Lessons Learned From Online vs. Paper-based Computer Information Students' Evaluation System

Jens Liegle, jliegle@gsu.edu
David S. McDonald, davemcdonald@gsu.edu
Computer Information Systems, Georgia State University
Atlanta, GA 30303, USA

Abstract

Many universities are offering online courses these days. What follows consequently is that instructors are being evaluated online as well, and due to – among other reasons –potential cost savings, even some traditional courses are being evaluated online now as well. This paper presents the results from a pilot test at a large south-eastern universities' Computer Information Systems department within the college of business of moving to online evaluations. The results show that some faculty did not like to be evaluated online due to fears of receiving lower scores or lower response rates; however, our study showed that there was no difference in the important instructor effectiveness question in paper vs. online evaluations, and – due to special circumstances – that online evaluations had an even higher response rate than paper based evaluations.

Keywords: pedagogy, evaluations, faculty effectiveness

1. Introduction

Student evaluations are, next to publications and to a lesser extend service, a major factor in determining the merit of faculty, especially during decision times such as hiring, promotion, tenure, and salary negotiations (Thorpe, 2002). Faculty, especially during their pretenure years, are therefore very interested in getting good student evaluations.

The use of student evaluations as a factor to evaluate faculty is not without problems. First, statistically one should not compute averages of Likert-scale data, since arguably Likert-scale data is not interval level (Thorpe, 2002). And second, the most common criticism of student evaluations seems to be that they are biased in that students tend to give higher ratings when they expect higher grades in the course (Rice, 1988, pp 335-6; Wilson, 1998; Greenwald and Gillmore, 1997, p1214), a problem that many believe is the main cause of grade inflation (Goldman, 1985; Sacks, 1986). Another factor influencing the scores are the potential effects of nonresponse bias if a limited number of students complete a course evaluation instrument (Thorpe, 2002). The response bias of traditional evaluation methods such as in-class paper based evaluation typically favor the instructor, since the timing of the evaluation is under the instructor's control; i.e. evaluations are typically conducted before the final exam and/or before term papers/final projects with low average scores are returned to the students. Further, since the evaluation takes place during class time, attendance is typically high, and nonresponse bias rarely an issue.

However, other means for evaluating teaching effectiveness are even worse in that they do not appear to be valid (Huemmer, 2004), and ratings by colleagues and trained observers are not even reliable (a necessary condition for validity) in that colleagues and observers do not even substantially agree with each other in instructor ratings (Marsh and Roche, 1997).

A new dimension of problems has been added by the attempt to conduct the evaluations online. Online course evaluations still see relatively limited use in higher education: Only two out of 200 institutions ranked as most "wired" by Yahoo reported institution-wide use of web-based evaluation systems (Hmieleski, 2000). The most cited reason for the lack of online-evaluations seems to be low response rates (Thorpe, 2002), with one study reporting 31% online response rates compared to 65% with in-class paper-based methods (Cummings and Ballatyne, 1999).

Yet, a number of potential benefits, most importantly instant results and cost savings, motivate more and more universities to look into online evaluations (Thorpe, 2002). This paper examines one such effort from the College of Business (CoB) at a large south-eastern University.

While faculty participating in a pilot were initially using the online system, acceptance dropped rapidly. Exact numbers are not available at the time for us, however, informal interviews with faculty revealed that many of them found that they had received lower scores when using electronic instead of paper evaluations. The administration, as a result, allowed faculty to now have online evaluations count against them for their annual review.

To identify the cause for this surprising result, considering that the CoB is by its very nature one of the most technology friendly pilot college one could think of, and in addition we are evaluation its CIS department – the most technology friendly department of them all, we conducted a preliminary round of informal interviews with faculty to find out what stopped them from ever using or from continued usage of the online system.

A number of perceived problems surfaced:

- Response rates seemed to be lower compared to the traditional way
- Responses were more negative and thus scores lower
- The fact that students in our case could evaluate instructors online after they took the final, while paper-based evaluations were conducted before the final exam only.

We found from the results of an analysis of the instructor evaluations that the means of evaluation made no difference, and that 'poorly' performing students were less likely to go online and evaluate instructors after the final.

2. Literature Review of Online vs. Paper based Evaluation

A number of potential benefits (See table 1) motivated the CoB at a large south-eastern University to be selected for a pilot program to test an online form of student evaluations starting in fall semester of 2002. Traditionally, the university had instructors hand out traditional paper and pencil based forms during the last two weeks of classes, which they had to hand out to students with enough class-time to fill out the survey, and then leave the classroom. A student volunteer would then administer the collection of the surveys and deliver them in a sealed envelope to a drop box. Overall, a number of problems have been associated with paper-based evaluations (see table 2).

Research comparing online vs. paper based surveys in general has found no significant difference in the results (Handwerk, Carson, and Blackwell, 2000; Matz, 1999; Sax et al, 2002). One factor influencing the scores are the potential effects of non-response bias if a limited number of students complete a course evaluation instrument (Thorpe, 2002). In terms of comparing respondents vs. non-respondents in online surveys, some demographic differences have been found (Underwood, Kim, & Matier, 2000; Tomsic, Hendel, & Matross, 2000) in terms of sex, GPA, and expected grade. Other factors influencing response rates for online surveys are familiarity with the Internet, the ease of completing the survey, and concerns for privacy and confidentiality (Dillman, 2000; Handwerk, Carson, and Blackwell, 2000).

Table 1: Advantages of online evaluation (Couper, 2000; Dillman, 2000)

Frees class time: Since students could take the survey at home, instructors no longer needed to reserve class time for administering it
Error free data entry
Reduced costs of online research (personnel, mailing, printing, etc)
Low administration costs
Rapid dissemination of results
Students/instructors can review evaluations from previous periods online
The ease of reaching representative samples of a population
The ability to validate data during collection

Table 2: Problems of Paper Based Evaluations

The risk of instructors forgetting to administer them altogether
Some would get lost because the student in charge of dropping them at the designated drop-box would not do so (cases of locked drop-boxes/could not find them)
Some instructors would administer the evaluations on an optional 'review' day or during student presentations, i.e. days with low class attendance in the hope that weaker students would not bother to come which would improve the instructor's scores
The risk of cheating on the instructor side, i.e. envelope stuffing
Since instructors are given the forms, but not the required number 2 pencils to fill them out, there were ongoing problems with the scanning/reading of the score cards due to the use of incorrect marking pens/pencils
The student comment sections needed to be manually copied for the instructor to read
And the associated time it took to compile the results and the related cost in terms of machinery, paper, and manpower were rather high.

Low response rates for web-based surveys were suggested as the primary issue that limits Institutions to in-class, paper-based course evaluation instruments. Cummings and Ballatyne (1999) describe the experiences of Murdoch University in piloting a web-based course evaluation process. The response rates were lower for the web-based system, 31% compared to 65% for the in-class paper-based method. A study by Thorpe (2002) compared web-based to paper based evaluations for three different courses. For both the math and the statistics course, instructor evaluations were better in the paper based than in the online version, while the computer science course had slightly better scores in the online version. However, the "level of interest in the topic" before the course began was reported higher in the CS course for online, while nearly identical in the math and statistics courses. Only the difference in the stat course was significant (p<0.05) with a paper-based score of 3.86 vs. an online score of 3.33.

3. Methodology

The CoB was selected for the pilot test of the system during the fall 2002 semester. This paper will report some findings based on the use of the online evaluation during its use by the Department of Computer Information Systems (CIS) for first year (fall semester of 2002, spring semester of 2003, summer semester of 2003, and fall semester of 2003).

The original paper-based evaluation form and subsequent online version, utilizes a 35-question evaluation form originally developed in-house by the Decision Science Department within the college. This tool was tested for both content and construct validity. Additionally, a factor analysis was used to determine 7 key dimensions thought most appropriate as an overall measure of instructors' performance.

Beginning in the fall semester of 2002, and continuing through the end of the fall semester of 2003 (inclusive of summer sections), 61 out of a total of 342 CIS department sections or 17.8% of class offerings were evaluated by students using the online evaluation methodology. Instructors received an email explaining them the rationale and strategy the College was taking with regards to student evaluations, and were encouraged to use the online evaluation system. However, it should be noted that instructors were given an option to continue to use the traditional system. It may then be surmised that negative instructor influence was kept to a minimum since each voluntarily chose the method of evaluation he or she preferred.

A critical issue in evaluations is a guarantor of privacy. For paper-based systems, students' evaluations may be influenced by a fear that there is a potential that an instructor may be able to identify their students by their handwriting, and thus fear revenge for poor evaluations if they have this instructor again. Online systems, on the other hand, by their very nature do not have this problem.

4. Analysis and Results

Instructors were concerned that few students would use the online evaluation to begin with, that poorly performing students would 'take revenge' on instructors, while good students might not go through the trouble to evaluate instructors at home. To examine these issues, we formulated the following hypotheses:

H1: Online evaluations have a lower response rate than paper-based systems
H2: Online evaluations have lower scores than paper based systems

To test these hypotheses, we selected sections during the Spring semester 2002 through the Fall semester of 2002 period and compared online vs. paper response rates and evaluations scores for the key question #34 (effectiveness of the instructor). Question 34 is used by the college as a primary evaluation variable for teaching effectiveness. An independent sample two-tailed *t*-test was used to compare the means between the response rates of 278 students using paper-based evaluations with 61 students using online evaluations (table 3). The response rate averaged 67.19% for those using paper-based evaluations compared to an 81.75% response rate for those students using online evaluations. Levene's test for equality of variance was used to test the validity of the hypothesis. No significance was found for populations with unequal variances (table 4). Furthermore, in using this independent *t*-test, three factors must be present (Zar, 1984, pp. 130-131): 1) normally distributed samples; 2) approximately equal sample sizes; and 3) equality of variance. For our testing, some of these conditions could not be met. However, numerous studies have shown that the *t* test is robust enough to stand considerable departures from these theoretical assumptions, especially when a two-tailed *t* test is employed (Boneau, 1960, Box, 1953, Cochran, 1947, Srivastava, 1958). Furthermore, if the underlying populations are markedly skewed, then one must be wary of one-tailed testing, and if there is considerable non-normality of populations then a very small significance levels (alph < 0.01) may not be depended upon, thus our rationale for using the two-tailed *t* test in the following analyses. Lastly, the power of the two-tailed *t* test is very little affected by skewness in the sampled populations, but there can be serious effects on a one-tailed test (Kohr and Games, 1974). Therefore, for completeness, table 5 shows a further analysis to determine the variance, skewness, and kurtosis of the population.

Thus, H1 is not supported. For this CIS department, the response rates were not significantly different between those students using online evaluations and those using the paper-based methodology.

One final observation…the total population included 339 students. Non-parametric testing would be required if significance was shown with any of the hypotheses that necessitated the use of highly unequal samples. These tests would be necessary to ensure a type I error was not committed (Zar 1984). However, since significance was not shown, non-parametric testing was deemed unnecessary.

To broaden our investigation of this phenomenon, we further examined the same scores for instructors who tried online evaluations vs. paper during the same period. Again, since the inequality of sample sizes, three distinct statistical analyses were performed. Table 6 shows for the same population, the average mean for paper-based evaluations on the likert-scaled question 34 (i.e., the effectiveness of the instructor) indicate those using the traditional evaluation method gave instructors an average score of 4.08 out of 5, while those students who evaluated instructors online evaluated their competence with a score of 3.96 out of 5. As with the previous analysis, tables 8 & 9 indicate the validity of this analysis. In particular, table 7 does not show support for H2.

To determine the underlying reasons for this phenomenon, we examined the online evaluation process closer.

A key factor that could possible influence the results was that some faculty in the past tried to discourage "weak" students from being present during evaluation by giving a break before administering them, scheduling them on a day where no exam related material is covered, during final presentation times, etc. These "tricks" would not longer work for online evaluations. As a result, we formulate the following two hypotheses:

> H3: Good students are less likely to fill out online evaluations than offline

> H4: Weaker students are more likely to fill out online evaluations than offline

Due to privacy issues, the dataset employed did not allow for individual grade analysis. However, data were available for the grade distribution of individual section. The data included a breakdown of the number of "A"s, "B"s, "C"s, "D"s, and "F"s given in each section offered. To normalize the data, these numbers were then converted into a percentage within each course. For this analysis, we determined a multi-year average of what constituted a course consisting of primarily "good" students. For the following analyses, an average of 76% of students in each section received either an "A" or a "B" for their respective courses constituted a surrogate of a course comprised primarily of "good" students. Using this as a cut-off point metric, for this analysis, sections with 24% of students or greater who did not receive an "A" or a "B" were deemed "weaker" students. Thus, table 10 compares the means percentage "good" students' response rate in 37 sections that used paper-based evaluations (71.7%) with the "good" student percentage in 29 sections of online evaluations (77.6%).

The *t*-test analysis shown in table 11 indicates no significant difference in response rates between the two groups of "good" students and therefore, there is no support for H3. However, when the data is examined using the dataset for "weak" students, significance is shown. Table 12 shows a marked drop in the response rate averages when compared with "good" students. I.e., on average, approximately three-fourths of good students filled out the evaluation form, whether paper-based or online compared with approximately 50% weaker students. Moreover, the response rate for the 32 sections of "weaker" students using paper evaluations was 59.7%, while only 43.3% of "weaker" students responded online in 14 different sections.

The results shown in table 13 are significant, but not as hypothesized. It was believed that weaker students would have a greater tendency to do online evaluations. However, the analysis indicates the opposite. Close to 60% of sections consisting of a greater-than-average number of "weaker" students filled out paper-based evaluations for their respective courses, while only 43% of their cohorts in classes using online evaluations bothered to fill out the form. Since these or CIS students, it may be plausible that weaker students in this major do not have the same affinity for technology as those that perform well;. To navigate through the necessary layers of menus to get to the evaluation form may have been too confusing for some of these weaker students. It is also likely that, although all evaluations are voluntary, the paper-based forms are filled out during a scheduled class period, where the online evaluations are accomplished at the students' leisure.

Finally, we continued to examine instructor evaluation characteristics of the same two cohorts used in H3 and H4. Paper-based evaluations were given during the last weeks of class, but before finals week. Students, therefore, did not know a significant portion of their grade, nor had they seen the final before being asked to fill out the paper evaluation form. Online evaluations, on the other hand, could be submitted during and after final exam week, and initially, even until after grades were posted.

A major concern of faculty was that students who did not like the final exam or their overall course grade would evaluate them lower than students who did not know yet have this vital information. Based upon these assumptions, we formulate the following hypotheses:

> H5: Students who expect a low grade are more likely to give low evaluations than students expecting a high grade.

> H6: Students who expect low grades are more likely to fill out a late online evaluation

For this analysis, we examined the average grade students and categorized them utilizing the same methodology described above. By the very nature of the paper-based evaluation process, all evaluations occurred *prior* to finals week and weeks before the end of the semester. However, students in sections that used online evaluations could evaluate their instructor starting with the same week the paper evaluations were handed out, but continuing into the next semester. The online analysis system time-stamped every evaluation form. Thus, we were able to determine which evaluations were filled out after the student took their final exam for the sections of course in question. It should also be noted that in order to increase the student response rate, in the summer semester, 2003, the university changed online evaluation methodology to make online grades available only to students who had already evaluated their instructors.

For completeness, our analysis goes beyond that needed to either support or deny the original hypotheses. Tables 14 and 15 indicate that amongst those sections consisting primarily of "good" students, the instructor's evaluation did not depend upon which evaluation methodology was utilized. Similarly, tables 16 and 17 indicate that amongst the sections with a larger than average number of "weaker" students, again, there does not appear to be any significant association between which evaluation methodology a student utilized and the evaluation score given to an instructor.

Tables 18 and 19 compare the means for instructors' performance given by 327sections with a greater-than-average number of "weaker" students with 275 sections where the evaluations were completed by a greater percentage of "good" students…regardless of the type of evaluation methodology utilized. As expected, the means for the "good" students are significantly higher in all cases (3.75 vs. 3.95).

Decomposing this data further demonstrates similar results when separating out for analysis sections with "good" vs. sections of "weaker" students. Table 20 shows this comparison only with those students using online evaluations with a 3.82 mean evaluation score given by "weaker" students as compared to the 4.28 evaluation given by "good" student.

Table 21 indicates these results are also significant at the $p < 0.00$ level. For completeness, tables 22 and 23 perform a similar comparison except with only those students using paper-based evaluations. Once again, for those students utilizing paper- based evaluations, the mean for "good" students was 4.19 as compared with 3.80 for "weaker" students.

For all three comparisons, overall evaluation regardless of evaluation methodology, students using online methodology only, and students using paper-based evaluation, each show significantly lower evaluations given by "weaker" students Therefore, H5 is supported.

Finally, tables 24 and 25 show the results of the analysis for testing H6 — students who expect low grades are more likely to fill out a late online evaluation. For this analysis, the mean resulting from the scores given by 1334 students on evaluation question #34, the effectiveness of the instructor, was compared with the mean score derived from 190 students who evaluated the instructor after taking their final exam. A significant difference was demonstrated, but not as the authors hypothesized. Those students evaluating instructors' performance *after* final exams showed a significant increase in the evaluation score given by those students who did not have the benefit of knowing their course grade. This is counter-intuitive to the premise posited in H6. Perhaps there are more negative consequences to faculty by students having a "fear of the unknown" than from those who know the grade they have earned. Often students received grades higher than they originally expected. Also, completing evaluation forms when students are highly stressed with a full slate of final exams looming before them may have a detrimental impact in their evaluations of their instructors. Regardless, these results necessarily lead to other interesting questions worthy of future study.

A second round of informal interviews with instructors showed that Instructors still felt that poorly performing students, once they took or had seen the final, would anticipate their low grade and undergo the effort to evaluate the instructor, while good students would not make the effort and just wait till their grades were mailed to them.

5. Summary and Discussion

Table 26 presents a summary of the hypothesis and the results, some of which were quite surprising.

Response Rate: H1 showed a higher response rate for online evaluations, which was contrary to previously reported results in the literature (Thorpe, 2002). Despite the much higher rate (82% online vs. 67% paper), the result was not statistically significant. A closer examination of the results revealed an anomaly: the standard deviation for the electronic evaluation was much higher than for the paper (99 vs. 21, see Table 3). Additionally, the distribution was not normal (see Table 4).

We can therefore assume that for the electronic evaluation, some other factor influenced the result. Our best assumption for the surprisingly high electronic response rate is the following: Since the evaluations were conducted for the CIS department, a large number of classes were in an electronic classroom with computer workstations for students, instructors had reserved a computer lab for multiple days including the evaluation day, or the course was online to begin with. Such instructors could be more likely to use online evaluations to begin with, since the infrastructure supported them. If these were the majority, and a few outliers did not have the infrastructure in place, and had to rely on students going online at their own convenience, then this could explain the high response rate combined with a large standard deviation, assuming that few students went home or to a lab and evaluated the instructor from there. Unfortunately, we do not have access to that information at this point.

Rating of Teacher Effectiveness: While a response rate is of concern to the administration, instructors are even more concerned about how they were ranked. In our pilot case, instructors were concerned that students who knew their grade (i.e. since it was posted on webct or known otherwise) and were unhappy with it would go and evaluate faculty poorly – which in our case they could, since online evaluations closed after the final exam period, while paper based closed before.

Hypothesis H2 tested for this scenario, and found no statistical significant difference between the means of evaluation. H3 tested if good students were more likely to go online and evaluate an instructor than poorly performing students. This was not the case. H4 tested if poorly performing students would evaluate online, a major concern of faculty that might have 'used tricks' before to ensure that poorly performing students were not present at the time of the evaluation. Surprisingly, poorly performing students were less likely to evaluate a faculty member online then on paper. We can only assume that they are either too lazy or technologically challenged to master the online evaluation process.

H6 tested for this scenario, looking at whether the students who evaluated an instructor after the final exam period would rate the instructor higher or lower. Surprisingly, the results show that post-final exam evaluations were even higher than pre-final exam ones (3.99 vs 3.78). Again, it seems that good performing students are more likely to reward a faculty, while 'bad' students are less likely to evaluate an instructor altogether online.

Hypothesis H5 tested how good/bad students ranked instructors online in anticipation of their respective grade. There was no difference among good students or 'bad' students regarding the mean of evaluation, i.e. it did not matter whether the evaluation was conducted online or on paper. Not surprisingly, for both online and paper, good students ranked faculty higher than 'bad' students (4.2 vs. 3.8 out of 5).

In summary, our study found that under the given conditions, faculty would actually benefit from conducting their evaluations online. The administration made some changes to the process before moving completely to online evaluations: Students can now only access their grade online after they evaluate the faculty. So while they still can wait till after the final, at least they wont know their grade, unless faculty posts it on WebCT etc.

References

Boneau, C.A., 1960, "The effects of violations of assumptions underlying the *t* test," *Psychology Bulletin* 57: 49-64.

Box, G.E.P., 1953, "Non-normality and tests on variances." *Biometrika* 40: 318-335.

Cochran, W.G., 1947, "Some consequences when the assumptions for analysis of variance are not satisfied." *Biometrics* 3: 22-38.

Couper, M. , 2000, Web surveys: A review of issues and approaches. Public Opinion Quarterly 64, pp 464-494.

Cummings, R. and C. Ballatyne, 1999, Student feedback on teaching: Online! On target? Paper presented at the Australisian Society Annual Conference, October, 1999. (http://cleo.murdoch.edu.au/evaluation/pubs/confs/aes99.htm)

Dillman, D., 2000, Mail and Internet Surveys: The Tailored Design Method. New York: John Wiley & Sons.

Greenwald, Anthony G. and M. Gerald Gillmore. 1997, "Grading Leniency Is a Removable Contaminant of Student Ratings," *American Psychologist* 11: 1209-17.

Goldman, Louis., 1985, "The Betrayal of the Gatekeepers: Grade Inflation," *Journal of General Education* 37 : 97-121.

Handwerk, P., Carson, C., and K. Blackwell, 2000, "Online vs. paper-and -pencil surveying of students: A case study," Paper presented at the 40 th Annual Meeting of the Association of Institutional Research, May.

Hmieleski, K., 2000, Barriers to online evaluation: Surveying the nation's top 200 most wired colleges. Report prepared by the Interactive and Distance Education Assessment Laboratory at Rensselaer Polytechnic Institute, Troy, NY.

Huemer, M., accessed 2004, "Student Evaluations: A critical review" (Available online at http://home.sprynet.com/~owl1/sef.htm).

Kohr, R.L. and P.A. Games,,1974, "Robustness of the analysis of variance, the Welch procedure, and a Box procedure to heterogeneous variances." *Journal of Experimental Education* 43: 61-69.

Marsh, Herbert W. and Lawrence A. Roche, 1997, "Making Students' Evaluations of Teaching Effectiveness Effective," *American Psychologist* 52: 1187-97.

Matz, C. ,1999,. "Administration of web versus paper surveys: Mode effects and response rates.," Masters Research Paper, University of North Carolina at Chapel Hill. (ERIC Document Reproduction Service ED 439 694).

Rice, L , 1988, "Student Evaluation of Teaching: Problems and Prospects," *Teaching Philosophy* 11: 329-44.

Sax, L., Gilmartin, S., J. Keup, A. Bryant, and M. Plecha, 2002, Findings from the 2001 pilot administration of Your First College Year (YFCY): National norms. Higher Education Research Institute, University of California. (Available online at http://www.gseis.ucla.edu/heri/yfcy/yfcy_report_02.pdf)

Sacks, Peter. *Generation X Goes to College* (LaSalle, IL: Open Court, 1986).

Srivastava, A.B.L., 1958, "Effect of non-normality on the power function of the *t*-test. *Biometrika* 45: 421-429

Tomsic, M., D. Hendel, R. Matross, 2000, A world wide web response to student satisfaction surveys: Comparisons using paper and Internet formats. Paper presented at the 40th Annual Meeting of the Association of Institutional Research, May 2000.

Thorpe, S.,2001, "Linking learning outcomes to student course evaluation," Paper presented at the 28 th Annual Conference of the Northeast Association for Institutional Research, Boston MA, November.

Thorpe, S., 2002, "Online Student Evaluation of Instruction: An Investigation of Non-Response Bias," Paper presented at the 42nd Annual Forum of the Association for Institutional Research Toronto, Canada, June.

Underwood, D., H. Kim, and, M. Matier, 2000, "To mail or to web: Comparisons of survey response rates and respondent characteristics," Paper presented at the 40 th Annual Meeting of the Association of Institutional Research, May.

Wilson, Robin, 1998, "New Research Casts Doubt on Value of Student Evaluations of Professors," *Chronicle of Higher Education* (Jan. 16, 1998): A12

Zar, J., 1984, Biostatistical Analysis, Second Edition, Prentice-Hall, Englewood Cliffs, N.J.

Table 3: Average response rate for Spring 2002 through Fall 2003

Group Statistics

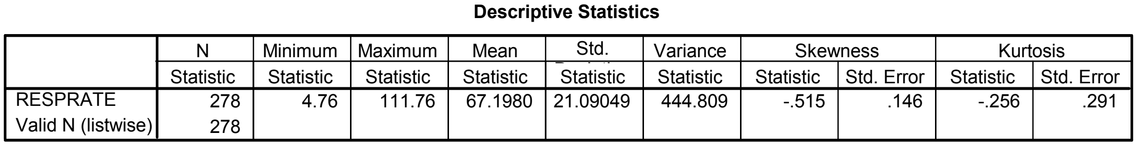
	0=paper, 1=electronic	N	Mean	Std. Deviation	Std. Error Mean
RESPRATE	.00	278	67.1980	21.09049	1.26492
	1.00	61	81.7528	99.70227	12.76557

Table 4: Independent *t*-test for the average response rate of Spring 2002 through Fall 2003

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
RESPRATE	Equal variances assumed	20.347	.000	-2.228	337	.027	-14.5548	6.53366	27.40666	-1.70286
	Equal variances not assumed			-1.135	61.183	.261	-14.5548	12.82808	40.20455	11.09503

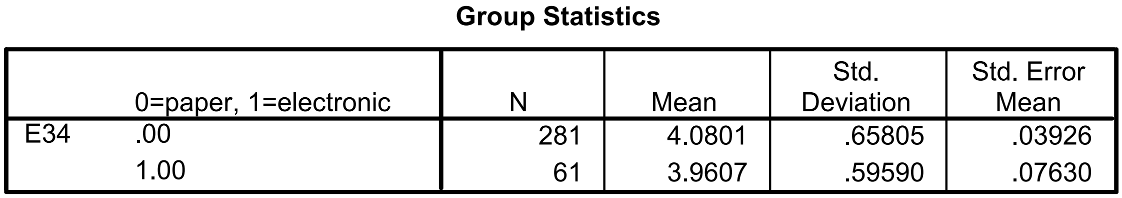
Table 5: Test for skewness and kurtosis of data



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
RESPRATE	278	4.76	111.76	67.1980	21.09049	444.809	-.515	.146	-.256	.291
Valid N (listwise)	278									

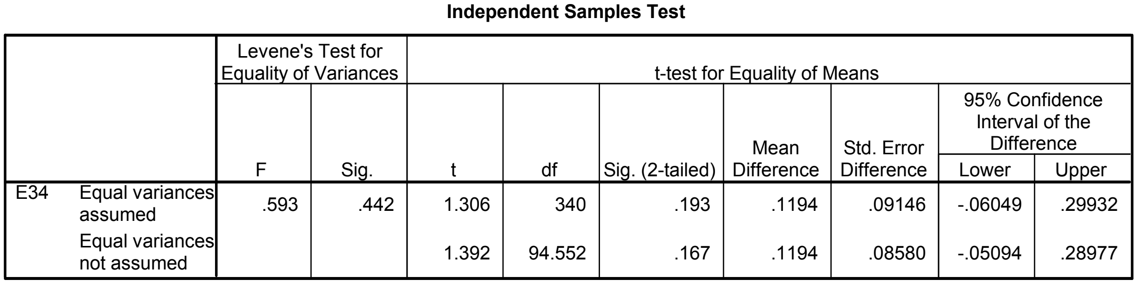
Table 6: Average score on question #34 for Spring 2002 through Fall 2003



Group Statistics

	0=paper, 1=electronic	N	Mean	Std. Deviation	Std. Error Mean
E34	.00	281	4.0801	.65805	.03926
	1.00	61	3.9607	.59590	.07630

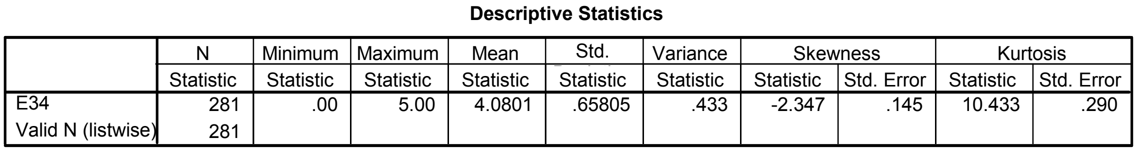
Table 7: Independent *t*-test on question #34 for Spring 2002 through Fall 2003



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
E34	Equal variances assumed	.593	.442	1.306	340	.193	.1194	.09146	-.06049	.29932
	Equal variances not assumed			1.392	94.552	.167	.1194	.08580	-.05094	.28977

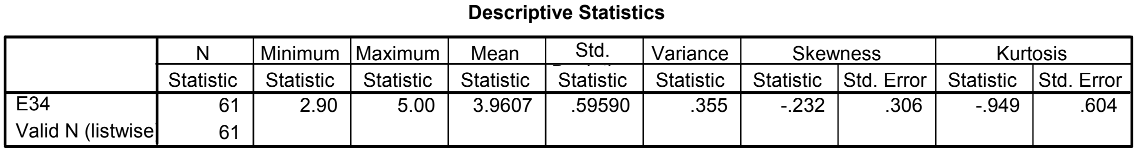
Tables: 8 Test for skewness and kurtosis of data for Question 34 (paper-based)



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
E34	281	.00	5.00	4.0801	.65805	.433	-2.347	.145	10.433	.290
Valid N (listwise)	281									

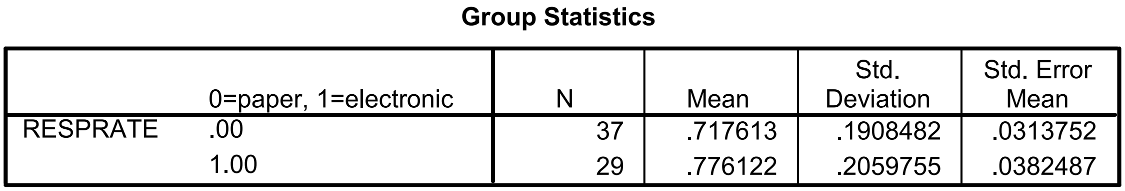
Tables: 9: Test for skewness and kurtosis of data for Question 34 (online)



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
E34	61	2.90	5.00	3.9607	.59590	.355	-.232	.306	-.949	.604
Valid N (listwise	61									

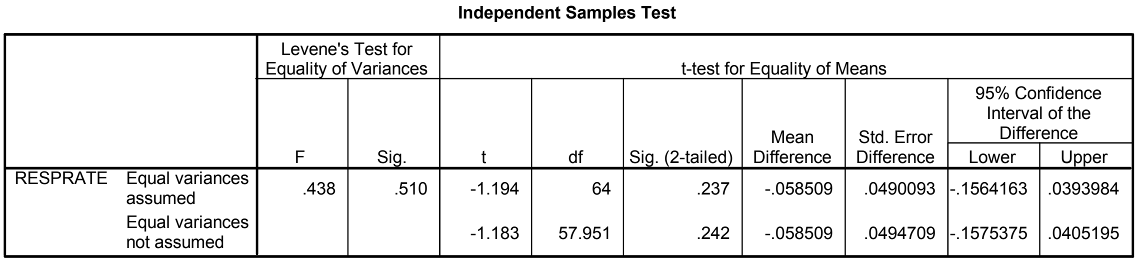
Tables: 10 Comparison of "good" students rating courses with paper vs. those rating online



Group Statistics

	0=paper, 1=electronic	N	Mean	Std. Deviation	Std. Error Mean
RESPRATE	.00	37	.717613	.1908482	.0313752
	1.00	29	.776122	.2059755	.0382487

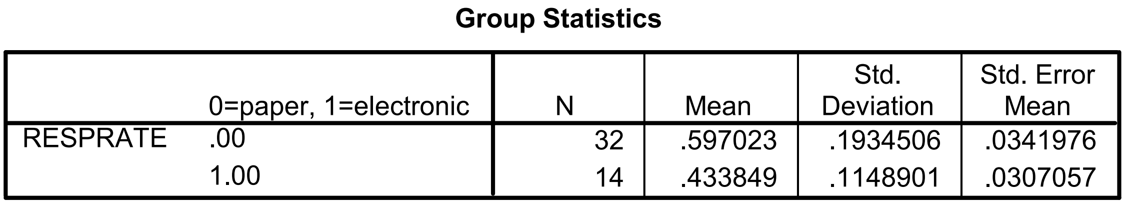
Table 11: *t*-test of "good" students' response rate using paper evaluations vs. the response rate of those using online evaluations.



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
RESPRATE	Equal variances assumed	.438	.510	-1.194	64	.237	-.058509	.0490093	-.1564163	.0393984
	Equal variances not assumed			-1.183	57.951	.242	-.058509	.0494709	-.1575375	.0405195

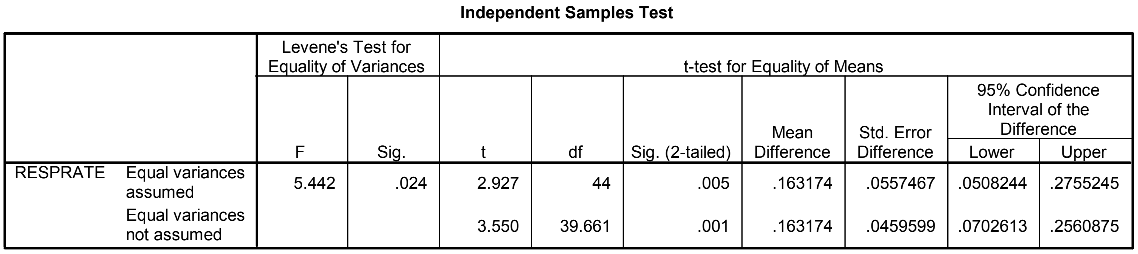
Tables 12: Comparison of "weaker" students' response rate in courses with paper vs. online evaluations.



Group Statistics

	0=paper, 1=electronic	N	Mean	Std. Deviation	Std. Error Mean
RESPRATE	.00	32	.597023	.1934506	.0341976
	1.00	14	.433849	.1148901	.0307057

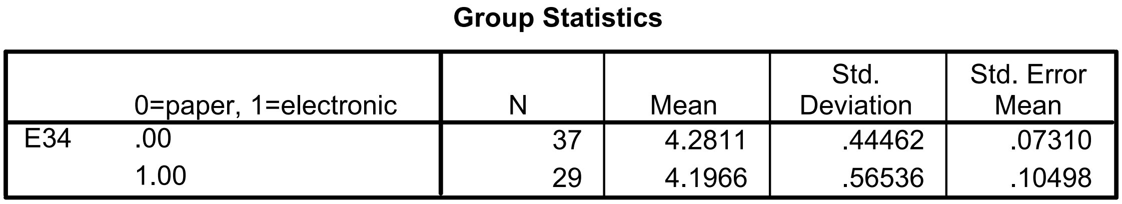
Table 13: *t*-test of "weaker" students' response rate in courses with paper vs. online evaluations.



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
RESPRATE	Equal variances assumed	5.442	.024	2.927	44	.005	.163174	.0557467	.0508244	.2755245
	Equal variances not assumed			3.550	39.661	.001	.163174	.0459599	.0702613	.2560875

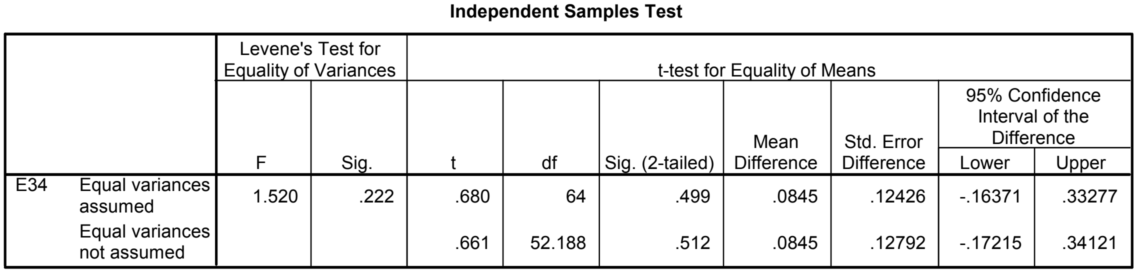
Table 14: Comparison of "good" students' instructor evaluation in courses with paper vs. online evaluations



Group Statistics

	0=paper, 1=electronic	N	Mean	Std. Deviation	Std. Error Mean
E34	.00	37	4.2811	.44462	.07310
	1.00	29	4.1966	.56536	.10498

Table 15: *t*-test of "good" students' instructor evaluations in courses with paper vs. online evaluations



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
E34	Equal variances assumed	1.520	.222	.680	64	.499	.0845	.12426	-.16371	.33277
	Equal variances not assumed			.661	52.188	.512	.0845	.12792	-.17215	.34121

Table 16: Comparison of "weaker" students' instructor evaluation in courses with paper vs. online evaluations

Group Statistics

	0=paper, 1=electronic	N	Mean	Std. Deviation	Std. Error Mean
E34	.00	32	3.8281	.48276	.08534
	1.00	14	3.8071	.60570	.16188

Table 17: *t*-test of "weaker" students' instructor evaluations in courses with paper vs. online evaluations

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
E34	Equal variances assumed	1.657	.205	.125	44	.901	.0210	.16730	-.31619	.35815
	Equal variances not assumed			.115	20.564	.910	.0210	.18300	-.36007	.40204

Table 18: Comparison of instructor evaluations by "weaker" students vs. "good" students

Group Statistics

	SMART	N	Mean	Std. Deviation	Std. Error Mean
E34	.00	327	3.7550	.50570	.02797
	1.00	275	3.9497	.56360	.03399

Table 19: Comparison of instructor evaluations by "weaker" students vs "good students"

Independent Samples Test

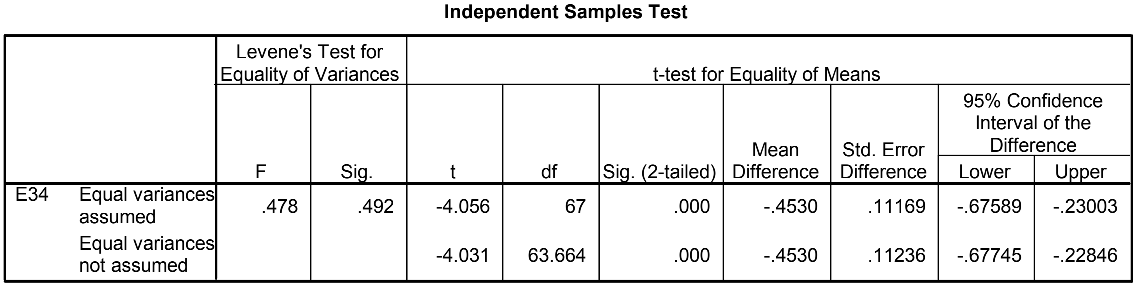
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
E34	Equal variances assumed	3.121	.078	-4.464	600	.000	-.1947	.04360	-.28030	-.10903
	Equal variances not assumed			-4.423	556.298	.000	-.1947	.04401	-.28112	-.10821

Table 20: Comparison of instructor evaluations by "weaker" students vs. "good" students (online methodology only)

Group Statistics

	SMART	N	Mean	Std. Deviation	Std. Error Mean
E34	.00	32	3.8281	.48276	.08534
	1.00	37	4.2811	.44462	.07310

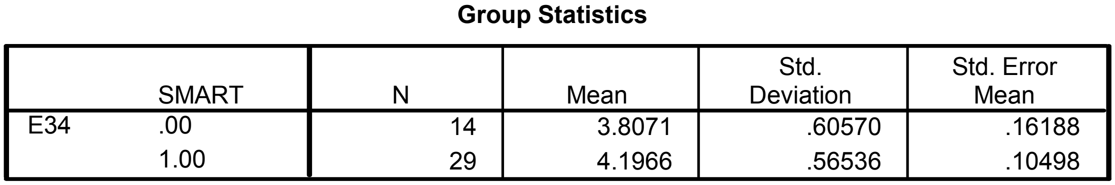
Table 21: Comparison of instructor evaluations by “weaker” students vs. “good” students (online methodology only)



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
E34	Equal variances assumed	.478	.492	-4.056	67	.000	-.4530	.11169	-.67589	-.23003
	Equal variances not assumed			-4.031	63.664	.000	-.4530	.11236	-.67745	-.22846

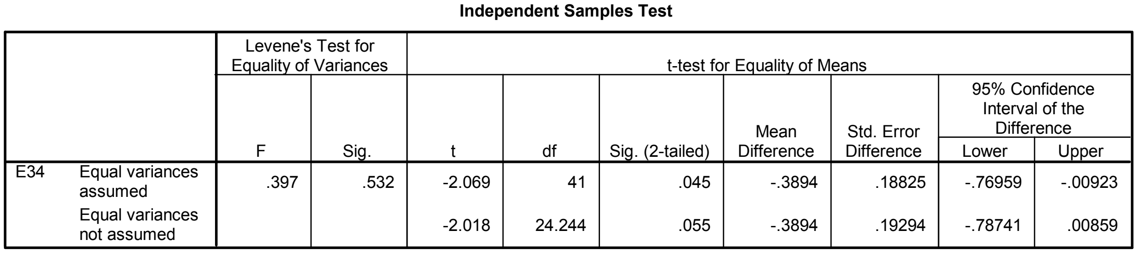
Table 22: Comparison of instructor evaluations by “weaker” students vs. “good” students (paper-based evaluations only)



Group Statistics

	SMART	N	Mean	Std. Deviation	Std. Error Mean
E34	.00	14	3.8071	.60570	.16188
	1.00	29	4.1966	.56536	.10498

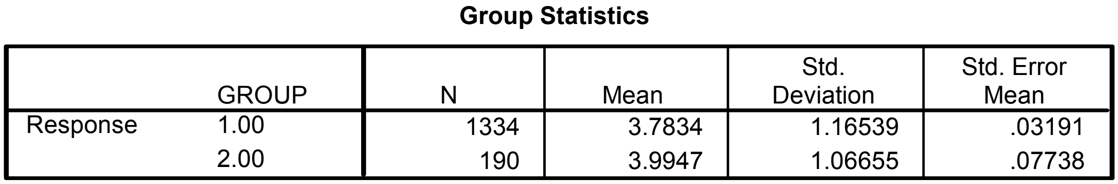
Table 23: Comparison of instructor evaluations by “weaker” students vs. “good” students (paper-based evaluations only)



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
E34	Equal variances assumed	.397	.532	-2.069	41	.045	-.3894	.18825	-.76959	-.00923
	Equal variances not assumed			-2.018	24.244	.055	-.3894	.19294	-.78741	.00859

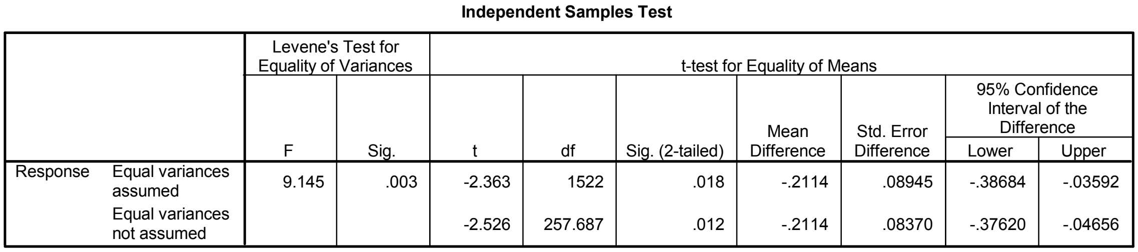
Table 24: Comparison of students completing an evaluation prior to final exams with those who used online evaluations during and after taking final exams



Group Statistics

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
Response	1.00	1334	3.7834	1.16539	.03191
	2.00	190	3.9947	1.06655	.07738

Table 25: T-test of scores given by students before taking final exams with those completing the instructor evaluation after taking the final exam.



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Response	Equal variances assumed	9.145	.003	-2.363	1522	.018	-.2114	.08945	-.38684	-.03592
	Equal variances not assumed			-2.526	257.687	.012	-.2114	.08370	-.37620	-.04656

Table 26: Summary of Hypothesis and Results

	Description	Test Description	N	Variable Means	p.
H1	Online evaluations have a lower response rate than paper-based systems	Comparison of student response rates of two evaluation methodologies	278 – Paper 61 - Online	Response Rate 67.19% 81.75%	.261
H2	Online evaluations have lower scores than paper based systems	Comparison of teacher effectiveness evaluations	281 – Paper 61 - Online	Teacher Effectiveness (Q34) 4.08 3.96	.167
H3	Good students are less likely to fill out online evaluations than offline	Likelihood of "good" students to fill out paper evaluations of instructor	37 – Paper 29 - Online	Response Rate 71.7% 77.6%	.242
H4	Weaker students are more likely to fill out online evaluations than offline	Likelihood of "weaker" students to use online evaluations of instructor	32 – Paper 14 - Online	Response Rate 59.7% 43.3% (opposite of H4)	.001*
H5	Students expecting a low grade are more likely to give low evaluations than students expecting a high grade	Comparison of "good" students use of evaluation methodology to rate instructors' effectiveness	37 – Paper 29 - Online	Teacher Effectiveness (Q34) 4.28 4.19	.512
H5	Cont.	Comparison of "weaker" students use of evaluation methodology to rate instructors' effectiveness	32 – Paper 14 - Online	Teacher Effectiveness (Q34) 3.82 3.80	.910
H5		Comparison of good vs. weak students' evaluations regardless of evaluation methodology	327 – Weak 275 - Good	Teacher Effectiveness (Q34) 3.75 3.94	.000*
H5		Use of online evaluation only	32 – Weak 37 - Good	Teacher Effectiveness (Q34) 3.82 4.28	.000*
H5		Use of paper evaluation only	14 – Weak 29 - Good	Teacher Effectiveness (Q34) 3.80 4.19	.055*
H6	Students who expect low grades are more likely to fill out a late online evaluation		1334 – Pre-final exam 190 – Post final exam	Teacher Effectiveness (Q34) 3.78 3.99	.012*

Summary of results
* = significance