The evolution of the ejournal
Disseminate your research immediately through a new online, open access, peer reviewed journal

Online Educational Research Journal

This paper was downloaded from

The Online Educational Research Journal

(OERJ)

www.oerj.org

OERJ is an entirely internet-based educational research journal. It is available to anyone who can access the web and all articles can be read and downloaded online. Anybody can submit articles as well as comment on and rate articles. Submissions are published immediately provided certain rules are followed.

Guidance, Regulator and a Rating System for Trials/Experiments in Social Sciences: Consolidated Conduction, Reporting and Evaluation of Trials Entity (CONCRETE).

Mirjan Zeneli

Durham University, School of Education, Durham, UK

This article provides new answers to an old debate, dealing with bias in experimental designs in social sciences and especially the field of Education. To begin with the article provides a short review of studies which have investigated the influence of experimental design on the magnitude of the effect size (ES) in various areas of social sciences. The studies reviewed here all possess one element in common: they illustrate that stricter experimental designs report lower ES. A phenomenon which is argued to be mainly due to multi-level biasing associated with weaker designs, such as quasi or single case trials. Consequently, the aim of this article is to introduce news ideas with the objective of reducing bias for trials/experiments in social sciences while borrowing from the field of Medicine. Specifically three new ideas for trials in social sciences, especially education, are presented: 1) a list of 31 recommendations to serve as guidance for future trials, 2) the recommendation of an independent regulator to supervise trials and work together with multiple researchers and actors, 3) a quality evaluation system to judge the trustworthiness of trials in social sciences. The benefits of these three proposals, which are missing in social sciences, are also discussed.

Key Words: Experimental Guidance, Trial Independent Regulator, Randomised Control Trial, RCT, Regulator, Quality Evaluation System.

Correspondence Address: School of Education, Durham University, Leazes Road, Durham, DH1 1TA, United Kingdom. Email: mirjan.zeneli@durham.ac.uk. Tel: 078 078 5 5353.

Acknowledgment: Special thanks for all their discussions on trials in social sciences to Professor Peter Tymms and Steve Higgins at Durham University in Durham, and Professor Allen Thurston at Queens University.

Introduction

Robust trials in social sciences are necessary due to the complex nature of the social world (Oakley, 2006; Trochim, 2012). Previous studies have established that robust research trials generally report lower effect sizes (ES) (Cheung and Slavin, 2015; Lipsey and Wilson, 1993; Slavin, Lake and Groffs, 2009; Zeneli, Thurston and Roseth, 2016). This conclusion has been established at various levels of social sciences from various researchers:

At the general level, Lipsy and Wilson (1993) investigated the relationship between trial design and the magnitude of the ES by exploring previous meta-analyses. By concentrating on diverse meta-analyses in areas such as psychological, educational and behavioural research, they categorised different trial designs and looked at their mean reported ES. They concluded that single case studies, i.e. only pre/post-test designs, reported an average ES of 0.76 (n=45), while for the comparison/controlled design group the mean ES was reported at 0.47, (n=45).

In a more specific area of social sciences, the field of Education, Cheung and Slavin (2015) investigated the relationship between trial designs and the magnitude of the ES. The authors reviewed some 645 studies in education in the areas of reading, mathematics and science, examined multiple research issues such as sample size, measurements, design and published vs. unpublished papers. They concluded that higher mean ES were evident for small scale, unpublished, non-independent measurements and quasi design studies. In terms of design, for quasi trials they report a significantly higher mean ES than for randomised trials, 0.23 (n=449) and 0.16 (n=196) respectively.

Research into the influence of research design on the magnitude of effect size has also been conducted with even more homogeneous population and areas. While Slavin et al., (2009) looked at the relationship between trial design and the magnitude of the ES for mathematics as a subject, Zeneli et al. (2016) investigated the relationship between trial design and the magnitude of the ES within an effective intervention in education, 'peer tutoring':

Slavin et al. (2009) investigated the relationship between research and the magnitude of ES of various research aspects. In terms of design they reported that randomised control trial (RCT) studies showed a mean ES of 0.05 (n=13) while quasi-experimental studies showed a mean ES of 0.13 (n=26).

The Zeneli et al. (2016) study also investigated multiple research characteristics and their relation to the magnitude of the ES for peer tutoring. They reported that single group studies have a higher mean ES 0.92 (n=17) than any category found within the controlled studies group; and within the controlled studies group quasi studies and RCT/1matched showed an ES of 0.66 (n=3) and 0.73 (n=17) respectively, while RCT/2matched and over category showed an ES of 0.23 (n=12). In other words, the more robust the trial design the lower the mean ES.

Consequently two questions are posed: Firstly, why does the design of the trial influence the magnitude of the ES? And, secondly, what can be done about this issue? The first part of the question can be answered by viewing RCTs as entities with elements that control not only bias that stems from the research field, but also control researcher intended/unintended bias. This explanation is put forward by the articles reviewed above. The next question, 'what can be done to reduce trial bias for social sciences?', is the purpose of this paper:

While enough work has been conducted to establish the influence of experimental design on the magnitude of the effect size in social sciences, the culture for ensuring robust trial is lacking behind. Specifically: a) the guidance for conducting trials and experiments in social

sciences is unclear and currently un-established, b) currently the field/industry is unregulated and a fully independent body in charge of standards does not exist, and c) at the moment a trial rating system or trial quality evaluation body that assesses the strengths and weaknesses of trials does not exist.

To begin with, in terms of guidance:

Firstly, most social scientists conducting trials and experiments borrow or follow the guidance set out for health and medicine in 'Consolidated Standards of Reporting Trials' (CONSORT) (Moher et al., 2010). Although many of the elements from the CONSORT list can also be applied to trials in social sciences, some elements, specifically, not informing the participants what treatment they are receiving (known as 'blinding') or controlling for placebo effects are inappropriate.

Secondly, the few elements of research guidance that have been reported are presented in the form of guidance on how to set out a clear and coherent research method as opposed to more specific guidance for trials, in particular the work by Gorard (2014a) which proposes how to evaluate the trustworthiness of research findings.

Finally, other types of work that have touched upon guidance on experimental trials in social sciences are those of Campbell and Staley (1963), Cook and Cambell (1979), Torgerson and Torgerson (2008), or in education: Hutchison and Styles (2010), and Slavin (2008a). However, these names have not provided a broad guidance check list; they have rather concentrated on how to design and carry out RCTS in general and the issues faced by weaker designs.

A fully independent regulator to monitor the trial process, or assess, rate and classify the quality of trials in social sciences is non-existent. Therefore, currently the researcher/s or individual institutions are in full control of every aspect of conducting a trial.

Consequently, at least three different, although related, tasks need to be established to improve trial standards in social sciences: a) the establishment of a list of guidance for running trials and experiments in social sciences, b) a fully independent regulatory/managing body, c) and a rating system for classifying/evaluating the quality of trials. This paper focusses on discussing the three points above, while drawing extensive examples from the field of pedagogy/education.

This paper proposes 'Consolidated Conduction, Reporting, and Evaluating of Trials Entity' (CONCRETE) as a means to reducing bias in the conduction of trials in social sciences. CONCRETE would be a fully independent regulator which would seek to ensure that trials in social sciences are of a strong quality. Many of the elements from CONCRETE guidance are similar to those of CONSORT for health and medicine. In addition to managing trial processes, CONCRETE would also deal with the end result, assigning points in the form of stars to trials as a means of trial evaluation, with the studies and trials meeting most of the trial guidance awarded the highest recognition. Such a system would paint a picture of their reliability and quality.

Consequently this paper concentrates on developing three ideas: a) trial guidelines in social sciences with cases drawn from education, b) trial regulators and processes, specifically proposing new researcher competences in the trial process, and c) a trial rating system for evaluating and categorising trial quality.

It will be concluded that the three ideas proposed here are interdependent on each other and one cannot fully function without the others. The establishment of such guidance, regulatory and trial rating systems would ensure not only that future research is of good quality, but would also produce high calibre academics, researchers and research institutions in general.

I) Trial Guidance:

The following 31 recommendations have been identified to aid researchers, regulators and trial quality assessors:

1) Trial registration

Trial registration is essential in ensuring transparency and avoiding publication bias when conducting trials. It is common knowledge that trials that are unsuccessful are rarely published (Torgerson, D., and Torgerson C., 2003). This influences the future of a particular idea being tested, with its impact overestimated as its true effect is hidden from the rest of the research community (BioMed Central, International Standard Randomised Controlled Trial Number (ISRCTN), 2016).

2) Funding body reported

This is essential in order to establish whether any conflict of interest exists. The researcher might be under pressure, or might bias the results in order to correspond to the funding body's interests and secure future funds (Roseman et al., 2012).

3) An effort to have an RCT or a strictly designed matched design.

A research paper needs to report efforts taken to secure an RCT or strictly designed study. Specifically, a paper would benefit if it reports: a) whether the researchers proposed to the participants an RCT or strictly matched study, b) if the benefits of a strong design were explained to the participants, c) if the proposals were rejected, where any alternatives were suggested.

4) Conduct strong systematic reviews of the topic under consideration to verify the study's originality.

A trial would also greatly benefit from the conduction of a systematic literature review. The reason for this would be to establish one of three points: firstly, whether there is theoretical originality, for example, testing an important idea which has not been tested from a new or old theory or whether the trial is pragmatic. Secondly, to establish methodical originality, in other words testing a theoretical idea that has been previously tested in a non-rigorous way. And finally, a good review is necessary to place the study within the theoretical and empirical context in relation to the rest of the literature (McMillan and Wergin, 2006).

5) The study needs to clarify the significance of its investigation.

A comprehensive research paper would benefit from placing the research direction into context and identifying the importance of the research (McMillan and Wergin, 2006). The article could identify the significance for the area of interest, – the significance for professionals, such as teachers, care workers or therapists, – it could identify the significance for policymakers and the academic community, by providing more information, opening a new debate or solving an old puzzle.

6) Outlining clear aims, objectives, research questions/hypotheses together with a structure.

In terms of appropriate reporting, a paper needs to have a clear section identifying the aims, objectives and the main hypothesis, identifying dependent and independent variables. In terms of aims, these would be the overall direction of the investigation or the main research question. The objectives would be sub-questions deriving from the overall aim, giving way to various hypotheses, with the latter identifying clearly what the dependent and the independent variables would be (Blaikie, 2009).

A structure should also be identified somewhere in the beginning of the article in order to guide the reader.

7) Clear justification is provided as to why the study has chosen those particular participants, and their characteristics explored.

A study also needs to report why or how it recruited the particular population and what their characteristics are (Moher, et al., 2010). It could be that a study needs to concentrate on a particular population due to it being ignored in the past or that the past research is not clear enough about the impact of an intervention on a specific population. Also, the paper would benefit by clearly illustrating the characteristics of the sample, specifically age, nationality, ethnicity, socio-economic background and any other important characteristics relevant to the research question at hand.

8) Ethical considerations are met.

The study needs to state clearly what the potential risks are to the participants and the ethical checks and approvals by the ethics committee of the institution running the trial, as well as the participants themselves (British Educational Research Association (BERA), 2011). In terms of approval by the ethics committee, the paper needs to state clearly the institution of the ethics committee in question and the date approval was given. In terms of ethical approval by the participants, the study needs to report the type of approval, for example opt-in (written consent provided) or op-out (participants have to opt out if they do not wish to take place in a trial) approval.

9) Conducted appropriate power analysis to ensure samples sizes are appropriate for expected power.

Power analyses are essential when conducting trials (Ellis, 2010). Power analyses identifying the required sample size need to be conducted and reported. The analyses would be informed by previous similar research on the findings and need to be conducted prior to the recruitment of the participant. All necessary coefficients relating to power analysis would need to be reported, and the name of the calculator or the procedure of the analysis would need to be identified. The use of such analysis would establish the idea that the report or study is not just searching for something meaningful or fishing for results.

10) Stratified random selection of participants.

In order to ensure that results can be generalised to a population, a simple randomisation of participants from the population sometimes is not enough. In the social sciences even a simple random selection of participants may be ignored, with researchers jumping instead to

generalising conclusions. The lack of random selection of participants has implications for external validity and very often it is not even necessary to report confidence intervals or indeed p values without random selection from the population (Gorard, 2014b). In terms of random selection of participants, stratified random selection is the highest external validity point. This refers to being able to randomly select people from different strata (groups, characteristics, backgrounds) so that everyone is represented in the sample (Blaikie, 2009).

11) Clusters of participants:

Rather than working with individual level participants, a study would have more legitimacy if conducted at the cluster level, taking into account social complexities. For example, in education, working at the cluster level, i.e. at the school level, reduces many limitations such as contamination within a school, classroom context or the local authority context if this is incorporated into the design (Thurston, 2008).

12) Outline the intervention to be tested and its frequency, and illustrate the effort made to ensure implementation integrity.

The study would also need to report the exact nature of the intervention, including the lengths and dosage of the trial (Moher et al., 2010). Specifically a paper would benefit from a dedicated section entitled 'intervention characteristics', outlining the history, previous applicability and any changes or newly applied ideas.

Also, a separate section outlining the steps taken to measure the level of fidelity should be provided. It would be unwise to make any statements regarding findings without first discussing the level of the intervention implementation. Consequently, the study would need to clearly report the procedure of the fidelity measurement, the researchers in charge, their characteristics, the actual fidelity criteria and their origin.

13) Ensure that the instrument is the appropriate one for the aims, objectives, and the intervention focus.

Although researchers should make an effort always to use standardised or well-established instruments (Slavin and Maden 2008), sometimes such instruments do not capture the impact of an intervention. Specifically, if the intervention is extremely short or only covers a few of the areas measured by the instrument, then the chances of downplaying the effectiveness of the intervention are increased as the instrument would be too broad.

14) Report the reliability of the instrument.

Regardless of whether the instrument is new or old, a study needs to state clearly the reliability and construct validity of the instrument (Moher et al., 2010). This can be done by showing either the Cronbach's alpha or Factor Analysis (FA) coefficients if the instrument is new, or simply by referring to previous publications outlining the reliability and construct validity of the instrument.

15) The programme should be implemented for at least 12 weeks to rule out Hawthorne effect.

The Hawthorne effect could influence the findings by showing positive effects. The 'Hawthorne effect' relates to an experiment conducted at the Hawthorne factory, in which

participants increased their output simply due to being watched as opposed to the intervention in place. In order to deal with the Hawthorne effect it is recommended that the studies last at least 12 weeks (Slavin, 2008a; Slavin and Maden, 2008). Studies have found that after week 8 the Hawthorne effect usually starts to reduce (Clarke and Sugrue, 1991).

16) The groups need to be strictly matched or strictly randomised i.e. via blocked randomisation techniques.

In order to deal with unequal group characteristics, researchers should aim for either a blocked RCT or, where this is not feasible, at least a matched design. The randomisation or the matching should be applied to the dependent variable based on the pre-test data. When trying to match groups, if an intervention in pedagogy, for example, seeks to measure its impact in mathematics performance, then matching should take place on mathematics performance data.

In terms of randomisation, blocked RCTs based on the dependent variable should take place, unless the sample size is large enough (at least 100), in which case randomisation theory promises that the groups will achieve equalisation (Torgerson, C., and Torgerson, D. 2013). For example, if we were to measure the impact of cooperative group learning on mathematics performance with only 60 participants, then it is necessary to conduct blocked randomisation (Lachin, Matts and Wei, 1988). This is done as follows:

- Rank the pre-test mathematic performance data for each participant from the highest performer to the lowest performer, 1, 2, 3, 4, 5, 6, 7, 8,…
- Create blocks of two participants, (1, 2) (3, 4) (4, 6) (7, 8).
- Then from each block randomise either to the control or to the intervention group.
- Blocked randomisation can take place either at the individual level, at the classroom, school or local authority level if enrolled throughout a country. The method ensures equalisation of groups for all unknown biases, as well as unequal group sizes.

17) The content of the topic in which participants are to be tested should be covered in all groups taking part in the trial.

This benchmark applies more to education studies in general. It is necessary that the content of the topic tested is also covered in the control group. For example: if an intervention concentrates on improving mathematics performance in education, but the mathematics topics measured by the instrument are not taught in the control group, then any positive finding can also be viewed as being the result of the control group not being introduced to the topic as opposed to the intervention having an impact (Slavin, 2008b).

18) The randomisation to be conducted by an independent body.

In order to minimise intention or unintentional researcher bias, it is crucial that randomisation of groups is conducted by an independent third party or authority. Such authority should not have a stake in the project and should possess the necessary randomisation experience and expertise (Moher et al., 2010).

19) The assessment to be conducted blindly by an independent third party.

Similarly, in order to minimise researcher bias it is also necessary that the assessment of the intervention is carried out by a third party with the necessary training and experience. The

concept of assessment here relates to a) independently choosing the instrument, b) measuring the extent of programme fidelity, c) as well as conducting statistical analysis of the main data (Moher et al., 2010).

20) Separation of independent bodies.

In order to further ensure that there is no contamination of results or conflict of interest it is necessary that those who randomise the participants, those who implement the intervention and those who assess it are keep separate from each other (Moher, et al., 2010). This is an extremely important process characteristic not only because it ensures legitimate results, but also for the identity of the intervention itself, resulting in future researchers not being able to make false claims for past interventions.

21) Clarifying in the pre-research protocol what the proposed statistical methods are, and justifying their appropriateness.

A study needs to clarify what statistical model is appropriate to answer the question at hand as well as justify its application. This is necessary in order to combat any concerns that the researchers are 'fishing for results' (Blaikie, 2009).

22) Clarify which participants entered the analysis and in what format.

This benchmark refers to the idea that once the groups have been established, every participant should enter the analysis regardless of whether they decided not to go ahead with the intervention. This is necessary in order to counteract selection bias. Although including participants who have dropped out of the intervention still influences the impact, treating the participants in their original groups ensures that the researchers do not tamper with the design once everything has been accepted. Hence, analysing the findings with the participants in their original groups is the way forward (Torgerson, D., and Torgerson, C., 2003).

23) An article should report whether all the required statistical requirements were achieved.

A study needs to clarify the statistical requirements to be achieved by a statistical method; whether the requirements were achieved in analysis of the findings, and if not what the researchers did to take into account the shortfalls. Statistical assumption investigations are necessary in order to create a clear picture of the credibility of the results and interpretations (Field and Hole, 2003; Howell, 2010).

24) The attrition rates should be reported for each group and the protocols used for missing data should be clear and explicit.

Attrition rates create bias (Torgerson, D., and Torgerson, C., 2003). Therefore, the attrition rate needs to be clearly stated (Moher et al., 2010). Also researchers should have a dedicated section explaining the implication of the attrition rate for results, both in terms of how it influences outcome and the degree to which it affects ability to generalise from them.

25) Reducing trial contamination.

In order to protect the credibility of an intervention, a piece of research would benefit from ensuring that the control group did not copy the experimental group. Such practices would result in reducing the importance of the intervention as the results would not be positive. One way to take this shortfall into account would be to conduct clustered trials (Thurston, 2008).

26) Report if schools/participants are implementing governmental reforms.

This can be carried out by the independent assessors. In order to ensure that the research context is appropriate for testing a trial it is essential to establish the *clarity* of the context by investigating whether any other interventions based on past best evidence practice are taking place simultaneously. Lemons, Fuchs, Gilbert, and Fuchs (2014) make a good case illustrating this point: When investigating the impact of a cooperative learning intervention via a longitudinal study they found that after several years the group applying cooperative learning no longer showed any positive impact. However, on closer inspection, they discovered that the control groups were adopting governmental recommendations based on the 'what works' literature. Hence, cooperative learning was not being compared to a true control group but to other interventions.

27) Reporting the effect size and placing it in context.

It is crucial that the size of the effect is presented. This can be done by conducting effect size calculations and reporting the type of statistical procedure used to calculate the effect size. This needs to be accompanied by a discussion as to what the magnitude of the effect size entails and how it can be understood (Coe, 2004). Reporting only the means for each group does not provide the full picture in terms of the effectiveness of an intervention (Blaikie, 2009; Howell, 2009). In order to provide a more enhanced picture, a study should provide the mean for each group, as well as standard deviations and sample sizes, so that future researchers can incorporate findings into meta-analyses by using different effect size calculation methods.

28) Reporting limitations and the degree to which the results can be generalised.

It is necessary to include a section outlining the study's limitations, both in terms of theoretical and methodological aspects. This is necessary in order to illustrate that the researchers are aware of their trial's quality, as well as accurately informing the public of the trial's weaknesses. In some fields such as education, many RCTs are already very good at identifying their limitations (Connolly, 2015).

29) Implication for future research.

Papers must put their findings into context and explain their meaning in terms of their relation to past research and findings, as well as making recommendations for future researchers. This helps to draw the boundaries of the research paper, making it accessible and more understandable to the readers.

30) Co-managing the research process with a fully independent regulator.

Carrying out the trial in a jointly with an independent regulator overseeing important decisions such as allocating the trial to independent bodies for randomisation or assessment, is the key to claiming trial process credibility. This is essential in order to increase the trustworthiness of the research.

31) Provide independent quality assessment of the trials.

In order to further establish the full credibility and worthiness of the trial, it is important that the study engages a fully independent regulatory body to provide explanations and rate the trial quality.

II) Trial Regulators and Processes:

Guidance and recommendations are an important step in the process towards producing higher trial quality, however, on their own they are insufficient. Without a fully autonomous institution, guidance and rules will always bend or not be fully respected; this is either due to researcher self-interest or institutional pressure. Consequently, trial guidance lacks implementation, especially considering that many of the guidance and rules are common knowledge and have been established for at least 20 years. However, very few of the important benchmarks, such as randomisation, proven instruments, and separation of competences, are consistently implemented.

Having an independent regulator involved when conducting trials is not a new idea when looking at medicine in the UK, where the government acts as an independent regulator for medicine trials (Medicine and Health Care Products Regulatory Agency, 2016).

Most of the guidance is completed by the researchers who are in charge of the project. Hence in order to enhance the trustworthiness of an RCT, a researcher would only be required to share competences in the following five points deriving from the 31 recommendations.

1. *The randomising or the matching of the participants into groups to be conducted by an independent body*. This would be necessary in order to ensure author bias does not emerge.

2. *The assessment to be conducted by an independent body.* The assessment body could cover four areas: a) selecting the measurement instrument, b) the fidelity measurement, c) an investigation on whether the context/control group is free of bias (i.e. the context is normal), d) analysing the final results.

3. *Ensuring total separation of powers between independent bodies:* The word ‘total’ here is important and there needs to be separation between those randomising, those assessing, those in charge of the project, and those providing quality assurance.

4. *Co-share the managing of the trial with an independent regulator*. In order to ensure the separation of powers between independent bodies it necessary for a researcher to co-share the managing of a trial with an independent regulator. The institution in charge of selecting the assessors (for example the independent regulator CONCRETE) would then need to ensure two issues: firstly, that the institution to be trusted with assessment of the trial does not hold subject-specific knowledge of the issues as there could be a clash of institutional interest. Secondly, that the assessors are not acquainted with the researchers in charge of the project or those who independently randomise the participants. These two issues could be resolved by the autonomous regulator (CONCRETE), which would withhold the information on the groups' nature to the assessors as well as maintain the randomisers anonymous.

5. *Quality assessment/rating of the trials to be conducted by an independent body*. This is necessary to ensure that any further author bias is eliminated. For this task CONCRETE could also act as the independent regulator to provide the quality assessment of the trial, since they would be in the best position to see the extent to which a trial has achieved fully independent research.

Consequently the training of the participants, the setting of the hypothesis and the writing of the findings are to be conducted by the researchers in charge of the project with the CONCRETE guidance points 1 to 31 in mind.

In an ideal procedure there would be a total of four main actors, or groups of actors: 1) the researchers managing the project, 2) an overarching regulator such as CONCRETE, independent of any university establishment, funding or governmental body, 3) an independent institution conducting the randomisation or the matching, and 4) an a different independent institution assessing the results.

The role of the overarching independent body would be to select the assessors, select the group of researchers conducting the randomisation or matching, and finally to assess the quality of the trial prior to the study being published, and rate it. Although most journals conduct their own quality assurance by means of the peer review process, many journals are interest-driven entities, and the peer review of a particular trial could very well have clashing interests or show favour to a particular idea tested by the trial at hand.

The lack of regulator autonomy as an issue is more evident and known in medicine where the 'Medicine and Health Care Product Regulatory Agency' in the UK has been diversely criticised, whether through satire (Goldacre, 2012) or report to British parliamentary enquiries (Health Care Committee: Evidence, 2005).

Therefore, apart from reducing intended bias, a fully autonomous regulatory body to manage separation of powers in social science would have the following additional benefits:

- *Reduced researcher unintentional bias:* Considering that research is a complex process, a helping hand would benefit the researcher and reduce unintentional bias, especially when considering that unintentional bias can derive from both the researcher and the field.

- *Pressure future researchers to implement trial guidelines*: The institution's simple existence would make researchers think twice about which route to take if they wish to produce trustworthy and high quality research trials, especially when seeing that other researchers take the route of trusting an independent regulator. It would be peculiar if a clustered trial worth millions of pounds was chosen not to go through the independent regulator.

- *Improving trust in findings.* The complete autonomy of the regulator from government, charities, research institutions and universities would assure researchers that their work is in safe hands. It would also assure journals and the public in general that a particular trial is worthy and of high quality.

- *The research process overall becomes fairer:* Regardless of who is conducting the trial, influential or non-influential researchers/institutions, the research field would be levelled and better ideas would truly be tested, resulting in system which is overall more fair and just.

III) Trial Rating System:

The absence of a trial rating system for the social sciences makes it harder for researchers without trial knowledge to make any sense of the quality of published trials. As mentioned above, names such as Gorard (2015) have proposed rating systems for research evaluation. However, these are only for judging the trustworthiness of general social research. Others have gone into more detail and provided benchmarks for assessing the quality of overall quantitative research in education (McMillan and Wergin, 2006) or more specifically trials in education (Education Endowment Foundation, 2014). Unfortunately, these benchmarks are limited and do not go as far the 31 benchmarks identified here.

Overall, standards/benchmarks which are harder to achieve need to be given more recognition. In other words, when assessing the quality of a trial it must be acknowledged that research activities such as reporting implications and limitations of a study take less effort than securing a trial which is conducted at the cluster level, or which is independently assessed. The system presented here is similar to that of Gorard (2014a), however, for RCTs as opposed to research in general.

Table 1, next page, serves as an example format to evaluate the degree to which a trial follows CONCRETE guidance. The quality assurance system can operate in a 7 star point system (see table 1). The evaluation checklist is an example which takes into account the notion that trials involving substantial effort in terms of design need to be rewarded and recognised accordingly; even though the specific importance of each benchmark can be debated.

Table 1. Trial evaluation checklist

	Guidance title	**Yes/No**
1	Systematic review *(half star).*	
2	Cluster *(half star).*	
3	Selection of participants: Blocked randomisation or randomisation with large sample *(half star).*	
4	Allocation of participants: Blocked randomisation or randomisation with large sample *(half a star).*	
5	The programme to be implemented for at least 12 weeks to rule out Hawthorne effects *(quarter star).*	
6	Independent randomisation allocation *(half star).*	
7	Independent assessment: Selecting the instrument, fidelity measurements and analysis *(three quarters of a star).*	
8	Independent assessment of the context *(quarter star).*	
9	Separation of independent bodies *(quarter star).*	
10	Independent regulator *(half star).*	
11	Independent quality assurance evaluated & reported *(half star).*	
The remaining 20 benchmarks to be awarded a total of two stars. (1/10 of a star for each benchmark)		
12	Trial Registration	
13	Funding body reported	
14	An effort to have an RCT or a strictly designed matched design.	
15	The study clarifies the significance of its investigation.	
16	The study outlines clear aims, objectives, research questions/hypotheses together with a structure.	
17	Clear justification is given as why the study has chosen those particular participants, i.e. explore their characteristics.	
18	Ethical considerations are met.	
19	Conducted appropriate power analysis to ensure samples sizes are appropriate for expected power.	
20	Outline the intervention to be tested and its frequency, and illustrate the efforts made to ensure that the method was implemented correctly.	
21	Ensure that the instrument is the appropriate one for the aims, objectives, and the area of intervention concentration.	
22	Report the reliability of the instrument.	
23	The content of the topic in which participants are to be tested should be covered in all groups taking part in the trial.	
24	Clarifying in the pre-research protocol what the proposed statistical methods are, and justify their appropriateness.	
25	Clarify which participants entered the analysis and in what format	
26	An article should report whether all the required statistical requirements were achieved	
27	The attrition rates should be reported for each group and the protocols used for missing data should be clear and explicit.	
28	Reducing trial contamination, i.e. social threats such as participants from the experimental group consulting with the control group.	
29	Reporting the effect size and placing it in context.	
30	Reporting the limitations and the level at which the results are generalizable.	
31	Reporting implication for future research.	
	Total:	**7 Stars**

The following constitute some additional interrelated benefits of a quality assurance/rating system:

1) *Motivating researchers to produce better work and to be more cautious:* The rating system motivates researchers, as it will clarify what research efforts are given more weight, hence fostering competition. Although many trial experts are aware what good research looks like, the rating system in place would remind and motivate them to conduct research at a higher level.

2) *Researchers become recognised for their work:* Researchers and institutions would be more renowned for their work, which would further contribute to their career and funding applications.

3) *A better way of judging which researchers produce higher quality of work:* For charities, governmental bodies or institutions hiring researchers it will be more evident which researchers provide higher quality work. Publishing in an international peer reviewed journal does not provide sufficient quality assurance.

4) *Researchers will be able to better self-evaluate and reflect:* This will allow researchers to see exactly what their strong and weak points are and reflect on their achievements and their future plans and directions.

5) *The system will lead to a better way of presenting data, processes and findings:* Having clear set benchmarks will also improve the presentation of how trials are reported and the processes followed during research. This is essential for readers and the research community in general.

6) *Helping future meta-analyses to streamline the process:* A rating system will benefit and speed up the meta-analysis or systematic review process, since it will make clear what studies to include or exclude from the review. In other words, methodical coding procedures will be reduced.

7) *Clearly show which journals hold the highest quality stock/papers:* The rating system will clearly show the true value of a journal. A journal with higher quality trials should be more influential and ranked higher. Although this is partially true currently, the rating system would further clarify the highest commitment to quality.

8) *Knowledge can be paid for appropriately.* A rating system is beneficial for both journals and their clients. A journal can better justify charging more for their higher quality trials, and the client can be confident that the product they purchase reflects the value they pay for.

9) *A generally better way of organising knowledge:* This point is crucial; not all knowledge is the same in terms of quality, therefore, within the trial industry it is only sensible that trials with more scientific effort are acknowledged. A rating system of trials would ensure that meta-analyses are not the only means to identify what works, or what trials have a strong or weak design.

Conclusion

This article is the product of past work related to the influence of experimental design on the magnitude of effect size, which has shown that robust designs have usually provided lower ES (Cheung and Slavin 2015; Lipsey and Wilson, 1993; Slavin, et al., 2009; Zeneli, et al., 2016). These previous studies have all explained, at various levels and areas of the social sciences, that such results are mainly due to bias deriving from multiple origins. In an effort to minimise bias in general when conducting RCTs in Social Sciences, this article has proposed three ideas:

The three ideas presented here, guidance, independent regulatory body, and quality assessment and rating, are all equally important in terms of improving the quality of trials in Social Sciences; *and none can fully exist without the other two*. In other words, guidance without an independent regulatory body can exist only on paper; and without any independent assessment to rate trials, the effort of following guidance and involving any independent regulator renders most of the effort redundant.

The strength of the new system is that it will reduce intentional and unintentional research bias emerging from the field, the researchers or funding bodies. The reduction of bias from these multiple sources will result in multiple benefits, some of which have been already mentioned:

Higher quality of trials: As a rule of thumb a trial which has undergone thorough investigation has a higher chance of quality, due to cooperation among expert researchers as well as checks and balances.

Quality of researchers: Such system provides higher calibre researchers, as better planning, thought and effort are necessary.

Improved trustworthiness: Politicians and policymakers, or even researchers among themselves, will gain more trust and respect for those conducting high quality trials and hold researchers accountable.

Improved organising of knowledge to benefit a wider community: the public, policymakers, journal and students of social sciences will find it easier to identify what trials are credible. Therefore, whereas the ideas proposed by Gorard (2015) help to identify the trustworthiness of research in Social Sciences in general, this article proposes evaluative benchmarks for RCTs.

Better evidence to hold politicians to accountability: Politicians too can be held accountable if they use policies for their own vested-interests, instead of implementing policies which are based on robust scientific research (Tymms, Merrel, and Coe, 2008).

Contribution to economy and society: This will result from: a) finances being spent more efficiently due to better knowledge of what works in a specific sector or area, b) the economic returns being higher when effective evidence-based ideas are rolled out by policymakers, c) society will benefit from adopting evidence-based ideas, without having to wait for politicians and policymakers to enrol effective programmes.

There are, however, limitations to the system outlined here, mainly for researchers:

Firstly, conducting research at the cluster level, or conducting research jointly with independent regulators would require more finance. This could lead to further advantages for institutions which already possess more financial backing. This is clear since institutions with more financial streams would be able to be engaged in clustered trials, which are lengthier and also be able pay for the independent bodies. Two research benchmarks which take more effort, however, are rated higher as a result.

Secondly, going through each of the benchmarks will require more time, and often researchers are very busy.

Finally, operating under such a regulated system would require more/better planning by researchers. The researcher in charge has to keep an eye on the timing and planning, since more steps are involved. Hence although trial quality would improve, trial quantity could slow down.

Overall the benefits outweigh the pitfalls. It is more fruitful to have few high quality trials, generating evidence and order, than too many low quality studies which can add to chaos.

For the above system to be successful it is essential that the overarching regulator, CONCRETE, is not affiliated to any journal, university, governmental or funding body, in order to both safeguard the research process and the final assessment and rating of the trials.

Finally, it is unjust that many aspects in our modern society are tightly regulated, yet the conduction of trials in social sciences research, which often inform policymaking and the public, are rule-less, unregulated, and without any rating system. The absence of such guidance, regulators and quality assurance mechanisms makes it hard to identify trustworthy researchers, institutions or knowledge for many researchers. Consequently, order needs to emerge in the chaos of RCTs in Social Sciences.

Bibliography:

BERA. 2011. *Ethical Guidelines for Educational Research.* Retrieved 12th December 2015, from: https://www.bera.ac.uk/researchers-resources/publications/ethical-guidelines-for-educational-research-2011).

Blaikie, N. 2009. *Analysing Quantitative Data: From Description to Explanation*. Saga Publications, London.

Campbell, D. & Stanley, J. 1963. *Experimental and Quasi-experimental Designs for Research*. Chicago, IL: Rand-McNally.

Cheung, A.C.K., Slavin, R.E. 2015. How Methodological Features Affect Effect Size in Education. *Best Evidence Encyclopaedia*. Retrieved on the 15th January, 2016 from: http://www.bestevidence.org/word/methodological_Sept_21_2015.pdf.

Clark, Richard E.; Sugrue, Brenda M. 1991. "Research on Instructional Media, 1978-1988". In G.J.Angl in. *Instructional technology: past, present, and future*. Englewood, Colorado: Libraries Unlimited. pp. 327–343.

Connolly, P. 2015. The Trials of Evidence-Based Practice in Education. *British Educational Research Association (BERA) Conference*. 15th September, Queens University, Belfast.

Cook, T. D., & Campbell, D. T. 1979. *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin Company.

Education Endorment Fondation. (EEF). 2014. *EEF Classification System*. Retrieved on the 23rd January 2016, from: https://educationendowmentfoundation.org.uk/uploads/pdf/Classifying_the_security_of_EEF_findings_FINAL.pdf.

Ellis, Paul D. 2010. *The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results.* Cambridge: Cambridge University Press.

Field, A. P., & Hole, G. J. 2003. *How to Design and Report Experiments.* London: Sage Publications.

Coe, R. 2004. Issues Arising From the Use of Effect Sizes in Analysing and Reporting Research. *But what does it mean? The use of effect sizes in educational research*. Schagen, I. & Elliot, K. Slough: National Foundation for Educational Research. 80-100.

Goldacre, Ben. 2012. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients.* London: Fourth Estat.

Gorard, S. 2014a. A Proposal for judging the Trustworthiness of Research Findings. *Radical Statistics*. 110: 47-59.

Gorard, S. 2014 b. The Widespread Abuse of Statistics by Researchers: What is the Problem and What is the Ethical Way Forward? *Psychology of Education Review.* 38(1): 3-10.

House of Commons Health Committee. 2005. The Influence of the Pharmaceutical Industry. *Fourth Report of Session 2004-05*. Volume 2.

Howell, D. 2009. *Statistical Methods for Psychology*. Wadsworth; International Edition.

Hutchison, D. and Styles, B. (2010). *A Guide to Running Randomised Controlled Trials for Educational Researchers*. Slough: NFER.

ISRCTN Register. 2015. *BioMed Central*. Retrieved on 11th of January 2016 from: http://www.isrctn.com/.

Lachin, J.M., Matts, J.P., Wei, L.J. 1988. Randomization in Clinical Trials: Conclusions And Recommendations. *Control Clinical Trials*, 9 (4), 365–74.

Lemons, C.J., Fuchs, D., Gilbert, J., Fuchs, L.S. 2014. Evidence-Based Practice in a Changing World: Reconsidering the Counterfactual in Education Research. *Educational Researcher*. 43(5), pp 242-252.

Lipsey, MW, Wilson, DB. 1993. The Efficacy of Psychological, Educational, and Behavioural Treatment: Confirmation from Meta-analysis. *American Psychologist*. 48(12). 1181-1209.

McMillan, J. H., & Wergin, J. F. 2006. U*nderstanding and Evaluating Educational Research.* Upper Saddle River, N.J: Pearson/Merrill Prentice Hall.

Medicine and Health Care Product Regulatory Agency. 2016. *Gov.UK*. Retrieved on the 11th of January, from: https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency/about.

Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P.J.,

Elbourne, D., Egger, M., & Altman, D.G., for the CONSORT Group. 2010. Explanation and Elaboration: Updated Guidelines for Reporting Parallel Group Randomised Trial. *BMJ*.340:c869.

Oakley, A. 2006. Resistances to 'New' Technologies of Evaluation: Education Research in the UK as a Case Study. *Evidence and Policy*, 2 (1), 63-87.

Roseman Michelle, Turner Erick H, Lexchin Joel, Coyne James C, Bero Lisa A, Thombs Brett D. 2012. Reporting of Conflicts of Interest from Drug Trials in Cochrane Reviews: Cross Sectional Study. *BMJ,* 345:e5155.

Slavin, R. E. 2008a. What works? Issues in Synthesizing Sducational Program Evaluations. *Educational Researcher*, 37(1), 5-14.

Slavin, R.E. 2008b. Evidence-based Reform in Education: Which Evidence Counts? *Educational Researcher*, 37 (1), 47-50.

Slavin, R. E., & Madden. N. A. 2008. Understanding Bias Due to Measures Inherent to Treatments in Systematic Reviews in Education. Retrieved July10th, 2012, From: http://www.bestevidence.org/methods/understand_bias_Mar_2008.pdf.

Slavin, R.E., Lake, C., & Groff, C. 2009. Effective Programs in Middle and High School Mathematics: A Best-evidence Synthesis. *Review of Educational Research*, 79 (2), 839-911.

Thurston, A. 2008. Cluster Randomised Controlled Trials: The Way Forward for Educational Research? *The Psychology of Education Review*, 32 (2), 21-23.

Torgerson, C. & Torgerson, D. 2013. *Randomised Controlled Trials in Education: An Introductory Handbook.* Educational Endowment Foundation.

Torgerson, D., & Torgerson, C. 2003 Avoiding Bias in Randomised Controlled Trials in Educational Research. *British Journal of Educational Studies.* 51:1, 36-45.

Torgerson, D. & Torgerson, C. 2008. *Designing and Running Randomised Trials in Health, Education and the Social Sciences.* Basingstoke, Palgrave Macmillan.

Trochim, W. 2012, Design. *The Research Method Knowledge Base*. Retrieved 14th January 2012 from: http://www.socialresearchmethods.net/kb/design.php.

Tymms, P.B., Merrell, C., & Coe, R.J. 2008. Educational Policies and Randomized Controlled Trials. *The Psychology of Education Review*, 32 (2), 3-7 & 26-29.

Zeneli, M., Thurston, A., Roseth , C. 2016. The Influence of Experimental Design on the Magnitude of the Effect Size - Peer Tutoring for Elementary, Middle and High School Settings: A Meta-analysis. *International Journal of Educational Research.* In Press, 21st January 2016.