A Roadmap to Cope with Common Problems in E-Learning Research Designs

Javier Sarsa and Tomás Escudero
Dept. of Educational Sciences, University of Zaragoza, Spain
jjsg@unizar.es
tescuder@unizar.es

Abstract: E-learning research is plenty of difficulties, as also research in education is. Usually, the high number of features involved in e-learning processes complicates and masks the identification and isolation of the factors which cause the expected benefits, when they exist. At the same time, a bunch of threats are ready to weaken the validity of the research, for example, disregard of previous research, use of small samples, absence of randomization in the assignment to groups, ineffective designs, lack of objectivity in the measuring process, poor descriptions of the research in publications (which implies few possibilities of replication), wrong statistical procedures, inappropriate inference of results, etc. All of these obstacles accumulate and are carried along the whole research, resulting in low quality studies or irrelevant ones. This theoretical paper suggests a roadmap in order to face the most common problems in e-learning research. The roadmap informs about some cautions which must be considered at each stage of the research and recommendations to increase the validity and reproducibility of results. The roadmap and conclusions included in this paper have been obtained from our experience in educational and e-learning research, also from our long path as reviewers in key journals of these fields, and from readings of significant research handbooks. This is not a strict guide but a set of milestones on which it is necessary to stop and reflect.

Keywords: e-Learning research, educational technology, research designs, e-learning effectiveness, methodology, validity.

1. Introduction: How is E-Learning Research Being Conducted?

The history of educational technology has much to do with unspecific and poorly conducted research. As Alexander *et al*. (2006) noticed, even after 50 years of e-learning (about 30 of them online) there is little evidence to support a strong body of knowledge in this field. Educational technology research has had a limited impact in transforming the use of technological tools in the classroom (Amiel and Reeves, 2008). Simultaneously, educational research is seen "as a practically sterile activity that has conspicuously failed to produce a rational base for educational policy and practice and is largely irrelevant to the needs of the educational policymakers and practitioners" (Carr, 2007, p.272). The lack of credibility in educational research and poor transferability of research results to the real world is a consequence of diverse issues, as the conflicting concept of teaching and learning effectiveness, incoherencies between definitions of research problems and methods to explore them, unmanageable amounts of data and others (Raffaghelli, Cucchiara and Persico, 2015). The same weaknesses are found in e-learning (Conole and Oliver, 2007), where a bunch of different methodologies coexist to treat a single problem: how to improve learning.

The reasons for that deficiency in research are multiple. But firstly, it must be answered, what is e-learning research for? As He Kekang (2014) said, "it is well known that the inherent characteristic of educational technology (*i.e.*, its qualitative prescription) is the use of various technologies to optimize the educational and teaching processes to achieve the goal of improving the effectiveness, efficiency, and benefits of education and teaching" (p.13). In a similar way, Masie (2008, p.379) defined e-learning as "the use of network technology to design, deliver, select, administer, and extend learning". Having in mind that the main objective of e-learning is obviously the improvement of learning, most of the research is based on how to tackle this issue. "Articles that report research studies with technology-based teaching strategies should begin by making it clear that they address a significant educational problem, as opposed to a proposed technology solution" (Roblyer, 2005, p.194–195). Nevertheless, recently, problems have shifted from the measurement of e-learning effectiveness to the study of teaching and learning practices and pedagogical methods (Hung, 2010).

In e-learning research, there is a large number of fields of study. In Hung's taxonomy of e-learning research, the most frequent themes in the groups related to learning were: e-learning communities and interactions; multimedia in e-learning; adaptation and usability; gaming in e-learning; simulation in e-learning; support system design and development; teaching and learning strategies; how to improve the effectiveness of e-

Reference this paper as: Sarsa J and Escudero T, "A Roadmap to Cope with Common Problems in E-Learning Research Designs" *The Electronic Journal of e-Learning Volume 14 Issue 5 2016*, (pp336-349) available online at www.ejel.org

learning and student motivation; large-scale national or state-level e-learning projects; and emerging technologies' impacts in educational fields.

For Haythornthwaite *et al.* (2016), the questions over which e-learning has evolved around have been: how to teach online; how to bring resources for distributed learners; and how to study and practice at the technology–learning interface; but nowadays e-learning questions are also about video-based resources, gamification, MOOC, virtual worlds, virtual communities, annotation of video lectures, dashboards, adaptive learning and about how to use the digital traces of engagement, interaction, communication, argumentation and learning which e-learning systems record (learning analytics), among others. For example, in virtual communities of learning, current research is raising questions on technical and social prerequisites to build and support them, and in gamification, about how to use strategic thinking to make choices or solve problems (Hillen and Landis, 2014).

In either case, any e-learning research should clearly define the area of application, because research in education usually deepens into a labyrinth of interactions.

> "Applied to e-learning, socio-technical perspectives draw attention to the complex of interacting elements that make up an e-learning case: the array of technologies; the individual and collective practices of teachers, learners, and educational institutions; the meaning associated with degrees, universities, and higher education; the technological readiness of stakeholders; the identity and accepted practice associated with the roles of teacher and student; and more". (Haythornthwaite *et al.*, 2016, p.6)

The high number of features involved in e-learning processes complicates and masks the identification and isolation of the intervening factors. The separation of the different contributions of each factor to variance, which is a key issue, is extremely complex. The students' backgrounds, their interactions, the tasks done and learning paths followed by each individual are different and uncontrollable, in spite of the existence of modern learning analytics, which record large data streams produced by the activity of the students within some learning management systems.

Therefore, an initial consideration about how current e-learning research is done has to do with the lack of definition and control.

On a separate level, the fast pace of technological developments does not help to scholars. Research on a particular technology or tool is quickly obsolete and the tool replaced by another one, more extended, accepted or trendy, independently of the educational potential of the replaced tool. And "we can assume that the e-learning landscape will become even more diverse" (Hillen and Landis, 2014, p.218). Within this context, an inconvenient is that "educational technologists are frequently more concerned with the possibilities of using a new technology (means), such as a newer course management system or the hottest wireless device, than seriously considering the ultimate aims of its use and its consequences" (Amiel and Reeves, 2008, p.33).

Indeed, e-learning is always evolving and of course media attributes too. Old conclusions are not always applicable to current technology: "…much of the research being conducted is designed for earlier forms of education resulting in no significant differences being found for new forms of education" (Spector, 2013, p.22). Media attributes in e-learning are now different from those found some years ago; for example, augmented reality using virtual glasses may facilitate the comprehension of complicated concepts, some systems of delivery of rich video lectures allow watching streaming video in combination with annotation and interactive activities, etc. These are emerging cognitive tools which support new forms of knowledge representation, acquisition and interaction, and would not have been possible with ancient media. Yang, Wang and Chiu (2014) stated that the tenets of Clark (the importance of instructional methods, 1994), Kozma (the importance of media attributes, 1991) and Mayer (the importance of the cognitive characteristics of the individual, 2008) should be taken into account all together in technology-enhanced learning and, of course, in e-learning research. Media attributes have to be considered at a similar level as instructional methods, since certain instructional methods can only be performed using a particular media or tool (*e.g.* 3D printer). Therefore, both technological and pedagogical-cognitive aspects should go hand in hand. As Rushby and Surry (2016) assert, "a key problem that particularly besets information and communication technologies (ICT) in learning is that the

champions tend to be well informed about technology itself but often less competent in the broader aspects of learning" (p.2).

Nomenclature does not help either. There are a myriad of names used for labeling e-learning derivatives (m-learning, u-learning, distance learning, web based learning…) and its associated tools and services (*e.g.* only in the case of video: digital video, streaming video, webcast, vodcast, webinar, vlog, video lecture, enhanced video lecture, etc.). It is hard to know what someone is referring to when there is a vague description of the study. Is e-learning equal to reading digital books, pdfs, interactive books, videoconference, video lectures, collaboration tools...? What kind of collaboration tools? Discussion forum, wiki construction, online shared documents? A big muddle, for example, is found with the concept of 'impact', which is used in very different ways: improvement in scores, better ICT competences of students, influence over the educational system, etc. (Balanskat, Blamire and Kefala, 2006). Some current e-learning research is carried out without taking into account the specific nature of each tool and the way of using it; again, a vagueness in the approaches can be noticed.

Another main consideration is largely determined by the heterogeneity of the studies, which hinders comparisons, and by their recurrent focus towards the gathering of (qualitative) subjective judgements. Indeed, differences in the way in which the problems are tackled is a difficulty for establishing research guidelines. Hung states that "Scholars in leading countries on e-learning development focused on its educational aspects. Scholars in early adopter countries tended to study e-learning from technical perspectives" (2010, p.10).

Since the statement of the four-level model of Kirkpatrick (1979) to evaluate e-learning, the first level, called reaction, has been commonly used. That means that many researchers evaluate mainly the participants' reactions to a specific e-learning program, tool or resource, that is, opinions, perceptions, motivation, satisfaction... about the program. After this, they draw conclusions related to its quality, its advantages, or surprisingly, they infer how much the students have learnt, instead of how much students think they have learnt. A typical question in this type of studies could be, do you think the inclusion of a quiz feedback helped you to learn more? This kind of findings is supported by subjective answers which lead to a dramatic decrease of the validity of results and contribute to a lack of credibility. Strother (2002) stated that pretest-posttest designs were not used by the majority of organizations in e-learning evaluation; only the satisfaction levels were considered. So, she concluded saying: "measuring learning requires a more rigorous process than a reaction survey" (p.5).

"It must be noted that the field of educational technology is a discipline characterized not by methodological unity but by methodological diversity" (Amiel and Reeves, 2008, p.404). But, in any case, and above all, methodological rigor must prevail. Both authors maintain that "educational technologists should not continue to simply investigate the impact or describe 'best cases' in *post facto* applications of technological devices" (p.33). Indeed, in educational settings, this kind of approach may result too much superficial.

Quantitative and qualitative methods are condemned to get along with each other, even after the U.S. Department of Education's (2003) decision "to give funding priority to research that adopts random sampling and experimental designs" (Randolph, 2008, p.13). The reason for adopting this resolution is the increasing need to find empirical evidences which can be replicated. Unfortunately, too many studies using experimental or quasi experimental designs also fail in one or more steps of the research process. Like in a high-fidelity system, if there is a low-quality part (speakers, amplifier room acoustics…), the whole system will sound bad.

Bulfin *et al.* (2014), from a survey carried out with 462 experts in educational technology, report "a preference for relatively basic forms of descriptive research, coupled with a lack of capacity in advanced qualitative data collection and analysis" (p.403). They also cite different authors to point out that, in educational research, methods and designs are usually under-specified and scientific design and accurate statistical analysis are often avoided. Perhaps, the reasons why this happens is due to the lack of knowledge of research designs or because of the statistical difficulty. On the opposite side, Hewson *et al.* (2003) stated that, while the majority of research on Internet has been making use of surveys and interviews, there have been many rigorous studies using experimental designs, in which individuals were randomly assigned to the experimental conditions.

The study carried out by Bulfin *et al.* (2014) showed that descriptive research was commonly used by a 50% of the surveyed researchers, 32% used collaborative research (as participatory methods), 25% comparative research, 21% experimental research, 15% design-based research, 15% ethnographic research, 14% longitudinal research, etc. Randolph (2008) found 41% of studies experimental/quasi-experimental, 26% quantitative, 16% qualitative, 12% correlational and 6% causal-comparative. In MOOC, Raffaghelli, Cucchiara and Persico (2015) identified almost every type of e-learning research methods, including quantitative, qualitative, mixed-methods, design-based research, literature review and theoretical-conceptual research. With reference to qualitative research, the top five inquiry methods are: narrative research, phenomenology, grounded theory, ethnography, and case study (Cresswell, 2013).

At the same time, in e-learning there is an increasing trend towards participatory/social and design-based methods (Kafai, 2005). But when individuals are involved in contexts which require a high social interaction, their autonomy can be diminished, and it hinders to take advantage of the benefits of digital media to advance at their own pace (Annand, 2007). For Kuboni (2013), requisites of social interaction enter into direct competition with the student autonomy.

Amiel and Reeves (2008) defend design-based approaches because of the complex interactions which take place in educational settings, where human, social and cultural aspects are linked to technical ones. In this way, a design-based method is developed within a specific educational context, contributing to create a connection with real-world problems. The two key aspects in design-based research are: on the one hand, the negotiation of the research goals between practitioners and researchers and, on the other hand, an iterative research process within a real educational context, in a cycle of investigations (*Ibid*.). So, classrooms become living laboratories (Sakai, 2005). Under this point of view, classical one-independent-variable studies would provide a very limited insight of the educational outcomes. Design-based research use mixed-methods, such as surveys, expert reviews, evaluations, case studies, interviews, retrospective analysis and formative evaluations (Wang and Hannafin, 2004).

In general, the way in which a lot of studies are conducted is too lax, especially regarding to the control of validity and reliability. This problem may happen at any stage of the research: definition, literature review, sample, gathering tools, statistical analysis, etc. Actually, it is one of the key problems which lead to the lack of credibility in e-learning research.

2. Common Errors in E-Learning Research

As a result of these difficulties, "e-learning evaluation is still deficient and does not have evaluation guidelines" (Tzeng, Chiang and Li, 2006, p.1040). Hence, this theoretical paper suggests a roadmap in order to face the most common problems in e-learning research. The roadmap informs about some cautions which must be considered at each stage of the research and recommendations to increase the validity of results. The stages on this map include warnings which correspond to the following phases:

1. Appropriate definition of the research problem and questions
2. In-depth reviewing of the specific state-of-the-art
3. Sufficient sample and assignment to groups
4. Appropriate methodology and powerful design
5. Variables and data gathering tools
6. Correct usage of statistics in the analysis of results
7. Cautions about inference
8. Detailed report generation and possibility of replication.

In this way, this roadmap can serve as a useful tool for beginners but also for researchers who like to check and avoid some risks at every step they take, mainly focusing on evaluative e-learning problems. This paper does not provide strict rules but some milestones on which to stop and reflect, always aiming to find out strong evidences.

2.1 Appropriate definition of the research problem and questions

The research problem and the questions derived from it are among the more important parts of the research. However, too often, questions are generic and imprecise. A typical but bad example of research question in e-

learning would be the following: did the online group do better than the face-to-face one? But, what means to do it better? Higher scores? More competences? Did they work more? Higher motivation? It is also necessary to wonder whether students spent the same time, or whether they interacted each other, if they used every tool, etc. Even though the researchers were able to measure accurately the key variables, there would not be many possibilities to disentangle the contribution of each factor because of the interactions among them. To avoid this complication, as mentioned above, many researchers choose to interrogate about the field of reactions, using questions like, do online students were more encouraged using such tool or technology?

Answers to general questions can be unreliable. Randolph (2008) recommends to broken down general research questions into more specific issues. An accurate problem and specific questions are mandatory in e-learning research. As one of Cohen's (1990) principles proposed, "more highly targeted issues" benefit research. In e-learning, another example of poor formulated question could be: can students learn more about medical contents using e-learning? A better and more specific question would be: do the retention of anatomy concepts improve when students can manipulate images in 3D? In this research question the retention of medical concepts will be the dependent variable and could be measured through a test, for example.

Furthermore, educational technology questions should always facilitate to provide answers from a cognitive perspective, *i.e.*, having into account the human cognitive architecture. That implies it is necessary to elaborate research questions which allow obtaining answers about the effects of an educational treatment (Is it working? How much is it working? What is happening?), but, at the same time, also consider how we learn and discover all the causes which are producing benefits or damages (Why is it working? How is it working?). The question writing should be specific enough to be able to detect the part of the contribution of each tool or methodology used in the treatment and why or why not it works. Sometimes significant effects can be found, but the mechanisms by which they happen remain unknown. To achieve this, qualitative approaches (qualitative questions), in which students give their opinions about why they think the treatment yielded an effect, are needed.

2.2 In-depth reviewing of the specific state-of-the-art

It is necessary an in-depth review of the previous research for each e-learning specific situation or problem studied. Rushby and Surry (2016) found that almost all the research questions about ICT had been already addressed and even answered during the two decades between 1980 and 2000 but, obviously, not using the current technology. They warned that many researchers only looked into articles published online, while they had saved time and resources if they had read earlier works before. Indeed, the progress in educational research is conditioned to the attention paid to established solid theories (Burkhardt and Schoenfeld, 2003).

In educational technology research, it is uncommon the replication of experiments aimed to confirm or reject previous results, and to advance in the consolidation of theories. Often, studies remain isolated; many studies start from scratch, leaving previous findings unattended. Starting e-learning research from scratch, and also the use of mutually incompatible research perspectives, does not lead to accumulation but to chaos; as Gros (2012) states, it seems we are always making the same questions. The cumulative criterion is one of the pillars to improve the relevance of e-learning research (Roblyer, 2005).

Moreover, in order to create a robust body of knowledge, reproducibility is mandatory; the most reasonable choice for e-learning progress would imply deepening into already existent research lines, taking previous studies as a reference and reproducing them. Only the consideration of previous research can lead to build a cumulative and robust body of knowledge. Mayer's Cognitive Theory of Multimedia Learning (2008) and Sweller's Cognitive Load Theory (2011) are two examples of theories which have been developed thanks to a well-driven set of studies and have achieved to set up a group of scientific and accepted rules.

Researchers must review what variables have been used previously to deal with equal or similar research problems. For example, in an e-learning program, can the time spent watching videos, the number of pictures used, the interactivity dose, the adequacy to students' learning styles, etc., improve the educational effectiveness of the materials? The unveiling of equivalent studies for each one of these variables can be a big help to find out other variables, to guide on the research design, to detect threats, etc.

2.3 Sufficient sample and assignment to groups

In e-learning research, participants use to be volunteers (convenience sample). Forcing students to participate is considered difficult and sometimes unethical. This problem can lead to get small and biased samples; samples which are not representative of the entire population. But "there is no sampling method which guarantees a simple is representative of the population under study" (Canal, 2006, p.122).

Some authors criticize sampling methods used in e-learning. Bulfin *et al.* (2014) point out that this self-selection sample method, lacking of randomness, is not suitable for obtaining statistical significance, even using non-parametrical testing.

However, as Hewson *et al.* argue (2016) although experimental designs rarely obtain probabilistic samples, statistical inference is justified if they are randomly assigned to conditions. Fortunately, in e-learning settings, many times it is possible to establish different groups freely and assign students to one or another group randomly. An advantage of this procedure (in those cases where random assignment to groups is possible) is that the most part of validity problems can be avoided. In any case, if there are doubts, a pretest is recommended to check the initial equivalence of the groups in the variable of study. When it is not possible that individuals are randomly assigned to groups, the study could be framed within quasi-experimental research.

Usually, in educational settings, to recruit a sufficient number of individuals, researchers have to resort to offer rewards in the form of additional points or other kind of gifts to convince the students to join. And even so, it must be envisaged that the sample mortality can be high. At least a number of 30 individuals per group is considered enough to perform parametrical analysis. "The number 30 seems to have arisen from the understanding that with fewer than 30 cases, you were dealing with 'small' samples that required specialized handling with 'small-sample statistics'." (Cohen, 1990, p.1304). In general, the bigger the sample is, the better reliability will have, and the more sophisticated statistical test will can be done (Cohen *et al.,* 2007). Of course, in qualitative studies, where generalization is not a goal, probabilistic samples are not necessary, since another kind of data, for example narratives, are sought.

2.4 Appropriate methodology and powerful design

Discussion between quantitative and qualitative focuses is still alive. Ramage (2002) argued that comparisons studies between the results of different groups of students remained among the best methods to determine the effectiveness of educational technologies. However, these methods, predominantly experimental or quasi-experimental, are neither free of problems, but on the contrary, they have to be carefully planned ahead to avoid the 'annoying' inconvenience of guaranteeing the validity and reliability of the results. Furthermore, higher rates of validity can be achieved using some form of triangulation. For Randolph (2008), "there is good reason for this mixing of methods of analysis... researchers can get a more holistic and valid view of a phenomenon by viewing and interpreting the phenomenon from different perspectives" (p.89).

Unfortunately, there are few guidelines which can be applied to all kind of research, since methods are context-dependent (Randolph, 2008), but all of them share the requirement of validity and reliability (although different for each one) and, in all of them, the combination of quantitative and qualitative explanations would be desirable. The important dilemma is "to decide what methods are most appropriate for what questions" (Bulfin *et al*., 2014, p. 404).

Cause-effect pure models tend to simplify realities in which many variables interfere with each other (Gros, 2012). This can justify why some experimental or quasi-experimental designs have been done in real-contexts, provided that the majority of strange variables and threats can be controlled. Experimental designs which are developed in real contexts gain in external validity (ecological validity) at the expense of internal validity and causal inferences, but they allow to explore what works better than artificial laboratory scenarios.

Even when experimental and quasi-experimental designs are highly valued by impact journals and educational administrations, mixed methods are gaining in acceptance (Randolph, 2008). Mixed methods combine quantitative and qualitative methods to obtain data which complement each other. Some questions cannot be answered using only quantitative data, because quantitative data seldom inform about the context. As an example, "most research into the use of recorded lectures by students has been done by using surveys or

interviews" (Gorissen, Van Bruggen and Jochems, 2013, p. 20), although reliable data about its effectiveness on learning cannot be evaluated in this way. In the particular case of MOOC data collection is mainly carried out through classical methods like surveys and analysis of learning analytics data (Raffaghelli, Cucchiara and Persico, 2015), which are often subjective or inaccurate. That is why the triangulation between quantitative data and qualitative data is so important. While quantitative studies try to approach objectivism and to find explanations or predictions, qualitative ones are interpretative and are seeking deep understanding (Lincoln, Lynham and Guba, 2011).

Traditionally, evaluative e-learning research has compared the effectiveness of a treatment in a face-to-face group versus an online second one (sometimes adding a third hybrid group). In this schema, the first group was the control group, while the second was the group who received the e-learning treatment or program. This kind of research is obsolete and plenty of difficulties to guarantee internal validity, due to the high number of differences between both scenarios, which make it difficult to identify which factors are responsible of the effects, when they exist. In this kind of distribution, the majority of students are involved in a sea of educational tasks which mask and complicate drawing reliable final conclusions. Usually, on the one hand, the face-to-face group is attending traditional classes with presentations, taking notes, making questions, participating in face-to-face discussions, etc., while on the other hand, the online group is studying at home, reading digital documents, videos, engaging in online discussions, etc. In order to understand whether a factor is causing an effect, groups have to be similar in all those possible characteristics different to the applied program itself. Abrami *et al.* (2011) suggest to compare an online treatment with another online treatment, using better research designs (controlling sample imbalances) and more precise measurements. As mentioned before, even though the sampling method is not random, the random assignment allows considering both groups as probabilistically equivalent and the use of experimental designs.

A 43% of the experimental studies in Randolph's (2008) meta-analysis used one-group posttest-only design, which is a pre-experimental design (Campbell y Stanley, 1963), and 19% a pretest-posttest without control group. Among the good designs to check the effectiveness of specific treatments in e-learning, it is possible to use the simple two-group posttest-only experimental design which allows using a t-test or one-way ANOVA for measuring statistical significance of posttest scores. Another better option is to use a two-group pretest-posttest design which allows to subtract the pretest effect using an ANCOVA test (using pretest scores as covariable).

As third option, but more difficult to tackle, the four-group Solomons' design (1949) –two with treatment and two for control, two with pretest and two without it– allows to compare the initial equivalence of the pretested groups, eliminate the pretest sensitivity effect using ANCOVA for the pretested groups, and check the effect of the independent variable in the pretested groups as well as in only-posttest groups. So, the Solomon design is among the most powerful experimental designs, because it enables the initial comparison between the pretested groups and the consideration of the posttest for all groups (statistical analysis was described by Braver and Braver, 1988). Experimental designs are able to eliminate all the threats to internal validity but it is not possible to guarantee external validity (Campbell y Stanley, 1963).

Quasi-experimental designs are very common in education and e-learning. Almost all non-equivalent-group designs contain many threats to validity. Since the assignment to groups is not random, it is not possible to affirm that both are initially equivalent (even using pretest). Besides, in real-world contexts, groups usually experience very different histories during the period of the study, apart from the treatment itself, like different teachers, materials, contacts with the other groups, etc. Many of these differences should be considered and planned ahead, or avoided, in e-learning contexts, before starting the research.

The application of mixed methods and design-based methods neither disqualify the realization of comparison between groups, nor the application of statistical tests. Quite the opposite, while quantitative methods are able to detect whether a treatment was significantly effective, qualitative methods, used in synergy with quantitative, allow strengthening the hypothesis, unveiling, especially, how and why the quantitative results took place. That is, they give important clues about the reasons which have produced the effectiveness or ineffectiveness of the treatment, and might have not been born in mind as factors initially.
The internal validity of many experiments is reduced by the various threats that, in too many occasions, become evident. Two of the most known threats in education and e-learning research are the experimenter expectancy effect or Rosenthal effect (Rosenthal and Jacobson, 1963) and the Hawthorne effect. The first one

takes place when the researchers applies a treatment or program which uses a novelty technology (it could be a new social network or an app of augmented reality). He or she introduces, consciously or unconsciously, an artificial encouraging component which modifies students' behavior. The Hawthorne effect occurs because individuals often modify their behaviors when they know that are being observed (*e.g.* a student who is evaluated using a participant observation method). Perhaps, these are some of the most difficult to control threats, although balanced designs can minimize it.

The control and elimination, if possible, of all kind of threats to internal validity should be the main concern for any e-learning researcher. It forces him or her to sit down, think meticulously and anticipate what might happen during the intervention, trying to achieve the treatment is the only difference between groups. For example, making a balance of participants between experimenters (when there are more than one) can contribute to avoid the Rosenthal effect. The experimental and control groups would 'suffer' the same teachers. Other useful designs in order to avoid internal validity problems are those which balance the treatment or program between groups, like the switching replications design. As third example, it is also possible to evaluate what happened in both groups in those parts of the course not subject to treatment. Did you find any difference? Of course, not.

2.5 Variables and data gathering tools

Data collection is one of the most important phases of e-learning research. In e-learning, individuals are usually distant and researchers cannot resort to the whole range of gathering tools. This limitation is more pronounced in 100% e-learning programs, because it is not possible to apply some face-to-face data collection tools in identical conditions (*e.g. in-situ* interviews).

With regard to quantitative research, the prevalent data were obtained from experimental and quasi-experimental measurements (23.8% expressed to have expertise) and visual/audio techniques (23.4%) (Bulfin *et al.* 2014). In the 21 experimental designs studied by Randolph (2008, p.67), data were obtained from questionnaires (91%), log files (29%), test (23%), interviews (23%), direct observation (21%), exercises (14%), teacher survey (10%), standardized test (10%), narrative analysis (10%), time on task (5%), focus groups (5%), etc.

In qualitative research, the gathering tools commonly used are interviewing (recordings), surveys (questionnaires) and observations (field notes). The study conducted by Bulfin *et al.* (2014) revealed that more than 50% of the participating researchers in educational technology had expertise with interviews, more than 40% with surveys and 39% with observations. However, in e-learning situations, while surveys can be easily applied online, interviews are very difficult to carry out and observations almost impossible.

Concerning the validity of the different tools used to collect data, it has to be noticed that, although all of them provide with valuable information, such information is often biased or inaccurate. Even experimental or quasi-experimental studies are not able to exclude all the possible alternatives to explain a particular effect, that is, they are unable to avoid the influence of all the strange variables which can affect the dependent variable (in particular, the researcher's influence and the applied instrumentation).

Regarding to variables, the "less is more" principle, defended by Cohen (1990, p.1304), entails the use of few independent variables and even fewer dependent variables in research. But Joy and García (2000) warn that many studies lack of validity because designs are wrong, some variables ignored and results questionable. They found, for example, errors in the sampling method, sample size, assignment to groups or ignorance of other influential variables (prior knowledge control, teacher effects, time spent in the task, learning styles, etc.). They highlight that is it very important to carefully control which factors are explaining the variance, whenever there is a suspicion that results might be influenced by one or more uncontrolled variables.

2.6 Correct usage of statistics in the analysis of results

Bulfin *et al.* (2014) have listed the percentage of familiarity of researchers with different methods of data analysis. First places were occupied by quantitative descriptive statistics (graphs and charts 41%, means and standard deviations 38%, frequencies 36%), followed by qualitative content analysis (29%), cross-tabulations (28%), comparison of means (t-test, or ANOVA, 28%), comparison of frequencies (chi-squared, or Mann-Whitney, 24%), correlations (Pearson, or Spearman, 24%), regressions (21%), discourse analysis and textual

analysis (18%), etc. However, many studies do not check variables requirements before applying parametric statistics (Kolmogorov-Smirnov, Shapiro-Wilk, Levene, etc.).

Continuing with Bulfin *et al.* (2014), they emphasize the absence of advanced data collection and analytical techniques such as data mining or learning analytics, which would provide with data for statistical and visual analysis. At the time learning management systems gather accurately user interactions, it will be possible to count with a huge amount of information that will allow to understand user's behaviors better and to derive more interpretations from them, including group comparisons (Giannakos, Chorianopoulos and Chrisochoides, 2015). With regard to MOOC, Raffaghelli, Cucchiara and Persico (2015) identified that the most commonly used analysis methods were traditional statistics and, recently, data visualization based on learning analytics.

In addition to P-values to determine the statistical significance of a variable, it is very necessary to use some indicator about effect size, which indicates how much the treatment or program is affecting. "The absolute effect size is the difference between the average, or mean, outcomes in two different intervention groups" (Sullivan and Feinn, 2012, p.279). Effect size is considered the substantial significance in a study, because P-values inform about whether an effect exists but not about how much it affects. Therefore, "the primary product of a research inquiry is one or more measures of effect size, not P values" (Cohen, 1990, p.1310).

Standardized indices of effect size make possible comparisons between studies, hence these indices (Cohen's d, Kappa, Odds ratio, correlation indexes, etc.) should be included in reports to advance in e-learning research. As an example, in the meta-analysis of Means *et al.* (2009) of empirical studies about the effectiveness of online learning, they found 176 of them using experimental or quasi-experimental designs, 99 had done at least one contrast test, but only 45 included effect size data.

Cohen (1990) also recommends "simple is better" (p.1305). He suggests starting descriptively but representing data in graphical ways when possible. For example, a frequency polygon or a Tukey stem and leaf diagram would be better than media, standard deviation, skewness and kurtosis values to describe a distribution; and a scatter diagram better than r coefficient (*Ibid*.).

2.7 Cautions about inference

One of the important limitations in any e-learning study is the difficulty to make inferences. Results are also heavily context-dependent and limited by the sample. Probably, one exception is the case of MOOC, whose possibilities to obtain large and demographically representative samples facilitate the extrapolation of results.

"External validity refers to the degree to which the results can be generalized to the wider population, cases or situations." (Cohen, Manion and Morrison, 2007, p.136). External validity of lab controlled e-learning experiments is always compromised, at least related to ecological validity, since the artificiality of the situation has not much to do with the real settings. "In education, ecological validity is particularly important and useful in charting how policies are actually happening 'at the chalk face'" (Brock-Utne 1996, p.617). Here, design-based methodologies allow to achieve a higher external validity, since interactions take place in real settings; however, internal validity is reduced whether strange variables cannot be controlled.

2.8 Detailed report generation and possibility of replication.

The need of a more detailed description of e-learning research is a common requirement. Often, there is no information about how the sample was obtained, how they were assigned to groups, detailed description of the intervention, clear relations between variables and hypothesis, whether parametric statistics could be applied (normality and homoscedasticity conditions), internal validity considerations (studying possible threats), why statistical test are used, effect size values, etc. Apart from these descriptions, it is always recommended to measure and describe the amount of time spent by each group in the research objective. What happens if the experimental group has spent a 300% more time using the treatment than the control group?

Abrami *et al.* (2011) demand a better description of all the treatments, both in face-to-face and e-learning, and the inclusion of full statistics. In the particular case of the treatments used, a superficial description of them could invalidate the research. What would happen if the experimental group received a treatment which included the questions that were going to be used in the posttest, while the control group would not? Results would be notoriously significant, but the research would not be valid.

"Without an adequate description of procedures, it is difficult or impossible to achieve replication – a cornerstone of science. Also, without an adequate description of settings and participants it is difficult or impossible to establish parameters for generalizing findings" (Randolph, 2008, p.68).

3. The Roadmap

This roadmap has been designed to provide researchers with a set of suggestions to detect and avoid some common errors in e-learning research designs. It has been designed from our experience in educational and e-learning research, also from our long path as reviewers in key journals of these fields, and from readings of significant research handbooks. Note that this roadmap is not a complete guide but a set of recommendations for a typical profile of e-learning research.

Stage 1: It is necessary an appropriate definition of the e-learning research problem and questions which must be relevant for improving learning. Generic questions should be broken down into more specific ones. Questions must respond to quantitative issues, like the effects of the program or treatment (*e.g.* has the application of video been significantly effective? How much did the students improve their grades? How long were they watching videos?), but questions should also be open to qualitative explanations of the reasons why those effects exist (including cognitive aspects), like, why do they think the application of videos took effect? What cognitive aspects were responsible? Are there alternative explanations?

Stage 2: The study of previous research (literature review) is useful for accumulation, which leads to the establishment of theories, and also for replication, which helps to configure a group of comparable studies.

Stage 3: A minimum sample of 30 individuals is mandatory to apply parametrical statistics (which are more powerful than non-parametric). When individuals have been randomly assigned to the groups, and the variables can be strongly controlled, designs can be considered as experimental or near-experimental nature. Nevertheless, it would be desirable to check the initial equivalence of the groups. When random assignment is not possible, it would be preferable a quasi-experimental consideration, even though equality in the initial observations was found.

Stage 4: Questions, methods and designs must be in tune. A mixed methodology would be desirable to answer both types of questions (quantitative and qualitative), that is, to check if there were significant effects and to carry out a detailed research of the reasons why the effects took place, considering alternative explanations. Triangulation through interviews, discussion and case studies are among the preferred options for qualitative inquiry.

A strong control of validity problems (threats to validity) is extremely important. It would be desirable to avoid pre-experimental designs and select two, three or four-group designs with pre and posttests, and elaborate carefully both tests (they must not be equal). Contamination and differential drop-out rates among the groups might jeopardize the internal validity. Conversely, balanced designs (especially balances whose aim is to find when the hypothesis is not satisfied) can serve as a way to confirm the hypothesis and gain in internal validity (for example, a balance of the treatment between groups).

Stage 5: Data gathering tools must be simple (few variables) and reliable (objective measures). Variables must be defined adequately (in accordance with questions and hypothesis) and without using artificial transformations (commonly found in ordinal scales which are treated as numerical ones). Sources must be reliable, because a non-reliable (manipulated or subjectively interpreted) gathering tool distort data and invalidate any posterior result.

Stage 6: Statistical analysis is an important phase. It must be explained in detail: why a statistical test has been selected, an in-depth description of the values (not only significance values, but also effect sizes), checking normality and homoscedasticity conditions, application of the right tests (parametric or non-parametric), etc.

Stage 7: Before making inferences it is necessary to reflect on what particular conditions of the research would be different in other contexts (students, teachers, subjects, e-learning technology…). In highly controlled

experimental settings, it should to be considered and explicitly expressed whether results would be the same in real settings (ecological validity).

Stage 8: A detailed description of every research phase must be included in the reports, especially concerning the sampling method, the statistical analysis, the validity of the gathering tools and data, possible alternative explanations for the results and replication possibilities.

Figure 1 shows the 8 stages considered −framed with round-corner boxes− and their associated recommendations −in rectangular boxes with the same color−.

4. Conclusions

Along this document, various recommendations to avoid some errors made in e-learning research have been given. Too many studies lacking of scientific rigor have caused that the establishment of generally accepted e-learning theories has been almost null. Among the reasons found are: unspecific problems and questions, shortage of theoretical foundations, disregard of previous research, insufficient or biased samples, application of unreliable methodologies, inappropriate designs, dismissal of threats, wrong statistics, few controls, incorrect generalization of results, incomplete reports, impossibility of replication, etc.

This roadmap provides novice researchers with a small guide which, although generic and flexible, is useful to avoid errors in typical e-learning research projects which are oriented to respond if a particular treatment, intervention or program is able to improve learning and, hence, helping to advance towards the ultimate goal of educational technology and e-learning: the improvement of learning.

Future works could be focused in two directions: 1) the development of more detailed sub-maps for each one of the stages considered here, and 2) the creation of different versions of roadmaps for specific types of e-learning research. The first work would allow to deepen into every phase, since the present roadmap is a simplified model. The second work could be focused on other types of research methods, like qualitative methods or design-based research; in any case, always searching how to obtain high rates of validity and credibility.

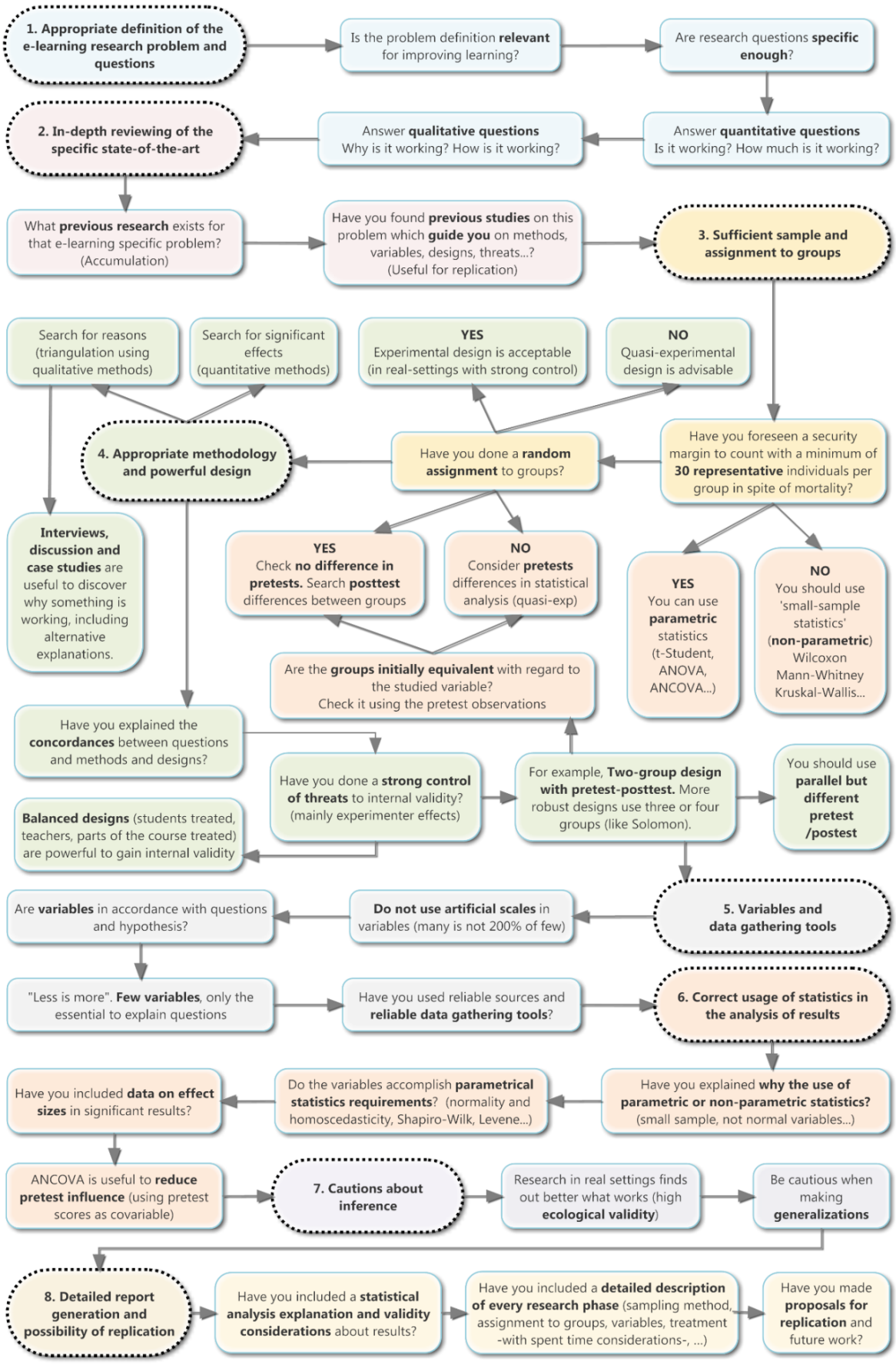

Figure 1: The roadmap of recommendations for effective e-learning research

References

Abrami, P.C., Bernard, R.M., Bures, E.M., Borokhovski, E. and Tamim, R. (2011). Interaction in distance education and online learning: Using evidence and theory to improve practice. *Journal of Computing in Higher Education*, 23, pp.82–103.

Alexander, S., Harper, C., Anderson, T., Golja, T., Lowe, D., McLaughlan, R., Schaverien, L. and Thompson, D. (2006). Towards a mapping of the field of e-learning. In: E. Pearson and P. Bohman, eds. *Proceedings of EdMedia: World Conference on Educational Media and Technology 2006*. Association for the Advancement of Computing in Education (AACE). pp. 1636-1642.

Amiel, T. and Reeves, T.C. (2008). Design-based research and educational technology: rethinking technology and the research agenda. *Educational Technology & Society,* 11(4), pp.29–40.

Annand, D. (2007). Re-organizing universities for the information age. *International Review of Research in Open and Distance Learning*, 8(3).

Balanskat, A., Blamire, R. and Kefala, S. (2006). *The ICT Impact Report: A Review of Studies of ICT Impact on Schools in Europe*. European Schoolnet, European Comission.

Braver, M.C. and Braver, S.L. (1988). Statistical treatment of the Solomon four-groups design: a meta-analitic approach. *Psychological Bulletin,* 104(1), pp.150–154.

Bulfin, S., Henderson, M., Johnson, N.F. and Selwyn, N. (2014). Methodological capacity within the field of "educational technology" research: an initial investigation. *British Journal of Educational Technology*, 45(3), pp.403–414.

Burkhart, H. and Schoenfeld, A. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher*, 32(9), pp.3–14.

Campbell, D.T. and Stanley, J.C. (1963). *Experimental and Quasi Experimental Designs for Research*. Hopewell, NJ: Houghton Mifflin Company.

Canal, N. (2006). Técnicas de muestreo. Sesgos más frecuentes. In: A. Guillén and R. Crespo, eds. *Métodos Estadísticos para Enfermería Nefrológica*. Madrid: SEDEN. pp.121–132

Carr, W. (2007) Educational research as a practical science. *International Journal of Research & Method in Education,* 30(3), pp.271–286,

Clark, R.E. (1994). Media will never influence learning. *Educational Technology Research and Development,* 42(2), pp.21–29.

Cohen, L., Manion, L. and Morrison, K. (2007). *Research Methods in Education*. New York, NY: Routledge.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), pp.1304–1312.

Conole, G. and Oliver, M. (2007). *Contemporary Perspectives in E-Learning Research: Themes, Methods and Impact On Practice*. London: Routledge

Cresswell, J.W. (2013). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. 3rd edition. Los Angeles: SAGE.

Giannakos, M.N., Chorianopoulos, K. and Chrisochoides, N. (2015). Making sense of video analytics: lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course. *International Review of Research in Open and Distributed Learning,* 16(1), pp.260–283.

Gorissen, P., Van Bruggen, J. and Jochems, W. (2013). Methodological triangulation of the student's use of recorded lectures. *International Journal of Learning Technologies,*8(1), pp.20–40.

Gros, B. (2012). Retos y tendencias sobre el futuro de la investigación acerca del aprendizaje con tecnologías digitales. *Revista de Educación a Distancia*, 32.

Haythornthwaite, C., Andrews, R., Fransman, J. and Meyers, E.M. (2016). *The SAGE Handbook of E-learning Research*. 2nd Edition. London: Sage.

He, K. (2014). Learning from and thoughts on the Handbook of Research on Educational Communications and Technology (3rd edition): Part 2 — insights in complexity theory, situational theory, and several other hot topics. *Journal of Educational Technology Development and Exchange*, 7(1), pp.1–18.

Hewson, C., Yule, P., Laurent, D. and Vogel, C. (2003). *Internet Research Methods. A Practical Guide For The Social And Behavioural Sciences*. London: Sage.

Hillen, S. and Landis, M. (2014). Two Perspectives on E-Learning Design: A Synopsis of a U. S. and a European Analysis. *The International Review of Research in Open and Distance Learning,* 15(4), pp.199–225.

Hung, J. (2012). Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics, *British Journal of Educational Technology,* 43(1), pp.5–16.

Joy, E.H. y García, F.E. (2000). Measuring learning effectiveness: a new look at 'no significant-difference' findings. *Journal of Asynchronous Learning Networks,* 4(1), pp.33–39.

Kirkpatrick, D. (1979). Techniques for evaluating training programs. *Training and Development Journal*, 33(6), pp.78–92.

Kozma, R.B. (1991). Learning with media. *Review of Educational Research,* 61, pp.179–212.

Kuboni, O. (2013). The preferred learning modes of online graduate students. *The International Review of Research in Open and Distance Learning*, 14(3), pp.228–250.

Lincoln, Y., Lynham, S. and Guba, N. (2011). Paradigmatic controversies, contradictions, and emerging confluences, revisited. In: N.K. Denzin and Y.S. Lincoln, eds. *The Sage Handbook of Qualitative Research*. 4th Edition. Thousand Oaks, CA: SAGE Publications. pp. 97–128.

Masie, E. (2008). Luminary perspective: what is the meaning of the e in e-learning. In: E. Biech, ed. ASTD Handbook for Workplace Learning Professionals. Alexandria, VA: ASTD Press. pp. 377–381

Mayer, R.E. (2008). Applying the science of learning: evidence-based principles for the design of multimedia instruction. *American Psychologist,* 63(8), pp.760–769.

Means, B., Toyama, Y., Murphy, R., Bakia, M. and Jones, K. (2009). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies*. Center for Technology in Learning, US Department of Education.

Raffaghelli, J., Cucchiara, S. and Persico, D. (2015). Methodological approaches in MOOC research: Retracing the myth of Proteus. *British Journal of Educational Technologies*, 46 (3), pp.488509.

Ramage, T.R. (2002). The "no significant difference" phenomenon: A literature review**.** *E-Journal of Instructional Science and Technology*, 5(1).

Randolph, J.J. (2008). Multidisciplinary Methods in Educational Technology Research and Development. Julkaisija: Hämeenlinna.

Roblyer, M.D. (2005). Educational technology research that makes a difference: Series introduction. *Contemporary Issues in Technology and Teacher Education*, 5(2), pp.192–201.

Rosenthal, R. and. Jacobson, L. (1963). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports*, 19, pp.115–118.

Rushby, N. and Surry, D.W. (2016). *The Wiley Handbook of Learning Technology*. Chichester, UK; Malden, MA: John Wiley and Sons.

Solomon, R.L. (1949). An extension of control group design. *Psychological Bulletin,* 46(2), pp.137–150.

Spector, J.M. (2013). Emerging Educational Technologies and Research Directions. *Educational Technology & Society,* 16 (2), pp.21–30.

Strother, J. (2002). An assessment of the effectiveness of e-learning in corporate training programs. *International Review of Research in Open and Distance Learning, 3*(1).

Sullivan, G. M. and Feinn, R. (2012). Using Effect Size - or Why the P Value Is Not Enough. *Journal of Graduate Medical Education,* 4(3). pp.279–282.

Sweller, J., Ayres, P. and Kalyuga, S. (2011). *Cognitive Load Theory*. New York: Springer.

Tzeng, G.H., Chiang, C.H. and Li, C.W. (2007). Evaluating intertwined effects in e-learning programs: A novel hybrid MCDM model based on factor analysis and DEMATEL. *Expert Systems with Applications*, 32, pp.1028–1044.

Wang, F. and Hannafin, M.J. (2004). Using design-based research in design and research of technology-enhanced learning environments. In: Annual Meeting of the American Educational Research Association, San Diego, CA.

Yang, K., Wang, T. and Chiu, M. (2014). How technology fosters learning: Inspiration from the “media debate”. *Creative Education,* 5*,* pp.1086–1090.