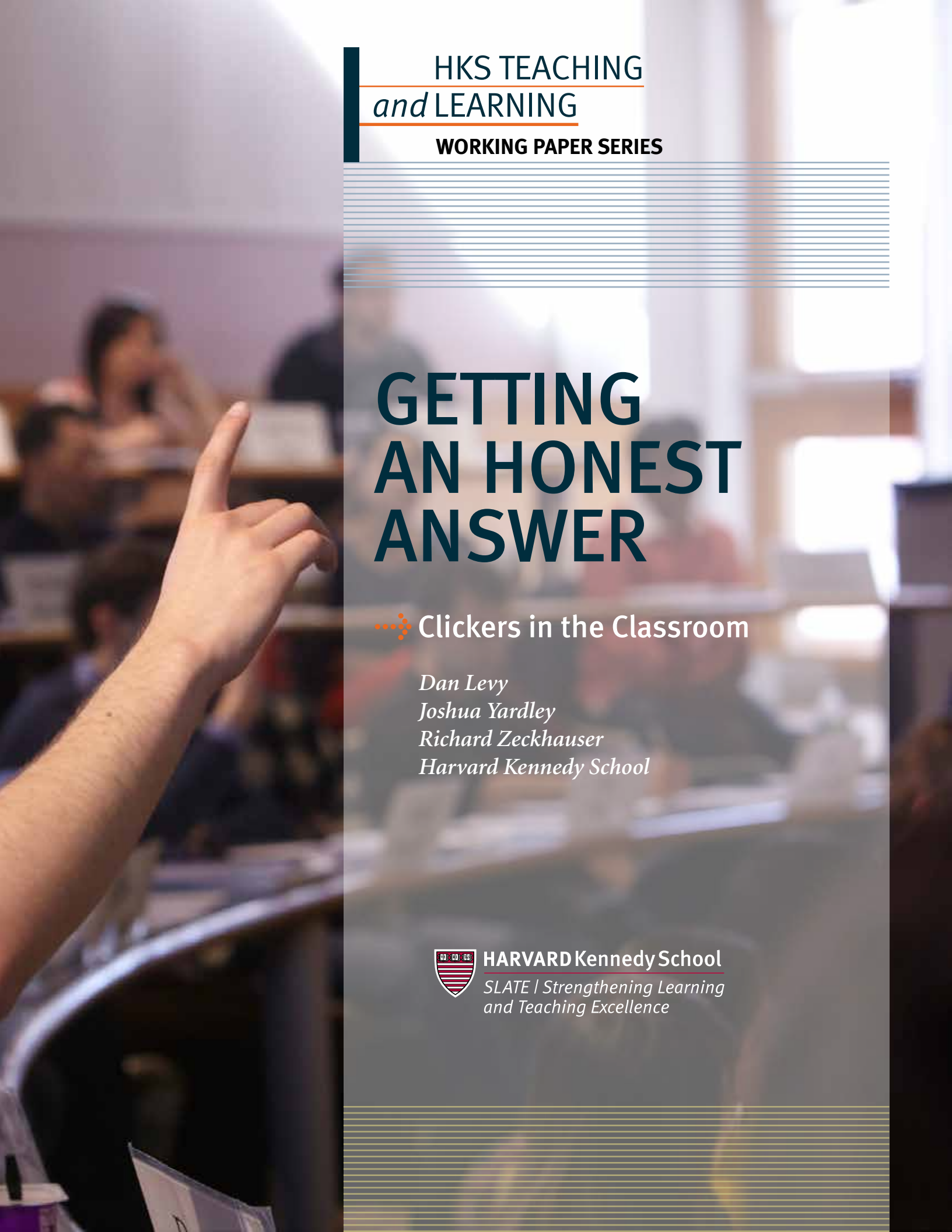


HKS TEACHING *and* LEARNING

WORKING PAPER SERIES

GETTING AN HONEST ANSWER

Clickers in the Classroom

Dan Levy
Joshua Yardley
Richard Zeckhauser
Harvard Kennedy School

HARVARD Kennedy School
SLATE | Strengthening Learning and Teaching Excellence

Getting an Honest Answer: Clickers in the Classroom

DAN LEVY[a]
JOSHUA YARDLEY[b]
RICHARD ZECKHAUSER[c]

Harvard Kennedy School
79 John F. Kennedy Street
Cambridge, MA 02138, USA

[a]Dan_Levy@hks.harvard.edu
[b]Joshua_Yardley@hks.harvard.edu
[c]Richard_Zeckhauser@hks.harvard.edu (Corresponding Author)

ABSTRACT

Asking students to raise their hands is a time-honored feedback mechanism in education. Hand raising allows the teacher to assess to what extent a concept has been understood, or to see where the class stands on a particular issue, and then to proceed with the lesson accordingly. For many types of questions, as the evidence here demonstrates, the tally from a public show of hands misrepresents the true knowledge or preferences of the class. The biases are predictable and systematic. Specifically, students raising their hands tend to herd and vote with the majority answer. Beyond impeding the teacher's ability to assess her class, such herding threatens to diminish learning by limiting the level to which a student engages with the questions posed by the teacher.

KEYWORDS

Audience response systems,[1] hand raising, herding, classroom feedback

1. INTRODUCTION

A teacher asks her class to work out a math problem. She then asks the students to raise their hands if they think the answer is greater than zero. A few confident hands shoot up while a greater number of less confident eyes dart around the room, collecting the relevant data necessary to "solve" the math problem. Slowly the rest of the hands go up. The teacher, satisfied that her class clearly understands the concept, moves on.

[1] Abbreviations used in this paper include ARS (Audience Response System) and DFE (Distance From Extreme). Each term is explained when it first appears in the paper.

This story has played out generation after generation, at every level of learning. It played out much the same way in a one-room schoolhouse in rural New Hampshire in 1750 as it did in the massive lecture halls of Harvard, and hundreds of other colleges and universities, last semester. The raising of hands as a means of responding to a teacher's questions is low-tech and low-cost; its simplicity and minimal resource requirements have made it the preferred approach in many settings, including the classroom. But is it reliable? And if not, is there a better way?

For students' responses to be useful in informing an instructor's teaching, the instructor must be confident that the responses she sees accurately reflect the students' knowledge, skills, or beliefs. This study shows that such confidence is misplaced when students respond with a show of hands. Relying on that technique, instructors are likely to overestimate students' grasp of a concept, making it difficult to target instruction to concepts or ideas that need additional elaboration. To make matters worse, herding behavior by students raising their hands may reflect not merely a decision to vote with the majority contrary to the students' own thinking, but also decisions to vote with the majority *instead of* thinking about the question on their own. If so, herding will reduce students' cognitive activity in response to teachers' questions. This keeps questions posed in the classroom, however well developed, from promoting meaningful student engagement and learning.

This paper examines the reliability of hand raising as a useful pedagogic tool by comparing it with another technique that aggregates student answers, a technique that does not allow students to see their classmates' responses. If hand-raising students do not employ other students' answers when selecting their own, these two aggregation techniques should yield similar results (up to the degree of sampling error). Yet, the experiments we conducted suggest that these two techniques lead to considerably different results. Moreover, the pattern of the differences in results suggests that students herd towards the most popular responses, an easy approach when there is a show of hands.[2]

The rest of this paper is organized as follows. Section 2 of this paper describes one class of techniques designed to address the challenge of soliciting accurate feedback from students in a classroom. Section 3 defines the study's research questions and describes the experimental design, and Section 4 explains the results of the experiments. Section 5 discusses the implications of these results for instructors in a classroom, and Section 6 concludes.

2. BACKGROUND

One relatively new technology that may be able to improve upon the ancient technique of hand raising is the audience-response system. Audience-response systems (ARSs) consist of electronic

[2] This is a simplification. If some students are known to be more knowledgeable, there may be herding toward their answers even when they are in the minority. Also, answers that are surprisingly popular, despite being in the minority, convey information and therefore may induce imitation. Thus, for example, students asked the capital of Australia, expecting either Sydney or Melbourne, may be influenced when 25% of their classmates correctly select Canberra.

handheld response keypads ("clickers") that allow students in a classroom to respond individually and anonymously to multiple-choice questions posed by the instructor. The responses are collected and aggregated, and the distribution of responses can be viewed by the instructor, and presented to the class in real time if she wishes.[3]

ARSs have been around since 1966, when Stanford University introduced an expensive, difficult-to-use version. ARSs did not become commercially available until the 1990s, but even then, the cost was prohibitive for most schools. Prices fell, and by the 2000s, ARSs began to be commonly used in secondary schools, colleges, and universities (Abrahamson, 2006). In the relatively brief period -- certainly relative to hand raising's existence -- that ARSs have been in classrooms, a number of studies began to establish the promise and the challenges of this technology for improving classroom outcomes.[4] We will briefly review the key findings of these studies before describing how our present study contributes to this literature.

2.1. Benefits of Clickers

In their 2009 review of the literature on ARSs, Kay and LeSage examined 67 papers and chapters on the use of this technology in classrooms, primarily in undergraduate classrooms. These studies found overwhelmingly that clickers in the classroom were popular. Of the 38 studies that measured student and/or teacher attitudes towards clickers, 36 reported respondents had positive views of the technology on average (Kay and LeSage, 2009). Since student satisfaction often correlates imperfectly with student learning, it is important to explore the ways in which clickers may impact learning specifically, in addition to the positive reviews they are getting from users.

A number of studies show results indicating that student attention (Siau et al., 2006; Latessa and Mouw, 2005) and engagement (Draper and Brown, 2004; Simpson and Oliver, 2007) are greater when clickers are used in the classroom. (See Kay and LeSage for a complete list of relevant studies.) Unfortunately, most of these studies fail to distinguish between two possible explanations for the source of this finding. The first is that some aspect of responding by using clickers, rather than by raising hands, promotes attention and engagement, for example being forced to think rather than merely mimicking the choices of others. The second is that encouraging students to respond to more questions during a class, regardless of how they respond, may be the primary driver of these effects. Since, in these studies, the introduction of clickers in classrooms was often accompanied by an increase in the questions that were asked during lectures, it is difficult to parse the explanations for these improvements in attention and

[3] In addition to clickers, a number of other products and technologies can be used to solicit feedback from a classroom, including products designed for students' smartphones. These products serve many of the same purposes as our clickers; they allow students to vote without seeing others' responses and teachers to aggregate the responses quickly and accurately. We expect many of the results discussed in this paper will also be applicable to these other technologies.

[4] For a great review of many uses of ARSs in the classroom, example clicker questions, and sample activities from a wide range of instructors, see Bruff (2009).

engagement. The many studies that show associations between clicker use and learning outcomes, such as test scores (Kennedy and Cutts, 2005; Preszler et al., 2007), suffer from the same difficulties. A notable exception is Mayer et al. (2008), whose research design included a control group with no questions and no clickers as well as a control group with questions but no clickers; Mayer et al. found that clickers use produced significant gains in test scores over both control groups.[5]

Two benefits of clickers are of particular interest to us here. First, unlike hand raising, clicker use can be anonymous. Student surveys have shown this anonymity to be important to some groups of students (Draper and Brown, 2004; Crouch and Mazur, 2001), and there is experimental evidence to suggest anonymity can positively affect classroom discussion and debate (Ainsworth et al., 2011). Second, assuming that students participate and give honest answers, clicked responses can be an important source of feedback for teachers. Thus, it allows for contingent teaching, whereby instructors adjust their teaching plans in real time in response to the feedback they receive from the clickers (Draper and Brown, 2004). Students can also get beneficial feedback, learning for example where their responses fall in the distribution of the responses by the class (Abrahamson, 2006).

2.2. Clickers and Response Reliability

Our study provides experimental evidence of the effect of the choice of response technique -- hand raising or clicking -- on student responses. In evaluating the benefits clickers bring to a classroom in terms of providing real-time feedback and allowing for contingent teaching, previous research has mostly focused on their ability to quickly collect, aggregate, and display student responses. Little work has been done to determine whether clickers may in fact *change* these responses, and if so how. We show evidence that clicked responses differ from raised-hand responses in predictable ways. Specifically, vote shares for responses given by hand raising are likely to be more extreme (closer to 100 or 0 percent, a result consistent with herding behavior) than vote shares for responses given by clicking.[6] We argue that clicked responses are more useful to a teacher in terms of feedback and contingent teaching than responses given by raised hands.

The process of responding by clicking differs from that of responding by hand raising in two important ways. First, raised-hand responses are immediately observable to the rest of the group, potentially influencing individuals who respond more slowly. Clicked responses are not observable until after the entire group has responded. Second, an individual can respond by

[5] One drawback of the Mayer et al. (2008) study is the small effective sample size. Each of the three groups (the clicker treatment group and the two controls) consisted of a single class offered in a given year. The treatment group was the class when it was offered in 2006. The control groups were that same class offered in 2005 (no clickers, no questions) and 2007 (no clickers, with questions).

[6] This is despite the fact that surprisingly large minority responses may also induce herding, but this would work against this finding. See footnote 2.

clicking, knowing that his response is anonymous and cannot be revealed publicly. (For expository ease in this essay, respondents are male and questioners female.) In theory, anonymity is achievable with hand raising. It requires that all the students close their eyes, and that all the students trust that all other students close their eyes. In practice, such an outcome is difficult to achieve in many settings.

We suggest that there are two primary reasons why respondents might be influenced by the answers of others when choosing their own responses. Both could explain why herding behavior might emerge in situations in which individual responses are elicited in a group setting.

First, respondents can learn from other students' answers. Such learning sometimes occurs in a flash, though anecdotal observation indicates that there is often a second or two delay as individuals scan the room to see the prior indications of others. When there is a right or a wrong answer to a question, observing before responding can clearly be a sensible strategy. A respondent who is poorly informed on a topic, wisely defers to the "wisdom of the crowd" and sides with the majority. Or if he can identify an expert (or relative expert) in the group, he may side with that person. If many respondents identify the same expert, this follow-the-expert strategy will produce aggregate response results that look very similar to a strategy of following the majority, even if only one person is being followed.

Second, respondents are concerned about preserving their reputations and avoiding embarrassment. Such concerns could apply to questions for which there is a single definite right answer. For instance, a respondent may think the majority is often wrong, but voting with it prevents him from any worries about being among a small dissenting minority that gets the answer wrong. Those providing erroneous answers love company. Clustering with others may also be helpful when there is no single right answer. Many people do not wish to be out of step with others in announced beliefs about political situations or ethical dilemmas, or even about something as simple as whether they liked a particular movie. Witness how groups coming out of a movie tend to concur in their ratings, though other similar groups produce quite different ratings. Observing others' responses allows individuals a chance to adjust their responses accordingly, based on the stated views of those who answer early.

To mimic or not to mimic, that is the question. Different people provide different answers. Some care little if they stand alone, even if they end up demonstrably wrong. A much larger number, our results suggest, would feel somewhere between slightly embarrassed to mortified to be in such a position. To determine how greatly individuals are influenced by the answers of others, we conducted experiments in which participants were asked to respond to questions -- factual, conceptual, and ethical -- at times using hand raising, enabling them to observe the responses of others, and at times using clickers, when they could not. Section 3 describes the design of these experiments.

3. RESEARCH DESIGN

The two key research questions we examine in this paper are the following:

(1) Do students give the same responses when using raised hands as when using clickers?
(2) If the answer to (1) is no, why not?

This section describes the research design we used to answer these two questions. It first describes the experimental protocol and indicating the types of questions we asked individuals in our sample. We then specify the key outcome variables of interest.

3.1. Experimental Protocol

To compare response outcomes using these two different techniques, hand raising and clicking, we conducted experiments on 22 different groups consisting of over 1100 participants in total.[7] Each group was divided into two, usually according to where the participants were seated,[8] and was asked one to four questions that each had only two possible responses. After the first question had been posed to the whole audience, the first half of the group was asked to respond anonymously using clickers. When the clicker half of the group had finished responding, the other hand-raising half was asked to respond with hand raising. Note that the two groups never knew the answers the other group had given prior to submitting their own responses. If the group was asked a second question, the two halves of the room switched roles. The original hand raisers responded first with clickers; then the original clickers responded with hand raising. Given that the two groups were likely to be similar to each other given the random or quasi-random assignment, the aggregate response results should not have been very different, or at least not consistently so. Yet for the most part, their results did differ in significant ways.

3.2. Question Types

The choice of questions used in these experiments is likely to be an important determinant both of whether any differences are observed in the aggregate responses from hand raising and clicking, and if so of what magnitude. We might expect greater differences for questions that are especially sensitive or for those requiring specific knowledge that might be gained from the answers by others. Thus, we might expect some individuals asked, "Would you support legislation to require the deportation of immigrants convicted of more than one misdemeanor?" would respond differently if their answer were public or anonymous. A student's response to a

[7] See Appendix A for a list of all the experiments.

[8] When logistically feasible, the facilitators randomized assignment to the clicking and hand-raising groups, rather than simply splitting the groups based on the seating arrangement at the time. Random assignment was implemented in nine of the 22 experiments. Quasi-random assignment (seating was assigned alphabetically by name) was implemented in an additional six experiments. The remaining seven experiment groups were formed according to the subjects' chosen seating arrangements. For these groups, random assignment should not be assumed. The tables reporting key results of these experiments are replicated in Appendices 2, 3, and 4, restricting the sample to experiments implemented with random and quasi-random assignment.

question requiring the solution of a probability problem might take guidance from the answer of a classmate known for his strong mathematical abilities. By contrast, we might expect little difference between clicked and hand-raised responses on the question, "Do you prefer vanilla or chocolate ice cream?" Identifying and understanding the variation in the questions used in our experiments will be important for interpreting the results presented in Section 4.

A total of 61 questions were asked in the 22 experiments. Each question fell into one of four italicized categories, which reflect the different incentives one might have for considering the responses of others before submitting a response.

No Single Right Answer, Sensitive Topic (27 of the 61 questions)

These questions explore a respondent's preference or belief about a potentially sensitive topic. Some of these questions are political in nature ("Would you support legislation to allow gay marriage in the state of New York?"). Others are ethical ("Would you reveal information about a mechanical problem of a car you are trying to sell?"). If made public, the responses to many of these questions might have reputational consequences for individuals, depending on the composition of the group. For instance, one question asks whether the individual plans to use the psychological principles and tools taught during the experiment to get her colleagues to agree to policies that differed significantly from what they would otherwise support. It is not clear that there is an ethically right or wrong answer to this, but one can imagine a roomful of colleagues scanning the crowd to note whose hands go up for the 'yes' vote.

No Clear Right Answer, Not a Sensitive Topic (14 of the 61 questions)

These questions also may pertain to respondent preferences, but about topics that are much less sensitive than those in the previous category. Some are less sensitive because the topics they cover are trivial. An example would be "Do you enjoy watching any reality TV shows?" In some instances, even apparently trivial preference questions might prove embarrassing. Some participants might be hesitant to admit that they like reality TV. Some questions that are not sensitive ask for answers that will eventually prove to be either right or wrong, outcomes that will not be known immediately ("Who do you think will win the Monday Night Football game?" and "Will the Dow Jones Average be up more than 6% per year over the next 3 years?"). We expect to observe fewer differences in the hand raising and clicking responses for such "innocuous" questions than for the other three categories. Unlike questions in the other three categories, these questions have no single right (or socially acceptable) answer that respondents may be tempted to glean from observing others.

Single Right Answer, Factual (10 of the 61 questions)

This category consists of questions about specific facts. Responding to these questions relies on factual knowledge and perhaps on some degree of general experience rather than on critical thinking. Examples include "Is the population of Turkey greater than 100 million?" and "Are

there more than 45 countries in Africa?" Hand-raising respondents unsure of the correct answer may be tempted to follow the "wisdom of the crowd" or to identify one or more knowledgeable individuals to help inform their response. For simplicity, our factual questions were yes/no questions. However, the same forces would be at play if we asked factual questions and offered multiple possible answers.

Single Right Answer, Conceptual (10 of the 61 questions)

The questions in this category were used mostly in graduate-level statistics classes and concerned conceptual material related to the courses. For most students, these questions could not be answered based simply by recalling memorized facts. Instead, working out the answers to these questions required drawing on knowledge as well as thinking critically about the problems. Some resembled trick questions. For instance, "You ask a woman you just met if she has any children. She says she has two. You ask if she has at least one girl. She says yes. Given this information, is the probability that both are girls equal to 50%?"[9] Other questions are less like riddles, but still require some degree of critical thinking ("Do all bivariate regressions suffer from omitted variable bias?"[10]). Some respondents will be tempted to identify experts in the class whose responses they might mimic, or to go with the crowd if they do not know who is expert. Conceptual or factual questions will tempt subjects to seek less guidance from others in areas where they are knowledgeable, and more guidance from others where embarrassment might attend error.

3.3. Outcome Measures

Two key outcome measures provide answers to the two research questions of interest. First, we computed the absolute differences in vote share to assess whether hand raising and clicking lead to different aggregate response outcomes for various types of questions.

Next, we constructed a variable called "distance from extreme" (DFE) to determine whether any of these differences provide evidence of herding behavior among the hand-raising group. Let *vote_share* be the proportion of participants responding with the first option.[11] Then DFE will be:

$$DFE = \min(vote\_share, 1 - vote\_share).$$

On questions for which the majority response for both hand raisers and clickers is the same, an outcome is consistent with herding to the majority when the DFE for hand raising is less than the DFE for clickers. In other words, when hand-raising vote shares are closer to the extremes (0 or

[9] No, it is not. Given this information, the probability she has two girls is only one-third.
[10] No, they do not. For example, in the context of a well-conducted randomized trial, a bivariate regression of the outcome variable regressed on the treatment indicator should not suffer from omitted variable bias.
[11] Since there are only two options for a given question, the *DFE* is the same no matter which response is arbitrarily chosen as the "first option."

1) than the clicker vote shares. If 90% of respondents using raised hands (DFE = 0.10) answered that they believed Africa has more than 45 countries, as opposed to 70% of respondents with clickers (DFE = 0.30), this would be consistent with an explanation that some hand raisers who would have otherwise guessed fewer than 45 countries chose instead to herd and vote with the majority.

On questions for which the majority response of hand raisers differed from that of clickers, defining what it means for results to be consistent with herd behavior is less obvious. We do note that if we take the clicker response to be an unbiased measure of herd-free behavior, then getting a result on the opposite side of 50% could reflect herding. In one experiment, for example, a majority of clicker responders (56%) indicated support for the US intervention in Libya taking place at the time, compared to only 25% of hand raisers, a difference significant at the 5% level. Hand raisers asked if they support this controversial intervention might be hesitant to admit that they do if they only see 25% of their classmates' hands go up in support.

An individual learning from others will have a personal tipping point that depends on his information. Herding that is triggered by a tipping point that differs from 50% is certainly a realistic possibility, indeed one to be expected in many situations, perhaps particularly on sensitive questions. Some students may only need to know that at least 40%, rather than a majority, of their classmates agree with their support of the intervention in Libya before they reveal their own support. We plan to explore more complex tipping behavior in future work. Since this paper is concerned primarily with herding to the majority, our herding analysis in the next section drops the few questions (7 out of the 61) where the majority responses of hand raisers and of clickers differed. The main results of the study are robust to including these seven questions.

4. RESULTS

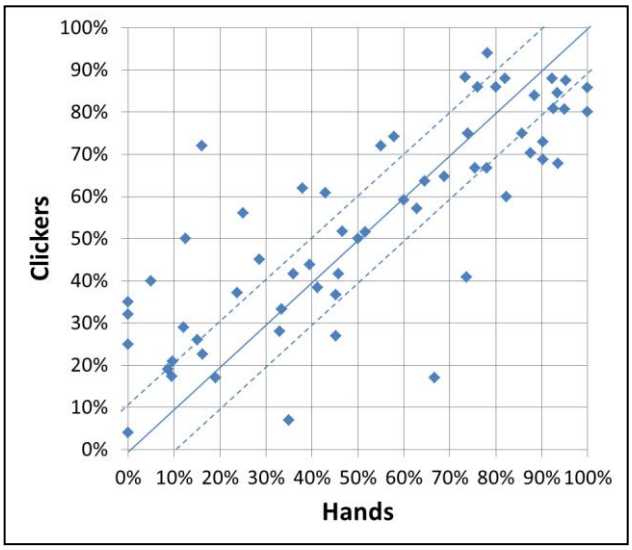
4.1. Hands and Clickers Lead to Different Responses

While we would expect hand-raising and clicking results to differ simply due to sampling fluctuations, the differences in these experiments prove to be far stronger than chance would produce. Thus, they provide strong evidence that the technique for eliciting choices (hand raising vs. clicking) will affect individual responses in group settings. On average, the absolute difference between the vote shares for a given response with hand raising and with clicking was 14.37 percentage points.[12] Fifty-seven percent (35 of 61) of the questions resulted in vote shares that differed by more than 10 percentage points between the two response techniques; over 40% of those (15 of 35) differed by more than 20 percentage points. Figure 1 plots the vote shares of

[12] The average absolute difference for the 25 questions posed during the nine experiments implemented with random assignment was 14.01 percentage points. This difference for the 44 questions posed during the 15 experiments with either random (nine experiments) or quasi-random (six experiments) assignment was 12.83 percentage points.

the first responses for hand raising and for clicking. If vote shares were consistently the same across techniques, the points would fall along the solid 45-degree line in Figure 1. Points outside the region bounded by the dotted diagonal lines in Figure 1 are questions for which the difference in vote shares was greater than 10 percentage points.

Figure 1 - Percent Polling First Response



Taken individually, 14 of the 61 differences (23%) in vote share were significant at the 5% level, assuming a normal sampling distribution for the differences in these proportions, as would be expected to emerge in a large sample. Using Fisher's method[13] to aggregate the results and test the null hypothesis that there are no differences in vote share across voting techniques yields a p-value virtually indistinguishable from zero.[14] However, given the small sample sizes of many of the individual experiments,[15] we chose not to assume a normal sampling distribution and instead used the more conservative Fisher's Exact Test.[16] Using this method, only 8 of the 61 differences (13%) proved significant at the 5% level. However, taking the experiments as a whole, the p-value associated with the null hypothesis that the two techniques produce the same vote shares is 0.0012. In short, it would be exceedingly unlikely for the experiments to yield results this extreme if raised hands and clickers did not produce different results.

These differences are driven mainly by sensitive questions that had no single right answers, and by factual and conceptual questions with single right answers, as Table 1 shows. Disaggregating by question type reduces statistical power, but even with the reduced power (and using the conservative Fisher's Exact Test to compute p-values for individual experiments) we see evidence of significant differences in aggregate responses between techniques for questions in

[13] Fisher's method aggregates p-values from *k* independent tests using the formula $X_{2k}^2 = -2\sum_{i=1}^{k}\ln(p_i)$, which has a chi-squared distribution with 2*k* degrees of freedom.
[14] 1.6×10^{-13} to be exact.
[15] Note that the average number of participants across all 22 experiments is about 50.
[16] Same Fisher as before, different method.

these three categories. Consistent with our priors outlined in Section 3.2, the evidence of differences between hand and clicker votes is much weaker for questions that have no single right answer and are not sensitive, the only category of question for which this difference is not statistically significant at any conventional level.

Table 1 – Response Differences by Question Type[17]

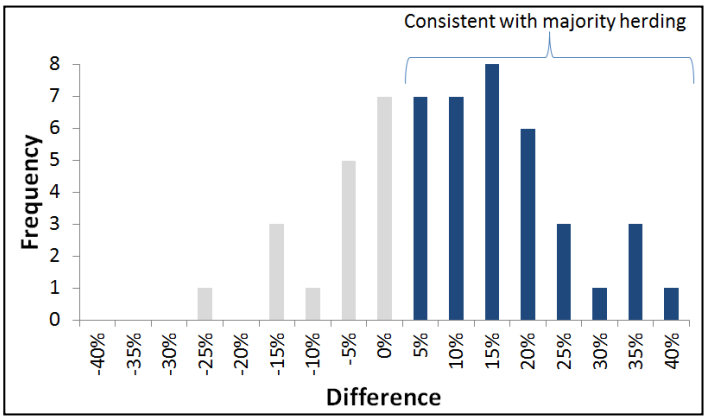
Question Type	Avg. absolute %-point difference between hand-raising and clickers	Fraction of differences greater than 10 %-points	p-value associated with null of no differences[18]
No Right Answer, Sensitive	13.30%	17 out of 27 (63.0%)	0.097*
No Right Answer, Not Sensitive	12.32%	7 out of 14 (50.0%)	0.256
Right Answer, Factual	17.32%	6 out of 10 (60.0%)	0.009***
Right Answer, Conceptual	17.18%	6 out of 10 (60.0%)	0.011**

** significant at 10% level, **5%, ***1%*

4.2. Differences in Hand and Clicker Responses are Consistent with Herding

36 of the 54 questions (or 67%[19]) for which the majority responses of hands and of clickers were the same had more hand-raising respondents choose the more common response than did the clicking respondents (see Figure 2), indicating the presence of majority herding. On average, the majority vote share of hand raisers was 6.46 percentage points greater than that of clickers. Since each question had only two responses, a 6.46 percentage point increase in the majority response also meant a 6.46 percentage point decrease in the minority response, implying a spread of nearly 13 percentage points between responses to a given question.[20]

Figure 2 – Difference in Vote Shares of Majority Responses (n = 54 questions)
(Hand raising – Clicking)



[17] See Appendix B for results with the sample restricted to experiments implemented with random and quasi-random assignment.
[18] Using Fisher's method to aggregate the p-values from individual experiments.
[19] Significantly different from 50% at the 5% level.
[20] These figures are robust to the inclusion of the seven questions for which the majority responses of hands and clickers were not the same. Including these seven questions, 40 of the 61 questions (66%) showed signs of herding among hand raisers, with an average increase in the majority response vote share of 6.46 percentage points, the same figure as without the seven questions.

Figure 3a updates the scatterplot shown earlier in Section 4.1 by shading the areas that represent data points consistent with majority herding behavior. Figure 3b plots the DFEs for these clicker and hand response data. Together, Figures 2, 3a, and 3b reveal the extent to which the questions that have the largest differences between hand-raising and clicking vote shares also tend to be those in which the signs of those differences tend to be consistent with herding.

Note from Figure 3b that in six of the 54 questions (11%), hand raising led to a unanimous response, indicated by a DFE of 0. No clicker vote was ever unanimous. Hand raising had 18 votes at 90% or above or at 10% or below. Clickers only had three such votes.

Figure 3a - Percent Polling First Response (n = 54 questions)
(Data points consistent with herding indicated by +)

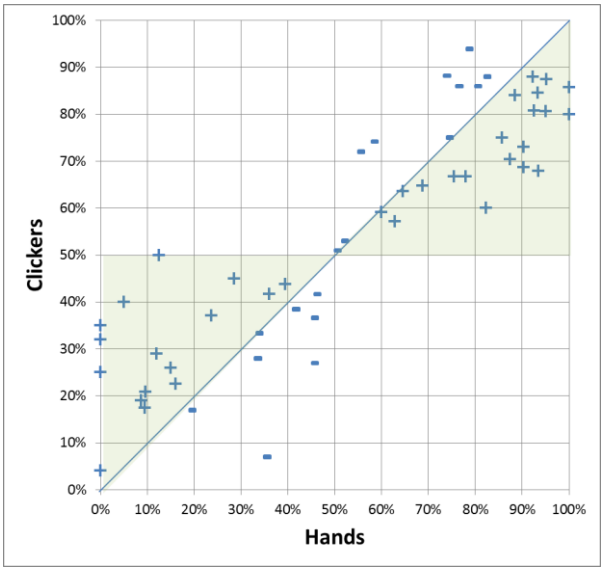
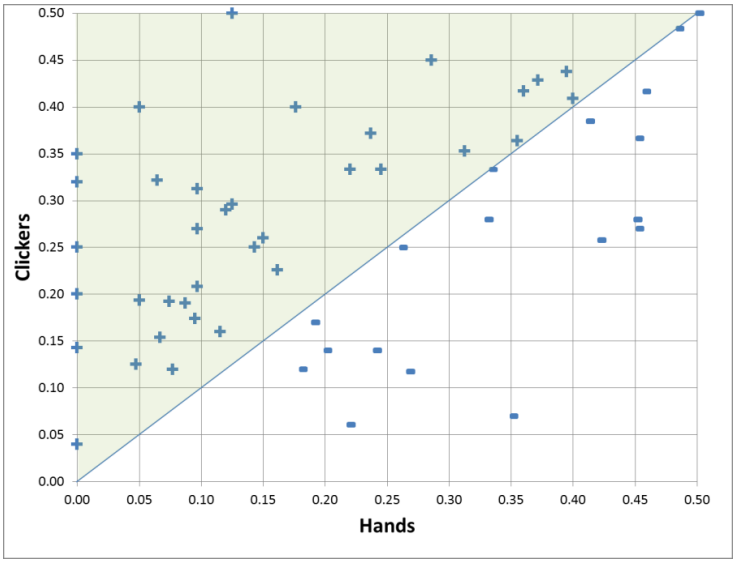


Figure 3b – Distance From Extreme – DFE (n = 54 questions)
(Data points consistent with herding indicated by +)



Of the 36 questions that showed signs of herding, in only six (or 17%) of those cases was this herding statistically significant at the 5% level; another three were significant at the 10% level. This is not surprising, given the small sample sizes for each question. However, aggregating the test statistics for the herding tests of each of the 54 questions (and allowing those questions in which herding did not take place to count against a conclusion of herding) yields an aggregate Z-score of 4.90,[21] which is statistically significant well beyond any conventional level.[22]

This evidence for herding hand raisers on average is strong and robust, persisting across a number of test specifications. Table 2 reports the results of eight models regressing the distance from extreme variable, or DFE, introduced in Section 3, on a dummy variable for responses given by clicker. The coefficients estimated for this dummy, *Clicker*, can be thought of as estimates of the degree to which herding is taking place among hand raisers. Specifically, it is the average difference between the vote share received by the majority response of the hand raisers and that of the clickers. According to the results in the first specification, responding with clickers is associated with a 6.2 percentage point increase in *DFE*. Analogously, hand raising is associated with majority response vote shares that are 6.2 percentage points greater than those associated with clicking. Note that this coefficient is significant at the 1% level for every specification.[23]

Table 2 –Regression Results of DFE on Clicker[24]

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Clicker	0.062***	0.062***	0.062***	0.067***	0.067***	0.067***	0.061***	0.065***
	(0.017)	(0.019)	(0.018)	(0.016)	(0.019)	(0.017)	(0.017)	(0.016)
Constant	0.224	0.224	0.224	-	-	-	0.213	-
	(0.012)	(0.021)	(0.020)				(0.013)	
Question Fixed Effects	No	No	No	Yes	Yes	Yes	No	Yes
Standardizing Weights	No	No	No	No	No	No	Yes	Yes
Standard Error Clustering	None	Question Level	Experiment Level	None	Question Level	Experiment Level	None	None

Standard errors in parentheses
** significant at 10% level, ** 5%, *** 1%*

[21] Inclusion of the seven dropped questions yields an aggregate Z-score of 4.95, also significant at any conventional level.
[22] We aggregate using the Stouffer Z-score method, in which the aggregated Z-score is computed by summing the Z-scores of the individual tests and dividing by the square root of the number of these tests.
[23] The significance of these coefficients is robust to the inclusion of the seven dropped questions, though including these seven questions results in slightly smaller point estimates (ranging from 0.055 to 0.060).
[24] See Appendix C for results with sample restricted to experiments implemented with random and quasi-random assignment.

After the data are disaggregated by question type, the evidence remains strongly consistent with herding, especially in both the sensitive and factual questions that we saw were driving the significant differences between hand-raising and clicker results in Section 4.1. Table 3 summarizes the herding evidence by question type.[25]

Table 3 – Herding by Question Type[26]

Question Type	Estimated herding (coefficient on Clicker in regression (1) from Table 2)	p-value (no clustering of SE)	p-value (SE clustered at question level)	Fraction of questions exhibiting herding
No Right Answer, Sensitive	0.071	0.002***	0.008***	17 out of 24 (70.8%)
No Right Answer, Not Sensitive	0.056	0.114	0.269	7 out of 13 (53.8%)
Right Answer, Factual	0.118	0.005***	0.014**	8 out of 9 (88.9%)
Right Answer, Conceptual	0.007	0.870	0.886	4 out of 8 (50.0%)

** significant at 10% level, **5%, ***1%*

In the 17 questions for which there was a single correct answer (the last two rows of Table 3), evidence of herding emerged in 12 of the questions. In a nod towards the "wisdom of crowds," in all but one of these 12 cases, the herding was towards the correct answer (hand raisers did better than clickers). In each of the five cases where there was no evidence of herding, the clickers did better than the hand raisers. But while hand raising may increase the number of students responding with the correct answer, that is hardly the goal if many students are merely following others. It seems likely that clicker responses give the teacher a better picture of where the students actually are in understanding a given concept or question.

5. DISCUSSION

To assess how these results apply in predicting the effectiveness of clickers in the classroom, it is important to consider why teachers do or should ask questions during class in the first place. Wittrock's generative theory of learning emphasizes that it is not a learner's behavioral activity (raising hands, clicking clickers) that causes learning, but rather the learner's cognitive activity that leads to their response. Students learn better when they are engaged in appropriate cognitive activity (Mayer and Wittrock, 2006; Wittrock, 1990). Specifically, Mayer outlines three cognitive processes that aid learning: selecting the relevant material, organizing that material into a coherent representation in working memory, and integrating that representation into long-term memory (Mayer 2001 and 2008). Well-designed questions in the classroom can spur students to engage in these cognitive processes as they identify the material being asked about, organize this material as they work towards an answer, and then incorporate this experience into their prior knowledge.

[25] Running this analysis while including the seven dropped questions yields similar results, with the exception of the p-values for factual questions with single right answers. For these questions, with no clustering of the standard error, the p-value is 0.066; with clustering of the standard error at the question level, it is 0.203.

[26] See Appendix D for results with the sample restricted to experiments implemented with random and quasi-random assignment.

Herding in response to questions in the classroom can short-circuit this process and thereby rob questions of their ability to spur this type of thinking. In fact, though our study was not designed to test this, it seems reasonable to believe that academically weaker students are more likely to herd. Students who probably have the greatest need for in-class generative learning processes are those most likely to simply vote with the majority. Since such a voting strategy requires only behavioral activity and little to no cognitive activity, the learning process is stunted.

Herding presents different additional problems for the teacher, especially since our results suggest that herding pushes more of the class towards the correct answer, on average. Teachers engaged in contingent teaching need a reliable means to ascertain what fraction of a class has understood a given topic or question. The hand-raising technique, given its herding correlate, is likely to exaggerate a class's grasp of the material to some unknown extent. This undermines the teacher's ability to engage in effective contingent teaching.

6. CONCLUSION

In response to the first key research question examined in this study (*Do students give the same responses when using raised hands as when using clickers?),* the experimental evidence provides a definitive *no*. Over half (57.3%) of questions show a difference in hand and clicker responses of over 10 percentage points, and nearly a quarter (24.6%) show a difference of over 20 percentage points. These correspond to differences in vote spreads of 20 percentage points and 40 percentage points, respectively. Taking all 61 questions together, we can reject at the 1% significance level the null hypothesis that the two techniques do not generate differences in vote shares. Statistically significant differences in vote shares mostly persist even when the data are split out by question type. The one exception is questions that have no single correct answer and do not involve sensitive issues. We hypothesize that these low-stakes questions may not be as likely to lead to students' reliance on the responses of other students.

The answer to the second key question (*Why don't students give the same responses when using raised hands as when using clickers*?) is more difficult to determine, since there are multiple possible explanations. One explanation that is consistent with our data is that when answers "matter" (that is, when the question has only one correct answer or is uncomfortably sensitive), students raising their hands have a tendency to herd and vote with the majority. Two-thirds of questions in our experiments show evidence of this type of herding, including the 11% of questions that elicited unanimous responses given hand raising. (No clicker response led to a unanimous outcome.) Majority-vote shares for hand raisers average 6.46 percentage points higher than for clickers, which corresponds to a difference in vote spread of nearly 13 percentage points. These results are robust to a number of alternative regression specifications and weighting schemes. While majority herding may explain only part of the differences between hand-raising and clicker responses, the results of the experiments in this study suggest it may be an important one.

The implications for instructors are clear. The large and persistent differences found in this study between the responses communicated by raised hands and those communicated by clickers should be considered by teachers who persist with hand raising when they interpret students' shows of hands. If a high proportion of hand raisers respond correctly to a question, it may mean that the class has understood the topic and the lesson can move forward. Unfortunately, it may also mean that although some students did indeed understand, others simply voted with the herd.

Questions in the classroom have two major purposes: First, to engage students in meaningful cognitive activity; and second, to enable the teacher to assess student capabilities, knowledge or preferences. Either purpose is defeated if students simply vote with the herd, as they frequently do with hand raising. By contrast, clickers, and other audience response systems, only allow students to respond on an individual basis. The use of clickers in the classroom enables questions to fulfill their intended pedagogic roles.

ACKNOWLEDGEMENTS

The authors would like to thank all of the instructors who allowed us to run these experiments in their classrooms.

REFERENCES

Abrahamson, L. (2006). A brief history of networked classrooms: Effects, cases, pedagogy, and implications. In D. A. Banks (Ed.), *Audience response systems in higher education* (pp. 1–25). Hershey, PA: Information Science Publishing.

Ainsworth, S., Gelmini-Hornsby, G., Threapleton, K., Crook, C., O'Malley, C., and Buda, M. (2011). Anonymity in classroom voting and debating. *Learning and Instruction*, 21(3), 365-378.

Bruff, Derek. (2009). *Teaching with classroom response systems: Creating active learning environments*. San Francisco, CA: Jossey-Bass.

Crouch, C. H., and Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, 69(9), 970–977.

Draper, S. W., and Brown, M. I. (2004). Increasing interactivity in lectures using an electronic voting system. *Journal of Computer Assisted Learning*, 20(2), 81–94.

Kay, R. and LeSage, A. (2009). Examining the benefits and challenges of using audience response systems: A review of the literature. *Computers and Education*, 53, 819-827.

Kennedy, G. E., and Cutts, Q. I. (2005). The association between students' use of electronic voting systems and their learning outcomes. *Journal of Computer Assisted Learning*, 21(4), 260–268.

Latessa, R., and Mouw, D. (2005). Use of audience response system to augment interactive learning. *Family Medicine*, 37(I), 12–14.

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.

Mayer, R. E. (2008). *Learning and instruction*. New York: Pearson Merrill Prentice Hall.

Mayer, R. E., and Wittrock, M. C. (2006). Problem solving. In P. A. Alexander and P. H. Winne (Eds.), *Handbook of educational psychology: Second edition* (pp. 287–304). Mahwah, NJ: Erlbaum.

Preszler, R. W., Dawe, A., Shuster, C. B., and Shuster, M. (2007). Assessment of the effects of student response systems on student learning and attitudes over a broad range of biology courses. *CBE-Life Sciences Education*, 6(1), 29–41.

Siau, K., Sheng, H., and Nah, F. (2006). Use of classroom response system to enhance classroom interactivity. *IEEE Transactions on Education*, 49(3), 398–403.

Simpson,V. and Oliver, M. (2007). Electronic voting systems for lectures then and now: A comparison of research and practice. *Australasian Journal of Educational Technology*, 23(2), 187–208.

Wittrock, M. C. (1990). Generative processes of comprehension. *Educational Psychologist*, 24, 354–376.

Appendix A

Number	Location	Date	Participants	Number of Participants	Number of Questions
1	Harvard	October 2010	Faculty from Various Universities	85	2
2	New York University	March 2011	NYU Faculty	36	2
3	Ohio State University	April 2011	Faculty from Various Universities	39	2
4	Harvard	March 2012	Harvard Kennedy School Faculty	90	1
5	Harvard	April 2012	Master's Students	37	2
6	Harvard	April 2012	Master's Students	44	3
7	Harvard	April 2012	Master's Students	56	3
8	Harvard	April 2012	Master's Students	65	3
9	Harvard	April 2012	Master's Students	51	4
10	Harvard	April 2012	Master's Students	36	2
11	Harvard	April 2012	Master's Students	73	4
12	Harvard	April 2013	Young Global Leaders (Executive Education)	63	4
13	Boston	March 2013	Cambridge Associates Investment Officers	52	4
14	Boston	April 2013	Fidelity Investments Employees	16	1
15	Fudan University (China)	May 2013	Conference Participants	47	3
16	Harvard	June 2013	Executive Education Participants	32	4
17	Harvard	July 2013	Executive Education Participants	65	3
18	Harvard	November 2013	Executive Education Participants	50	3
19	Tilburg University (Netherlands)	August 2014	Seminar Participants	91	2
20	Harvard	August 2014	Executive Education Participants	33	3
21	Harvard	March 2015	Executive Education Participants	37	2
22	Ohio State University	March 2015	Faculty and students from OSU	51	3

Appendix B

Response Differences by Question Type (only random assignment)

Question Type	Avg. %-point difference between technologies	Fraction of differences greater than 10 %-points	p-value associated with null of no differences
No Right Answer, Sensitive	14.24%	6 out of 8 (75.0%)	0.259
No Right Answer, Not Sensitive	10.63%	3 out of 7 (42.9%)	0.740
Right Answer, Factual	35.00%	2 out of 2 (100%)	0.002***
Right Answer, Conceptual	11.48%	4 out of 8 (50.0%)	0.528

** significant at 10% level, **5%, ***1%*

Response Differences by Question Type (only random and quasi-random assignment)

Question Type	Avg. %-point difference between technologies	Fraction of differences greater than 10 %-points	p-value associated with null of no differences
No Right Answer, Sensitive	13.86%	14 out of 20 (70.0%)	0.065*
No Right Answer, Not Sensitive	9.81%	4 out of 11 (36.4%)	0.897
Right Answer, Factual	17.52%	3 out of 5 (60.0%)	0.027**
Right Answer, Conceptual	11.48%	4 out of 8 (50.0%)	0.528

** significant at 10% level, **5%, ***1%*

Appendix C

Regression of DFE on Clicker (only random assignment)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Clicker	0.043*	0.043	0.043	0.054**	0.054*	0.054*	0.043*	0.055**
	(0.026)	(0.027)	(0.024)	(0.026)	(0.028)	(0.025)	(0.026)	(0.026)
Constant	0.250	0.250	0.250	-	-	-	0.240	-
	(0.018)	(0.032)	(0.022)				(0.018)	
Question Fixed Effects	No	No	No	Yes	Yes	Yes	No	Yes
Standardizing Weights	No	No	No	No	No	No	Yes	Yes
SE Clustering	None	Question Level	Experiment Level	None	Question Level	Experiment Level	None	None

Standard errors in parentheses

** significant at 10% level, ** 5%, *** 1%*

Regression of DFE on Clicker (only random and quasi-random assignment)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Clicker	0.055***	0.055**	0.055**	0.061***	0.061***	0.061***	0.050**	0.058***
	(0.020)	(0.021)	(0.019)	(0.021)	(0.021)	(0.018)	(0.020)	(0.020)
Constant	0.231	0.231	0.231	-	-	-	0.227	-
	(0.013)	(0.025)	(0.022)				(0.014)	
Question Fixed Effects	No	No	No	Yes	Yes	Yes	No	Yes
Standardizing Weights	No	No	No	No	No	No	Yes	Yes
SE Clustering	None	Question Level	Experiment Level	None	Question Level	Experiment Level	None	None

Standard errors in parentheses

** significant at 10% level, ** 5%, *** 1%*

Appendix D

Herding by Question Type (only random assignment)

Question Type	Estimated herding (coefficient on Clicker in regression (1) from Appendix C)	p-value (no clustering of SE)	p-value (SE clustered at question level)	Fraction of questions exhibiting herding
No Right Answer, Sensitive	0.067	0.113	0.142	4 out of 7 (57.1%)
No Right Answer, Not Sensitive	0.012	0.817	0.782	4 out of 6 (66.7%)
Right Answer, Factual	0.346	0.000***	0.019**	2 out of 2 (100%)
Right Answer, Conceptual	0.007	0.870	0.886	4 out of 8 (50.0%)

** significant at 10% level, **5%, ***1%*

Herding by Question Type (only random and quasi-random assignment)

Question Type	Estimated herding (coefficient on Clicker in regression (1) from Appendix C)	p-value (no clustering of SE)	p-value (SE clustered at question level)	Fraction of questions exhibiting herding
No Right Answer, Sensitive	0.084	0.001***	0.016**	13 out of 18 (72.2%)
No Right Answer, Not Sensitive	0.012	0.781	0.717	5 out of 10 (50.0%)
Right Answer, Factual	0.174	0.002***	0.044**	4 out of 5 (80.0%)
Right Answer, Conceptual	0.007	0.870	0.886	4 out of 8 (50.0%)

** significant at 10% level, **5%, ***1%*