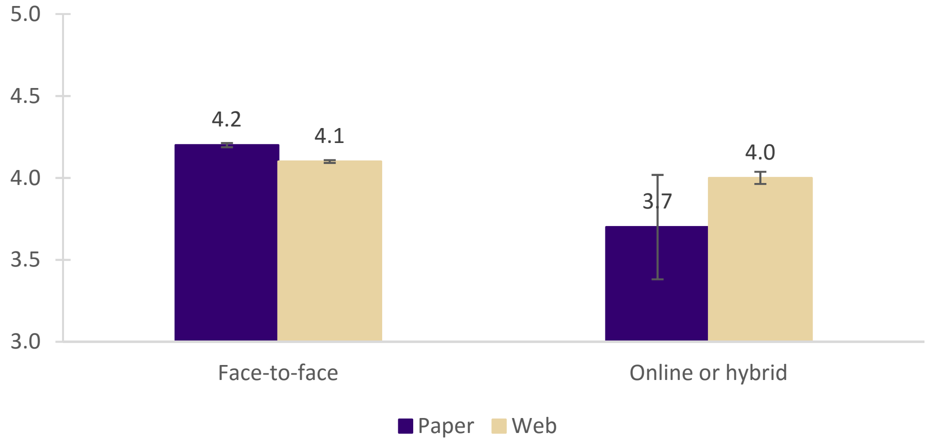

Figure 3. Average global median by instruction mode and evaluation mode

Response Rate

A decline in student response rate is a consistent concern at institutions transitioning from paper to online course evaluations. To evaluate the impact of response rate on average ratings, we computed zero-order correlations between response rate (operationalized as the number of student responses divided by official enrollment) and median ratings. The average correlation was $\bar{r}$ = .20; in other words, as response rate increased, so did median rating. Furthermore, response rates differed by evaluation mode. Higher response rates were observed in classes that used paper evaluations (*M* = 74%) than in those that used online evaluations (*M* = 58%). These two findings together suggest that the lower average ratings observed with online evaluations might be accounted for (at least partially) by response rate. To test this hypothesis, we divided response rate into quintiles and conducted a two-way ANOVA for each of the five measures.

There were two main findings. First, we found that for all five measures the effect for response rate was much larger than the effect for evaluation mode (see Table 3). For example, for ratings of "the course as a whole," the difference between paper and online ratings was *M* = .05, but the difference between the lowest and highest response rate quintiles was *M* = .37.

Table 3. Average *IASystem*™ summative medians by evaluation mode and survey response rate

		Pap			Online			
Measure	Response %	Mean	SD	n	Mean	SD	n	*F*-tests
Median01	<42%	3.94	0.63	547	3.88	0.71	5224	
	42-55	3.95	0.63	819	4.02	0.66	5351	Evaluation mode: 8.80, *p* = .003
	56-69	3.99	0.62	1524	4.09	0.66	4895	Response rate: 183, *p* < .001
	70-82	4.11	0.58	2079	4.17	0.59	3636	Mode x Rate: 11.16, *p* < .001
	83+%	4.27	0.57	3147	4.24	0.58	3551	
Median02	<42%	3.97	0.60	547	3.91	0.66	5224	
	42-55	3.97	0.57	819	4.03	0.60	5351	Evaluation mode: 6.78, *p* = .009
	56-69	3.99	0.58	1524	4.09	0.62	4895	Response rate: 142, *p* < .001
	70-82	4.10	0.55	2079	4.15	0.57	3636	Mode x Rate: 13.14, *p* < .001
	83+%	4.24	0.55	3147	4.21	0.56	3551	

Table 3. Average *IASystem*$^{\text{TM}}$ summative medians by evaluation mode and survey response rate (continued)

					Online			
Measure	Response %	Mean	SD	n	Mean	SD	n	*F*-tests
Median03	<42%	4.10	0.69	547	4.05	0.74	5224	
	42-55	4.14	0.67	819	4.21	0.68	5351	Evaluation mode: 9.91, *p* = .002
	56-69	4.20	0.64	1524	4.29	0.66	4895	Response rate: 223, *p* < .001
	70-82	4.35	0.57	2079	4.39	0.58	3636	Mode x Rate: 6.42, *p* < .001
	83+%	4.47	0.56	3147	4.47	0.55	3551	
Median04	<42%	3.99	0.75	547	3.95	0.79	5224	Evaluation mode: 16.58, *p* < .001
	42-55	3.99	0.74	819	4.09	0.73	5351	Response rate: 190, *p* < .001
	56-69	4.06	0.71	1524	4.17	0.73	4895	Mode x Rate: 6.74, *p* < .001
	70-82	4.22	0.64	2079	4.27	0.64	3636	
	83+%	4.35	0.63	3147	4.35	0.62	3551	
Global	<42%	4.01	0.64	547	3.96	0.69	5224	
	42-55	4.02	0.63	819	4.10	0.64	5351	Evaluation mode: 13.93, *p* < .001
	56-69	4.07	0.62	1524	4.17	0.64	4895	Response rate: 201, *p* < .001
	70-82	4.20	0.57	2079	4.26	0.57	3636	Mode x Rate: 9.54, *p* < .001
	83+%	4.34	0.55	3147	4.33	0.56	3551	

Second, the analyses detected an interaction between response rate and evaluation mode. In general, at the lowest and highest response rates, average online ratings were either the same or lower than paper ratings, but in the middle groups online ratings were significantly higher than paper ratings. This interaction on the global median is depicted in Figure 4. (The reason the grand mean for online ratings was less than the grand mean for paper ratings, despite the above finding, was that there was a far greater percentage (23%) of low-response-rate cases among online ratings, relative to paper ratings (6%).)

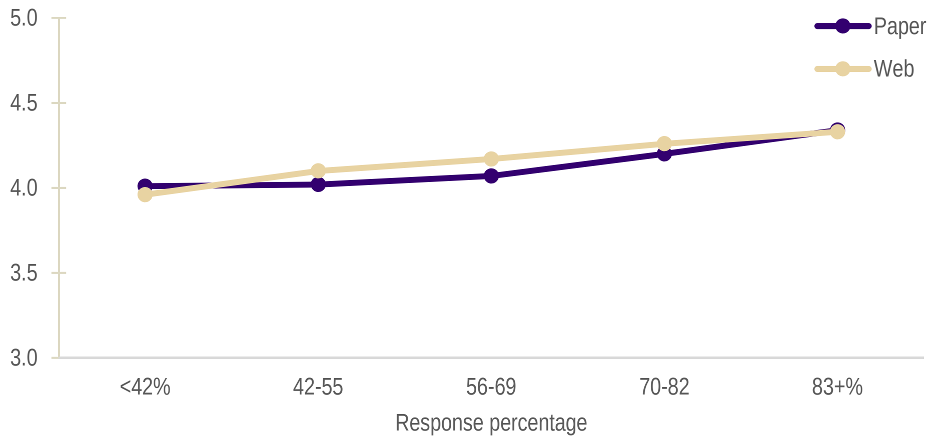

Figure 4. Average global median by response rate and evaluation mode

DISCUSSION

The primary finding of this study is that course ratings are slightly (but statistically significantly) higher when courses are evaluated using paper evaluation forms than when courses are evaluated online, but (1) this effect is very small, and (2) the size of the effect is influenced by other factors, in particular response rate.

Size of Differences

Although the differences between online and paper ratings are statistically significant, there is very little "practical" significance. *IASystem*™ reports item medians to the nearest single decimal and the observed differences are smaller than that for three of the four summative items. Differences for the fourth item, "instructor's contribution to the course", are somewhat larger and likely reflect the real difference in the role of the instructor in online courses (nearly always evaluated online) versus face-to-face classes (which may be evaluated either online or on paper).

The lack of practical significance between results of evaluations conducted online versus on paper is also supported by the results of regression analyses. The magnitude of the evaluation mode effect in these analyses is significantly smaller than those of existing correction variables, and not large enough to warrant addition of this variable to the current correction for bias.

Effect of Response Rate

We have found that classes with higher response rates receive more positive ratings than do classes with low response rates. The relationship between response rate and average rating is much stronger than the effect of evaluation mode. In analyses reported above, the difference between the lowest and highest response rate quintiles was $M = .37$, as compared to a difference of $M = .05$ between paper and online ratings. For both types of ratings, the fewer students responding, the lower the average rating. It may be that those students with a negative view of a class are the ones most motivated to provide an evaluation, and thus ratings are higher when respondents are more representative of all enrolled students. This finding is strong support for the need for deliberate practices on the part of both faculty and departments to ensure that evaluations are completed by a large (70%), representative portion of students in each class.

[1] D. McGhee and N. Lowell. (2015). Effects of Course Delivery Mode and Course Evaluation Mode on Student Ratings of Instruction, *OEA Report 15-02*.