British Journal of Educational Technology *Vol 47 No 4 2016* 710–720
doi:10.1111/bjet.12258

Does pressing a button make it easier to pass an exam? Evaluating the effectiveness of interactive technologies in higher education

Jose I. Castillo-Manzano, Mercedes Castro-Nuño, María Teresa Sanz Díaz and Rocio Yñiguez

Jose I. Castillo-Manzano is a main researcher in the Applied Economics & Management Research Group at the University of Seville. Mercedes Castro-Nuño is a researcher in the Applied Economics & Management Research Group at the University of Seville. María Teresa Sanz Díaz is a researcher in the Economic Theory and Political Economy Group at the University of Seville. Rocio Yñiguez is a researcher in the Economic Theory and Political Economy Group at the University of Seville. Address for correspondence: Prof. Jose I. Castillo-Manzano, University of Seville, Applied Economics & Management Research Group, Avda. Ramón y Cajal N° 1, Seville 41018, Spain. Email: jignacio@us.es

Abstract

The aim of this paper was to evaluate how audience response system (ARS) technology may increase improvements in academic performance in higher education, using the first year of the Administration and Business Management degree course at the University of Seville (Spain) as a case study. The experiment assesses whether the use of ARSs increases the likelihood that students will pass the final examinations in the subject of Principles of Economics. An econometric model is applied to a sample of 119 students in an intervention group, with a control group of 322. The statistically significant results show that at the very least, ARSs improve performance in the theoretical examination, albeit with certain limitations. It is concluded that ARSs should be used frequently to optimize outcomes, not just as a sporadic event during the course.

Introduction

European universities are undergoing a restructuring in the current context of the European Higher Education Area with teaching excellence emerging as a key part of the so-called *value creation process* (Barile & Polese, 2010) and students being actively involved in an interactive learning process.

New teaching strategies and the use of innovative tools are required to promote the development of generic and specific competences in students, who are both autonomous and able to undertake independent learning (Salas Velasco, Sánchez Martínez & Rodríguez Ferrero, 2012). Information and communication technology (ICT) seems destined to play a major role in innovative teaching in this respect, and, more specifically, one kind of ICT tool called audience response systems (ARSs) or *clickers* (Moss & Crowley, 2011).

ARSs were first used in higher education in the USA in the 1960s, but it was in the mid-1990s that their use spread to universities (Arenas-Márquez, Machuca & Medina-López, 2012) as an innovative tool to address limitations of traditional lecture classes: students distracted and unmotivated, poor interaction with professors, lack of anonymity and slowness in modes of involvement such as hand-raising and the request for volunteers. ARSs-based teaching seems to provide several advantages (Salemi, 2009): it allows anonymous responses, feedback is immediate and

Practitioner Notes

What is already known about this topic

- Innovative technologies, such as audience response systems (ARSs), are useful tools in current university teaching strategy.
- ARSs have several advantages (allow anonymous responses, enable all the students to respond, immediate feedback and very quick information processing), but also some disadvantages (the cost of clickers, technical malfunctions).
- Few and contradictory results are known about the role of ARSs in students' learning from an empirical point of view. Previous literature seems to be very heterogeneous, usually based on qualitative methodologies, and does not allow solid conclusions to be drawn about the improvements achieved.

What this paper adds

- We use econometric methodology and a broad control group compared with the intervention group that allow for robust conclusions regarding the influence of ARSs on students' outcomes in a final examination with both theoretical and mathematical papers.
- The likelihood of passing the theoretical examination paper rose by a mean of 14% for students who used ARSs (and improves with the frequency with which they are used), although there is no empirical evidence at all regarding their influence on solving mathematical problems.
- A student who has done all three intervention tests using ARSs has a real 15% greater chance of passing the theoretical paper, while a student who has only done one of these tests only sees his or her chances increased by slightly less than 5%, and the student who has done two tests, by under 10%.

Implications for practice and/or policy

- ARSs should not be used sporadically because their effects are optimized if frequent use is made of them.
- Further research is required to determine what the optimal number of tests is and what the interval of the number of tests should be to maximize the likelihood of passing the examination.
- Any plan to implement ARSs at a university should specifically take into account the type of teaching that they are to be used for, especially whether it is theory or practice.

information processing is very quick and easy. However, researchers also state that ARSs may have some disadvantages: the cost of clickers (King & Robinson, 2009) and the possibility of technical malfunctions (Guse & Zobitz, 2011). Authors like Champagne (2013) and Ng'ambi (2013) analyze some resources to overcome these drawbacks, such as the implementation of new software that uses the Internet, laptops, cell phones, tablets and computers.

Relatively little is known about the role of ARSs in students' achievements, and only a limited number of studies have been carried out in the social sciences (see eg, systematic review by Kay & LeSage, 2009). Research has focused on students' experiences relating to aspects of e-learning courses (Ginns & Ellis, 2009), interaction with an instructor or learning with a specific management system (Paechter, Maier & Macher, 2010), and students' general satisfaction (López-Pérez, Pérez-López & Rodríguez-Ariza, 2011).

Although the majority of recent studies on the effect of ARSs on academic achievement conclude that ARSs may have a positive impact, there seems to be a certain degree of heterogeneity, both with regard to the methodologies used (before–after analysis, such as Hancock, 2010; qualitative techniques, such as Mollborn & Hoekstra, 2010; econometric regressions, such as Desrochers & Shelnutt, 2012 and Marshall & Varnon, 2012; or even multicriteria decision, such as Crossgrove & Curran, 2008) and sample characteristics (with or without control groups that can be used as counterfactuals to the intervention groups).

Our study can complement the previous literature, which uses econometrics and includes a better intervention group/control group ratio than has been used in other studies (1/3 ratio in our case versus 1/1 in Desrochers & Shelnutt, 2012 and 1/2 in Marshall & Varnon, 2012).

We investigate a case study of the Principles of Economics course at the University of Seville (Spain) to evaluate the effectiveness of ARSs as a tool for improving academic performance in higher education. ARSs have gained popularity in almost all scientific fields, although following Bojinova and Oigara (2011), there has been limited research in economics regarding the impact of clickers on examination scores.

Our aim is to assess the hypothesis as to whether using ARSs increases the likelihood that students will pass their final examinations. To be specific, we have used an econometric model framed by statistical causal inference (see Appendix or Castillo-Manzano & Sánchez-Braza, 2011) to test the effect of the introduction and use of clickers, and to analyze whether ARS technology improves exam scores on the course. For this, the students were randomly divided into two groups, the first provided with clickers and the second without.

Research design

Participants

The population consisted of all the students who regularly attended class in the subject: 441 students in 9 groups during the first semester of the first year of the Administration and Business Management degree course. The subject analyzed, namely Principles of Economics, was taught to all these groups by different professors using basic methodology: theoretical and practical classes based on mathematical exercises. Students were given a final written examination with a theoretical section containing questions where they had to elaborate on their answers and solve a mathematical problem. At the beginning of the semester, a questionnaire on personal characteristics (socio-demographic and academic matters) was conducted (variable and descriptive statistics in Table 1 [database is available from the authors upon request]).

Two of the groups, with a total of 119 students, were randomly chosen as the sample for the ARSs intervention (treatment group). The remaining 322 students formed a control group that had no access to ARSs but had to sit exactly for the same final examination.

One interesting aspect of this research is that students in their first year at the School of Economics were chosen because it was virtually impossible for them to have had any previous experience of ARSs, having just come from senior high school; as far as is known, ARSs are not being used at any high school in the proximity of the university at this time.

Professors at the Economics Department have detected a lack of student involvement and participation in previous years, particularly in large classes with students beginning their first degree course. Clicker-based active learning methodologies have therefore been implemented to try to engage students in active learning, increase their participation and attention in the classroom, and thus enable them to achieve better scores in the final written examination.

Experimental intervention

A procedure described by Bartsch (2013) was followed to minimize any potential bias due to students' expectations (the "Hawthorne effect," namely, research outcomes might be influenced

Table 1: Covariates and descriptive statistics

Variable	*Description*	*No. of observations*	*Mean*	SD
Gender	1 if male; 0 if female	202	0.483	0.500
Age	Student's age	—	19.000	3.337
Sevillian	1 if Seville was the student's place of residence before going to university; 0 otherwise	261	0.595	0.492
Worker	1 if the student has a paid job; 0 otherwise	30	0.073	0.260
Place of residence	1 if not living in family home (hall of residence or student flat); 0 if living in family home	132	0.301	0.459
Freshman	1 if student's first year at university; 0 if repeating academic year	390	0.931	0.254
First option	1 if student obtained a place on his or her first choice degree course at university; 0 otherwise	357	0.890	0.313
University access examinations	Mark or grade that the student achieved in the national university access examinations, which also takes in his or her mean high school mark/grade: on a scale of 1 (lowest) to 14 (highest)	—	9.581	1.655
Vocation	Student's personal assessment of his or her vocation for the degree course that he or she is on: scale of 1 (lowest) to 4 (highest)	—	3.569	1.014
Family pressure	Student's personal assessment of family pressure on him or her to study at university: on a scale of 1 (lowest) to 4 (highest)	—	1.630	0.926

SD, standard deviation.

by increased attention being placed on the intervention groups) or to greater interest shown by the professor-researchers that may distort the validity of outcomes: first, students were not informed of the purpose of the experiment and at the end of semester were assessed using the same written examination as the students in the control group and, second, the research phases were distributed in a "triangular" fashion; professors who performed the experiment using clickers were not the same professors that designed the tests, developed the database, applied the econometric model and obtained the results.

Methodologically, the experiment was designed according to the principles of random sampling based on social sciences standards. As Table 1 shows, 10 covariates were included to homogenize the characteristics of the individuals in the control (non-clicker) and treatment groups. This was done by calculating the propensity score (see Appendix) with these covariates to correct for the effect of the treatment—in this case, the use of clickers—for any bias that could be attributed to their use.

The variables were chosen to address individual differences in student profiles (gender, age, lifestyle) and their personal circumstances (income level, expectations, motivational and family aspects) following previous literature (Edens, 2008; Kay & Knaack, 2009) and considering ideas brainstormed by the professors of the Principles of Economics course according to their expertise.

Three tests of knowledge were conducted using clickers in treatment group sessions. These tests corresponded to the three parts into which the subject program was divided (Fundamentals of Economics and Economic Thought, Macroeconomics and Microeconomics). The tests were given through PowerPoint presentations after each block of the program had been taught in class.

The structure of the tests was the same in all three cases: 20 multiple choice questions with three possible answers each, only one of which was correct. The software corrected the tests instantly upon their completion and students were informed about their correct answers and their mistakes. In the case of the latter, professors explained the questions and supplied the correct answers underneath. However, these tests did not count toward the final mark in the subject. The main objective was to explore students' comprehension levels for the different parts of the program and compare them with the other students in the control group by means of a proxy based on students' scores in the final written examination. As stated previously, the aim was to determine whether students who did these tests with clickers during the semester achieved better scores in the final examination compared with those that were not subjected to the intervention.

Results and discussion

Table 2 summarizes the results of the propensity score estimation (Appendix, Model 1) in the context of the 10 covariates in Table 1. A logit specification was chosen as it presents the best statistical properties. The logit is the standard discrete choice model used to analyze dichotomous choices, ie, when there are only two alternatives to choose between, which in our case is belonging to either the control group or the treatment group.

The resulting coefficients indicate the degree to which each of the 10 covariates contributes to the *propensity* score. The propensity score can be defined as the likelihood that individuals will receive the treatment, in this case, using clickers in their learning, given the values of their covariates. Its main goal is to make the individuals from the treatment and the control groups as homogeneous as possible as far as the 10 covariates are concerned.

A bivariate probit specification was then used to estimate the causal effects (see Appendix). A bivariate probit model is a discrete choice model used to examine the determinants of two closely linked dichotomous options, in our case, passing (or failing) the theoretical and practical parts of the examination.

The results are shown in Table 3. In Model 1 (columns 2 and 3), D_i takes a value of 1 for students in the group that used ARSs and 0 for students in the control group. In Model 2 (the last two columns), D_i is a variable that increases from 0 to 3 depending on the number of tests that the

Table 2: Probit estimation of the propensity score

Covariate	*Coefficient*
Gender	−0.198 (0.285)
Age	−0.071 (0.143)
Sevillian	−0.736 (0.308)**
Worker	−0.562 (0.820)
Place of residence	−0.022 (0.336)
Freshman	0.270 (0.626)
First option	1.381 (0.571)**
University access examinations	−0.224 (0.084)***
Vocation	0.049 (0.121)
Family pressure	−0.127 (0.155)
Constant	1.343 (3.204)
No. observations	340
Log pseudo-likelihood	−177.017
Pseudo R^2	0.052
Wald chi-square	16.99

Note: Standard errors robust to heteroscedasticity (when the variance of the dependent variable varies across the data) in brackets in the coefficient column. One, two or three asterisks indicate that the coefficient is statistically significant from zero at 10%, 5% and 1% respectively. One percent is the optimum situation and indicates the greatest significance.

Table 3: Bivariate probit estimation of relevant causal effects

	Model 1		*Model 2*	
Variable	*Theory pass*	*Practical pass*	*Theory pass*	*Practical pass*
Clicker (D_i)	0.383 (0.178)**	0.022 (0.174)		
Number of clickers (D_i)			0.129 (0.064)**	0.031 (0.062)
Constant	0.272 (0.203)	−0.062 (0.205)	0.276 (0.203)	−0.069 (0.203)
$\hat{\varepsilon}(x_i)$	−0.319 (0.826)	−1.444 (0.851)*	−0.312 (0.828)	−1.471 (0.856)*
$(\hat{\varepsilon}(x_i)-E[\hat{\varepsilon}(x)])D_i$	1.871 (1.666)	2.658 (1.611)*	1.913 (1.663)	2.658 (1.611)

Note: The considerations are the same as in the note for Table 2. $\hat{\varepsilon}(x_i)$ is the estimated value of the propensity score for the *i*-th individual. $(\hat{\varepsilon}(x_i)-E[\hat{\varepsilon}(x)])D_i$ is the difference between this value and the mean of this value for the whole of the sample multiplied by 1 or 0 depending on whether the individual belongs to the treatment or the control group respectively. One or two asterisks indicate that the coefficient is statistically significant from zero at 10 percent and 5 percent, respectively.

Table 4: Marginal effect of D_i *variables at the mean*

	Clicker (Model 1)	*Number of clickers (Model 2)*
Theory pass	0.1401 Δ 14.01%	0.0489 Δ 4.89%

student has done with ARSs prior to the final examination. Furthermore, the aim of including the propensity score in the estimations of the treatment effect—in this case, the use of clickers (see the last and last-but-one rows in Table 3)—is to correct for the effect of any bias caused by the 10 covariates used.

The positive coefficient sign and the significance of both the “Clicker” (Table 3, column 2) and “Number of clickers” (Table 3, column 4) variables at the 5% level lead to the conclusion that ARSs have a positive effect on the likelihood of a student managing to pass the theoretical examination paper, and that the more tests he or she does with ARSs, the greater the likelihood is (according to Mun, Hew & Cheung, 2009). However, the lack of significance of these two variables (columns 3 and 5) for the practical exercise implies that ARSs are ineffective when it comes to improving students’ academic performance in this type of examination. In this regard, Lin, Liu and Chu (2011) conclude that ARSs improve performance in theoretical concepts, but are not significant for practical concepts.

We think that the immediate feedback afforded by this technique provides students with the chance to see what their shortcomings/gaps in their knowledge are before their final examinations, and therefore serves as a guide as to what they need to study more.

The estimation of the marginal effect at the mean of the students who received instruction with the use of ARSs is included to quantify the effects of the “Clicker” and “Number of clickers” variables on increasing the likelihood of students passing the theory paper (Table 4).

Conclusions

Most Spanish universities have made major efforts to boost innovation during the implementation of the European Higher Education Area, building international campuses of excellence and reformulating teaching techniques for the efficient incorporation of new degree courses (Ion & Castro, 2012). During this process, they have developed ICT based on tools such as ARSs and these are being used as a didactic aid for lectures and to keep track of attendance (Novo-Cortí, Varela-Candamio & Ramil-Díaz, 2013).

The consideration of recent studies that analyze the utility of ARSs shows that they might have a positive impact on the learning process in general terms. However, previous literature seems to be very heterogeneous and does not allow any consensus to be reached on the improvements achieved. Therefore, new empirical research based on large samples and the use of control groups is required to provide concrete results regarding the real utility of ARSs.

The findings of the present study show clear statistical evidence at 5% that ARSs are a significant aid to improving academic performance, with clicker groups apparently feeling more engaged and active within their classes. From a pedagogical point of view, this attitude is due to students' familiarity with the use of technology, even though they have not used these particular resources previously, and as a result the novelty appeals to them. This experiment can also provide them with immediate feedback and reinforce their learning, which enables them to be aware of where they have to concentrate their learning and even spark a class discussion with other students and the professor in order to clarify certain issues.

However, this support is not a universal panacea and does have some clear constraints. First, as mentioned above, it focuses on a theoretical examination. Specifically, the likelihood of passing the theoretical examination paper rose by a mean of 14% for students who had the opportunity to use ARSs or clickers. But there is no empirical evidence at all that they are an aid to solving mathematical problems, even though the tests that the students using ARSs had to do included questions of this type.

Second, the help that ARSs give students clearly depends on the frequency with which they are used by the said students. In other words, ARSs should not be used sporadically, and the outcome of ARS use is optimized when they are used frequently. According to the results of Model 2, the likelihood of passing the theoretical examination paper improves in accordance with the number of tests done with clickers. More specifically, a student who has done all three tests using ARSs has a real 15% greater chance of passing the theoretical paper, while a student who has only done one of these tests only sees his or her chances increased by slightly less than 5%, and a student who has done two tests, by under 10%.

On the other hand, the database is formed of human elements, and therefore the research may be exposed to several unavoidable limitations, ie, the positive outcomes might be influenced by increased attention being given to the students in the intervention groups or by a greater interest shown in them by professor-researchers. The research design section describes the attempt made to minimize this circumstance, but these factors might bias this type of research.

Despite the findings being statistically significant in terms of academic performance, indicating possible benefits from the use of clickers, it should also be pointed out that, following previous research, the effectiveness of ARSs may be limited by other external factors. These include the way that the instructor makes use of this technology (sporadic assessments or summative assessment, the type of questions used, for initiating discussion in class, for peer evaluation) and the discipline for which clickers are being used.

More research is needed in the future to be able to state whether clickers can enhance student learning and motivation compared with traditional teaching. It would be logical to suppose that tests done with ARSs will have a decreasing marginal utility, as otherwise the absurd point would be reached where all the teaching time during the whole 4 months of the term would be taken up by tests. The effect of clickers also needs to be studied in the long term, not just in a specific subject or academic year, as in this case. It would then be possible to rule out any positive bias that comes from the initial enthusiasm that occurs when a new and attractive technology like ARS is introduced into the teaching for the first time. The next question to answer is, therefore, what the optimal number of tests is or, failing that, what the interval between tests should be for students to maximize the likelihood of their passing the examination.

However, in the main, future research should be directed at accumulating further case studies to confirm these findings. This becomes an even greater imperative if it is taken into account that relevant recommendations on educational planning will result. Any plan to implement ARSs at a university should specifically take into account the purpose of the tool and should be carried out in an ambitious way, with a sufficient number of remotes to enable students to use them frequently and not just on the odd occasion.

Acknowledgments
The authors are grateful to Nick Rushby and the three anonymous reviewers for their very helpful comments.

Statements on open data, ethics and conflicts of interest

a. The entire database is fully available from the authors without restriction upon request.

b. The authors have operated with an ethic of respect for the privacy of all the students involved in the research. On the one hand, the authors took voluntary informed consent as the condition in which participants understood and agreed to their participation in the experiment; thus, all participants in the research understood the process in which they were to be engaged, including how it would be used, and how and to whom it would be reported.

On the other hand, as we were working with aggregate data from such a large sample, 441 students, no personal information was revealed; it could not be otherwise, as Spanish Organic Law 15/1999 of 13th December on the Protection of Personal Data is entirely restrictive in this regard. In any case, to ensure maximum accuracy in student responses to the initial survey in which we asked for their baseline characteristics, we guaranteed the confidential and anonymous treatment of personal data: their private data (name, identity card number or student identification) were not added to the final database and so there was nothing in it that the authors could identify the students by. We hired an outside person to do all this (by means of the grant that we received from our university for this research project) and to take charge of the questionnaires directly and build the database, linking data from the initial questionnaires with test scores, ie, the authors have not had access to students' personal data (such as family or socio-economic status).

c. The authors declare that there is no potential conflict of interest in the work.

References

Arenas-Márquez, F. J., Machuca, J. A. D. & Medina-López, C. (2012). Interactive learning in operations management higher education. *International Journal of Operations & Production Management*, *32*, 1395–1426.

Barile, S. & Polese, F. (2010). Linking the viable system and many-to-many network approaches to service-dominant logic and service science. *International Journal of Quality and Service Sciences*, *2*, 23–42.

Bartsch, R. A. (2013). Designing SoTL studies—part I: validity. *New Directions for Teaching and Learning*, *2013*, 136, 17–33.

Bojinova, E. & Oigara, J. (2011). Teaching and learning with clickers: are clickers good for students? *Interdisciplinary Journal of E-Learning and Learning Objects*, *7*, 169–184.

Castillo-Manzano, J. I. & Sánchez-Braza, A. (2011). Evaluation of the establishment of a taxi flat rate from city to airport: the case of Seville. *Urban Studies*, *48*, 1909–1924.

Champagne, M. V. (2013). Student use of mobile devices in course evaluation: a longitudinal study. *Educational Research and Evaluation*, *19*, 636–646.

Crossgrove, K. & Curran, K. L. (2008). Using clickers in nonmajors- and majors-level biology courses: student opinion, learning, and long-term retention of course material. *Life Science Education*, *7*, 146–154.

Desrochers, M. N. & Shelnutt, J. M. (2012). Effect of answer format and review method on college students learning. *Computers and Education*, *59*, 946–951.

Edens, K. M. (2008). The interaction of pedagogical approach, gender, self-regulation, and goal orientation using student response system technology. *Journal of Research on Technology in Education*, *41*, 161–177.

Ginns, P. & Ellis, R. A. (2009). Evaluating the quality of e-learning at the degree level in the student experience of blended learning. *British Journal of Educational Technology*, *40*, 652–663.

Guse, D. M. & Zobitz, P. M. (2011). Validation of the audience response system. *British Journal of Educational Technology*, *42*, 985–991.
Hancock, T. M. (2010). Use of audience response systems for summative assessment in large classes. *Australasian Journal of Educational Technology*, *26*, 226–237.
Hirano, K. & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services & Outcomes Research Methodology*, *2*, 259–278.
Ion, G. & Castro, D. (2012). Governance in Spanish Universities: changing paradigms. *Journal of Hispanic Higher Education*, *11*, 336–350.
Kay, R. & Knaack, L. (2009). Exploring individual differences in attitudes toward audience response systems. *Canadian Journal Of Learning And Technology*, *35*. Retrieved May 2014, from http://www.cjlt.ca/index.php/cjlt/article/view/509/239
Kay, R. H. & LeSage, A. (2009). Examining the benefits and challenges of using audience response systems: a review of the literature. *Computers and Education*, *53*, 819–827.
King, S. O. & Robinson, C. L. (2009). Pretty lights and maths! Increasing student engagement and enhancing learning through the use of electronic voting systems. *Computers and Education*, *53*, 189–199.
Lin, Y.-C., Liu, T.-C. & Chu, C.-C. (2011). Implementing clickers to assist learning in science lectures: the clicker-assisted conceptual change model. *Australasian Journal of Educational Technology*, *27*, 979–996.
López-Pérez, M. V., Pérez-López, M. C. & Rodríguez-Ariza, L. (2011). Blended learning in higher education: students' perceptions and their relation to outcomes. *Computers and Education*, *56*, 818–826.
Marshall, L. L. & Varnon, A. W. (2012). An empirical investigation of clicker technology in Financial Accounting Principles. *Journal of Learning in Higher Education*, *8*, 7–17.
Mollborn, S. & Hoekstra, A. (2010). A meeting of minds: using clickers for critical thinking and discussion in large sociology classes. *Teaching Sociology*, *38*, 10, 18–27.
Moss, K. & Crowley, M. (2011). Effective learning in science: the use of personal response systems with a wide range of audiences. *Computers and Education*, *56*, 36–43.
Mun, W. K., Hew, K. F. & Cheung, W. S. (2009). The impact of the use of response pad system on the learning of secondary school physics concepts: a Singapore quasi-experiment study. *British Journal of Educational Technology*, *40*, 848–860.
Ng'ambi, D. (2013). Effective and ineffective uses of emerging technologies: towards a transformative pedagogical model. *British Journal of Educational Technology*, *44*, 652–661.
Novo-Cortí, I., Varela-Candamio, L. & Ramil-Díaz, M. (2013). E-learning and face to face mixed methodology: evaluating effectiveness of e-learning and perceived satisfaction for a microeconomic course using the Moodle Platform. *Computers in Human Behavior*, *29*, 410–415.
Paechter, M., Maier, B. & Macher, D. (2010). Students' expectations and experiences in e-learning: their relation to learning achievements and course satisfaction. *Computers and Education*, *54*, 222–229.
Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.
Salas Velasco, M., Sánchez Martínez, M. T. & Rodríguez Ferrero, N. (2012). Developing generic competences in the European Higher Education Area: a proposal for teaching the principles of economics. *European Journal of Education*, *47*, 463–476.
Salemi, M. K. (2009). Clickeconomics: using a classroom response system to increase student engagement in a large- enrollment principles of economics course. *The Journal of Economic Education*, *40*, 85–404.

Appendix

Details of experimental intervention

Table A1 shows the steps followed in the experimental intervention in the two treatment groups and control groups. In all cases, the professor explained the full three parts of the program, the students did microeconomics practical exercises with the professor and, at the end of term, all students sat for the same final exam: three long theoretical questions (eg, analysis and graphical explanation of the market demand curve) and one microeconomics practical exercise similar to those done in class.

Once each part had been explained; a test was done only in the two treatment groups. The software corrected the tests instantly and the professor answered any questions the students had about each of the test questions. The test questions were designed to reinforce the theoretical concepts explained in class (eg, indifference curves are: a) falling, concave and never intersect; b) falling, convex and never intersect; c) falling, convex and can intersect in exceptional circumstances).

Table A1: Experimental intervention: steps in treatment and control groups

	Treatment groups	*Control groups*
Explanation of first part of program, Fundamentals of Economics and Economic Thought	**X**	**X**
Test on basic theoretical concepts of first part of program (clickers)	**X**	
Instant correction of test on first part of program and explanation of questions and answers in the test	**X**	
Explanation of second part of program, Macroeconomics	**X**	**X**
Test on basic theoretical concepts of second part of program (clickers)	**X**	
Instant correction of test on second part of program and explanation of questions and answers in the test	**X**	
Explanation of third part of program (theory and practice), Microeconomics	**X**	**X**
Test on basic theoretical concepts of third part of program (clickers)	**X**	
Instant correction of test on third part of program and explanation of questions and answers in the test	**X**	
Final examination on full program with three long theoretical questions and one practical exercise	**X**	**X**

Statistical method

The methodology used in our study is framed by statistical causal inference. It is based on the estimation of the causal effect that a specific measure or fact can have on one or more relevant variables (see Castillo-Manzano & Sanchez-Braza, 2011). This methodology allows consistent estimators of the effects of the evaluated measure to be obtained by determining and isolating the possible impact of additional contaminating variables. This methodology is based on the typical method used to test the effectiveness of new medicines before they are approved for their release onto the market. Specifically, two groups are used, the treatment group, to which the new medicine is applied, and the control group, which continues to be treated in the traditional way. The ultimate aim is to quantify any increase in the likelihood of a patient recovering, thanks to being treated with the new medicine, over patients that receive the traditional medication.

In detail, we begin with an N-size random sample, and the binary variable D is defined that indicates the observation corresponding to a student who has used audience response systems (ARSs) ($D_i = 1$) or a student in the control group who received traditional instruction, essentially based on lecture classes ($D_i = 0$). Thus, our N observations were divided into N_1 and N_0 observations (using ARSs or clickers vs. traditional instruction). In our case, N_1 stands for the 119 students who used ARSs or clickers, while N_0 represents the remaining 322 students.

The outcome variables Y_{ij} and Y_{ik} are students passing the theoretical examination paper and the mathematical problems respectively. From this, the average treatment effect is obtained (following Hirano & Imbens, 2001) to determine the influence of the use of ARSs on the likelihood of passing the theoretical examination paper (Equation 1):

$$\alpha_j = E[Y_{ij}(1) - Y_{ij}(0)] = \frac{1}{N}\sum_{i=1}^{N}[Y_{ij}(1) - Y_{ij}(0)] \tag{1}$$

A K-dimensional vector of observed covariates was also defined as X. A triad could therefore be observed for each individual (D_{ij}, Y_{ij}, X_{ij}).

The observations were considered for students who had used ARSs ($D_i = 1$) first and subsequently the observations for the other students who had followed the traditional teaching method ($D_i = 0$). Using this process, the *propensity score* was estimated. This is defined by Rosenbaum and

Rubin (1983) as the conditional probability of "participating in the evaluated measure," given a vector X of observed covariates. The binary response model (logit or probit) that maximized the log pseudo-likelihood was used to estimate the propensity score:

$$\varepsilon(X) = P(D=1|X) = F(\beta X) \tag{2}$$

where β is the vector of parameters associated with X. In our case, X was composed of the 10 covariates.

The second step was to calculate the measure's causal effect on the response variable, in this case, the probability of passing both the theoretical and practical examination papers.

In this case, the outcome variables Y_{ij} and Y_{ik} are discrete choice variables, which take the value of 1 if the student has managed to pass the theoretical and practical papers respectively. A priori, there should be a relatively strong correlation between the two as the factors that affect whether the student passes the theoretical and practical papers can be anticipated to be similar in both cases. For this reason, a bivariate probit model was used. This model category is specially designed for cases where two questions with very closely linked binary answers need to be answered.

The bivariate probit formula is (Equation 3):

$$\begin{gathered} L = \sum \ln \Phi_2 (q_1 (X_i \tau)^{\tau}, q_2 (Z_i \gamma)^{\gamma}, \rho_i^*) \\ q_1 = \begin{cases} 1 & if \quad y_1 \neq 0 \\ -1 & if \quad y_1 = 0 \end{cases} \quad q_2 = \begin{cases} 1 & if \quad y_2 \neq 0 \\ -1 & if \quad y_2 = 0 \end{cases} \end{gathered} \tag{3}$$

where Φ_2 is the cumulative bivariate normal distribution function. In our case, $X_i = Z_i$ for the variables that deal with feature that vary from one student to another. Also, following Hirano and Imbens (2001), we have:

$$x_i' \tau_m = \tau_{m0} + \alpha D_i + \tau_{m1} \hat{\varepsilon}(x_i) + \tau_{m2} (\hat{\varepsilon}(x_i) - E[\hat{\varepsilon}(x)]) D_i + u_{ij} \tag{4}$$

As in binary outcome models, in a bivariate probit model, only the mathematical sign of the coefficient can be interpreted directly. Thus, a positive coefficient in the bivariate probit model means that as the regressor increases, there will be a greater likelihood that the event in question will come to pass, which in this case means that a student will be more likely to pass the theoretical or practical examination papers. For this reason, the marginal effects of each of the explanatory variables have been calculated at the mean. The marginal effects of variable D_i will therefore give us the increase in the likelihood that students using ARSs will pass both the theoretical and the practical papers, ie, it will give us a relatively accurate quantitative estimation of the contribution made by the use of ARSs in the specific teaching of Principles of Economics.