Computers & Education 96 (2016) 109–121

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Measuring the effect of ARS on academic performance: A global meta-analysis

José I. Castillo-Manzano [a, *], Mercedes Castro-Nuño [a], Lourdes López-Valpuesta [a], María Teresa Sanz-Díaz [b], Rocío Yñiguez [b]

[a] Applied Economics & Management Research Group, Universidad de Sevilla, Avda. Ramón y Cajal, 1, 41018 Seville Spain
[b] Facultad de Ciencias Económicas y Empresariales, Universidad de Sevilla, Avda. Ramón y Cajal, 1, 41018 Seville Spain

ARTICLE INFO

Article history:
Received 23 June 2015
Received in revised form 17 February 2016
Accepted 19 February 2016
Available online 23 February 2016

Keywords:
Audience response systems (ARSs)
Meta-analysis
Interactive learning environments
University disciplines

ABSTRACT

An increasing number of studies have addressed the impact of Audience Response Systems (ARS) on academic performance at all stages of education, although the evidence does not seem conclusive. With the aim of shedding light on the extent and diversity of the research outcomes, we conduct a meta-analysis of studies worldwide on this topic to assess whether the exam scores of students included in ARS experiments achieve better results than others taught using more conventional teaching tools. From an initial sample of 254 studies, data from 51 papers published between 2008 and 2012 (involving 14,963 participants) that set academic quality criteria, were extracted and analyzed following technical protocols for meta-analyses. Their high degree of heterogeneity shows that the effect of ARS on exam scores seems to be moderated by specific features. So, through a random-effects model, our results provide a positive, although moderated pooled effect of ARS on examination scores that is much greater in experiments performed in non-university contexts (Hedges' g = 0.48; S.E. = .2665) than at the university level (Hedge's g = 0.22, S.E. = .0434). Specifically, the categories of university disciplines in which ARS interventions are implemented seem to influence their usefulness for achieving better academic marks, being more effective when either Pure Soft Sciences or Applied Hard Sciences are considered. These findings might provide guidance for governments, researchers and educators into the effectiveness of learning based on the new interactive technologies.

1. Introduction

Today the role that the so-called Information and Communication Technologies (ICTs) have played as an innovative teaching support in university education is undisputed (Caird & Lane, 2015). According to researchers such as Lantz (2010) and Lara, Lizcano, Martínez, Pazos, and Riera (2014), in Europe ICTs have helped overhaul teaching methods in the wake of the implementation of the European Higher Education Area by responding to the new needs of the teaching-learning process and optimizing their adaptation to the economic-social environment.

* Corresponding author.
E-mail addresses: jignacio@us.es (J.I. Castillo-Manzano), mercas@us.es (M. Castro-Nuño), lolopez@us.es (L. López-Valpuesta), mtsanz@us.es (M.T. Sanz-Díaz), ovando@us.es (R. Yñiguez).

http://dx.doi.org/10.1016/j.compedu.2016.02.007
0360-1315/

The academic literature illustrates perfectly what ICTs have achieved with respect to student learning (Padilla-Meléndez, Águila-Obra, & Garrido-Moreno, 2015; Webster & Son, 2015), and demonstrates the limitations of traditional teaching approaches, including less active learner (university student or school pupil) involvement in the classroom (Stowell, Oldham, & Bennett, 2010) and the lack of periodic feedback on the degree to which the knowledge conveyed has been assimilated (Gok, 2011).

Among the ICT-based tools used to support teaching that can be highlighted are the so-called Audience Response Systems (hereinafter, ARS). Following DeSorbo, Noble, Shaffer, Gerin, and Williams (2013: 531), an ARS can be defined as "*an interactive electronic tool used to survey an audience for their responses to specific questions*".

The current paper uses a meta-analysis to summarize the prior evidence for the effectiveness of ARS for improving the academic score of students by education level and academic disciplines compared to other learning methods.

Following Fernández-Alemán, Sánchez, López, and López (2014), ARS were first used in US universities in the nineteen-sixties, although it was not until the nineteen-nineties that their use began to become widespread. From the technical point-of-view, ARS are normally based on a receiver system connected to learners' computers (which are in turn permanently linked to a projector), combined with a series of small hand held response devices that learners use to send a signal to the receiver system (King & Robinson, 2009). Students can thus use their devices to respond to the questions that the instructor brings up on the screen (or an interactive whiteboard). These are usually in the format of a Power Point presentation or similar. Following Sutherlin, Sutherlin, and Akpanudo (2013), the questions can be multiple choice (in which the learner has to choose one answer as correct); questions in a true–false format; fill in the blank or numerical questions and, as Desrochers and Shelnutt (2012) also consider, short word answers. Once the question has been answered by the learners, it is processed by the software in the instructor's computer, which checks the results of the test in real time (DeSorbo et al., 2013; Tlhoaele, Hofman, Naidoo, & Winnips, 2014).

The main advantages of this tool are, first, that it enables learners to be anonymous and so provides an incentive to participate for any learners who are usually too shy to do so (Bojinova & Oigara, 2013), reducing conformity, shame and anxiety in class (Stowell & Nelson, 2007; Stowell et al., 2010); and, second, that it enables information to be processed quickly and easily, both for instructors and for learners (Duncan, 2007; Hill & Babbit, 2013; Wood, 2004), thus enabling any concepts that have not been properly assimilated to be emphasized (Vana, Silva, Muzyka, & Hirani, 2011).

However, there are also some disadvantages to the ARS method, particularly the following: malfunctions (Desrochers & Shelnutt, 2012) produced by technology failures (White, Syncox, & Alters, 2011) – mainly in signal transmission (Guse & Zobitz, 2011); economic barriers, due to the high initial cost of acquiring the system, as stated by Desrochers and Shelnutt (2012) and Fernández-Alemán et al. (2014). King and Robinson (2009) and Rothman (2014) state that this means that on some occasions the cost has to be passed directly on to learners. Other authors, such as White et al., (2011), also consider the possibility of cheating as one of the most worrying disadvantages of ARS use, although for Roberson (2009) this technology has the opposite effect, limiting cheating during examinations.

As Arnesen, Sivertsen Korpås, Hennissen, and Birger (2013) and Costello (2010) suggest, technological advances, especially the spread of Wi-Fi networks to campuses and, to a lesser extent, high schools, combined with the success of smartphones, enables the main disadvantages of ARS to be overcome. For example, today it may be enough to have the computer connected to the projector and use smartphones (Arnesen et al., 2013; Stav, Nielsen, Hansen-Nygård, & Thorseth, 2010), tablets (Bryfczynski et al., 2014) or laptops (Costello, 2010) as response devices. Downloading ARS applications to these devices or using free web-based software thus makes it possible for the use of ARS to be generalized at a much lower cost (Fernández-Alemán et al., 2014; Hecht, Adams, Cunningham, Lane, & Howell, 2013).

From an academic point-of-view, as Blasco-Arcas, Buil, Hernández-Ortega, and Sese (2013) and Rothman (2014) state, research on the use of ARS applications in teaching is quite recent, although studies can be found that cover different academic levels and a range of disciplines. By education level, most studies focus on the university environment, where experiments have been carried out in a number of disciplines in the Experimental Sciences (Addison, Wright, & Milner, 2009; Nicol & Boyle, 2003), the Health Sciences (Levesque, 2011; Patterson, Kilpatrick, & Woebkenberg, 2010) and the Social Sciences (Castillo-Manzano, Castro-Nuño, Sanz Díaz, & Yñiguez, 2015; Stowell & Nelson, 2007). Studies have also been carried out at pre-university levels (Wash, 2012) for both primary education (DeSorbo et al., 2013) and secondary education (Barnes, 2008; Kay, 2009).

Broadly-speaking, these studies pursue objectives that, on the one hand, focus on learners' attitudes and perceptions regarding the application of ARS and on the effects that they have on the learning process; and, on the other hand, there are studies that analyze the impact of these tools in terms of academic performance.

The methodology used in the majority of the studies in the first group is usually interview- and survey-based. Using a Likert scale and implemented among the learners involved in the experiment, they are used to build a database and the data are analyzed descriptively and statistically (detailing indicators, such as the mean, mode frequency analysis, etc.). These studies use a wide range of outcomes to analyze the effects of ARS on learning, whether they aim to evaluate the learner's involvement in the classroom or active learning (Gauci, Dantas, Williams, & Kemm, 2009), or one or various related aspects, such as increases in attendance (Lattesa & Mouw, 2005; Preszler, Dawe, Shuster, & Shuster, 2007; Schackow, Chavez, Loya, & Friedman, 2004), attention levels (Burnstein & Lederman, 2001; Cue, 1998; Gachago, Morris, & Simon, 2011); and engagement (Cain, Black, & Rohr, 2009; Freeman, Haak, & Wenderoth, 2011; Preszler et al., 2007; Stowell & Nelson, 2007), classroom interactivity and participation (Bright, Reilly Kroustos, & Kinder, 2013; DeBourgh, 2008; Denker, 2013; Siau, Sheng, & Nah,

2006; Sprague & Dahl, 2010); the fostering of peer discussion (Berry, 2009; Nicol & Boyle, 2003; Smith & Rosenkoetter, 2009) and peer instruction (Fies & Marshall, 2006); and even formative assessment (Beatty & Gerace, 2009).

While ARS are widely advocated for their capacity to enhance learner motivation and engagement overall for large classes (Cheong, Bruno, & Cheong, 2012; Draper & Brownw, 2004; Gok, 2011), it is not so clear that they lead to improved learner academic performance in terms of exam scores (Arnesen et al., 2013; Berry, 2009; Fortner-Wood, Armistead, Marchand, & Morris, 2013; Han, 2014). In this environment, research is not conclusive. On the one hand some studies, such as Bright et al. (2013), Levesque (2011) and Mostyn, Meade, and Lymn (2012), report some degree of positive correlation between the use of ARS and perceived learning through specific exams results. Conoley, Croom, Moore, and Flowers (2007) and Mostyn et al. (2012) associate these better results achieved with ARS with the feedback that learners receive with this technology. However, and to the contrary, Gauci et al. (2009) note that the better results achieved in grade marks with ARS use are linked to greater learner participation.

However, other studies state that the effect of ARS use is not sufficiently significant in terms of learners' examination scores, and that the evidence found is not sufficiently robust to state that groups included in an ARS experiment achieve better results than others that are taught using more conventional alternative teaching techniques. This is the case of studies such as Anthis (2011) and Patterson et al. (2010), which compare the academic results obtained to those using a raised hand response, when exactly the same questions are asked.

Among the possible explanations for this unclear effect we find studies that state that the benefit of ARS in terms of learner scores is dependent upon specific factors linked directly to an individual's intrinsic characteristics, such as prior academic performance (Addison et al., 2009) or gender (Kang, Lundeberg, Wolter, DelMas, & Herreid, 2012). In others, such as FitzPatrick, Finn, and Campisi (2011), the outcomes regarding the scores are different depending on the subject in which the experiment is done.

Finally, Elicker and McConnell (2011) analyze the effect of ARS on participation in class and examination performance and compare it to the effects of other, more traditional teaching techniques (hand-held flashcards, and hand raising). They conclude that although learners like the technological method more, no differences can be seen in examination results.

Authors such as Arneja, Narasimhan, Bouwman, and Bridge (2009) offer false positives as another explanation for the lack of consistency in the findings when evaluating the effectiveness of ARS in score terms, highlighting the bias introduced by the favorable attitude of instructors and learners alike to new technologies, dubbed the “Hawthorne effect”.

The debate aroused around the question of the effectiveness of ARS in teaching is also brought to light in the various literature reviews published on the topic in recent years, including those by Caldwell (2007); Fies and Marshall (2006); Han (2014); Judson and Sawada (2002); Kay and LeSage (2009); Keough (2012) and Simpson and Oliver (2007). To be specific, on the one hand Kay and LeSage (2009) state that the use of ARS improves learning performance, although they recognize that research combining qualitative and quantitative methods needs to be done to analyze this aspect. On the other hand, Han (2014) states that although the use of ARS is positive in learners' perceptions of their engagement and learning, the effect of their use in enhancing learner performance is not so evident, as a result of which he considers that quantitative research methods need to be applied that help close gaps in the literature on this topic.

These criticisms seem to be confirmed by the lack of studies that have applied a meta-analysis to analyze the effectiveness of ARS as a teaching tool. As far as we know, there are only two obvious studies to date. The first of these, by Léger, Bourque, and Richard (2010), performs a meta-analysis to test the effectiveness of ARS for academic outcome. Said study, which finds a positive relationship between ARS use and learner outcome, has the limitation that it uses a small sample composed of only eight scientific articles (with a total of 17 estimates), the last of which is for 2009. A subsequent meta-analysis by Nelson, Hartling, Campbell, and Oswald (2012) seeks to determine the effects of ARS use on teaching outcomes in the area of the Health Sciences. Said meta-analysis looks at 21 studies up to 2010 and concludes that ARS use can improve knowledge outcomes in the short and long term.

However, meta-analysis is a methodology that has been used in a number of studies in the area of teaching and education. As an example, both more general meta-analyses that study the impact of computer technology on education (Archer et al., 2013; Liao, 2007; Schmid et al., 2014; Tamim, Bernard, Borokhovski, Abrami, & Schmid, 2011) can be cited, and also others that analyze the use of more specific traditional educational tools and methods in the learning process, such as concept and knowledge maps (Nesbit & Adesope, 2006); bilingualism (Adesope, Lavin, Thompson, & Ungerleider, 2010); mobile devices (Wu et al., 2012); verbally redundant presentations (Adesope & Nesbit, 2011); virtual patients in medical education (Consorti, Mancuso, Nocioni, & Piccolo, 2012); instructional support in game-based learning (Wouters & Van Oostendorp, 2013); intelligent tutoring systems (Ma, Adesope, Nesbit, & Liu, 2014); virtual reality technology-based instruction (Merchant, Goetz, Cifuentes, Keeney-Kennicutt, & Davis, 2014), and argumentation in computer-supported collaborative learning (Wecker & Fischer, 2014).

The present article seeks to offer the most comprehensive analysis to date of the effectiveness of ARS for learner academic performance at a time when the cost of using the system has fallen dramatically. Specifically, our research contributes to the literature by performing a meta-analysis of the effects of ARS use based on previous empirical studies on this topic worldwide. In this respect, taking all the foregoing into account, our study is justified both by the advances and extraordinary progress achieved by publications on the topic in recent years, and the lack of application to date of a robust methodology that enables a quantitative synthesis to be carried out of their results.

2. Methodological issues

2.1. Research questions

We aim to respond to the following research questions:

- are better examination scores achieved by learners who use ARS, with its immediate feedback, to prepare content and as an evaluation method, than learners who do not use ARS in any phase of the learning-teaching process?
- do the new teaching technologies based on interactive involvement with ARS achieve better, worse or similar academic results to other methods?
- is the effectiveness of ARS dependent on education level, high school compared to university, for example? And, within the university, does the effectiveness of ARS vary from one academic discipline to another?

2.2. Search strategy

During the process to find relevant studies for our meta-analysis two researchers exhaustively consulted typical electronic bibliographic databases using a double-blind process and without having any contact with one another. The databases included Web of Science, Google Scholar, Science Direct and Scopus, as well as databases specifically devoted to educational affairs, such as ERIC (Education Resources Information Center), Education Research Databases and ProQuest® Education Journal, for the time period of 1970 to March 2015, since, as commented in the introduction, ARS first started to be implemented in the nineteen-sixties.

During a second phase, searches were done of the references sections in the papers found in the above-described search to find additional references. Similarly, articles with systematic reviews of the preceding literature and two previous meta-analyses found in the databases, as commented in the previous section, were also reviewed in depth in order to find any possible further papers on the subject.

Twenty search terms were used[1] that fitted in with the definition of ARS given in the Introduction, irrespective of the commercial brand, to ensure that all the relevant studies on the topic were considered.

The initial search process returned 254 studies.

2.3. Screening and codification

Using the results of this initial search as our basis, a series of criteria were implemented to select the studies for consideration in the meta-analysis. These had to be studies that specifically:

1. Contained original data based on experiments done on learners (not on teachers) and that isolated and quantified the effectiveness of ARS. Any other studies that used ARS in combination with other innovative teaching tools and did not enable the individualized result of ARS use to be extracted were excluded.
2. Evaluated the effectiveness of ARS objectively using learners' examination scores as a measure of academic performance (and that had precision measurement systems in place that enabled the rigor of the findings to be clearly seen). Studies that considered other outcomes related to the learning process with ARS use, such as attitudes, perceptions, learner satisfaction, ability to retain knowledge, feedback, stimulus for peer-discussion among learners, level of engagement or motivation, opinions and any other factor that was subjective, were all rejected. In this regard, we know that scores may not capture the full process of learner learning, but we focus on exam scores because, following Mohr (2013), they are the most direct measure of academic performance, specifically when innovative strategies are applied in large sized classes.
3. Compared the scores of learners who had used ARS as a learning or assessment tool with those of learners who had not used the system or had used a different learning or assessment system. To put it another way, studies that included control groups.
4. Were published in accordance with internationally accepted quality criteria; in other words, papers published in journals indexed in the Journal Citation Reports (of both Science Citation Index and the Social Science Citation Index).
5. Were published after 2008 and up to March 2015. The year 2008 was chosen as the beginning of the time period due to the fact that Kay and LeSage (2009) stated in their exhaustive systematic review of studies on ARS that a meta-analysis was not feasible with the literature available up to said year because of the lack of a formal statistical evaluation approach in the majority of the studies.

[1] The specific keywords used were: audience-response devices, audience-response systems, audience response technology, BYOD (Bring your own device), classroom communication systems, classroom performance system, classroom response systems, clickers, electronic classroom voting systems, electronic response systems, electronic student response technology, group response systems, interactive classroom communications systems, keypad, personal response systems, personal response units, real-time polling, student response systems, voting machines and wireless course feedback systems.

In short, two other researchers in our research group independently analyzed the initial 254 papers and rejected any that did not comply with the previously established criteria, noting the reasons for their exclusion. Any issues that had arisen during the codification process were discussed and settled by consensus.

The two researchers in question codified information about these studies in keeping with the following parameters: (a) place where the experiment had taken place; (b) level of education (university or non-university); (c) field of study, following a classification based on the proposals of Becher (1989), Biglan (1973) and Lam, McNaught, Lee, and Chan (2014); (d) duration of the experiment; (e) intervention group size; (f) intervention group profile; (g) control group size; (h) experiment design; (i) statistical evaluation methods; (j) way that the results were expressed, and (k) accuracy of the results.

2.4. Sample characteristics

After applying our selection criteria, 33 papers remained to be coded, with a total of 14,963 participants in the intervention groups. The sample estimates for our meta-analysis were taken from these papers, and totaled 53 in all. Of these, 41 were independent estimates obtained from experiments performed on different samples and, in some cases, even from the same study (e.g., Kang et al., 2012 with 8 independent estimates from different samples). The remaining estimates (12) up to a total of 53 for the sample were combined estimates of dependent outcomes (59 dependent estimates in total), i.e., taken from experiments that measured the effectiveness of ARS taking into account the same groups (e.g., Christopherson, 2011 provided 6 dependent estimates that result in 1 combined estimate).

The sample size was greater than that of the other two meta-analyses cited in the preceding section on ARS, which only contained 17 (Léger et al., 2010) and 21 (Nelson et al. (2012) estimates, respectively.

Focusing on the characteristics of the studies considered in our sample, the following special features can be highlighted:

(i) by geographical area, most studies (20) were carried out in North America (19 exclusively in the United States and one for learners in the United States and Canada); followed by Europe (5 studies: 3 in Spain and 2 in the United Kingdom); Asia (4 studies, specifically in Taiwan, Singapore, China and Japan); and finally, Australia (2 studies) and Africa (2, one in South Africa and the second in Nigeria).

(ii) according to the level of education, university and post-university learners at any stage were the most frequent research populations in our sample (in 28 of the 33 studies considered), compared to lower levels of education (in the 5 remaining studies, specifically Agbatogun, 2012; Barnes, 2008; DeSorbo et al., 2013; Mun, Hew, & Cheung, 2009; and Vital, 2012).

(iii) by field of education, as previously commented, we used the codification of our selected studies according to a classification of disciplines into 2 dimensions and 4 categories inspired by the well-known Becher (1989) and Biglan (1973) approaches, developed by Lam et al. (2014):

Dimension 1: Pure Sciences
Category 1.1 Pure Hard: natural sciences and mathematics.
Category 1.2 Pure Soft: humanities and social sciences.
Dimension 2: Applied Sciences
Category 2.1 Applied Hard: science-based professions and engineering fields.
Category 2.2 Applied Soft: social professions, education, social work and law.

Following this classification our sample, grouped by discipline, is relatively balanced, with 9 studies in the pure soft category; 8 in applied hard; 6 in pure hard, and 5 in applied soft.

(iv) all the experiments consisted of using ARS in treatment groups compared to control groups that did not use the device. The type of voting method used for comparison stood out in some cases, such as raising hands (Fernández-Alemán et al., 2014); paddles (Brady, Seli, & Rosenthal, 2013); mobile polling (Sun, 2014) or holding up response cards (Desrochers & Shelnutt, 2012). Regarding the type of test, while some studies in our sample used ARS in lecture classes and tested the results in a final examination (e.g., Castillo-Manzano et al., 2015), in many the final grade was usually a result of tests done using ARS (e.g., Morling, McAuliffe, Cohen, & DiLorenzo, 2008; Patterson et al., 2010; Tlhoaele et al., 2014; and Vital, 2012). In other cases, students were awarded points for attending classes in which ARS were used (Mayer et al., 2009) or were given a grade based on their involvement, regardless of whether the answer was correct or not (FitzPatrick et al., 2011). If we analyze the question format, most of the studies in our sample used some type of multiple choice questions (e.g., Castillo-Manzano et al., 2015; Desrochers & Shelnutt, 2012; Millor et al. 2015; Morling et al. 2008; and FitzPatrick et al., 2011) or even true/false questions (Christopherson, 2011).

(v) from the perspective of the different methodologies used and the subsequent way of reporting the results, the majority of the studies considered expressed their findings in a comparison of the treatment group and the control group, basically in the form of means differences with post-data (as in e.g., Liu, Gettig, & Fjortoft, 2010 and Rothman, 2014), or pre-post data for both groups (as in e.g., FitzPatrick et al., 2011, and Lin, Liu, & Chu, 2011). In certain cases, such as in the Levesque (2011) study, where the results of an experiment were expressed in more than one way (e.g., both means differences and correlations) the decision was taken to select either the result that was most similar in methodological

terms to the other studies considered (so as to facilitate the calculation of the size of the effect), or the overall result, should one exist, that synthesized the foregoing partial results (e.g., Patterson et al., 2010).

3. Results and discussion[2]

3.1. Extracting data and calculating effect sizes

As stated above, from the 33 studies that remained after applying our inclusion/exclusion criteria, we obtained a summary of 53 estimates (in fact, our meta-analysis sample), as several of the studies provided multiple outcomes.

Using Comprehensive Meta-Analysis (CMA) software—version 3.3.070 (www.meta-analysis.com), we transformed each effect of ARS on academic scores from all individual studies into a common measure of effect size (ES). Taking into account that most of the estimates from primary studies were reported as the difference in means between the ARS treatment and the control groups, we opted to consider a typical summary measure for continuous data from the so-called '*d-family*' in terms of standardized mean difference (see Lakens, 2013). The measure that we chose to consider was the common *Hedges' g* instead of the popular *Cohen's d*. Both of these measures consist of the standardized estimate of the difference in means divided by the pooled standard deviation, although following scholars such as Adesope and Nesbit (2011), Lipsey and Wilson (2001) and Hedges and Olkin (1985), *Cohen's d* may bias effect sizes because of differential sample sizes across primary studies. As this bias can be reduced by a correction proposed by Hedges (1981), we calculated the effect sizes for our sample of estimates from this unbiased estimate or the *Hedges' g* standardized mean difference to minimize scale differences.

When estimates from primary studies were provided in the form of other statistics (e.g., point estimates, such as in Castillo-Manzano et al., 2015), the CMA software was applied to obtain equivalent transformations in term of *Hedges' g* ES (see Borenstein, Hedges, Higgins, & Rothstein, 2009).

Table 1 shows the Hedges' g and basic descriptive statistics for all the studies considered in our meta-analysis, according to Adesope et al. (2010), Adesope and Nesbit (2011) and Wecker and Fischer (2014), among many others. In our case this unbiased estimator suggested by Hedges and Olkin (1985) could be interpreted as a standardized ES mean difference between treatment-control groups, i.e., the difference in mean examination scores or achievement between learners who used ARS for any part of the teaching-learning process (e.g., as a tool for exam response or to analyze partial tests on subject content), compared to other learners who, instead of ARS, used other or traditional teaching tools (hand-raising, pencil and paper), divided by the pooled standard deviation of the two groups.

In our sample, 11 studies provided independent experiments (based on interventions over different treatment and control groups) that led to multiple and independent outcomes that we have represented with a subscript number added to the study name in Table 1 (e.g., Kang et al. 1, Kang et al. 2, Kang et al. 3, and so on). Other studies evaluated the effects of ARS using similar constructions of both treatment and control groups (e.g., the same intervention group with different control groups), resulting in dependent outcomes. In order to resolve this statistical dependence and generate a real sample of independent ES, following previous meta-analyses, such as Merchant et al. (2014) and Nesbit and Adesope (2006), we transformed these multiple dependent outcomes into a single combined estimate through a weighted average ES. This was the case for studies such as Castillo-Manzano et al. (2015), Gauci et al. (2009), Morling et al., (2008) and Vital (2012), for example. Table 1 presents in total 11 studies with independent estimates, and 12 combined estimates derived from dependent results from different studies.

Bearing in mind that positive ES would indicate a favorable impact on examination scores when learners have used ARS during the teaching-learning process, and using the definition of ES considered by Cohen (1988) to assess the obtained ES (large ES if > 0.8, moderate ES if between 0.5 and 0.8, and small ES when between 0.5 and 0.2), we can see in Table 1 that 77.35% of estimates reported a positive ES; 18.87% (10 out of 53 estimates) gave a negative ES, and only 2 estimates provided a neutral ES (Kang et al. 4, 2012 and Roberson 2, 2009). Specifically, the Hedges' g varied in a range between the value for DeSorbo et al. (2013) ($g = +35.8810$, $SE = 1.9127$), and the ES for FitzPatrick et al. 2 (2011) ($g = -1.0183$, $SE = 0.2748$).

3.2. Statistical procedure and heterogeneity diagnosis

We synthesized the overall 53 Hedges' g from Table 1 into a Summary Effect (SUE) computed by the weighted mean of ES for a 95% confidence interval of statistical significance, where, following typical meta-analytic, technical and practical procedures (see e.g., analytic expressions by Adesope et al., 2010; Borenstein, Hedges, Higgins, & Rothstein, 2010; Castillo-Manzano & Castro-Nuño, 2012; Castro-Nuño, Molina-Toucedo, & Pablo-Romero, 2013; Glass, McGaw, & Smith, 1981; Lipsey & Wilson, 2001), each ES is computed in the SUE inversely weighted by its precision (statistical weight). The resulting values are given in Table 2, where we also include the assessment of variability in ES through a heterogeneity analysis.

Following Huedo-Medina, Sánchez-Meca, Marín-Martínez, and Botella (2006) and Lipsey and Wilson (2001), we tested the assumption of homogeneity between estimates using the Q statistic following a chi-squared distribution with k-1 degrees of freedom (k = number of ES, 52 in our case). The excessive Q obtained ($Q = 705.815$) led us to reject (at level $p < .05$) the null hypothesis of homogeneity, indicating that the variability between ES was higher than could be expected from pure

[2] All the outcome graphics from our meta-analysis are available from the authors upon request.

Table 1
Descriptive statistics of overall sample of studies and estimates.

Study name [a]	No. of combined multiple dependent outcomes	Treatment group size	Education level	Category of university disciplines	Hedges' g (effect Size)(ES)	Standard error (SE)	Lower and upper confidence interval (95% level)		Z-Value	p-Value
Agbatogun, 2012		41	Non-university		1.5493	0.2816	0.9973	2.1013	5.5012	0.0000
Barnes 1, 2008		25	Non-university		0.3119	0.3053	−0.2864	0.9103	1.0217	0.3069
Barnes 2, 2008		17	Non-university		0.3325	0.3186	−0.2919	0.9569	1.0438	0.2966
Brady et al., 2013		83	University	Pure soft	0.3914	0.1606	0.0765	0.7062	2.4364	0.0148
Castillo-Manzano et al., 2015	2	119	University	Applied soft	0.4230	0.3919	−0.3452	1.1912	1.0793	0.2804
Christopherson, 2011	6	21	University	Pure soft	0.4360	0.3197	−0.1907	1.0626	1.3637	0.1727
DeSorbo et al., 2013	5	105	Non-university		35.8810	1.9127	32.1322	39.6298	18.7594	0.0000
Desrochers & Shelnutt, 2012		35	University	Pure soft	0.1142	0.2366	−0.3495	0.5779	0.4827	0.6293
Elashvili et al. 1, 2008	4	40	University	Applied hard	0.4059	0.2355	−0.0556	0.8673	1.7237	0.0848
Elashvili et al. 2, 2008	4	37	University	Applied hard	0.1359	0.2286	−0.3121	0.5839	0.5946	0.5521
Fernández-Alemán et al., 2014		58	University	Applied hard	0.5010	0.1781	0.1519	0.8502	2.8127	0.0049
FitzPatrick et al. 1, 2011		58	University	Applied hard	−0.0140	0.1853	−0.3771	0.3491	−0.0756	0.9398
FitzPatrick et al. 2, 2011	2	32	University	Applied hard	−1.0183	0.2748	−1.5570	−0.4796	−3.7050	0.0002
Gauci et al., 2009	4	169	University	Applied hard	0.7431	0.1198	0.5082	0.9780	6.2009	0.0000
Gebru, Phelps, & Wulfsberg, 2012		38	University	Pure hard	0.1009	0.2101	−0.3109	0.5127	0.4803	0.6310
Jones et al. 1, 2013		24	University	Pure soft	0.1245	0.2815	−0.4272	0.6762	0.4423	0.6583
Jones et al. 2, 2013		21	University	Pure soft	0.7520	0.3260	0.1131	1.3909	2.3070	0.0211
Kang et al. 1, 2012		1540	University	Pure hard	0.1829	0.0435	0.0976	0.2683	4.2007	0.0000
Kang et al. 2, 2012		1238	University	Pure hard	0.2555	0.0512	0.1550	0.3559	4.9855	0.0000
Kang et al. 3, 2012		549	University	Pure hard	0.0895	0.0742	−0.0559	0.2350	1.2064	0.2277
Kang et al. 4, 2012		850	University	Pure hard	0.0000	0.0529	−0.1037	0.1038	0.0008	0.9993
Kang et al. 5, 2012		1675	University	Pure hard	−0.2707	0.0454	−0.3598	−0.1817	−5.9583	0.0000
Kang et al. 6, 2012		1726	University	Pure hard	−0.0303	0.0472	−0.1228	0.0623	−0.6407	0.5217
Kang et al. 7, 2012		1538	University	Pure hard	0.0180	0.0427	−0.0656	0.1016	0.4218	0.6732
Kang et al. 8, 2012		1840	University	Pure hard	0.0756	0.0437	−0.0101	0.1612	1.7296	0.0837
King & Robinson, 2009		145	University	Pure hard	0.0645	0.1441	−0.2179	0.3469	0.4476	0.6545
Levesque 1, 2011		30	University	Pure hard	0.7538	0.2530	0.2580	1.2496	2.9800	0.0029
Levesque 2, 2011		33	University	Pure hard	0.8089	0.2449	0.3289	1.2890	3.3026	0.0010
Lin et al., 2011	2	50	University	Pure hard	−0.2702	0.1563	−0.5766	0.0362	−1.7282	0.0839
Liu et al., 2010	2	88	University	Applied hard	0.1687	0.1494	−0.1241	0.4615	1.1291	0.2588
Mayer et al., 2009		111	University	Pure soft	0.3829	0.1289	0.1303	0.6355	2.9708	0.0030
Millor et al., 2015		87	University	Applied hard	0.3921	0.1520	0.0942	0.6899	2.5797	0.0099
Morling et al., 2008	2	560	University	Pure soft	0.1949	0.0625	0.0724	0.3173	3.1197	0.0018
Mun et al., 2009		35	Non-university		0.8493	0.2471	0.3651	1.3335	3.4376	0.0006
Patterson et al., 2010		38	University	Applied soft	0.2280	0.2381	−0.2386	0.6946	0.9578	0.3382
Roberson 1, 2009		110	University	Applied soft	0.0779	0.2048	−0.3234	0.4793	0.3806	0.7035
Roberson 2, 2009		112	University	Applied soft	0.0000	0.1994	−0.3909	0.3909	0.0000	1.0000
Rothman, 2014		235	University	Pure soft	−0.1560	0.0796	−0.3120	−0.0001	−1.9607	0.0499
Shaffer & Collura, 2009		42	University	Pure soft	0.4577	0.2103	0.0455	0.8698	2.1765	0.0295
Stowell et al.1, 2010		10	University	Pure soft	0.7350	0.3295	0.0891	1.3808	2.2305	0.0257
Stowell et al.2, 2010		20	University	Pure soft	0.8033	0.2437	0.3257	1.2810	3.2962	0.0010
Stowell et al.3, 2010		40	University	Pure soft	0.6041	0.1833	0.2448	0.9634	3.2955	0.0010
Sun 1, 2014		13	University	Applied soft	−0.0057	0.3735	−0.7377	0.7264	−0.0152	0.9879
Sun 2, 2014		19	University	Applied soft	−1.0058	0.3234	−1.6396	−0.3721	−3.1106	0.0019
Tlhoaele et al., 2014		36	University	Applied hard	2.5710	0.3189	1.9460	3.1960	8.0630	0.0000
Tregonning, Doherty, Hornbuckle, & Dickinson, 2012		106	University	Applied hard	0.2335	0.1367	−0.0344	0.5014	1.7081	0.0876
Vana et al., 2011		78	University	Applied soft	0.1254	0.1752	−0.2180	0.4689	0.7157	0.4742
Vital 1, 2012	13	66	Non-university		−0.0926	0.3206	−0.7210	0.5357	−0.2890	0.7726
Vital 2, 2012	13	45	Non-university		−0.1759	0.3321	−0.8268	0.4751	−0.5295	0.5965
		366	University	Pure hard	0.0302	0.0786	−0.1239	0.1843	0.3843	0.7007

(*continued on next page*)

Table 1 (*continued*)

Study name [a]	No. of combined multiple dependent outcomes	Treatment group size	Education level	Category of university disciplines	Hedges' g (effect Size)(ES)	Standard error (SE)	Lower and upper confidence interval (95% level)		Z-Value	p-Value
Voelkel and Bennett 1, 2014										
Voelkel and Bennett 2, 2014		326	University	Pure hard	0.0383	0.0796	−0.1177	0.1942	0.4810	0.6305
Voelkel and Bennett 3, 2014		206	University	Pure hard	0.4968	0.1087	0.2837	0.7098	4.5693	0.0000
Voelkel and Bennett 4, 2014		77	UniversitY	Pure hard	0.5788	0.1573	0.2704	0.8871	3.6783	0.0002

Note: a = superscript numbers indicate independent estimates of each study.
*, ** and *** = statistical significance at 0.1, 0.05 and 0.01, respectively.

randomness (*within-study variance* due to sampling error). Because of the low power of this test, highlighted by Takkouche, Cadarso-Suárez, and Spiegelman (1999), we introduced as a complement the I^2 statistic proposed by Higgins, Thompson, Deeks, and Altman (2003), which indicates the proportion of the variation between studies (*between-studies variance*) in the total variation; that is, the proportion of total variation due to heterogeneity, with $I^2 < 0.25$ representing low heterogeneity, $I^2 < 0.50$ moderate heterogeneity, and $I^2 > 0.75$ high heterogeneity.

Table 2 reveals an $I^2 = 0.92633$ for the overall sample of estimates. Thus, according to Borenstein et al. (2009), estimated ES of the included estimates are only a random sample of all those possible and do not estimate a common population mean, but differ from each other. Therefore, we included in Table 2 SUE pooled both according to the *Fixed Effects Model*, or FEM, (where the inference is made by the *inverse variance weighted method*, under the premise that there is homogeneity among ES and the only determinant of the weight of each study would be its own variance), and the *Random Effects Model*, or REM, (which considers that the ES are only a random sample of all those possible and that there are two sources of variance within the studies or random error, and between studies or true dispersion) (see Castro-Nuño et al., 2013 for equations). Taking into account the variability detected by the above heterogeneity analysis, and considering the great differences between experiments and interventions with ARS implemented in primary studies, REM may be a more accurate technique (see Borenstein et al., 2009; 2010) with the weighted average of the ES used to minimize any possible bias caused by the considerable variance in the ES in our estimate sample (Consorti et al., 2012).

Thus, the resulting Hedges' g using REM, applying the *variance-weighted method*, in Table 2 was g = +0.288 (SE = 0.058; C.I. = 0.175–0.401), evidencing a greater value than Hedges' g for FEM (g = 0.099) (both $p < .001$), but nevertheless, a positive moderate-small effect of ARS on examination scores.

With the goal of achieving a more homogenous analysis, we used the CMA software to search for potential outlier studies, removing one to one estimates from the initial sample and recalculating the SUE and meta-analysis outcomes. After examining our sample, we saw that, as stated above, the study by DeSorbo et al. (2013) clearly presented an overstated upper ES than the remaining sample of estimates (nearly 35 times greater), so it was considered to be an outlier and removed from the distribution. Said study presented special features, as the subject being assessed to analyze the effectiveness of ARS in school pupils of 9–11 years of age was not academic, but consisted of health education programs. Thus, the ES for the remaining sample of 52 estimates ranged from g = 2.571 (SE = 0.319) for Tlhoaele et al. (2014) to g = −1.018 (SE = 0.275) for FitzPatrick et al. 2 (2011). Although the SUE (by REM) varied from g = 0.240 (SE = 0.043, C.I. at 95% = 0.1548–0.3245) to (by FEM) g = 0.097 (SE = 0.014, C.I. at 95% = 0.070–0.124), the recalculated results showed that the hypothesis of homogeneity had again been rejected and did not produce a more homogeneous solution ($Q(51) = 355.8172$; $p < .001$; $I^2 = 85.667\%$).

3.3. Moderator variables analysis and discussion

Considering the categorization and coding variables for our filtered studies already mentioned in Section 2, we explored the variability for the new sample of estimates (excluding DeSorbo et al., 2013). Specifically, we conducted a new meta-analysis combining the estimates within subgroups and, following Adesope et al. (2010) and Borenstein et al. (2009), implementing a *Mixed-Effects Model* (MEM) for a moderator variable analysis. MEM summarizes ES for each subgroup using

Table 2
Summary effect (SUE) and meta-analysis outcomes for overall estimates.

Meta-analysis I	No. of estimates	Fixed Hedges' g/Z-value	Random Hedges' g/Z-value	Heterogeneity measures		
				Q value	I^2 (%)	Tau^2
Overall estimates	53	0.099/7.127***	0.288/4.980***	705.815***		0.134

*, ** and *** = statistical significance at 0.1, 0.05 and 0.01, respectively.

REM and calculates the difference between the subgroups with FEM. The results for this second meta-analysis framework are given in Table 3.

First, we conducted a meta-analysis for a categorization based on a first moderator: *the level of education*, separating studies which considered the effects of ARS on learners' scores at university level (with a subsample of 46 estimates) versus other intervention groups with ARS in the non-university context (with a subsample of 6 estimates). Comparatively, the impact of ARS on examination marks is positive but still moderate-small for both scenarios, although clearly greater for non-university studies (under FEM, g = 0.5496, SE = 0.1209; under REM, g = 0.4801, SE = 0.266) than for university experiments (under FEM, g = 0.0909, SE = 0.0140; under REM, g = 0.2156, SE = 0.0434). In short, our results highlight a differential assessment of ARS depending on the education level, in the sense that an inverse effect of ARS on examination performance was found with more advanced education levels.

From the homogeneity point of view, a higher I^2 statistic was observed for the university group ($I^2 = 0.8583$) than the non-university group ($I^2 = 0.7910$); i.e., for university studies, 85.83% of the variance was due to between-studies variance (with a tau-square of 0.0569). Therefore, we decided to focus on the specific effect of the *university discipline categorization* on the SUE, and a second moderator analysis was conducted by MEM to further determine if the academic field might explain the variability in ES within the university group. As can be seen in Table 3, we used the codification into 2 dimensions and 4 categories already explained in Section 2.

Table 3 shows that the Hedges' g resulting from REM are more appropriate for our categorizations of academic university disciplines, insomuch as the heterogeneity analysis of studies was still moderate-high, as shown by a wide dispersion of the selected studies, especially the Hard categories within both dimensions. Resulting SUE were positive in all cases except for Applied Soft Sciences (with a g = −0.0018, SE = 0.1366), although this average effect was statistically non-significant. When SUE was compared in the remaining science dimensions, we noted that it was rather weak for the Pure Hard disciplines (g = 0.1180, SE = 0.4999, p < .01), while both Pure Soft and Applied Hard sciences showed a clearly positive and statistically significant pooled effect of ARS on examination scores greater than the effect seen in meta-analysis I (overall estimates, Table 2).

Beyond variations in teaching approaches across disciplines evidenced by authors such as Lindblom-Ylänne, Trigwell, Nevgi, and Ashwin (2006), we particularly highlight the difference found between the two categories of Pure sciences, with a higher effect of ARS on learners of the Pure Soft sciences type. This may be supported by the hypothesis introduced by Laird, Shoup, Kuh, and Schwarz (2008) to the effect that learners from Soft disciplines are more likely to be amenable to active learning, which, as is known, is one of the characteristic features of ARS use.

In this line, the greatest effectiveness of ARS for the Applied Hard disciplines (with a g = 0.3921, SE = 0.1808, p < .05) might be supported by the study by Neumann (2001), who claimed that multiple choice questions (the usual tool for ARS implementation) are more likely favored in applied than in pure fields.

3.4. Assessment of publication bias

Using CMA software and following the methodological indications of Borenstein et al. (2009), an analysis of publication bias based on quantitative measures was carried out in our study to determine whether a potential bias could modify the validity and robustness of our findings in different scenarios.

Statistical assessment for three common tests of publication bias is shown in Table 4: the so-called 'classic fail-safe *N*′ test that determines the number of studies that might have been missed or not included that would nullify the SUE found; the

Table 3
Summary effect (SUE) and meta-analysis outcomes for moderator analysis excluding the DeSorbo et al. (2013) outlier.

Meta-analysis II	Scenario I: Education level		Scenario II: Categories of disciplines at *university level*			
			Scenario II.A. Pure sciences		Scenario II.B. Applied sciences	
	University	Non-university	Hard	Soft	Hard	Soft
No. Estimates	46	6	17	12	10	7
Fixed effects Hedges' g	0.0909	0.5496	0.0522	0.1982	0.3790	0.0250
Z-value	6.5129	4.544	3.3154	5.0193	6.9071	0.2765
p-value	0.0000***	0.0000***	0.0009***	0.0000***	0.0000***	0.7821
Standard Error (SE)	0.0140	0.1209	0.0157	0.0395	0.0549	0.0904
Random effects Hedges' g	0.2156	0.4801	0.1180	0.3520	0.3921	−0.0018
Z-value	4.9634	1.8017	2.3637	3.7449	2.1690	−0.0131
p-value	0.0000***	0.0716*	0.0181**	0.0002***	0.0301**	0.9895
Standard Error (SE)	0.0434	0.2665	0.0499	0.0940	0.1808	0.1366
Heterogeneity analysis						
Q test	317.6924***	23.9266***	130.0387***	42.1937***	91.5698***	12.3390*
I^2 (%)	85.8354	79.1027	87.6960	73.9297	90.1714	51.3735
Tau^2	0.0569	0.3354	0.0314	0.0639	0.2859	0.0634

*, ** and *** = statistical significance at 0.1, 0.05 and 0.01, respectively.

'Egger's linear regression' test that compares the size of the treatment effect through the slope of a regression line (B_1) with the potential bias represented by intercept (B_0); and the 'Duval and Tweedie trim and fill' method (see Duval & Tweedie, 2000) that, under a REM, detects missing studies to the left/right of the mean effect and recomputes the combined effect.

As Table 4 shows, except in the case of the meta-analysis II scenario, moderator variable category discipline Applied Soft (which, as stated above, is not statistically significant), these tests would suggest the absence of a relevant publication bias that might significantly undermine the validity of the current meta-analyses.

4. Conclusions

Technological advances have overcome the main barrier to the widespread introduction of ARS into classrooms. Wi-Fi on the majority of campuses in a large number of countries, combined with the mass consumption of smartphones by young people, means that implementing ARS is not a utopia but a reasonable scenario.

In this new context, it makes sense to synthesize quantitatively all recent evaluations of the impact of these devices on academic performance. Our paper specifically offers the most complete meta-analysis to date of studies of the effectiveness of audience response systems at all stages of education. From 254 initial studies, our meta-analysis focuses on a sample of 51 papers that conform to both the formal (e.g., the reputation of the journal in which the results are published) and the underlying (e.g., studies that apply a rigorous methodology and carry out the experiment with a control group) academic quality criteria set.

The high degree of heterogeneity found for the sample as a whole (see Table 2) shows that the effect of ARS on academic performance seems to be moderated by the characteristics of the different combinations of subjects, especially by the educational environment into which ARS are introduced. A clearly significant effect (at 1% and positive) can be seen in both university and non-university studies. However, the effect is evidently much greater in experiments performed in non-university contexts, with a 0.48 Hedges' g statistic that is very close to the 0.5 threshold that would allow us to speak of an effect in the Cohen scale "middle band" (see Table 3). On the other hand, the 0.22 value for the university level statistic (see Table 3) implies an effect which, despite being highly significant, is not so great and is marked by the high heterogeneity of the combination of subjects.

If a second moderating variable is introduced based on the type of scientific discipline, more specific results are obtained by field of knowledge. In short, although our meta-analysis reports the favorable but limited effectiveness of ARS on academic scores compared to other traditional formats of teaching-learning, our findings also provide evidence that suggest that certain underlying factors may be affecting these modest results. Specifically, the educational context and the category of discipline in which ARS interventions are implemented seem to influence their usefulness for achieving better academic marks. Moreover, these systems seem to be more effective for lower levels of education and, within the university context, when both Pure Soft Sciences (such as anthropology, sociology, psychology, etc.) and Applied Hard Sciences (mainly medicine and engineering) are considered.

To summarize, this variability in the effects of ARS on academic performance suggests that any institutions that wish to introduce the system into its teaching should behave prudently. Its introduction should be limited at first and followed up with an assessment of the effect that the system's use has, and only if the assessment is positive should this teaching tool be rolled out universally. Special attention should be paid to test and test content design when assessing the effectiveness of ARS on academic performance in order to ensure that the final grade is really due to the use of this new technology and not to other factors that affect subject content and pedagogical methodology. For example, the simplification effect should be excluded that results from converting subject content into immediate response questions and adapting the teaching staff's teaching model to focus more on favoring student interactivity, and concepts that have not been fully grasped by students should be reinforced.

Some of these results may well appear disappointing at first, but if we return to the current context in which the cost of using ARS in teaching is much lower than in the past, in general terms the introduction of the system would present a clearly

Table 4
Publication Bias Analysis for all scenarios of meta-analysis performed.

Publication bias measures	Meta-analysis I: Overall estimates	Meta-analysis II: Education level moderator variable		Meta-analysis II: University disciplines moderator variable			
		University	Non-university	Pure hard	Pure soft	Applied hard	Applied soft
Classic fail-safe N (No. of missing studies that would raise p-value>.05)	2.184	1.012	212	43	116	106	0
Egger' s regression intercept (B_0)	2.511 (SE = 0.736)	1.713 (SE = 0.607)	18.864 (SE = 4.251)	2.255 (SE = 1.403)	2.185 (SE = 0.863)	0.073 (SE = 3.833)	−1.290 (SE = 2.061)
Duval and Tweedie Trim and Fill (No. of missed studies under REM)	0 left of the mean 16 right of the mean	0 left of the mean 2 right of the mean	10 left of the mean 0 right of the mean	0 left of the mean 0 right of the mean	3 left of the mean 0 right of the mean	0 left of the mean 3 right of the mean	2 left of the mean 0 right of the mean

positive social cost-benefit analysis for the academic institution in question. Therefore, ARS should be expected to have a promising future.

Acknowledgements

The authors would like to express their gratitude to the University of Seville for its support through the University's Second Own Teaching Plan Aid Program.

References

Addison, S., Wright, A., & Milner, R. (2009). Using clickers to improve student engagement and performance in an introductory biochemistry class. *Biochemistry and Molecular Biology Education, 37*(2), 84–91.
Adesope, O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*(2), 207–245.
Adesope, O., & Nesbit, J. (2011). Verbal redundancy in multimedia learning environments: a meta-analysis. *Journal of Educational Psychology, 104*(1), 250–263.
Agbatogun, A. O. (2012). Exploring the efficacy of student response system in a Sub-Saharan African country: a sociocultural perspective. *Journal of Information Technology Education: Research, 11*, 249–267.
Anthis, K. (2011). Is it the clicker, or is it the question? Untangling the effects of student response system use. *Teaching of Psychology, 38*(3), 189–193.
Archer, K., Savage, R., Sanghera-Sidhu, S., Wood, S., Gottardo, A., & Chen, V. (2013). Examining the effectiveness of technology use in classrooms: a tertiary meta-analysis. *Computers and Education, 78*, 140–149.
Arneja, J. S., Narasimhan, K., Bouwman, D., & Bridge, P. D. (2009). Qualitative and quantitative outcomes of audience response systems as an educational tool in a plastic surgery residency program. *Plastic and Reconstructive Surgery Journal*, 2179–2184. December.
Arnesen, K., Sivertsen Korpås, G. S., Hennissen, J. E., & Birger, S. J. (2013). Experiences with use of various pedagogical methods utilizing a student response system – motivation and learning outcome. *The Electronic Journal of e-Learning, 11*(3), 169–181.
Barnes, L. J. (2008). Lecture-free high school biology using an audience response system. *The American Biology Teacher, 70*(9), 531–536.
Beatty, I., & Gerace, W. (2009). Technology-enhanced formative assessment: a research-based pedagogy for teaching science with classroom response technology. *Journal Science Education Technology, 18*, 146–162.
Becher, T. (1989). *Academic tribes and territories*. Buckingham: SRHE and Open University Press.
Berry, J. (2009). Technology supports in nursing education: clickers in the classroom. *Nursing Education Perspective ProQuest Health and Medical Complete, 30*(5), 295–298.
Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology, 57*(3), 195–203.
Blasco-Arcas, L., Buil, I., Hernández-Ortega, B., & Sese, F. J. (2013). Using clickers in class. The role of interactivity, active collaborative learning and engagement in learning performance. *Computers and Education, 62*, 102–110.
Bojinova, E., & Oigara, J. (2013). Teaching and learning with clickers in higher education. *International Journal of Teaching and Learning in Higher Education, 25*(2), 154–165.
Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex: John Wiley and Sons, Ltd.
Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*, 97–111.
Brady, M., Seli, H., & Rosenthal, J. (2013). "Clickers" and metacognition: a quasi-experimental comparative study about metacognitive self-regulation and use of electronic feedback devices. *Computers and Education, 65*, 56–63.
Bright, D. R., Reilly Kroustos, K., & Kinder, D. H. (2013). Audience response systems during case-based discussions: a pilot study of student perceptions. *Currents in Pharmacy Teaching and Learning, 5*(5), 410–416.
Bryfczynski, S. P., Brown, R., Hester, J., Herrmann, A., Koch, D. L., Cooper, M. M., et al. (2014). uRespond: iPad as interactive, personal response system. *Journal of Chemical Education, 91*, 357–366.
Burnstein, R., & Lederman, L. (2001). Using wireless keypads in lecture classes. *Physics Teacher, 39*, 8–11.
Cain, J., Black, E. P., & Rohr, J. (2009). An audience response system strategy to improve student motivation, attention, and feedback. *American Journal of Pharmaceutical Education, 73*(2). Article 21.
Caird, S., & Lane, A. (2015). Conceptualising the role of information and communication technologies in the design of higher education teaching models used in the UK. *British Journal of Educational Technology, 46*(1), 58–70.
Caldwell, J. E. (2007). Clickers in the large classroom: current research and best-practice tips. *CBE- Life Sciences Education, 6*(1), 9–20.
Castillo-Manzano, J. I., Castro-Nuño, M., Sanz Díaz, M. T., & Yñiguez, R. (2015). Does pressing a button make it easier to pass an exam? Evaluating the effectiveness of interactive technologies in higher education. *British Journal of Educational Technology*. http://dx.doi.org/10.1111/bjet.12258.
Castillo-Manzano, J. I., & Castro-Nuño, M. (2012). Driving licenses based on points systems: efficient road safety strategy or latest fashion in global transport policy? A worldwide meta-analysis. *Transport Policy, 21*, 191–201.
Castro-Nuño, M., Molina-Toucedo, J. A., & Pablo-Romero, M. P. (2013). Tourism and GDP: a meta-analysis of panel data studies. *Journal of Travel Research, 52*(6), 745–758.
Cheong, Ch, Bruno, V., & Cheong, F. (2012). Designing a Mobile-app-based collaborative learning system. *Journal of Information Technology Education: Innovations in Practice, 11*(1), 94–119.
Christopherson, K. M. (2011). Hardware or wetware: what are the possible interactions of pedagogy and technology in the classroom? *Teaching of Psychology, 38*(4), 288–292.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Conoley, J. W., Croom, D. B., Moore, G. E., & Flowers, J. L. (2007). Using electronic audience response systems in high school agriscience courses. *Journal of Agricultural Education, 48*(3), 67–77.
Consorti, F., Mancuso, R., Nocioni, M., & Piccolo, A. (2012). Efficacy of virtual patients in medical education: a meta-analysis of randomized studies. *Computers and Education, 59*, 1001–1008.
Costello, P. (2010). A cost-effective classroom response system. *British Journal of Educational Technology, 41*(6), 153–154.
Cue, N. (1998). A universal learning tool for classrooms?. In *Paper presented in the first quality in teaching and learning conference (Hong Kong, SAR, China, December 10-12, 1998)*.
DeBourgh, G. (2008). Use of classroom "clickers" to promote acquisition of advanced reasoning skills. *Nurse Education in Practice, 8*, 76–87.
Denker, K. J. (2013). Student response systems and facilitating the large lecture basic communication course: assessing engagement and learning. *Communication Teacher, 27*(1), 50–69.
DeSorbo, A. L., Noble, J. M., Shaffer, M., Gerin, W., & Williams, O. A. (2013). The use of an audience response system in an elementary School–Based health education program. *Health Education and Behavior, 40*(5), 531–535.
Desrochers, M. N., & Shelnutt, J. M. (2012). Effect of answer format and review method on college students learning. *Computers and Education, 59*, 946–951.
Draper, S., & Brownw, M. (2004). Increasing interactivity in lectures using an electronic voting system. *Journal of Computer Assisted Learning, 20*, 81–94.

Duncan, D. (2007). Clickers: a new teaching aid with exceptional promise. *Astronomy Education Review, 5*(1), 70–88.
Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463.
Elashvili, A., Denehy, G. E., Dawson, D. V., & Cunningham, M. A. (2008). Evaluation of an audience response system in a preclinical operative dentistry course. *Journal of Dental Education, 72*(11), 1296–1303.
Elicker, J., & McConnell, N. (2011). Interactive learning in the classroom: Is student response method related to performance? *Teaching of Psychology, 38*(3), 147–150.
Fernández-Alemán, J. L., Sánchez, A. B., López, M. J., & López, J. J. (2014). Examining the benefits of learning based on an audience response system when confronting emergency situations. *CIN: Computers, Informatics, Nursing, 32*(5), 207–213.
Fies, C., & Marshall, J. (2006). Classroom response systems: a review of the literature. *Journal of Science Education and Technology, 15*(1), 101–109.
FitzPatrick, K. A., Finn, K. E., & Campisi, J. (2011). Effect of personal response systems on student perception and academic performance in courses in a health sciences curriculum. *Advances in Physiology Education, 35*, 280–289.
Fortner-Wood, Ch, Armistead, L., Marchand, A., & Morris, F. (2013). The effects of student response systems on student learning and attitudes in undergraduate psychology courses. *Teaching of Psychology, 40*(1), 26–30.
Freeman, S., Haak, D., & Wenderoth, M. (2011). Increased course structure improves performance in introductory biology. *CBE Life Science Education, 10*(2), 175–186.
Gachago, D., Morris, A., & Simon, E. (2011). Engagement levels in a graphic design clicker class: students' perceptions around attention, participation and peer learning. *Journal of Information Technology Education, 10*, 254–269.
Gauci, S., Dantas, A., Williams, D., & Kemm, R. (2009). Promoting student-centered active learning in lectures with a personal response system. *Advances in Physiology Education, 33*, 60–71.
Gebru, M. T., Phelps, A. J., & Wulfsberg, G. (2012). Effect of clickers versus online homework on students' long-term retention of general chemistry course material. *Chemistry Education Research and Practice, 13*, 325–329.
Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverley Hills, CA: Sage.
Gok, T. (2011). An evaluation of student response systems from the viewpoint of instructors and students. *The Turkish Online Journal of Educational Technology, 10*(4), 67–83.
Guse, D. M., & Zobitz, P. M. (2011). Validation of the audience response system. *British Journal of Educational Technology, 42*(6), 985–991.
Han, J. H. (2014). Closing the missing links and opening the relationships among the factors: a literature review on the use of clicker technology using the 3P model. *Educational Technology and Society, 17*(4), 150–168.
Hecht, S., Adams, W. H., Cunningham, M. A., Lane, I. F., & Howell, N. E. (2013). Student performance and course evaluations before and after use of the classroom performance system™ in a third-year veterinary radiology course. *Veterinary Radiology Ultrasound, 54*(2), 114–121.
Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 107–128.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557–560.
Hill, A., & Babbit, B. (2013). Examining the efficacy of personal response devices in army training. *Journal of Information Technology Education: Innovations in Practice, 12*, 1–12.
Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index? *Psychological Methods, 11*(2), 193–206.
Jones, S. J., Crandall, J., Vogler, J. S., & Robinson, D. H. (2013). Classroom response systems facilitate student accountability, readiness, and learning. *Journal of Educational Computing Research, 49*(2), 155–171.
Judson, E., & Sawada, D. (2002). Learning from past and present: electronic response systems in college lecture halls. *Journal of Computers in Mathematics and Science Teaching, 21*(2), 167–181.
Kang, H., Lundeberg, M., Wolter, B., DelMas, R., & Herreid, C. (2012). Gender differences in student performance in large lecture classrooms using personal response systems ('clickers') with narrative case studies. Learning. *Media and Technology, 37*(1), 53–76.
Kay, R. H. (2009). Examining gender differences in attitudes toward interactive classroom communications systems (ICCS). *Computers and Education, 52*, 730–740.
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.
Keough, S. M. (2012). Clickers in the Classroom: a review and a Replication. *Journal of Management Education*, 1–26. http://dx.doi.org/10.1177/1052562912454808.
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.
Laird, T. F. N., Shoup, R., Kuh, G. D., & Schwarz, M. J. (2008). The effects of discipline on deep approaches to student learning and college outcomes. *Research in Higher Education, 49*(6), 469–494.
Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 1–12.
Lam, P., McNaught, C., Lee, J., & Chan, M. (2014). Disciplinary difference in students' use of technology, experience in using eLearning strategies and perceptions towards eLearning. *Computers and Education, 73*, 111–120.
Lantz, M. E. (2010). The use of "Clickers" in the classroom: teaching innovation or merely an amusing novelty? *Computers in Human Behavior, 26*, 556–561.
Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). A system for knowledge discovery in e-learning environments within the European higher education area – application to student data from Open University of Madrid, UDIMA. *Computers and Education, 72*, 23–36.
Lattesa, R., & Mouw, D. (2005). Use of an audience response system to augment interactive learning. *Family Medicine, 37*(1), 12–14.
Léger, M., Bourque, J., & Richard, J. (2010). Influence des télévoteurs sur le résultat scolaire: une méta-analyse. *International Journal of Technologies in Higher Education, 7*(2), 35–47.
Levesque, A. (2011). Using clickers to facilitate development of problem-solving skills. *CBE—Life Sciences Education, 10*, 406–417.
Liao, Y.-K. (2007). Effects of computer-assisted instruction on students' achievement in Taiwan: a meta-analysis. *Computers and Education, 48*, 216–233.
Lindblom-Ylänne, S., Trigwell, K., Nevgi, A., & Ashwin, P. (2006). How approaches to teaching are affected by discipline and teaching context. *Studies in Higher Education, 31*(3), 285–298.
Lin, Y.-C., Liu, T.-C., & Chu, C.-C. (2011). Implementing clickers to assist learning in science lectures: the Clicker-Assisted conceptual change model. *Australian Journal of Educational Technology, 27*(6), 979–996.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
Liu, F. C., Gettig, J. P., & Fjortoft, N. (2010). Impact of a student response system on short- and long-term learning in a drug literature evaluation course. *American Journal of Pharmaceutical Education, 74*(1), 1–5.
Ma, W., Adesope, O., Nesbit, J., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: a meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918.
Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., et al. (2009). Clickers in college classrooms: fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology, 34*, 51–57.
Merchant, Z., Goetz, E., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: a meta-analysis. *Computers and Education, 70*, 29–40.

Millor, M., Etxano, J., Slon, P., García-Barquín, P., Villanueva, A., Bastarrika, G., et al. (2015). Use of remote response devices: an effective interactive method in the long- term learning. *European Radiology, 25*, 894–900.
Mohr, T. M. (2013). iClickers and student performance. *International Review of Economics Education, 14*, 16–23.
Morling, B., McAuliffe, M., Cohen, L., & DiLorenzo, T. M. (2008). Efficacy of personal response systems ("Clickers") in large, introductory psychology classes. *Teaching of Psychology, 35*, 45–50.
Mostyn, A., Meade, O., & Lymn, J. (2012). Using audience response technology to provide formative feedback on pharmacology performance for non-medical prescribing students – a preliminary evaluation. *British Medical Education, 12*(113), 2–8.
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*(5), 848–860.
Nelson, C., Hartling, L., Campbell, S., & Oswald, A. (2012). The effects of audience response systems on learning outcomes in health professions education. A BEME systematic review, guide no. 2. *Medical Teacher, 34*, e386–e405.
Nesbit, J., & Adesope, O. (2006). Learning with concept and knowledge maps: a meta-analysis. *Review of Educational Research, 76*, 413–448.
Neumann, R. (2001). Disciplinary differences and university teaching. *Studies in Higher Education, 26*(2), 135–146.
Nicol, D. J., & Boyle, J. T. (2003). Peer instruction versus class-wide discussion in large classes: a comparison of two interaction methods in the wired classroom. *Studies in Higher Education, 28*(4), 457–473.
Padilla-Meléndez, A., Águila-Obra, R., & Garrido-Moreno, A. (2015). Using moodle in teaching-learning processes in business management: the new profile of EHEA student. *Educación XXI, 18*(1), 125–145.
Patterson, B., Kilpatrick, J., & Woebkenberg, E. (2010). Evidence for teaching practice: the impact of clickers in a large classroom environment. *Nurse Education Today, 30*, 603–607.
Preszler, R. W., Dawe, A., Shuster, C. B., & 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*, 29–41.
Roberson, D. (2009). Using a student response system to reduce academic cheating. *Nurse Educator, 34*(2), 60–63.
Rothman, S. (2014). A study of twitter and clickers as audience response systems in international relations courses. *PS: Political Science and Politics, 47*(3), 698–702.
Schackow, T., Chavez, M., Loya, L., & Friedman, M. (2004). Audience response system: effect on learning in family medicine residents. *Family Medicine, 36*(7), 496–504.
Schmid, R., Bernard, R., Borokhovski, E., Tamim, R., Abrami, P., Surkes, M., et al. (2014). The effects of technology use in postsecondary education: a meta-analysis of classroom applications. *Computers and Education, 72*, 271–291.
Shaffer, D. M., & Collura, M. J. (2009). Evaluating the effectiveness of a personal response system in the classroom. *Teaching of Psychology, 36*, 273–277.
Siau, K., Sheng, H., & Nah, F. (2006). Use of a classroom response system to enhance classroom interactivity. *IEEE Transactions on Education, 49*(3), 398–403.
Simpson, V., & 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.
Smith, D. A., & Rosenkoetter, M. M. (2009). Effectiveness, challenges, and perceptions of classroom participation system. *Nurse Educator, 34*, 156–161.
Sprague, E. W., & Dahl, D. W. (2010). Learning to click: an evaluation of the personal response system clicker technology in introductory marketing courses. *Journal of Marketing Education, 32*(1), 93–103.
Stav, J., Nielsen, K., Hansen-Nygård, G., & Thorseth, T. (2010). Experiences obtained with integration of student response systems for iPod touch and iPhone into e-learning environments. *Electronic Journal of e-Learning, 8*(2), 179–190.
Stowell, J., & Nelson, J. M. (2007). Benefits of electronic audience response systems on student participation, learning, and emotion. *Teaching of Psychology, 34*(4), 253–258.
Stowell, J. R., Oldham, T., & Bennett, D. (2010). Using student response systems ("Clickers") to combat conformity and shyness. *Teaching of Psychology, 37*, 135–140.
Sun, J. C.-Y. (2014). Influence of polling technologies on student engagement: an analysis of student motivation, academic performance, and brainwave data. *Computers and Education, 72*, 80–89.
Sutherlin, A. L., Sutherlin, G. R., & Akpanudo, U. M. (2013). The effect of clickers in university science courses. *Journal Science Education Technology, 22*, 651–666.
Takkouche, B., Cadarso-Suárez, C., & Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology, 150*, 206–215.
Tamim, R., Bernard, R., Borokhovski, E., Abrami, P., & Schmid, R. (2011). What forty years of research says about the impact of technology on learning: a second order meta-analysis and validation study. *Review of Educational Research, 81*(3), 4–28.
Tlhoaele, M., Hofman, A., Naidoo, A., & Winnips, K. (2014). Using clickers to facilitate interactive engagement activities in a lecture room for improved performance by students. *Innovations in Education and Teaching International, 51*(5), 497–509.
Tregonning, A. M., Doherty, D. A., Hornbuckle, J., & Dickinson, J. E. (2012). The audience response system and knowledge gain: a prospective study. *Medical Teacher, 34*, 269–274.
Vana, K. D., Silva, G. E., Muzyka, D., & Hirani, L. M. (2011). Effectiveness of an audience response system in teaching pharmacology to baccalaureate nursing students. Computers. *Nursing Informatics, 29*(6), 326–334.
Vital, F. (2012). Creating a positive learning environment with the use of clickers in a high school chemistry classroom. *Journal of chemical Education, 89*, 470–473.
Voelkel, S., & Bennett, D. (2014). New uses for a familiar technology: introducing mobile phone polling in large classes. *Innovations in Education and Teaching International, 51*, 46–58.
Wash, P. D. (2012). The power of a mouse! *SRATE Journal, 21*(2), 39–46.
Webster, T. E., & Son, J. B. (2015). Doing what works: a grounded theory case study of technology use by teachers of English at a Korean university. *Computers and Education, 80*, 84–94.
Wecker, Ch, & Fischer, F. (2014). Where is the evidence? A meta-analysis on the role of argumentation for the acquisition of domain-specific knowledge in computer-supported collaborative learning. *Computers and Education, 75*, 218–228.
White, P., Syncox, D., & Alters, B. (2011). .Clicking for grades? Really? Investigating the use of clickers for awarding grade-points in postsecondary education. *Interactive Learning Environments*. http://dx.doi.org/10.1080/10494821003612638. Available at:.
Wood, W. B. (2004). Clickers: a teaching gimmick that works. *Developmental Cell, 7*, 796–798.
Wouters, P., & Van Oostendorp, H. (2013). A meta-analytic review of the role of instructional support in game-based learning. *Computers and Education, 60*, 412–425.
Wu, W.-H., Wu, Y., Chen, C. H.-Y., Kao, H.-Y., Lin, C. H.-H., & Sih-Han Huang, S.-H. (2012). Review of trends from mobile learning studies: a meta-analysis. *Computers and Education, 59*, 817–827.