Technology talks: Clickers and grading incentive in the large lecture hall

Shannon D. Willoughby[a] and Eric Gustafson
Department of Physics, Montana State University, Bozeman, Montana 59717

(Received 5 June 2008; accepted 15 October 2008)

Two sections of an introductory astronomy class were given different grading incentives for clicker participation for two consecutive semesters. In the high stakes classroom points were awarded only for correct answers, in contrast to the low stakes classroom in which points were awarded simply for participating. Self-formed groups of four students each were recorded in both sections several times during the spring 2007 semester and their conversations were transcribed and categorized into nine topics to analyze the variations between the sections. Performance on clicker questions and tendency to block vote were correlated with class grades and gains for the pre- and post-test scores on the Astronomy Diagnostic Test. © *2009 American Association of Physics Teachers.*
[DOI: 10.1119/1.3013542]

I. INTRODUCTION

Personal response systems have been used for many years,[1] but have only recently gained widespread use. General overviews of clickers and their use in the classroom have been given recently.[2,3] Beatty *et al.*[4] outlined methods for developing effective clicker questions and others have detailed the effect of awarding points for choosing the correct answer versus awarding points simply for participating.[5,6] Another study explored giving a clicker to each group versus giving a clicker to each student, and developed a set of three questions of increasing difficulty for each topic to meaure the level of student understanding.[7] Several groups have studied the use of clickers in physics classrooms.[8,9]

James' study[5] involved two classes taught by different instructors, each using different grading rubrics for clicker questions. The larger enrollment course was a high stakes classroom in which clicker use counted for 12.5% of the class grade and incorrect responses earned one third the points of a correct response. In the low stakes classroom the clicker portion was 20% of the student's overall grade, and all clicker responses earned the same number of points, regardless of their correctness. In both sections students were allowed to pair up and discuss the clicker questions. The conversation bias within student groups was defined as the difference between the percentage of words spoken by each person. James found that conversation bias in the low stakes classroom was small (mean of 14.8%), with each student expressing ideas equally. The high stakes classroom showed a larger conversation bias (mean of 33.2%) especially when there was a large gap in student knowledge. Conversations in the high stakes classroom tended to focus on the dominant partner's choice of answer, instead of each student expressing ideas in the more balanced manner seen in the other classroom. Students' voting habits were also analyzed, with a focus on whether or not partners chose the same answer or not. In the low stakes classroom students voted differently 36.8% of the time, but voting differently occurred only 7.6% of the time in the high stakes classroom. The conversation bias and block voting seen in the high stakes classroom suggest that the degree of apparent student understanding may be inflated in this environment and not reflect what students actually think. When student conversations are dominated by one partner, the use of clickers to spark discussion and interest does not occur in a manner that most instructors would prefer.

Another study of grading incentives used a slightly different method to encourage student discussion. Len[6] asked two types of clicker questions: introductory and review. Points were awarded for both types of questions for clicking any answer. The points for review questions were doubled when at least 80% of the class chose the correct answer. Len divided students into self-described groups of self-testers or collaborators on introductory questions (all but one student described themselves as collaborators on review questions). Self-testers chose an answer on their own for introductory questions, whereas collaborators considered other students' views while voting. Len examined each group's attitudes toward astronomy, its perceived affect, relevance, difficulty, and student's cognitive competence by a pre- and post-test. He found that collaborative students' cognitive competence decreased over the course of the semester, as did their perceived value of the subject. Students who were identified as self-testers showed opposite results for both, with perceived difficulty of material and effect remaining relatively unchanged for both groups.

The purpose of this study is to explore how the grading of clicker questions affected students' tendency to block vote on the questions, overall course grade, and learning gains. Although attitude testing in high and low stakes classrooms would be useful, we have not done so, and hence comparisons to Ref. 6 are not yet relevant.

II. METHODS

A. Study design

Mysteries of the Sky is a popular course at Montana State University with each of the two sections offered every semester capped at 200 students. The vast majority of students in this course are non-science majors, and sections are typically half male and half female. During this study (done during the spring and fall semesters of 2007) the four sections had the same instructor (Willoughby); the classes were taught with the same materials, textbook, and lectures, and nearly identical clicker questions. Clickers were required in both sections of the course for both semesters.[10] The only difference between the sections was how points were awarded for clicker participation. This participation was worth 4% of the students' overall grade to encourage students to purchase a clicker, but was not a large enough percentage of their grade to be onerous. Only during the first day of class was the point breakdown discussed by the in-

Table I. Average gains, grades, block voting and average percentage of correct clicker responses by section and semester.

Semester, section	Average gain $\langle g \rangle$		Final grade		Block voting (%)		Average correct responses (%)	
Spring, high stakes classroom	0.207	$p=0.674$	0.827	$p=0.124$	69.5	$p<0.0005$	56.6	$p<0.0005$
Spring, low stakes classroom	0.217		0.810		45.4		49.5	
Fall, high stakes classroom	0.161	$p=0.499$	0.785	$p=0.556$	59.2	$p=0.331$	60.3	$p=0.7$
Fall, low stakes classroom	0.144		0.776		55.2		59.4	

structor, after which it was not mentioned again. One point was awarded for a correct answer (no points for the incorrect answer) in the high stakes classroom and one point was awarded for clicking any answer in the low stakes classroom. Self-formed groups of four students each were recorded several times during the spring 2007 semester with digital voice recorders while they discussed clicker questions presented to the class. Conversation topics from recorded sessions were placed into nine categories, and total word counts were tabulated for each group.

On the first day of classes and during the last week of the semester students were given the Astronomy Diagnostic Test,[11] a reliable and validated exam on general astronomy knowledge usually taught in high school science courses. Gains calculated from the Astronomy Diagnostic Test were tabulated along with the overall course grade for each student.

B. Data collection

In both recording and nonrecording phases of the study clicker questions were asked during every class. Discussions were encouraged among group members and students were reminded regularly to do so. The first semester we used digital recorders several times throughout the course to record learning groups while they discussed a particular clicker question. These qualitative data added richness to the statistical portion of our study and allowed us to probe possible differences between the nature of conversations in the low versus high stakes classroom. Students were given informed consent forms in which they were given the choice to be recorded during the semester (possibly more than once) while discussing clicker questions.[12] As an incentive to opt in, three gift certificates to a popular store were raffled at the end of the semester, regardless of whether or not they had ever actually been recorded. Groups in which all members had signed consent forms were chosen at random and asked to record themselves while discussing the clicker questions.

This study continued in the fall semester, but we did not record students during this semester. The syllabi again made explicit the manner in which points were to be awarded, with each section being awarded points in the same manner as in the spring semester. Ninety two percent ($n=703$) of the students purchased and regularly used their clickers in class.

C. Data analysis

The discussions were placed into nine categories (similar to the rubric used in Ref. 5) to quantize the ideas and suggestions put forth by each student. Students were asked approximately 58 questions during the semester.

Several hundred student responses were sampled to determine how often group members in each section chose the same answer as each of the other group members (that is, block voted). Data were taken from nine clicker questions asked during each of the semesters. Students had the option of block voting only if at least two group members were present during each of the three days from which data was sampled, giving an upper limit for the number of possible cases of block voting of 450 votes per section.

Pre- and post-test scores were matched so that a more robust statistical analysis could be performed and gains were calculated as

$$\text{gain} = [\text{postscore} - \text{prescore}]/[\text{postscore} + \text{prescore}]. \quad (1)$$

These gains were used to calculate the average gain per section.

III. RESULTS AND DISCUSSION

During the first semester of the study, the differences between the low stakes and high stakes classrooms were much more pronounced than during the second semester. Students in the spring high stakes classroom chose the correct answer 56.6% of the time, whereas in the low stakes classroom the percentage was 49.5%, a statistically significant difference ($p<0.0005$). This difference was not observed for the fall semester when both sections had an average correct response percentage near 60%.

In the high stakes classroom (spring semester) learning groups block voted 69% of the time, in contrast to the low stakes classroom in which learning groups block voted 45% of the time, with a t-test verifying that the difference is statistically significant with $p<0.0005$. During the fall semester there was not a statistically significant difference between the block voting habits of each section. As can be seen in Table I, students in each section (and in each semester) showed no difference in overall class averages or gains on the Astronomy Diagnostic Test. It is possible that the block voting habits of students during the first semester of the study were skewed due to the Hawthorne effect.[13] Although clicker grading incentive was not discussed by the instructor, the use of digital recorders throughout the semesters provided the students with a visual reminder that they were being studied, specifically with regard to their clicker usage. Because this repeated visual reminder may have altered student behavior, the study is being repeated this academic year (2008/2009) with recorders only being used in one section and not at all in the other section. (Low stakes will be recorded during the fall semester and high stakes during the spring semester.) Because the block voting habits of the students varied so

Table II. Number of recorded discussion statements by section.

Category	High stakes classroom	Low stakes classroom
Restate question	15	22
State answer preference	49	75
Provide positive information	33	39
Provide negative information	8	3
Articulate new question	2	6
State agreement	9	9
State disagreement	3	2
Ask for clarification	12	22
State uncertainty	14	8
Total statements	145	186

much from the first part of the study to the second part, the possible effect of recorders on student behavior must be studied in greater detail.

The results of the spring semester are consistent with the results of Ref. 5 in that the overall level of student understanding as measured with clicker questions may be inflated in the high stakes classroom. In Ref. 5 it was found that students tended to block vote much more often in the high stakes classroom (student pairs block voted 92.4% of the time in the high stakes classroom versus 63.2% in the low stakes classroom), values that are higher than in the current study. One key difference between these studies is that in Ref. 5 student pairs were studied, whereas we examined groups of four students. It is likely easier for a student to convince one peer to vote similarly than it is to convince three peers. In Ref. 5 the portion of the grade allotted to clicker usage was significantly higher than in our study (12.5% in high stakes, 20% in low stakes, versus 4% for both sections in our study), which could have a strong influence on students' propensity to block vote. James' study was done only with recorders in the classroom and has not been repeated without the presence of recorders.

Even though students in the high stakes classroom chose the correct answer to clicker questions more frequently than students in the low stakes classroom, differences in overall knowledge gained were not apparent in either the average course grade or gains on the Astronomy Diagnostic Test. This performance on the latter is a strong indication that students in the high stakes classroom are not actually learning more than their counterparts in the low stakes classroom, as may be concluded from analyzing responses to clicker questions alone.

Overall class grades were 3% higher during the spring semester ($p<0.0005$) and a difference of 0.059 was seen in gains for the spring versus fall semesters as well ($p<0.0005$) (see Table I). It is unclear why student performance varied as a function of semester.

Transcription of the recordings revealed differences in the nature of the conversations students had in each section while discussing clicker questions. Table II lists the categories into which the discussions were placed and the total number of statements recorded in each section (145 and 186 in the high and low stakes classroom, respectively). The differences between the types of discourse in the low stakes and high stakes classrooms can be seen by studying the types of statements made by the students. Students in the low stakes classroom had many more instances of both stating an answer preference and of asking for clarification than students in the other section, which suggests that students were more comfortable expressing their choice or their lack of understanding when they knew that incorrect responses were weighted the same as correct answers. There were also many more instances of students in the low stakes classroom restating the question and articulating a new question, implying that they were trying to frame the questions in their own words and to understand material related to the questions, in contrast to the students in the high stakes classroom, who spoke 20% less than their low stakes peers in the study. Students in the high stakes classroom had more instances of stating uncertainty and less instances of either asking for clarification or stating an answer preference, indicating that students in this classroom were less comfortable speaking their mind or making it obvious that they did not understand the question or concept. Many instructors use clicker questions to stimulate classroom discussion and to spark interest in their students, but we conclude from our analysis of the conversations that the use of a high stakes rubric for grading responses will not lead to an increase of frank discussion among the students. Students in the high stakes classroom had more statements than the other classroom only in the categories of providing negative information (that is, why not to choose one answer) and stating uncertainty. These two categories are not likely to be considered as improving student discussion.

Our study raises the possibility that student behavior (as measured by tendency to block vote) may be altered by the presence of digital recorders in the classroom. Because block voting behavior of students as a function of grading incentive was quite pronounced when the recorders were present (over 2/3 of the students regularly block voted in the high stakes classroom, compared to less than 1/2 the students in the low stakes classroom), but was almost identical during the nonrecording phase of the study, more research should be done to determine how student behavior changes as a result of having recorders in the classroom. Our analysis shows no differences in learning between the classrooms, but the transcripts reveal that a high stakes grading incentive has a somewhat chilling effect on student discussions.

[a]Electronic mail: willoughby@physics.montana.edu

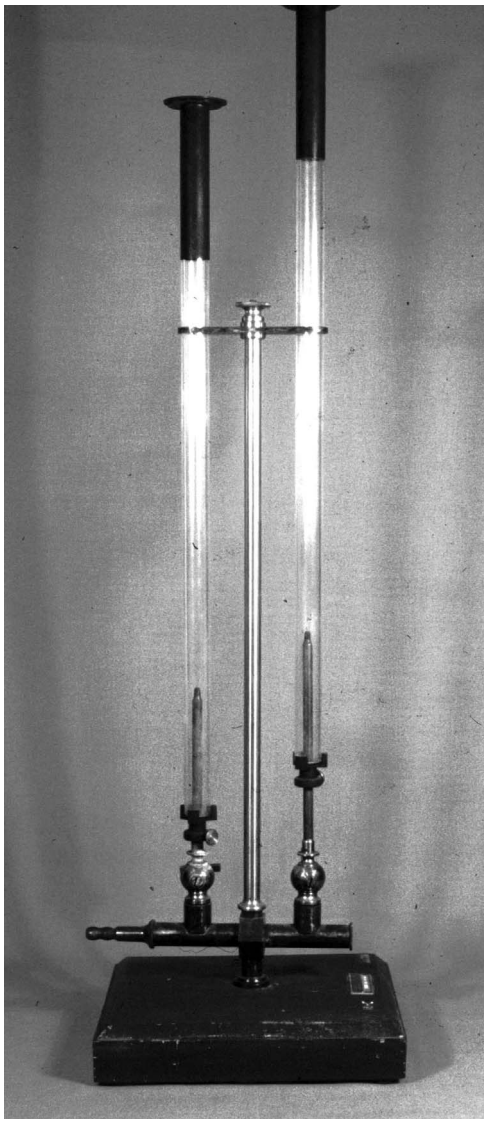
[1]E. Judson and D. Sawada, "Learning from past and present: Electronic response systems in college lecture halls," J. Comput. Math. Sci. Teach. **21**, 167–181 (2002).
[2]D. Duncan, "Clickers: A new teaching aid with exceptional promise," Astronomy Educ. Rev. **5**(1), 70–88 (2007).
[3]J. Caldwell, "Clickers in the large classroom: Current research and best practice tips," CBE Life Sciences Educ. **6**, 9–20 (2007).
[4]I. D. Beatty, W. J. Gerace, W. J. Leonard, and R. J. Dufresne, "Designing effective questions for classroom response system teaching," Am. J. Phys. **74**, 31–42 (2006).
[5]M. James, "The effect of grading incentive on student discourse in peer instruction," Am. J. Phys. **74**, 689–691 (2006).
[6]P. M. Len, "Different reward structures to motivate student interaction with electronic response systems in astronomy," Astronomy Educ. Rev. **5**(2), 5–15 (2006).
[7]N. W. Reay, L. Bao, P. Li, R. Warnakulasooriya, and G. Baugh, "Toward the effective use of voting machines in physics lectures," Am. J. Phys. **73**, 554–558 (2005).
[8]E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997).
[9]C. Wieman and K. Perkins, "Transforming physics education," Phys. Today **58**(11), 36–41 (2005).

[10]Montana State University has adopted I-clicker, ⟨iclicker.com⟩, as the campus wide classroom response system.

[11]G. L. Deming, "Results of the astronomy diagnostic test national project," Astronomy Educ. Rev. **1**(1), 52–57 (2002). Version 2.0 was used.

[12]The study was performed with the review and approval of the Institutional Review Board for the Protection of Human Subjects at MSU Bozeman.

[13]The Hawthorne effect (coined in 1955 by Henry A. Landsberger) leads people to behave differently when they know they are being studied.



Chemical Harmonica. The chemical harmonica is a singing flame: a gas flame burning in a glass tube is set into oscillation when enclosed in a glass tube, thereby producing a loud sound of definite pitch. Faraday proposed that the flame was extinguished and rekindled by the hot burner at the same frequency as the singing sound. Wheatstone and Tyndall used a rotating mirror to show that Faraday's hypothesis was, indeed, correct. The image of the flame, viewed in the rotating mirror which supplied a time base, could be seen to oscillate up and down. The fundamental wavelength is twice the length of the glass tube, after making the necessary end corrections. This example is at Oberlin College in Ohio. (Photograph and Notes by Thomas B. Greenslade, Jr., Kenyon College)