EVALUATING THE IMPACT OF MULTIMEDIA LECTURES ON STUDENT LEARNING AND ATTITUDES ®

Sterling C. Hilton and Howard B. Christensen
Brigham Young University
USA

This paper presents the results from a large, randomized, controlled experiment conducted in the introductory statistics course at Brigham Young University. The purpose of the study was to assess the impact of multimedia lectures on student learning and attitudes. A randomized complete block design was implemented to evaluate the treatment that had two levels: multimedia versus overhead transparencies. Data was collected over four semesters on 5,603 students. Several student characteristics were measured and controlled for in the analyses. Our findings indicate that the multimedia lectures did not improve student learning or attitudes compared to the control group. However, our research also indicates that large, randomized, controlled experiments can be implemented in educational research. We advocate their use as the standard method of evaluation for educational innovations.

INTRODUCTION

With advances in technology occurring at a rapid pace, there is an existing movement to incorporate higher levels of technology into the educational arena; however, there is no clear evidence that advanced technologies improve student learning. This push to use the latest technologies in the university classroom is simply symptomatic of a larger problem in education research: there is too little rigorous evaluation of educational innovations. Our sustained inquiry into existing educational literatures has led us to this conclusion. A similar conclusion was reached by Mosteller, Light and Sachs in 1996 after their sustained inquiry into the empirical evidence for the common practice of grouping students according to skill level for certain subjects. They state, “Our exploration reveals that too little sustained evaluation of current practice and innovations is now being carried out [in education]” (p798).

Over the past decade, our department has experienced increasing enrollments in our introductory statistics course. For the past three years we have taught about 1500 students each semester. Pressured by increasing enrollment, our department sought to better utilize departmental resources and to offer a more uniform and improved learning experience for students taking introductory statistics. To this end our department embarked on what it called the Learning Research Initiative in October of 1997. We also sought a better understanding of how using more technology in the presentation of course material would impact on student outcomes. Therefore, we designed an experiment that would allow us to collect data that could be used to address this issue.

METHODS

The primary introductory statistics course taught by our department can be used to satisfy a university general education requirement and is a required course for several majors on campus. Over the course of the data collection phase of our experiment, class size ranged from 160 to 200 students with either 6 or 8 classes taught per semester. Classes met three times a week for a 50 minute lecture, and students also enrolled in an one hour weekly recitation session (lab) consisting of 20 students. Due to personnel and space constraints, our class size is large which required the use of a lecture format to present material to the students. While not all universities and departments have these same limitations, we feel that our situation is not unique (Wild, Triggs, & Pfannkuch, 1997).

We chose to assess the impact of advanced technologies in the classroom on student outcomes for a variety of reasons. First, our university administration has made a concerted push to incorporate advanced technologies in the classroom. Second, the classroom is still the major point of contact between the instructor and student in our department. Third, there is some evidence in the literature that learners prefer group-based multimedia to individualized multimedia (Smith, Hsu, Azzarello, & McMichael, 1993).

The multimedia treatment consisted of 42 PowerPoint presentations that contain animation to get and hold student attention, video clips and sound to motivate statistical concepts using real-world situations, and computer applets to dynamically and graphically illustrate statistical concepts and principles using real data. The control level of the treatment consisted of overhead transparencies that were created from the multimedia presentations. In this way, the content of the lectures and the presentation sequence were very similar for the two treatment levels. This is important, since it is important to avoid confounding content with presentation method in order to assess the impact of the multimedia technology on learning.

The experimental design used in this study was a randomized complete block design with instructor as the blocking factor. Because the treatment variable is presentation method, the experimental unit in this study is the class, not the student; therefore, each semester, every instructor was randomly assigned to teach one class using the multimedia presentation and one class using the overhead presentation. The following constraints were imposed on the randomization scheme: 1) there had to be at least a two hour break between the classes which meant that each instructor taught one class in the morning (9:00 or 10:00) and one class in the afternoon (1:00 or 2:00), and 2) the method of lecture presentation was balanced with time of day which meant that two of the morning classes were taught using multimedia and two were taught using overhead transparencies.

Given the constraints of the university enrollment system, students were not randomly assigned to the class. Even though the student is not the experimental unit, there are at least reasons for wanting to randomly assign students to the class. The first would be to avoid potential selection bias occurring when students enroll in a class. In an effort to eliminate selection bias, students were only given information regarding the time and location of the class during the enrollment process; they were blinded to the instructor and the method of lecture presentation. Even though students can transfer classes during the first two weeks of a semester, data collected from one semester indicated that only 4% of the students actually transferred classes and only 1% of students transferred from one treatment level to the other.

The second reason for wanting to randomly assign students to classes is to improve the likelihood of the treatment and control groups being similar with regard to relevant student characteristics; however, we anticipated that due to the diversity of students' schedules and the large numbers of students in the study, the two treatment groups would be comparable with regard to relevant student characteristics. Table 1 presents descriptive statistics of the students enrolled in our introductory course over the study period. The information is stratified by treatment group in order to better assess the comparability of the two groups. Hypothesis tests of equality between the two groups were done (t-tests for quantitative variables and chi-square tests for categorical variables) with the resulting p-values are reported in Table 1.

Based upon the information presented in Table 1, it is fairly clear that the multimedia and overhead transparencies groups are quite similar with respect to student characteristics. The distributions of gender, major, previous statistics experience and most recent math class taken are all comparable for the two groups, as are the distributions for ACT composite and math scores, and university GPA. Statistically significant differences were found in the distributions of age and academic class. These two variables are clearly associated with each other since a majority of students enter the university at age 18 and take sufficient credit to move from one academic class to the next each year. Despite being significantly different, the observed differences have little practical significance, e.g. the average age in the control group is only 2.4 months higher than the average age in the multimedia group.

Student learning and student attitudes are the primary outcomes of the experiment. Scores on the final and three mid-term examinations are the primary measures of student learning. Exams were revised each semester using the following guideline: improve the quality of questions while maintaining a high level of content consistency in the questions across semesters. Common questions across the four semesters of the study were identified from each of the four exams. The percentage of common problems ranged from 74% for mid-term 1 to 90% for mid-term exam 2. The percent correct of the problems common across the four semesters for each of the four exams are the primary measures of student learning used in this report.

Table 1
Mean and Percent Differences for Student Characteristics Stratified by Treatment Group

Variable	Multimedia Mean (s.d.)	Control Mean (s.d.)	p-value
Age	21.3 (2.6)	21.5 (2.6)	.02
Math ACT	25.1 (4.4)	25.2 (4.3)	.68
Composite ACT	25.3 (3.7)	25.2 (3.7)	.66
Overall GPA	3.25 (0.67)	3.28 (0.63)	.12
	%	%	
Gender	N=2850	N=2749	.37
Male	50.6	51.8	
Female	49.4	48.2	
Academic class	N=2850	N=2749	.0001
Freshman	12.0	8.8	
Sophomore	35.7	32.0	
Junior	28.6	32.8	
Senior	23.7	26.4	
Previous Statistics course	N=2768	N=2653	.67
Yes	15.4	14.9	
No	84.6	85.1	
Declared major	N=2768	N=2753	.31
Biology and Agriculture	7.7	9.0	
Business Management	24.7	24.3	
Education	5.7	5.1	
Engineering	5.5	5.5	
Fine Arts/Communications.	8.5	7.4	
Health & Human Performance	9.9	11.0	
Humanities	3.0	2.8	
Nursing	3.4	3.8	
Physical. & Math. Science	3.3	2.6	
Social Science	15.7	15.1	
Open	12.6	3.4	
Visual/Verbal Learning Style	N=2766	N=2655	.43
Strong Visual	23.9	25.7	
Moderate Visual	34.8	32.9	
Balanced		35.8	36.3
Moderate Verbal	4.5	4.3	
Strong Verbal	0.9	0.8	

Student attitudes were measured using the *Survey of Attitudes toward Statistics* (SATS) (Schau, Stevens, Dauphinee, & Del Vecchio, 1995). Attitudes were measured at the beginning and end of the semester. The primary attitude outcomes are the differences between the pre- and post-semester scores on the SATS. he SATS is a 28 item, 7 point Likert instrument that has four sub-scales: affect-positive and negative feelings concerning statistics, cognitive competence-attitudes about intellectual knowledge and skills when applied to statistics, value-attitudes about the usefulness and worth of statistics in personal and professional life, and difficulty-attitudes about the difficulty of statistics as a subject. Each sub-scale has a range of 1(strong negative feelings or attitudes) to 7 (strong positive feelings or attitudes).

Exam scores and changes in attitudes were analyzed using linear, mixed models. Random factors included in the final model are semester, instructor, class, lab section and student. Fixed factors included in the final model are treatment, gender, major and previous course in statistics. Covariates included in the final model are learning styles, ACT math score, overall university GPA, lecture attendance, learning styles, and a pre-semester, self-reported measure of confidence in their ability to master the course material.

RESULTS

Results from the analyses of the student learning outcomes are presented in Table 2. All reported differences are the estimated average effects after controlling for all other variables in the model. There is no significant treatment effect for any of the exam scores. Estimated average differences in exam scores between the multimedia and control groups were never larger than 0.6%. There is a consistent and significant gender effect. The estimated average differences in exam scores between males and females were roughly 3% in favor of males. Having had a previous statistics course was statistically significant across the four exams, although the estimated effect declined over the four exams. On average, those students with previous statistics experience performed 4% better on exam 1 and 2% better on the final compared to those students without previous statistics experience.

Table 2
Results from the Analyses of Student Learning Outcomes

	Exam 1	Exam 2	Exam 3	Final
		F-tests and p-values		
Fixed effects				
treatment	0.80 (0.38)	0.00 (0.99)	0.59 (0.45)	0.00 (0.95)
gender	11.86 (<.01)	35.28 (<.01)	15.17 (<.01)	23.89 (<.01)
major	0.76 (0.67)	4.40 (<.01)	2.01 (0.03)	2.79 (<.01)
major*gender	1.34 (0.20)	0.92 (0.52)	1.62 (0.09)	1.74 (0.07)
previous stat course	22.92 (<.01)	11.22 (<.01)	7.61 (<.01)	6.58 (<.01)
		Coefficients and standard errors		
Covariates				
ACT math	0.64 (.05)	0.64 (.05)	0.73 (.06)	0.79 (.05)
overall GPA	3.17 (.23)	3.23 (.24)	4.11 (.30)	4.39 (.25)
class attendance	0.09 (.01)	0.13 (.01)	0.14 (.01)	0.14 (.01)
self confidence	1.11 (.06)	1.30 (.13)	0.99 (.17)	1.42 (.14)
learning style				
active/reflective	-0.18 (.06)	-0.23 (.06)	-0.26 (.07)	-0.14 (.06)
sequential/global	0.12 (.06)	0.06 (.06)	0.43 (.08)	0.21 (.06)
		Estimates and standard errors		
Random terms				
semester	3.43 (2.98)	4.85 (4.30)	5.57 (4.81)	3.68 (3.22)
instructor	0.65 (0.52)	0.74 (0.70)	0.10 (0.34)	0.0
class	0.0	1.11 (0.56)	0.78 (0.59)	1.07 (0.53)
lab	1.71 (0.67)	0.0	0.52 (0.94)	0.26 (0.70)
student	99.09 (2.10)	101.51 (2.10)	159.09 (3.42)	116.32 (2.51)
Effect		Least square means		
treatment				
multimedia	85.29	83.51	78.32	78.95
control	85.59	83.50	78.71	78.92
Gender				
female	84.50	81.87	77.16	77.48
male	86.38	85.14	79.87	80.39

As expected, students with better math skills as measured by their ACT math score did significantly better on all four exams. The estimated effect (0.7% increase for every 1 point increase in ACT math score) was generally consistent across the four exams. Also, students with higher GPA at the university performed significantly better on average than students with lower university GPA.

There was a significantly positive association between class attendance and exam scores; however, it isn't very practically significant: students who attend 10% more often, score on average 1.3% higher on their exams. The student-to-student variability in these data is orders of magnitudes greater than the variability across semesters, instructors, classes, or lab sections. In fact, there is very little variability in the student learning outcomes that can be attributed to instructor, class or lab section.

Results from the mixed model analyses of student attitudes are not presented in this paper due to space constraints. However, there was no significant treatment effect on any of the four sub-scales. Gender and major were significant across all four sub-scales; overall GPA and previous statistics experience were significant on some of the sub-scales. Despite the statistical significance of some of the fixed factors and covariates, none of the effects had any practical importance.

Table 3 presents the marginal distributions of the post-semester SATS sub-scales. The sub-scales were categorized into three categories: negative (1-3.0), neutral (3.1- 4.9) and positive (5.0 - 7). For example, at the end of a semester, 25% of the students have positive feelings towards statistics, 52% have neutral feelings and 23% have negative feelings. The most salient findings with respect to changes in student attitudes are presented in Table 4, which summarizes the two-way contingency tables between pre- and post-semester sub-scales with regard to change. While the majority of students do not change their attitudes regarding statistics over the semester, for those students that do, a significantly larger proportion of students' attitudes become more negative than positive. This finding is significant for each sub-scale. A negative shift in attitude is defined as either moving from positive on the pre-semester sub-scale to neutral or negative on the post-semester or moving from neutral on the pre-semester to negative on the post. These findings are not necessarily undesirable outcomes. Since 85% of the students have never had any formal exposure to statistics, it is quite possible that the pre-semester attitudes reflect either optimism, naivete or both. Post-semester attitudes, however, may be more realistic, i.e., the students may have been educated.

Table 3
Marginal Distributions of Post-Semester SATS Sub-Scales

Sub-scale	Negative (1-3) N (%)	Neutral (3.1-4.9) N (%)	Positive (5-7) N (%)
affect	1347 (25.6)	2542 (48.3)	1374 (26.1)
cognitive competence	443 (8.4)	2175 (41.3)	2645 (50.3)
value	464 (8.8)	2403 (45.7)	2396 (45.5)
difficulty	1434 (27.2)	3441 (65.4)	388 (7.4)

Table 4
Summarization of two-way Contingency Tables between pre- and post- Attitudes for SATS Sub-Scale

Sub-scale	% negative shift	% positive shift	% no shift
affect	34	14	52
cognitive competence	25	13	62
value	30	10	60
difficulty	23	12	65

CONCLUSIONS

In summary, our research indicates that a high-level multimedia lecture presentation does not improve student learning nor student attitudes in an introductory statistics course compared to a lecture presentation using overhead transparencies. Therefore, justification for adopting a higher level of technology in the introductory statistics classroom must be found elsewhere. Of course, these findings are only generalizable to situations similar to ours; however, they may lead one to question whether higher levels of technology used in smaller classes or in more interactive ways would improve student learning and attitudes.

Of equal or greater importance to the substantive findings of our research, are the methods we used to conduct and evaluate our study. Our research of the past four years demonstrate that a large randomized, controlled experiment is feasible, albeit extremely difficult. Admittedly, rigorous evaluation of educational innovations is difficult because it involves many teachers and numerous students, and there are a myriad of factors that influence learning; however, there are parallels with the evaluation of medical innovations. Medical researchers have to grapple with doctors and patients and a number of other factors that influence health. The medical community has adopted clinical trials as the standard for evaluating medical innovations. A clinical trial is simply a randomized, controlled experiment. We feel that a similar standard should be adopted by the educational research community, especially in the statistical education research community. Only in this way will be able to ascertain which innovations really improve student outcomes.

REFERENCES

Mosteller F., Light R.J., & Sachs, J.A. (1996). Sustained inquiry in education: lessons from skill grouping and class size. *Harvard Educational Review*, *66,* 797-842.

Schau, C., Stevens, J., Dauphinee, T.L., & Del Vecchio, A. (1995). The development and validation of the Survey of Attitudes toward Statistics. *Educational and Psychological Measurement*, *55*, 868-875.

Smith, P.L., Hsu, S. Azzarello, J., & McMichael, J. (1993). *Group-based multimedia research conclusions and future questions*. Paper presented at the Association for Educational Communications and Technology Convention, New Orleans, LA.

Wild, C., Triggs, C., & Pfannkuch, M. (1997). Assessment on a budget: using traditional methods imaginatively. In I. Gal and J.B. Garfield (Eds.), *The assessment challenge in statistics education*, (pp. 205-220). Amsterdam; IOS Press,