Experimental research into teaching innovations: responding to methodological and ethical challenges

Accepted for publication in ***Studies in Science Education***

Keith S. Taber

Faculty of Education, University of Cambridge

kst24@cam.ac.uk

Submitted: January 2018

Revised: October 2018; May 2019

Accepted: August 2019

Experimental research into teaching innovations: responding to methodological and ethical challenges

Abstract

Experimental studies are often employed to test the effectiveness of teaching innovations such as new pedagogy, curriculum, or learning resources. This article offers guidance on good practice in developing research designs, and in drawing conclusions from published reports. Random control trials potentially support the use of statistical inference, but face a number of potential threats to validity. Research in educational contexts often employs quasi-experiments or natural experiments rather than true experiments, and these types of designs raise additional questions about the equivalence between experimental and control groups and the potential influence of confounding variables. Where it is impractical for experimental studies to employ samples that fully reflect diverse populations, generalisation is limited. Series of small-scale replication studies may be useful here, especially if these are conceptualised as being akin to multiple case studies, and complemented by qualitative studies. Control conditions for experimental studies need to be carefully selected to provide the most appropriate test for a particular intervention, and considering the interests of all participants. Control groups in studies that replicate innovations that have been widely shown to be effective in other settings should experience teaching conditions that reflect good practice and meet expected teaching standards in the research context.

Key words:

educational experiments; random control trials; control conditions; teaching interventions; replication; ethical comparison conditions

Introduction

It is common for educational innovations, such as teaching approaches, new curricula, or new learning resources, to be evaluated by an experiment where learning gains or other desired outcomes are compared between an experimental condition involving the innovative experimental 'treatment' and some comparison condition where the treatment being evaluated is absent. A small selection of published studies of this kind are listed in Table 1 to give a sense of the potential range of research foci. Such experimental approaches can be very powerful, although there may sometimes be a range of alternative explanations for research outcomes apart from the superiority, or otherwise, of the innovation being tested.

[Table 1 about here]

Table 1: A sample of published experimental studies testing teaching innovations

The present article offers a thematic review of some key issues and challenges that arise in the design and interpretation of experimental studies in education, drawing upon selected illustratory examples of published studies. It is intended that this review will be useful both as guidance for those looking to undertake experimental studies of teaching innovations, and also for those seeking to be informed by reading research reports of such studies. The article considers the particular practical challenges of carrying-out experimental studies in education. This analysis highlights some inherent limitations in many small-scale experimental studies which cannot be assumed to generalise to other contexts. The article considers notions of generalisability and replication to both argue for how such studies can best be understood to contribute to our understanding of teaching and learning and to suggest how individual studies can be best designed to usefully add to the literature. Particular attention is given to the selection of the most informative 'control' conditions with which experimental treatments may be compared. The article suggests guidelines for best practice in establishing control conditions for studies that will be both ethical and informative.

The use of random control trials in education

Teaching is a complex and challenging process, and a core focus of educational research is in informing effective teaching (Pring, 2000). Such research draws upon a wide range of theoretical perspectives, and adopts a spread of different methodologies. Different studies address quite different research questions, and so different methods (collecting and analysing different kinds of data) are appropriate in different studies. As the U.S. National Research Council's Committee on Scientific Principles for Educational Research (2002, p.

3) has noted "methods can only be judged in terms of their appropriateness and effectiveness in addressing a particular research question" and so a "wide variety of legitimate scientific designs are available for educational research" (p.6). From this perspective, *experimental* designs are very suitable for some educational studies, but are not indicated for others (Taber, 2014b). Particular research techniques have specific requirements, without which they are not strictly valid, and a research design that fails to meet the prerequisite conditions of its component techniques may not support robust conclusions. In this article the challenges of undertaking informative experimental research is discussed. Inevitably, then, this review emphasises the limitations of experimental work, and the practical issues that arise in designing valid studies and generalising from them. This is not intended to suggest such studies do not make an important contribution, but rather offers guidance for evaluating such studies, and, indeed, for considering when experimental research can be productively complemented by other forms of enquiry.

Experimental research and units of analysis

The adoption of an experimental approach is intended to avoid falsely inferring that a treatment brings about an outcome, by employing the most appropriate comparison conditions. An important term used in discussing experimental research is 'unit of analysis'. An experiment may, for example, be comparing outcomes between different learners, different classes, different year groups, or different schools (see Table 1 for some examples). It is important at the outset of an experimental study to clarify what the unit of analysis is, and this should be explicit in research reports so that readers are aware what is being compared.

A random control trial (RCT) is an experiment where the units of analysis are randomly assigned to different conditions, and statistical methods are used to determine whether any overall difference in the measured outcomes in those conditions is (probably) due to the intervention. Statistics can only indicate how likely a measured result would occur by chance (as randomisation of units of analysis to different treatments can only make uneven group composition unlikely, not impossible). The usual convention is that a result is statistically significant when its probability (p) of occurring by chance is less than 5 percent (i.e., $p<0.5$). The precise statistical test(s) chosen depend upon the research question(s). A null hypothesis (that there is no difference between the treatments, which is refuted by a finding that *either* of the treatments is more effective) is not simply the inverse of the hypothesis that the experimental treatment will be *more* effective, and researchers should set out the specific question to be tested before designing the research.

A RCT is referred to as a 'true experiment' because there is randomisation of the 'units of analysis' (people, classes, schools, etc.) to conditions. Ben Goldacre, in a position paper on using research evidence in schools that was commissioned by the UK Department for Education, offers a caricature of this type of study:

> Where they are feasible, randomised trials are generally the most reliable tool we have for finding out which of two interventions works best. We simply take a group of children, or schools…; we split them into two groups at random; we give one intervention to one group, and the other intervention to the other group; and then we measure how each group is doing, to see if one intervention achieved its supposed outcome any better (Goldacre, 2013, p. 8)

The 'where feasible' proviso here is important, and a number of potential challenges in undertaking this kind of study are discussed in this article. RCT are sometimes difficult to arrange in education and other social contexts. 'Simply' taking a sample of children or schools and splitting them into two groups at random often raises practical difficulties - and later in this article studies that do not meet the requirements of being a 'true experiment' (such as most of those in Table 1) are discussed.

Randomisation cannot ensure equivalence between groups (even if it makes any imbalance just as likely to advantage either condition) so "while a substantial imbalance is unlikely to occur in a very large trial, small trials may well be subject to sufficient differences between groups to affect the overall result of the trial" (Moore, Graham, & Diamond, 2003, p. 683). Researchers therefore sometimes seek to classify units (e.g., schools) in a sample into similar groupings and randomise from each of these clusters or 'blocks' rather than the complete pool (Moore et al., 2003; Ruthven et al., 2016). This so-called randomised block design requires both identifying what characteristics are pertinent to judging similarity in a particular study (e.g., school size?; location?; curriculum?; selectivity of intake?; gender / ethnic / socio-economic composition of pupils?; etc.) and having accurate measurements of these qualities.

Research reports from small-scale studies (such as those comparing outcomes in two classes, see examples in Table 1) rarely inform readers how the randomisation was achieved, and it has been reported that authors sometimes seem unable to provide such information when asked (by journal editors, for example). It has therefore been recommended that the technique for making a random selection should be briefly reported in methodology sections of reports along with other details of techniques used in the study (Taber, 2013c).

If the units of analysis are schools, it may be difficult to enrol a large enough number of schools into the sample for the statistical methods to be used - especially in those national contexts that rely on schools responding to invitations to volunteer (this is less of a problem when research access is granted at regional/ district or state level). Ruthven and colleagues (Ruthven et al., 2016) report a project (*Effecting Principled Improvement in STEM Education* - '*epiSTEMe*') undertaken in England. The project team were based at a prestigious university that also had extensive and long-standing networks with schools in its region. The research was part of an initiative (the *Targeted Initiative on Science and Mathematics Education*) funded by a national research funding agency (the Economic and Social Research Council) in partnership with the

Gatsby Charitable Foundation, the Institute of Physics and the Association for Science Education. Despite these indicators of status, it proved difficult to recruit schools at the level hoped for,

> The intention was to recruit 30 schools to participate, together providing 60 teachers/ classes in each [of science and mathematics], so as to yield a structured sample of sufficient size to afford a hierarchical analysis of adequate statistical power. …In particular, while the original stipulation was that schools should nominate two science teachers and two mathematics teachers, it became clear that insisting on this would result in far too few schools participating in the trial. Consequently, both the two-subject and teacher-pair requirements were relaxed. … This yielded 25 participating schools: 12 in the intervention group and 13 in the control. Thus, while the number of schools participating came close to original intentions (25 rather than 30), as a result of the relaxation of participation requirements noted above the number of teachers/classes fell well short (34 in mathematics, 36 in science, rather than 60 in each). (Ruthven et al., 2016, pp. 25-26)

In practice, most published studies are based on a much smaller number of classes, and indeed many are based on comparisons between one intervention class and one control class (see the examples in Table 1).

Potential threats to the validity of findings from RCT

The simplest type of RCT will compare two conditions, and often the treatment in one condition will be an innovation (a new teaching approach, or curriculum, or set of learning resources, etc.) to be compared with a treatment that is some form of 'standard' or 'typical' or 'traditional' alternative - for example, in the Ruthven study cited above, "teaching via established methods" (Ruthven et al., 2016, p. 26). The choice of different forms of comparison condition (i.e., no educational input versus customary teaching versus recognised good practice) is considered in a later section of this review.

Where a RCT has been carefully designed and carried out, and when the actual treatment learners experience reflects the intended treatment - that is, that there is a high degree of 'intervention fidelity' (O'Donnell, 2008) - then it is concluded that the teaching innovation gives superior results to the comparison condition if the extent of greater learning gains (or a more positive shift in attitudes, or whatever the desired outcome was) in the innovative condition reaches statistical significance. As the units of analysis were randomly assigned to conditions, this is *unlikely* to be due to a difference in the composition of the two groups (e.g., that the higher attaining students, or the better-behaved classes, were assigned to the intervention/experimental treatment). However, randomisation cannot allow for systematic differences introduced by other aspects of the study design. If fifty students were randomly assigned to two different classes in the same school - 25 to the experimental group experiencing some teaching innovation, and 25 to the control group experiencing typical teaching - but the classes were taught by two different teachers, in different classrooms, with lessons at different times during the week's timetable, then there are clearly

differences in the treatments (i.e., important variables not controlled so they are the same in both treatments) that can potentially confound any effect of the innovation, despite randomisation. Whilst it may seem obvious that the 'teacher variable' needs to be controlled (that is, the same teacher should teach both classes), this excludes controlling other variables (e.g., the same teacher cannot teach two classes simultaneously) and, as discussed later, the same teacher may not be equally experienced, competent and comfortable in different conditions.

25 students in the same class (even if assigned to the class randomly) cannot be considered to be independent learners as they interact and influence each other's learning - so student outcomes within a class tend to be more similar than if the students were not taught together, leading to clustering of measurement outcomes within classes (Dorman, 2012). Such variables are less relevant in large scale RCT as there are many different classes in each condition. This is why studies comparing two, or a small number of classes, may not be especially informative individually, even if randomisation of students to classes is possible, as findings may not be generalisable beyond the specific experiment. (How such studies may be seen as part of a programme of research building up a wider picture of an intervention is considered later in this review.)

Even if studies have large enough samples for such issues to be likely to only produce 'noise' in the data (such that statistical significance testing can reveal a true 'signal' above that noise), there may also be systematic differences that simply cannot be avoided as they are inherent to the way human beings relate to innovative experiences (regardless of the qualities of the innovations themselves). Some such common threats to validity are discussed in this section. These will not be relevant to all RCT, but they are all likely to be potentially pertinent to many experimental studies testing innovations in teaching (including quasi-experiments and natural experiments, discussed further below, where the randomisation required for true experiments is not feasible). It is good practice in research reporting for such issues to be acknowledged, as this helps those looking to learn from the research to consider whether these issues undermine confidence in drawing conclusions about a direct unmediated causal link between an innovation and positive study outcomes. Often, the reader will judge the findings robust regardless, and transparent reporting supports an informed evaluation.

Participants' expectations can influence outcomes

A key issue that often arises in studies with human participants, is that the outcomes in treatments may in part depend upon participants' expectations. This is important because of the demonstrated effect of expectations in producing changes in measured outcomes. In medicine, patients may have a lot invested in the promise of a new drug treatment, and those receiving an experimental treatment may be looking for any

small sign that the medicine is working for them - whilst those assigned to a control condition may feel disappointed, having enrolled in a trial in the hope of getting the new experimental drug. If clinicians are optimistic about the new drug, their expectations may be inadvertently communicated to patients, or may bias their measurements of effect when rating subjective reports of symptoms for example. This is readily avoided if neither patent nor doctor knows who is getting which treatment (a situation known as double-blind), and the analysts are working with anonymised data.

Similar threats to validity are at work in educational settings. This was demonstrated in a study where teachers in a school were told that tests on the children had identified those - 'growth-spurters' - who were likely to make higher levels of progress in the following school year (Rosenthal & Jacobson, 1968; Rosenthal & Jacobson, 1970). These predictions came true: statistically, the identified children did significantly better in school than their classmates after their teachers had been told of their status as growth-spurters. Actually these children had been assigned this label at random, so the results were either an unlikely chance event, or were somehow the outcome of teachers' expectations mediating classroom processes. That is, either by chance the students identified just happened to be those who were indeed going to make better than average progress in the next school year (and this is not logically ruled out by the statistics, but rather just shown to be very unlikely) or there was a substantive effect due to teachers knowing who had been identified as about to make good progress. The students that teachers expected to do well actually tended to do well even though they had been selected purely by chance.

A great many other studies have since replicated effects of this kind (Rosenthal & Rubin, 1978). It is unlikely that such an experiment would be considered ethically acceptable today (British Educational Research Association, 2018). Deceiving study participants (in this case teachers were lied to) should be avoided, and now this 'Pygmalion effect', or self-fulfilling prophecy, is well established it would be considered unfair to those children not identified as likely to make progress (i.e., those in the control condition).

Researchers and teachers may be optimistic about some new teaching approach or curriculum materials and this could bias their judgements, and change their classroom behaviour. Teachers may subtly communicate their expectations to learners who may also respond to a teacher's additional enthusiasm for, and commitment to, an intervention, even if they are not directly aware that the teaching is in some way different from the norm.

This is clearly a major issue in experimental research in science education. If researchers strongly expect co-operative learning, or a flipped classroom, or enquiry-based [1] teaching (e.g., see the discussion of 'rhetorical' experiments, below) - or indeed, for that matter, rote learning and drill exercises, or potentially even starting lessons with a ten-minute nap - to be more effective, then this expectation is likely to have an influence even when the intervention (of itself) may not have otherwise been effective. The response to

such a threat used in drug trials - doing the research double blind - is seldom an option in education as it is usually obvious to researchers, teachers, and even learners, when they are part of an experimental treatment condition.

Participants can respond to perceived novelty

Students experiencing innovative teaching treatments may well be aware there is something unusual going on. If the intervention only involves an individual teacher changing their teaching sequence or activities in a particular topic, then the students in the class may not be aware that things are being done differently compared with the teacher's previous practice. Yet, when the intervention involves an obvious change from what has gone before (e.g., an abrupt shift from teacher-centred teaching and silent individual desk work, to activity-based enquiry learning in groups) then they will be aware something unusual is happening, and may simply respond to the novelty.

Perhaps some learners are less comfortable with changes of routine, but when students are familiar with a routine that makes classes seem mundane, anything unusual is likely to make them more attentive and alert, and so likely to influence learning, simply because of its novelty. There is a tendency built into our cognitive systems to be aware of anything unusual and to pay it special attention, so we would expect students to pay more attention than usual when there is a change in the way things are carried out. This is a consideration in some of the science education studies discussed below where it is claimed that students in the population sampled normally experience teacher-centred instruction where they are largely passive, and by contrast the intervention involves enquiry-based practical activities, group-based discussion work, creative activities, co-operative learning, and so forth. For example, a study of 'active learning' teaching methods reported that "regular instruction in this high school is commonly teacher-centered with a lecture-type format and students passively participate in the learning process. They only listen to their teacher, write notes, and use textbooks as a learning material" (Sesen & Tarhan, 2011, p. 209).

Moreover, if students are involved in theory-directed research (Taber, 2013a) initiated by external researchers (rather than context-directed enquiry undertaken by a single teacher or department as part of the usual ongoing review and development of teaching) then they (and/or their parents) are likely to have been asked to give informed consent for their participation; they may possibly have been involved in completing official looking tests or questionnaires; and their classroom may well have been visited by strangers carrying out observations or making recordings of some kind. All of this is likely to prime students to be more attentive to what is going on in that class.

The joint influence of novelty effects and expectancy effects may in part explain why many interventions that seem effective on first testing, may seem to lose their efficacy once they are 'rolled-out' on a larger

scale to become part of normal ways of doing things (Barab & Luehmann, 2003). It seems that when carrying out educational experiments we have to consider that any apparent outcome may be the result of the combination of *the particular intervention* being tested plus the simple fact of participants experiencing *an intervention*. That applies even when the research is relatively large-scale, involving a large number of classes in different schools working with different teachers, and randomly assigned to one of two conditions: when one condition reflects the status quo, and the other condition something noticeably unusual, then a large sample size and the apparent 'objective' nature of the outcomes of statistical tests offer no way of separating any effect of (i) the special nature of the novel treatment, from (ii) that of the experience of novelty itself.

Despite novelty and expectancy effects being well-recognised, many experimental studies make no reference to these potential threats to validity. One exception is a study that looked at "the effect of reflective science journal writing on students' self-regulated learning strategies" (Al-Rawahi & Al-Balushi, 2015, p. 367). This did acknowledge that "students in the experimental group spent extra time doing something different or new… This was not the case in the control group" and suggested this could have been mitigated by "a second experimental group … given extra time to do something new" (pp.377-378). The same study also acknowledges the potential of an expectancy effect, but suggests this was "controlled for" (p.378) because teachers in both groups took opportunities to offer formal feedback to their students. Yet, it is likely teacher expectations are often communicated in more insidious ways (Rosenthal, 2003). In any case, similar opportunities to express their expectations would only be helpful if it was shown that teachers in both conditions had similar expectations of outcomes from their teaching.

Fair testing should involve teachers in different treatment groups having comparable levels of experience of their assigned teaching 'treatment'

An important variable in research into the effectiveness of teaching innovations is the teacher. Teachers have different levels of skill and experience, different strengths and attitudes, different teaching styles and levels of comfort with different pedagogical approaches, and so forth. Outcomes in two different treatments taught by two different teachers will likely be as much influenced by the 'teacher variable' as the 'treatment variable'. Two approaches to addressing the teacher variable might be to either have the same teacher teach in both conditions or to have a sufficiently large sample so that a diverse range of teachers are employed in each condition.

Whilst employing the same teacher in different conditions may seem to control for the 'teacher effect' a particular teacher's skill set or pedagogic style may suit them to working more effectively in one way, where the opposite may be the case for another teacher. That is, there will be interactions between the teacher

variable and the treatment variable such that having the same teacher in different conditions (whilst, all other things being equal, preferable to comparing across different teachers) does not completely eliminate the teacher variable when seeking to generalise findings from a study context to other teaching contexts (an issue discussed in a later section). In large scale studies there may be enough variation within conditions to allow both for differences between teachers themselves, and the ways particular teachers may engage with different treatments. The approach is likely to be especially valuable when comparing between different treatments that are equally familiar to the teachers in the study.

One variable that may be relevant in many educational experiments that seek to investigate teaching innovations is the level of teacher experience with the innovation. This could undermine even a true experiment that uses a randomisation process. One might consider that the experimental treatment is a new teaching approach, or a new curriculum, or new teaching resources, and the comparison condition comprises of a traditional alternative. The hypothesis being tested is that the innovation will support more effective teaching and so greater learning (that being the motivation for the innovation).

One could imagine a large-scale trial where perhaps 100 suitable teachers (that is, those teaching the appropriate year group and topic) volunteered to take part, and a randomisation process was used to create two groups: a group of 50 teachers in the intervention group and 50 teachers in the comparison condition. Now it may be that the teachers involved in the study, and the classes they are to teach, and the schools where they work, are diverse in terms of teacher skills, student ability, school catchment area, and indeed any number of other potentially relevant variables. As the teachers (and their classes, and schools) have been assigned to conditions randomly it can be assumed that these factors *are likely to* cancel out and so inferential statistics that show statistically significant differences between the treatments are probably not confounded by these variables.

However, this logic may be undermined by a systematic difference between the two groups. The comparison group consists of teachers who generally have experience of teaching in the way they will teaching during the experiment - they will generally have taught this topic in the same way to classes of this age several times before. Yet, typically, the teachers in the intervention group are given some materials and training, and then teach using the innovation for the first time. Generally, when teachers first teach in a new way, or using a new scheme of work or new teaching materials, they do not do so in an optimum way. Teacher Pedagogic Knowledge is to some extent context specific (Park & Oliver, 2008), and usually teachers need to run through a new approach several times before optimising their practice - honing timings and identifying foci for emphasis, finding out how students respond to aspects of the innovative teaching, determining when and how much structure and guidance should be offered during activities, and so forth. Despite whatever prior professional development is offered, teachers teaching in an innovative way for the first time will be learning through the process (van Driel, Beijaard, & Verloop, 2001) and cannot be fairly compared with experienced

teachers working in their customary way. There is also a potential interaction effect here with teacher expectancies (discussed above), as teachers' self-efficacy will usually develop with increasing experience. A teacher who is confident in working in an innovative way may have high expectations for learning outcomes - a teacher who is still adjusting their practice to a new way of working may not.

Now, *in principle*, there is an easy response to this challenge. In this kind of research, data should be collected over several school years and outcomes in the two conditions monitored. It is quite likely that outcomes in the second implementation of the innovation will be better than the first; and outcomes in the third implementation better than the second - but eventually performance will plateau: at which point a comparison between conditions will be fairer. In practice this means running the experiment and collecting and analysing data over a much longer period, which is why this precaution is seldom taken. This approach is also subject to greater potential experimental attrition (where participants drop-out), especially in those teaching contexts where teachers typically only remain in post for a few years before leaving a school.

Participants may make gains during a study due to maturation

Just as research into teaching interventions needs to take account of how teachers develop their skills in applying particular teaching treatments through cycles of implementation, there are parallel consideration about the nature of learners and learning. One issue is the possibility of maturation. As people mature they acquire new cognitive abilities (Goswami, 2008; Piaget, 1970/1972) and so can be expected to achieve more on tests of scientific understanding. One well-known project in science education was known as *Cognitive Acceleration in Science Education* (Adey, 1999), and involved providing regular teaching inputs designed to help facilitate a shift in intellectual development that lower secondary age students (e.g., 11-13 years olds) were expected to be undergoing. So, in this programme, which in educational terms can be considered a long-term intervention (over several school years) the participants would have been expected to be undergoing changes regardless of the intervention. Therefore, simply reporting that students at the end of the programme appeared to show cognitive development compared with the outset would not have been informative. This was recognised in the reference to cognitive 'acceleration' in the programme title.

Rather, when evaluating the effectiveness of the programme, what was tested was whether the intervention encouraged faster cognitive development than would otherwise be the case, by comparing the results of school examinations (taken by participants some years after the intervention) for participants with those of comparable groups who had not experienced the intervention. The argument was that effective cognitive acceleration would support more effective student learning over the remainder of their secondary school career, which could be detected in terms of general academic performance at the end of compulsory schooling (Adey & Shayer, 2002).

In that innovation, maturation was a focus of the study. In other research it is possible that gains measured after an intervention could be due to maturation *rather than* the specific intended teaching input. This is more likely to be the case when an intervention takes place (i) over an extended period, and/or(ii) with young learners who are developing relatively quickly. An example would be a study into the effectiveness of curriculum designed to help young pupils from age 4 learn about floating and sinking (Leuchter, Saalbach, & Hardy, 2014). Leuchter and colleagues controlled for possible maturation by testing for changes in a comparison group of similar ages to their intervention group.

Learning in many areas has been shown to follow a 'U-shaped curve' such that learner performance on objective measures actually dips first, before it subsequently improves (Siegler, 2004). Observing gains in such cases may then depend very much on the time-span between initial and final testing, with effective strategies potentially leading to gains, no change, or even losses, depending when the final measurement is made. In such a situation, mean post-test results for an experimental group that are not significantly better than mean pre-test performance might still represent a positive outcome if control group learners are found to show decreasing performance from pre-test to post-test.

Participants may learn from pre-tests

Pre-tests, then, offer a benchmark by which to compare post-test measurements. So, for example, a study may involve a pre-test, an intervention, and a post-test. The pre-test and post-test are intended to test the same variable of interest (e.g., knowledge, understanding, attitude, skills) that the teaching intervention is intended to impact on. It is important that the instruments actually test what is intended if they are to offer valid measures. Choices also have to be made about how to construct the pre-test and post-test so they are testing the same features. One extreme is to use precisely the same test on both occasions, as then the equivalence of the two tests is assured. An alternative approach is to develop alternative items intended to be equivalent: something that can (where resources allow) be checked by testing the items with a suitable sample of learners from the same general population as the study participants.

The process of completing a pre-test can potentially be a learning experience. Thinking about questions and attempting to provide suitable answers on a pre-test can of itself make it more likely that a person will be more successful on a post-test, especially if precisely the same test items are used on both tests. Even if parallel, but non-identical, items are used, being tested on the first set of questions may trigger thinking processes that lead to learning, that then supports a better performance on the post-test items.

This is a particular issue because of the nature of how learning about science occurs - the changes that may be triggered by a learning experience are not necessarily immediate, but may continue for some time (days, weeks, or longer) after the initial experience (Taber, 2013b). The brain may continue to process experiences

at a preconscious level, which can lead to new (conscious) insights some time later. The use of a control condition, where learners undertake the same pre-tests and post-tests, can go some way to allowing for this effect. If the experience of undertaking the pre-test directly primes students to do better on the post-test, then this should be experienced in both the experimental and control conditions.

There can however also be indirect effects, due to interactions between the experience of taking the pre-test and the subsequent teaching. Current understanding of memory suggests that each time a memory is activated it is reinforced (Dudai & Eisenberg, 2004) so this may happen if the teaching intervention causes students to bring to mind thinking triggered by the pre-test. If pre-test items do not *directly* lead to any new science learning, it is still possible they may prime more effective learning from teaching that follows the pre-test. The education psychologist Ausubel (1978) discussed the notion of an advance organiser, presented before material to be taught, which can help structure the later learning experience. One experimental study of advance organisers in science lessons, that used pre-test items (Gidena & Gebeyehu, 2017, p. 2234) suggested that such advance organisers "can take many shapes" (p.2230). In teaching perspectives informed by the developmental/learning theory of Vygotsky (1934/1986, 1978), some types of learning scaffold (Wood, 1988) are employed to help learners bring to mind relevant prior learning, and to orientate them to the scope of the forthcoming teaching (Taber, 2018). Pre-test items that act in this way can indirectly influence post-test scores by facilitating learning from the intervening treatment.

This can be a concern in research designs that compare an intervention with a comparison condition that does not offer a parallel treatment (the rationale for such a design is discussed later in the section on different forms of control group). Such a study may indicate that the intervention is effective, but strictly the pre-test may need to be considered part of the intervention. It is also possible that any interactions between a pre-test and subsequent teaching may occur differentially in experimental and control conditions that involve different ways of teaching the same topic, given that these are inherently different teaching inputs. As good teaching practice includes the testing of prior learning at the start of a unit, these issues could be somewhat countered by designing and making available pre-test instruments that teachers can then access and adopt as part of their normal teaching (so implementation will reflect this aspect of the tested innovation). Here the development of suitable pre-tests as research instruments has the useful consequence of supporting good teaching practice through the provision of resources.

Deciding when learning is best measured

The issue of the timescale of the learning process, referred to above, also raises the question of when post-tests should best be undertaken. If student consolidation of learning, due to normal brain processes,

continues for some time after teaching, then it may be more informative to test students with a deferred post-test rather than one taken immediately after teaching.

Less optimistically, studies have also shown that measured immediate gains may not be maintained. So, interventions to challenge common alternative conceptions may bring about immediate changes in student thinking; but then apparent levels of conceptual change may appear to diminish if measured some weeks later - at which point students' responses may reflect their initial conceptions (Gauld, 1989). Similar effects occur more generally when learners are no longer actively studying topics, where they may revert to patterns of thinking that dominated before learning took place (Taber, 2003). Teachers are primarily interested in learning that is long-lasting, suggesting deferred post-tests may be more informative than immediate post-tests. However, the greater the delay in measurement, then the more (uncontrolled and unknown) additional learning opportunities participants could have experienced in the interim. Post-tests some weeks, but not longer, after a teaching intervention may offer a sensible compromise here.

Measurement instruments may be considered to be biased towards one treatment

There is potential for the tests used to measure experimental outcomes to be biased towards (or indeed against) the experimental intervention, unless existing standard tests are used that are recognised as valid measures of focal learning outcomes. As an example, in the *epiSTEM*e project teaching modules and assessment instruments were prepared (for 11-12 year old learners) in two science topics - forces (Howe et al., 2014) and electricity (Taber et al., 2015). As well as incorporating principles adopted across the project, in particular a dialogic approach to teaching (Mortimer & Scott, 2003; Ruthven et al., 2016), each module had its own specific features. Within the electricity module there was a focus on teaching about aspects of the nature of science, in particular the use of models and analogies in science, alongside teaching circuit principles (Taber et al., 2016). The forces module had a focus on teaching about proportional relations, something not usually emphasised in teaching physics to this age group. The project included a measure to check for potential test bias towards the intervention condition in relation to the comparison classes studying the same school curriculum topics. Class teachers rated the suitability of module test items "for this class given its experience of the topic this school year".

In the electricity module the test items only examined understanding of circuit properties and no items on the nature of science teaching objectives were included as it was inappropriate to assume teachers in the control condition would emphasise these ideas. (Nature of science objectives were included the official curriculum for 11-14 year old students, but they were not linked to specific teaching topics and could have been introduced at any point over three school years). In the forces module, where there had been

emphasis on the use of proportional relations in teaching the physics concepts, learning of this aspect was tested. It was found (Ruthven et al., 2016) (a) that there was (on average) no more learning of circuit principles in the experimental condition than in the control condition when studying electricity - but it was not possible to know if students had developed a better understanding of the nature of science through studying the intervention module as this was not tested; whereas (b) in the forces module there was significantly more progress in learning about forces in the experimental condition, but teacher ratings suggested the tests measuring this were biased towards learning in that condition.

Judgement is needed in deciding whether such bias is problematic or, perhaps, to be welcomed. If traditional teaching is considered to be ineffective in meeting some particular established curriculum aims, and a teaching intervention is intended to address this, then instruments biased towards testing those specific aims may well be appropriate. However, when tests are biased to objectives or outcomes that researchers particularly value, but which do not represent existing official curriculum aims and are not shared by teachers, then such bias may be considered to undermine the findings of the experiment among the teaching community.

Other potential confounds

This section has discussed a number of issues that may complicate experimental research designs by admitting uncontrolled (and unintended) differences between experimental and comparison conditions. These issues may be pertinent in true experiments (where randomisation is used, as discussed above) as well as in the types of experiment designs discussed in the next section - quasi-experiments and natural experiments - as they operate systematically regardless of randomisation of units of analysis to conditions - as for example when being randomly assigned to an innovative learning condition may tend to increase student engagement.

There are of course many other possible interactions between the experimental teaching input and other experiences that can seldom be controlled - learners may do self-directed reading, watch documentaries, visit science museums and the like, alongside the teaching inputs. This can happen in both experimental and comparison conditions (and is not something science educators would wish to discourage), and generally such effects should not lead to a systematic difference between the two conditions - at least not in RCT where there has been randomisation of the units of analysis to the different conditions. However, not all experimental studies of teaching innovations are RCT, and where randomisation of the units of analysis (e.g., students) to the learning condition is not feasible then it cannot be assumed that the groups in the different conditions are equivalent at the outset, making it more difficult to interpret measured differences at the end of the study.

Quasi-experiments and natural experiments employed when randomisation is not plausible

When experimental research explores classroom teaching in schools, and the units of analysis are individual learners, it is seldom possible (and may not be educationally desirable) to break up existing classes to randomise individual students into new groups for the research. One study included in Table 1 took place in a school designated as a 'laboratory charter school' where randomisation "was part of the school's research mission" (Yin, Tomita, & Shavelson, 2013, p. 538), but more often creating new groupings is not feasible when working with school classes.

So one might consider 50 students who were to be part of a study where it was intended to use individual student test results as a measure of learning to explore whether some teaching approach brought about greater learning than some other teaching approach. If it is possible to randomly assign the 50 students into two groups of 25, then there are 25 'units of analysis' in each group. However, if the researchers are required to work with existing classes then the most randomisation that is possible is to assign whole classes to the two conditions. This would mean the units of analysis were whole classes (one in each condition). To consider this a true experiment (meeting the requirement of randomisation, see Figure 1) there would need to be one measure of learning from each class (cf. Figure 5), but it would be difficult to use statistics to infer anything useful when comparing just two values.

Quasi-experiments

In practice, in such situations, researchers tend to treat the individual learners within intact classes as the units of analysis, in order to collect enough data to be able to undertake statistical testing - but as the units of analysis are not randomly assigned (see the examples in Table 1) it is not possible to draw meaningful conclusions simply by calculating how likely the study outcomes are *by chance* (and compare this with a cut-off such as $p<0.05$), as the students were *not assigned to conditions by chance*. In a quasi-experiment (see Figure 1) then, it is not possible to draw general conclusions by simply comparing the measured outcomes in the two conditions.

Natural experiments

Another term often met in educational research is that of a natural experiment. A natural experiment takes advantage of differences in conditions that already occur, rather than being based on experimental manipulation (see Figure 1). This may be especially useful where researchers are interested in the possible

detrimental effect of some condition, and it would be unethical to create that condition and assign participants to it to test the effect (consider for example a study to find out if victims of bullying make less progress in their science classes - such a study would look to - sensitively - enrol existing students identified as victims rather than experimentally create new victims).

[Figure 1 about here]

Figure 1: Experimental designs may be categorised as true experiments, quasi-experiments and natural experiments

Sometimes 'natural experiments' are possible due to some particular set of circumstances that happen to provide the type of comparison researchers are interested in studying. For example, in many countries, schools run through an annual cycle starting at a particular time of year, and students start formal schooling at the start of the school year following a particular birthday. In this situation it is possible to compare the younger and older students in a year group (Morrison, Smith, & Dow-Ehrensberger, 1995) who have experienced the same educational contexts and experiences, but beginning at a different age (i.e., at the earliest grade levels a child starting school at, say, 5 years and 1 day old is substantially younger - and so typically less developed - than a classmate starting school in the same class on the same day, but at, say, 5 years and 351 days old).

A natural experiment might be possible where some innovative teaching approach, curriculum, or learning resource is already being adopted by some teachers offering researchers a 'natural' opportunity to test its effectiveness against some other more routine or traditional treatment. As, again, there is no random assignment to conditions, it is not possible to simply compare outcomes in the two conditions to infer a possible difference in effectiveness (as it may be, for example, that teachers adopting innovative approaches tend to be atypical in terms of any of more teaching experience, more skills, more confidence, working with more cooperative classes, having better rapport with their students, etc.)

Testing for equivalence between groups

In quasi-experiments or natural experiments a more complex design than simply comparing outcome measures is needed. For example, researchers have to either demonstrate that despite the lack of randomisation, the distribution of 'units of analysis' in the conditions can be considered equivalent prior to the treatment (something often checked even in RCT as a random process cannot *assure* equivalence); or that a difference in outcome seems to be due to the focal variable despite this non-equivalence. In either case this means identifying and measuring any relevant variables.

For example, if (hypothetically) prior knowledge was judged the only relevant variable influencing learning in some study, then a suitable pre-test (see above) might be used to test whether prior learning could be considered equivalent across the two conditions. Often, however, there are other variables which it is recognised could have an effect, other than the dependent variable: 'confounding' variables. If the social class of students and reading age were also considered relevant then it would need to be shown that valid measures of these were also equivalent. This raises the question of what should be considered as 'equivalent'.

Equivalence is more than a lack of significance difference

Considering the prior learning variable, if students in two classes were given the same instrument considered to be a valid test of relevant prior learning, and if the mean scores and standard deviations in the scores in the two classes were found to be identical, then this might be considered convincing evidence for equivalence. This is also extremely unlikely to happen (so much so that such a result could look suspiciously convenient). The question then becomes how much of a difference between the measurements of prior learning in the two groups is so small that it can be assumed to make no practical difference.

The account of one study of enquiry-based learning reports,

> In order to ensure the equivalence at experimental and control groups, students' previous year graduate points of achievement (GPA), intelligence fields, the number of students at the groups and pretest results were taken into account. It was found that experimental group was statistically equal to control group. (Abdi, 2014, p. 37)

Most of this data was not reported in the paper, but the "statistically equal" mean pre-test scores were 2.95 for the control group and 3.15 for the experimental group (p.40). In a study testing the use of advance organisers in physics teaching (Gidena & Gebeyehu, 2017), three parallel groups were pre-tested, and the two that were reported to "have equivalent means" were selected for comparison and assigned as experimental (mean score = 6.61) and control (mean score = 6.26) groups.

Although statistical tests can offer some guidance on what counts as equivalent, they need to be interpreted differently than when looking for a statistically significance difference in the outcomes of the experiment (see Figure 2). An initial difference which is substantial, but statistically non-significant, may be sufficient to explain outcome differences that do reach statistical significance (Taber, 2013a, p. 85: Fig. 4.3). If statistical tests are applied to the starting conditions using the usual $p<0.05$ criterion then they will only flag up differences between the two groups which are very unlikely to be due to chance differences. However, what should be looked for is *evidence of close similarity*, rather than the *absence of evidence of improbable differences*. (One might say that testing for equivalence pre-intervention, and for experimental effects post-intervention,

involve looking at different tails of a distribution.) Two classes with differences between them that are at a level *quite unlikely* to occur by chance are certainly *not* equivalent (at least in the sense that the word is generally employed).

[Figure 2 about here]

Figure 2: Evaluations of equivalence between different groups should be more rigorous than simply excluding differences reaching statistical significance

As an example of good practice here, in their study of the effect of cooperative learning strategies on understanding electrochemistry concepts, Acar and Tarhan (2007), compared treatments in two intact classes of students. Although they only randomised intact classes to conditions, they treated each of the 41 students in the study as a separate unit of analysis (that is, the individual units of analysis were not randomly assigned) and so could not consider this a true experiment. They used a pre-test to compare across the two conditions and reported that "independent t-test analysis showed that there was no statistically significant difference between the mean scores of the experimental and the control groups with respect to ($t=0.199$, $p>.05$) the pre-test" (p.360). They quote a probability value, p, of approximately 0.84 (p.361), which suggests the measured initial differences between the patterns of attainment in the two groups is at a level that would be likely to occur by chance (see Figure 2).

However, Koksal and Berberoglu (2014, p. 66) report a study designed "to investigate the effectiveness of guided-inquiry approach in science classes…", where evidence of equivalence was much weaker. In this study, the treatment group comprised of five classes in one school, and the control group was composed of nine classes in six other schools "to prevent any interaction between the control group and experimental group students" (p.70). They sought to demonstrate "equivalency of schools" as "evaluated with respect to socio-economic characteristics of the students" (p.70), and they reported that the measure used "did not indicate any significant difference" (p.70). Koksal and Berberoglu quote a value of p of 0.21 which is indeed >0.05 (see Figure 2), but means the degree of difference found is large enough to only be likely to occur on about one of five occasions by chance. That is, the differences in socio-economic backgrounds between the two conditions were not so great as to reach statistical significance, but could not be considered small enough to be at a level of just 'noise' in the data.

Using statistics to respond to non-equivalence

When groups in different treatments cannot be considered equivalent, then it is not sufficient to simply compare output measures at the end of the intervention. Rather, some kind of mathematical model (such as

the 'hierarchical analysis' alluded to in the quotation from Ruthven and colleagues above) is needed, in order to allow for how those differences in starting points for the two groups will influence outcomes. Then it can be judged whether any measured differences after the experiment can be considered as due to the difference in treatment, rather than differences in the measured values of confounding variables.

Koksal and Berberoglu characterise their study (see above) as a "non-equivalent control group quasi-experimental design" (p.69). They explain the variables measured:

> "Guided-inquiry approach was the independent variable. While guided-inquiry teaching and learning was implemented in the experimental group, traditional teaching and learning was followed in the control group during the 'Reproduction, Development, and Growth in Living Things' (RDGLT) unit. In both groups, the students' academic achievement, science process skills, and attitudes toward science were defined as dependent variables. And the unit Achievement Test (RDGLT), Science Process Skills Test (SPS), and Attitudes Toward Science Questionnaire (Att) were administered to both experimental and control groups prior to and after the treatment." (p.69).

Given the quasi-experimental design, the researchers used analysis of variance to interrogate the various measures made before and after the intervention in both the experimental and comparison conditions.

Choosing comparison conditions

Whilst all experimental designs have certain commonalities, there are considerable differences in the kinds of educational activity considered appropriate for the control or comparison groups in different studies. Table 2 sets out a simple typology of three levels depending upon the nature of the educational input provided for the learners in a control or comparison group. The activity undertaken with a group of learners that could potentially be educative is here referred to as a 'treatment'. In experimental work the experimental/intervention group is subject to a treatment that differs in some well-characterised way from the treatment of the control or comparison group. The three levels suggested in Table 1 set different tests for the experimental treatment. These are, in effect,

- does it have any educational value? (level 1);
- is it better than standard educational provision? (level 2);
- how does it compare to what is already recognised as good practice? (level 3).

[Table 2 about here]

Table 2: Distinct levels of control in experimental designs according to the nature of the educational 'treatment' experience by the control or comparison group.

As with most typologies used to analyse complex phenomena, it is not suggested that all relevant studies will fit clearly within one of the three categories, but rather that the typology offers a useful starting point for thinking about this aspect of studies. Some examples of studies that might be categorised according to these levels are summarised in Table 3, and discussed below.

[Table 3 about here]

Table 3: Examples of different 'levels' of control condition

Does the experimental treatment have any educational effect?

The first level of experimental design suggested in Table 2 simply looks to see if outcomes on some educational measure are better after some treatment than in a matched group of learners who did not experience any treatment. This level of design is potentially useful when the research question concerns whether there is any value in introducing some new educational provision or resource that would be additional to current provision. That is, this type of study is not concerned with doing something differently, but rather whether there is sufficient value in committing additional resources to do something *extra*, that is not currently done, to consider recommending it should be *added* to existing educational provision.

One example of this type of study (see Table 3) was reported by Moore, Graham and Diamond (2003) who conducted "a randomised controlled trial to test the effectiveness of a teacher-led intervention to improve teenagers' knowledge of emergency contraception" (p.673). The intervention comprised a lesson to be delivered to 14-15 year old students following a two-hour teacher development input. This intervention was to be given as something additional to existing sex education provision:

> the chosen control group treatment for the emergency contraception trial was that control group schools would be asked to continue with their existing sex education provision, whilst those randomised to the intervention group would be asked to continue with normal sex education, and *to additionally receive* the in-service training and deliver the emergency contraception lesson (p.681, *emphasis added*)

For Moore and colleagues this was a principled decision:

> that based on what is known at the start of the trial, control group participants should not be offered something known to be less effective than
>
> (i) what the intervention group receive, or

(ii) what they would have received if the trial were not taking place (p.681)

The decision to frame this research in terms of (what is described here) as a level 1 study means that all Moore and colleagues could test was whether the additional lesson added value *over and above* existing provision. However, as it was considered that existing provision was deficient (i.e., that students were not effectively learning about an important topic in their standard sex education provision) and so some kind of additional input on this topic was needed to augment standard provision, this was a sufficient and suitable test.

Another 'level 1' type experimental study was reported by Hong, Lin, Chen, Wang and Lin (2013). They implemented a 24-hour intervention programme of "inquiry-based aesthetic science activities" over twelve weeks (see Table 3). No special curriculum activity was offered in parallel for the students in the comparison group, so positive outcomes reported by Hong and colleagues (in terms of 'learning goal orientation' and attitude to science, p.231) reflect the value added by the intervention as an additional extra-curriculum opportunity.

The study cited earlier by Leuchter, Saalbach and Hardy (2014) testing a curriculum intervention in the topic of floating and sinking (see Table 3) included a "control group that participated in a pre- and post-test, but not in an implementation of the curriculum on floating and sinking" (p.1758). In that study, teachers "were asked to follow their usual curriculum [but not] offer any curriculum on floating and sinking between pre- and posttests" (p.1762). Leuchter, Saalbach and Hardy reported positive results for their study. The group of learners who experienced the learning experiences provided by the curriculum intervention showed significantly better outcomes than the group of learners who had not been provided with any relevant learning experiences. This kind of design can be valuable where there might be theoretical grounds to doubt whether an educational intervention could have any significant effect. In the context of Leuchter, Saalbach and Hardy's study such arguments might be that learners of this age are too young to benefit from educational experiences of this kind, or that teachers of the lowest age grades generally lack the necessary specialist knowledge or skills to support learning of abstract scientific concepts. The control condition here acts as a check against the possibility that measured gains in the treatment could be explained by such possible effect as learning from the pre-test, spontaneous learning from general experience, or general cognitive development due to maturation (factors discussed earlier in this review). A useful feature of the report of this study, in common with the work of Moore, Graham and Diamond (2003), is that the report offers a clear rationale for why this level of control ('level 1', cf. Table 2) was chosen.

Does the intervention represent an improvement on current practice?

The second type of experimental design represented in Table 2 concerns the testing of an innovation which is conjectured to offer an *improved* form of educational provision in relation to some specific educational aim(s). In this situation the innovation is compared with what is considered the 'standard' or 'normal' form of provision. An example of this type of study would be that of Grooms, Sampson and Golden (2014) where the use of enquiry-based undergraduate laboratories was compared to what the researchers considered a traditional ("cookbook") approach (see Table 3). Grooms, Sampson and Golden compared outcomes ("the quality of students' arguments", p.1416) after two groups of students had experienced a semester of laboratory work. The raters who scored the student responses to the instruments used as pre- and post-tests were not aware which sets of responses related to each of the two conditions, an appropriate precaution to avoid any unconscious bias in the analysis. (This was then 'single blind': the students themselves would have been aware whether or not they were being taught in a novel condition, as would the teaching staff). In another study that can be characterised as having a level 2 control group (see Table 3), Bramwell-Lalor and Rainford (2013) ensured that the use of concept maps in the experimental treatment was balanced by equivalent time spent on more customary learning activities in the control condition.

When Yin, Tomita and Shavelson (2013) investigated learning progression-aligned formal embedded formative assessment (see Table 3), they set up the teaching in the comparison condition to be as close to that in the experimental condition as possible, to the extent of having the same teacher teach the same activities to both groups. They even included additional "curriculum-specific extension activities" (p.531) for the comparison students to offer a relevant learning activity to substitute for the formative feedback activities undertaken by the experimental group. Arguably, Yin, Tomita and Shavelson's study somewhat exceed the characteristics of a level 2 study (level 2+, perhaps) and approaches the next level, both because it ensured the comparison group were taught as similarly as possible to the innovative treatment group, and as it provided relevant additional learning opportunities for the comparison group learners to balance the specific intervention-relevant activity in the experimental group.

How does an innovation compare with currently recognised good practice?

Yin, Tomita and Shavelson's study design approaches the third type of experimental design in Table 2 that sets a higher standard for an innovation to be measured against. Where at the first level researchers seek to find out if some educational treatment has some effect in comparison to no treatment at all; and at the second level researchers look to see if an innovative approach has a more positive effect than standard provision; at the third level a comparison is made with educational provision considered to reflect good

practice. In effect, researchers are asking if an innovation is as good as, or even better than, something that is already considered to be effective.

An example of this type of design would be a study reported by Bunterm, Lee, Ng Lan Kong, Srikoon, Vangpoomyai, Rattanavongsa, et al. (2014) who compared learning in classes instructed according to the model of enquiry learning recommended by the Thai national ministry (see Table 3). The researchers provided lesson plans according to this model which were adapted according to either structured or guided enquiry. That is, the treatments varied in the extent to which the teacher directed student decision-making during the exploration, explanation and elaboration phases of the enquiry cycle. Here, then, the authors compared different implementations of recommended good practice.

Another study which might be understood as falling in this category was carried out by Chen and colleagues with undergraduates in electronics (Chen, Chang, Lai, & Tsai, 2014). In this study (see Table 3) both treatments involved (i) using a pre-test to diagnose student's misconceptions relating to diodes; (ii) providing them feedback from the pre-test; (iii) providing training in using electronic teaching materials designed to address such misconceptions; and then (iv) use of those learning materials. The difference was in the form the instructional materials took: in one case directly providing remedial information, and in the other engaging students in the P-O-E (Predict-Observe-Explain) sequence in working through the same content. All students experienced aspects of good teaching practice: a diagnostic exercise to check perquisite learning and instructional materials designed to address identified misconceptions.

Guidance on selecting control conditions: logical considerations

The choice between (a) level 1 control conditions where a teaching innovation is compared with a treatment without teaching (or where standard teaching that is supplemented by an additional teaching input is compared with only the standard provision) and (b), level 2 and 3 control conditions that offer an equivalent level of teaching input intended to meet the same educational objectives as the innovatory treatment, will derive from the motivation for the study. In many teaching contexts there will be existing provision which, even if not considered effective, will be assumed to bring about learning objectives to some extent. In these situations, a level 1 control condition is of limited use as such a study will simply show that the tested teaching treatment produces some level of learning - something that is to be expected (as even mediocre teaching is likely to facilitate some level of learning), and, without a meaningful comparison with existing practice, offers little guidance for teachers.

The choice between levels 2 (the comparison treatment being standard provision) and 3 (the comparison treatment being recognised good practice), may depend upon what the innovation is hoped to provide. If existing provision is considered to draw upon too high a resource level, or is found to have some

undesirable effects, then seeking an alternative that is just as effective may be well-motivated. So, a hypothetical school level biology course using animal dissection might lead to satisfactory levels of learning of anatomy, but lead to a minority of students declining to take part. In such a situation an experiment to test an alternative to dissection may only be seeking to find an approach that produces learning outcomes that are *as good as* in the comparison condition. In this situation, current standard practice provides an effective comparison condition and there is a sensible rationale for a 'level 2' control (see Table 2).

Many published studies argue that the innovation being tested has the potential to be more effective than current standard teaching practice, and seek to demonstrate this by comparing an innovative treatment with existing practice that is not seen as especially effective. This seems logical where the likely effectiveness of the innovation being tested is genuinely uncertain, and the 'standard' provision is the only available comparison. However, often these studies are carried out in contexts where the advantages of a range of innovative approaches have already been well demonstrated, in which case it would be more informative to test the innovation that is the focus of the study against some other approach already shown to be effective.

These different situations are summarised in Table 4.

[Table 4 about here]

Table 4: Guidance on the logic of selecting control conditions

Guidance on selecting control conditions: ethical considerations

Education has values at its core, and educational researchers should always pay particular attention to research ethics: the potential consequences that their actions could have for others. Participants (and suitable gatekeepers, when participants are children) in educational research studies should always give voluntary, informed, consent - but researchers retain a major responsibility for the ethics of experiments as participants cannot be assumed to fully understand the background and nature of the research in the way the researchers do. Teachers and educational researchers should in particular seek to avoid doing anything that is likely to harm those they are working with (Taber, 2014a). In most educational research experiments of the type discussed in this article, potential harm is likely to be limited to subjecting students (and teachers) to conditions where teaching may be less effective, and perhaps demotivating. This may happen in experimental treatments with genuine innovations (given the nature of research). It can also potentially occur in control conditions if students are subjected to teaching inputs of low effectiveness when better alternatives were available. This may be judged only a modest level of harm, but - given that the whole

purpose of experiments to test teaching innovations is to improve teaching effectiveness - this possibility should be taken seriously.

This leads to two general recommendations:

Firstly, often there will be some scope for interpretation in deciding, on the basis of the logic of a study, whether to set up a research study with level 2 or level 3 control (see Table 3). Where this choice is unclear, the ethical imperative would suggest seeking to set up a level 3 study as this has the most potential to benefit participants. In general, participants in comparison conditions should never be treated merely as sources of data.

Secondly, it is good practice to seek to offer an innovation to the control condition where possible. This may either mean offering this to those assigned to the control condition after the study (Moore et al., 2003; Ruthven et al., 2016), or setting up a design where participants all experience the experimental condition at some point in the study (e.g., see Figure 3). Such a design has methodological as well as ethical strengths. For one thing it offers two discrete tests of the treatment being investigated. It also somewhat mitigates any uncontrolled differences between the two groups. If, by chance, one group would learn more effectively across a wider range of conditions, then this design avoids that group being exclusively either the experimental or comparison group.

[Figure 3 about here]

Figure 3 - A compensatory research design where both groups experience the innovation

One of the questions raised in designing a study is whether the innovation can reasonably be *expected* to be effective. By the nature of an experimental test this should be unknown at the start of the study, and in the natural sciences 'bold' conjectures are said to be potentially the most informative (Popper, 1989). Yet, clearly, it would be ethically questionable to set up a large-scale study to test a genuine innovation were there not some good grounds to hypothesise this would lead to positive outcomes. There needs to be a balance of considerations between the risks of carrying out experiments with untested teaching approaches based on overly bold conjectures, and of setting up experimental 'tests' that will only demonstrate what has already become well accepted to be the case.

The former situation risks poor educational outcomes in the experimental treatment. The latter situation uses valuable resources ineffectively, and inconveniences participants despite having little scope for developing new knowledge. Yet, most new studies of teaching innovations are to some degree looking to

replicate findings from existing published studies. This reflects the key issue of the extent to which it is possible to generalise from the results of educational experiments.

Generalising from experimental studies

The issues considered so far in this article have in particular concerned the question:

How can we be confident that the difference in measured outcomes from an educational experiment reflects differential effectiveness of the treatments compared, rather than some other factor(s)?

In this section a rather different question is considered:

Assuming we are confident that the difference in measured outcomes from an educational experiment reflects differential effectiveness of the treatments in the context studied, how can we also be confident the differential effectiveness would be found in other contexts?

That is, how can we know that the result of an educational experiment can be generalised beyond its original context, to justify recommending that the evaluated innovation should be adopted more widely.

Reporting effect sizes

Even when an educational experiment offers statistically significant results that indicate that an innovation was effective in bringing about desired educational outcomes, this may not be a good enough reason to suggest wider implementation. Innovations tend to have resource costs - such as retraining teachers or publishing and disseminating new resources - and so it must also be judged that any gains will be 'cost-effective'. It is in the nature of statistical significance that although it indicates a difference between treatments which is unlikely to occur just by chance, this does not mean the difference is substantial. It is possible (especially where the samples include large numbers of the units of analysis) for a difference that is modest in absolute terms to reach statistical significance. In "education research studies that compare different educational interventions, effect size is the magnitude of the difference between groups" (Sullivan & Feinn, 2012, p. 279), and it is good practice for reports of educational experiments to quote an effect size for statistically significant results. As one example, in the study by Koksal and Berberoglu (2014) discussed above, the researchers reported a significant effect of the treatment on student achievement, process skills, and attitudes, but also report that although these effects all reached significance, "the effect size in achievement measure is small" (p.75).

We can consider a hypothetical educational experiment in some specific classrooms and schools, to test some teaching innovation which has been well designed and carried out, and which has reported statistically significant effects with large effect sizes. This suggests the intervention resulted in a substantial effect which seems unlikely to be a statistical fluke: but poses questions of potential generalisation.

- On what basis can we assume that the results are relevant to *other* classrooms and schools?
- Is it sensible to recommend changes in other teaching contexts that may be quite different from those involved in the study on the assumption that the same effect will be observed?

The assumption that the results of an experiment should apply beyond the specific sample tested can be based on assumptions about the kind of entities the units of analysis are; or on statistical inference; or may rely on comparisons of similarity between contexts. Each of these possibilities is considered below.

Natural kinds and theoretical generalisation

In the natural sciences, the units of analysis are usually examples of what are called 'natural kinds' (LaPorte, 2004), such that in terms of certain 'essential' qualities it can be assumed that what is found with one specimen applies to any other specimen of that kind. Science text books and data books reflect this assumption when they report ionisation enthalpies for different elements, electrical conductivities of different metals, the charge on any electron, the skeleton structure of (any) frog, and so forth. This is a kind of *theoretical* generalisation where what is found to be the case for some particular specimen or sample is considered to apply to other specimens based on theoretical considerations about what makes these different specimens to be of the same kind.

Life scientists may expect more variation within a natural kind (say, a species) than physical scientists, but even here techniques may be used that work with particular 'strains' or genetic lines (Knorr Cetina, 1999) so that different specimens of the same type are very similar in their responses to experimental interventions. It may still be inappropriate to assume that what is found with one mouse or one bacterium can be generalised to all, so a larger number of specimens may be randomly assigned to experimental and control conditions and statistical techniques used to compare outcomes across the two conditions - which is superficially similar to many of the educational studies discussed in this review.

In the natural sciences, then, theoretical considerations allow us to assume that certain measurements made on one specimen will apply to others of the same kind, or at least (in the life sciences) that average differences between conditions would apply to other samples of specimens of that kind. What are considered as natural kinds and which properties are essential qualities of such kinds have to be determined. For example, in many ways the chemical elements and compounds offer prototypical examples

of natural kinds. Yet, for certain particular purposes, samples of elements must be considered as mixtures of several kinds (isotopes). The failure to recognise two different kinds in the drug thalidomide (that is, assuming two different enantiomers could be considered to be the same natural kind for the purposes of drug production) led to tragic outcomes (Fabro, Smith, & Williams, 1967). In general, however, this kind of generalisation has been very effective. Just one of myriad examples would be that once *the composition and geometry* of *ammonia molecules* has been established, this can be assumed to apply to all *ammonia molecules*.

In educational studies, however, the units of analysis are not considered to be natural kinds that can be taken to share common properties to this extent. Social kinds, such as learners, teachers, classes, and the like, differ from each other in a great many ways, so there are few useful common properties that once measured on one specimen or sample can be assumed to apply more generally across that social kind.

Statistical generalisation

Research in education (and the social sciences more widely) cannot usually assume the units of analysis can be treated as natural kinds: what is found out about this particuloar 15 year old learner, this physics class, this novice science teacher, cannot be assumed to apply to any 15 year-old learner, any physics class, or any novice science teacher. It is known that learners, classes, or teachers vary across a whole range of variables that may impact on teaching and learning - so theoretical generalisations (e.g., something was found to be the case with one biology class so it will be the case for all biology classes) cannot be made based on the basis of social kinds given such diversity within the 'same' kind (be that biology classes, university chemistry teachers; children attending primary school science clubs, etc.).

Instead, a form of statistical generalisation is often used, where the results of an educational experiment tell us something about what is typically the case with, say, 15 year old learners, physics classes, or novice science teachers. Results therefore offer guidance on what is *likely to* be the case more generally, more often than not, rather than what has been shown to *always* be the case with these kinds. Moreover, as explained below, such forms of generalisation strictly rely upon following particular procedures.

When the design of an educational experiment cannot support statistical generalisation, then there is greater doubt over whether the results of an educational experiment can offer guidance beyond the specific samples involved in the study to other samples of the same kind. However, in these cases it may be possible to offer what is known as 'reader generalisability' supporting what is sometimes labelled analytical generalisation. This will be considered below (see 'Replication studies'), where the issue of the role of replication of experiments in generalising results is discussed.

Strict conditions for statistical generalisation

The importance of randomisation of units of analysis to the different conditions in true experiments was explained above, as this gives an assurance that differences identified in outcomes are unlikely to be due to chance differences in the make-up of the different groups. Even if such conditions are met, this does not ensure that valid results from a specific trial are relevant beyond the sample involved in the research.

Where statistical generalisability is intended, researchers need to:

a) identify a specific population that the trial is intended to be relevant to

b) ensure that those selected for the study experiment comprise a fair sample of the wider population

If the implications of studies are to be clear, it is good practice for research reports to be explicit about precisely what population was sampled. One of the studies listed in Table 1 reports that "the population of the study consists of all 397 pre-service science teachers studying at a state university in Turkey, 121 of which participated in the study making the sample 30% of the population" (Taşlidere, 2013, p. 147). However, many studies have titles or research questions implying a broad population (e.g., 'students') where the sample is drawn from a very particular context (see Table 1). Often, it is left to reader to infer the population that results are intended to generalise to.

Sampling

Ideally statistical generalisation is supported by selecting a random sample of the population of interest, which gives the strongest grounds for considering results from the trial to reflect a general pattern that would be found across the wider population (see Figure 4). Selecting the units of analysis at random from the population (so each unit that is part of the population has an equal chances of being part of the study) avoids the need to understand the diversity of the population (what the relevant variables are, and how they are distributed in the population) in a parallel way to how randomly assigning units to conditions avoids the need to characterise and then show equivalence between the groups in the different conditions.

[Figure 4 about here]

Figure 4: When an experiment tests a sample drawn at random from a wider population, then the findings of the experiment can be assumed to apply (on average) to the population

However, it is often not feasible to be able to identify all units in a population, let alone ensure they are potentially included in a sample. So, whilst this would be the ideal situation, few educational trials achieve it. Alternatively, statistical generalisation could be supported by an argument that a non-random sample is representative of the wider population on those variables most likely to be relevant to outcome - based for example on findings from surveys of the population. As there may be a range of potentially relevant factors, which may interact, building a representative sample can be challenging.

In many small-scale studies that only involve a few classes or schools (cf. Table 1), an inherently weaker design is often employed, where units of analysis are chosen to be fairly typical of the wider population, to avoid obvious 'outliers', but this does not strictly allow statistical generalisation to a wider population. An example would be Chen and colleagues study (see Table 1) where they located their study in a school "ranked around 14 of 28 high schools in Taipei" (Chen et al., 2014, p. 915). Other studies may report having used 'convenience' sampling, i.e., where researchers can easily access the research site and necessary permissions are readily forthcoming, such as Yin and colleagues' work in a "a laboratory charter school [that] includes a focus on educational research as part of its charter" (Yin et al., 2013, p. 538). Access to research sites can be elusive, so convenience sampling may be justified, but this approach may not offer the most informative samples (see below).

Variation within a population

Even when statistical generalisation is possible, this does not imply that a teaching innovation found to be advantageous in the experiment would also be *universally* advantageous if implemented throughout the population sampled, only that it would *on average* be expected to produce better outcomes (see Figure 4). So, the implications are probabilistic. If a certain approach to teaching natural selection was found to give greater learning outcomes in a RCT based on a random sample of the population of secondary age classes in Florida, then this suggests that if the approach was implemented across Florida, it would (subject to the various caveats discussed earlier in the article) improve average learning outcomes in the state. A teacher in a particular school in Florida working with a particular class cannot be confident the innovation would improve learning gains in her class, but in the absence of any other direct evidence, she could reasonably assume that introducing the innovation will *probably* lead to greater learning gains. Where probabilistic evidence is all that is available, it can be the best guide for informing action.

One discussion of a large-scale education intervention programme for disadvantaged children in the United States ('Follow Through') reports how the programme evolved into "a series of 'planned variations' of education" that allowed 17 models of schooling for disadvantaged children to be compared (Guthrie, 1977, p. 240). Thirteen of the models offered sufficient data for comparisons to be made based on a "battery of

tests…to encompass basic skills, cognitive/conceptual development, and affective factors", and effectiveness was "judged by whether a model surpasses its control group in a particular site on a particular category of outcome test" (p. 241-2). This allowed the most and least effective programmes to be identified. Even though this enabled a form of overall ranking to be produced, it was noted that

> We should be alerted to the fact that no program was successful everywhere it was tried…All of the programs were successful in at least one location on at least one class of outcome, indicating that local effects are extremely important (Guthrie, 1977, p. 243)

It was noted above that in the *epiSTEMe* project the experimental classes who studied the electricity module did not outperform the comparison classes: indeed the mean of class average learning gains (deferred post-test - pre-test) was slightly greater in the control cognition, albeit by a non-significant amount (Ruthven et al., 2016). What is perhaps more noteworthy is the range of outcomes in the two conditions - as Figure 5 shows, there was a wide range of learning gains in both conditions. Indeed, this was wider (including two classes showing *reductions* in average test score after teaching) in the intervention condition where all the classes were intended to follow the same scheme of work, including prepared teaching slides and common learning activities supported by the same printed learning resources (Taber et al., 2016). Perhaps the most reasonable conclusion to be drawn in this case is that the independent variable (the teaching scheme for studying the topic) appeared to be much less critical for determining learning than other factors that varied between the classes, and their teachers and schools.

[Figure 5 about here]

Figure 5: Results from a randomised trial showing the range of within-condition outcomes (Taber et al., 2016)

Replication studies

It seems that then that: (a) it may be difficult to set up experimental studies that meet the requirements to allow statistical generalisation of study findings to the wider population of interest, as random sampling of broad populations is seldom feasible, and building representative samples of broad populations (e.g., of secondary schools in England; of graduate chemistry teachers; of freshers on engineering degrees in Australia, etc.) is also challenging (see Figure 6); and (b) there may be such diversity within social kinds such as schools, teachers, or classes, that even when statistical inference is possible in general terms, it is likely that what is true on average for some identified population will not apply to all its members.

[Figure 6 about here]

Figure 6: Many educational experiments do not meet the conditions that allow statistical generalisation to a wider population

It is also useful to bear in mind that given that statistical significance only implies that an experimental outcome was *unlikely* to be due to chance, and there is always the possibility of false positives, as a small proportion of statistically significant results will have occurred by chance. A school or teacher considering changing practice in the light of an innovation that has been shown in an experiment to give statistically significantly better outcomes can be assured that, as $p<0.05$, this result is *probably* not just a fluke (although even then it could be due to systematic effects that could not be controlled for, as discussed earlier). However, inevitably, a small proportion of positive experimental outcomes are simply due to chance effects that are never absolutely ruled out by the statistics. Choosing a more stringent confidence level as the criterion for significance (e.g., $p<0.01$) would reduce the incidence of false positives (see Figure 7), but would also lead to more genuine effects not reaching the cut-off (i.e., more false negatives). Given these various challenges to generalising from educational experiments, replication studies can be informative in building up the evidence-based to support research-based practice.

[Figure 7 about here]

Figure 7: Choice of confidence level reflects a balance between admitting false positives (due to chance events) and false negatives (where real effects are not distinguished from chance events)

Replication in the natural sciences

There is a general principle in scientific research that experimental results need to be replicated before they are widely accepted. As suggested earlier, natural science studies so-called 'natural kinds' (LaPorte, 2004) where it is possible to generalise based on theoretical considerations. Millikan (1999, pp. 48-49) explains that

> in the case of many sciences, observations need to be made of only one or a very few exemplars of each kind studied in order to determine that certain properties are characteristic of the kind generally. If I have determined the boiling point of diethyl ether on one pure sample, then I have determined the boiling point of diethyl ether. If the experiment needs replication, this is not because some other sample of diethyl ether might have a different boiling point but because I may have made a mistake in measurement.

Replication in science then is in part concerned with whether the published report fairly describes the work: was sufficient care taken in carrying out and reporting the research such that readers can take the

published account as an accurate description of what happened, and therefore what will happen if the experimental conditions are recreated.

Educational studies might seem to be facing additional challenges, given, as discussed above, researchers cannot automatically assume that findings with one social kind (say, classes of 13-14 year old learners studying mechanics) are generalisable across the kind (from classes studying in Sweden, say, to classes studying in Singapore). Learning can be influenced by a wide range of factors, and teaching contexts vary considerably. Teachers looking to adopt evidence-based teaching practice work in very different institutions with their different norms, with students of different ages, and spreads of attainment (not to mention levels of interest and motivation), in a range of language and cultural contexts. Research that shows a particular technique, approach, or resource, seems to be effective in one classroom cannot be assumed to necessarily imply it should be adopted in other classrooms, with other teachers, working with different groups of students. Testing replicability across teaching contexts is therefore valuable.

This seems, *prima facie*, quite different from the rationale for undertaking replication in the natural sciences. Yet research into scientific practices actually suggest that replication in science is usually subtler than the notion of simply attempting to precisely repeat the original experiment. It has been argued, based on both the examination of historical cases, and observations of contemporary scientific research, that follow-up studies are seldom straight replications (Collins, 1992; Shapin & Schaffer, 2011). Indeed, simple replications may be perceived as lacking the originality expected for reporting in top journals (Franco, Malhotra, & Simonovits, 2014). In practice, it seems replication in science does not necessarily require precise replication of conditions. In the natural sciences, certainly the physical sciences, replication is more about extending and developing the original findings: can they be reproduced with modified apparatus, or with a wider range of materials, or under broader conditions. This offers a strong parallel with the situation in education.

Replication in a local educational context

Studies undertaken in education to replicate published experimental studies may be of two kinds, which have been labelled as theory-directed and context-directed (Taber, 2013a). As these labels suggest, theory-directed research is primarily intended to contribute generalisable knowledge to the research literature, whereas context-directed studies are concerned with improving the situation in a specific teaching context. Such context-directed studies are often carried out by teachers in their own classrooms, to address recognised issues and problems and improve some aspect of teaching and learning - perhaps using action research approaches (Hammersley, 2004).

In context-directed studies, teachers may often adopt ideas from published (i.e., theory-directed) research to test out whether recommendations are transferable to the specific local context - asking questions of

the form 'would that work in this school?'; '…with this class?'; '…in teaching this topic?', etcetera. As may be appreciated, the 'burden of proof' (i.e., the strength of a case argued from evidence built from the analysis of systematically collected data) is somewhat less demanding when the aim is to see if something works well in a particular teaching context, rather than seek to argue that it can be assumed to be likely to be effective more widely across a wide range of contexts. In particular, in context-directed research there is no need to make a case for the representativeness or typicality of the classroom(s) where the study was carried out.

Some of the challenges to validity discussed earlier in this article cease to be relevant in context-directed studies. For example, if a teacher is enthusiastic about an innovation, believing it has great potential to improve teaching and learning, then this might bias the outcomes of any trial. However, in that particular context, any positive outcomes from a trial of the innovation reflect the actual conditions where practice will be informed by the trial - and as long as the teacher remains enthusiastic for the innovation, any positive gains observed may well be maintained. The particular context may be atypical - it may comprise mainly of gifted learners, or of a high proportion of students studying science in a second language, or of learners in a special unit for school refusers, or of long term medical patients being schooled in hospital wards…: but what matters is whether an innovation is effective in *that* context, rather than how likely it is that any results can suggest what might happen elsewhere.

Programmes of replication across diverse contexts

Studies that are theory-directed are intended to contribute to the research literature and seek to offer generalisable findings. Such studies are set up to go beyond finding out if something works in the particular context where the research was undertaken, to instead make a case for the specific findings being relevant more widely. As was suggested above, generalisation beyond the research site can never be simply assumed, but it is possible to design studies to strengthen the case that findings are of wider relevance.

When there is a series of studies testing the same innovation, it is most useful if collectively they sample in a way that offers maximum information about the potential range of effectiveness of the innovation. There are clearly many factors that may be relevant. It may be useful for replication studies of effective innovations to take place with groups of different socio-economic status, or in different countries with different curriculum contexts, or indeed in countries with different cultural norms (and perhaps very different class sizes; different access to laboratory facilities) and languages of instruction (Taber, 2012). It may be useful to test the range of effectiveness of some innovations in terms of the ages of students, or across a range of quite different science topics. Such decisions should be based on theoretical considerations.

Given the large number of potentially relevant variables, there will be a great many combinations of possible sets of replication conditions. A large number of replications giving *similar results* within a small region of this

'phase space' means each new study adds little to the field. If all existing studies report positive outcomes, then it is most useful to select new samples that are as different as possible from those already tested. However, if replication contexts all simultaneously vary across a large number of factors, and outcomes *vary widely* (the innovation being more or less or not effective in different studies) this may also offer limited guidance to teachers hoping to learn from the research. When existing studies suggest the innovation is effective in some contexts but not others, then the characteristics of samples/context of published studies can be used to guide the selection of new samples/contexts (perhaps those judged as offering intermediate cases) that can help illuminate the boundaries of the range of effectiveness of the innovation. Progress in the field will then be best facilitated by a principled programme that complements existing studies by deliberately seeking to build systematically upon published studies when selecting the contexts of further replications.

Guidelines for supporting analytical or reader generalisation

This leads to two general guidelines for those seeking to undertake replications into innovations that have already been shown to be effective in published studies. The first concerns the theoretical justification for the importance of the study. So, for example, if an experimental study has already suggested that 11th grade students in one particular geographical location benefit from cooperative learning strategies when studying the topic of electricity (Acar & Tarhan, 2007), then researchers carrying out a replication study in the same city with 9th grade students studying the topic of metallic bonding (Acar & Tarhan, 2008, see Table 1) might be expected to discuss in theoretical terms why this modest degree of shift in the context is likely to be informative.

A second recommendation is that contexts need to be well-characterised. If researchers carefully consider the results of previous trials of an innovation in relation to the specific contexts of those studies when planning their own research, then the community of researchers can collectively build up a body of research which incrementally explores the range of effectiveness of different innovations. For this to occur, it is important that reports of teaching experiments are sufficiently detailed, not just in terms of technical matters, but also in terms of the specific teaching and learning contexts where the work takes place.

Given that such programmes can only explore the multidimensional extent of the range of effectiveness of a particular innovation incrementally, offering detailed contextual background to such studies can also support what has been labelled reader generalisability. Teachers reading research reports that offer 'thick description' (Geertz, 1973) of the research context are put in a strong position to answer the question 'how similar is the context of this study to my own teaching situation?' which may inform a decision about whether to try

out the innovation in the teacher's own classroom (a context-directed study). This is referred to as reader generalisation (Kvale, 1996).

This is a point often made in discussions of studies analysing qualitative data, and in particular case studies (Stake, 1995), which do not offer traditional forms of generalisability (Taber, 2000). Part of the inherent logic of the selection of case study methodology is that each case is unique (an idiosyncratic constellation of positions on a wide range of interacting variables) and embedded in a wider context, and so an examination of a single case detailed enough to explore interactions between features can be informative. Where cases are reported in detail, reader generalisation is supported - and the use of carefully selected multiple cases allows comparisons that may reveal general patterns (Stake, 2006).

The argument here then is that large scale RCT that use representative samples from populations of interest are necessarily rare in education. What are more common are individual small-scale experiments that cannot be considered to offer highly generalisable results. Despite this, where these individual studies are seen as being akin to case studies (and reported in sufficient detail) they can collectively build up a useful account of the range of application of tested innovations. That is, some inherent limitations of small-scale experimental studies can be mitigated across series of studies, but this is most effective when individual studies offer thick description of teaching contexts and when contexts for 'replication' studies are selected to best complement previous studies.

Planning ethical comparison conditions in replication studies

This article has reviewed some key themes relating to the challenges in designing experimental studies into teaching innovations. It is clear that whilst experimental studies can be very informative, researchers have to make a wide range of decisions in setting up an experimental study, and justify these decisions when publishing reports of their work. Considering the range of potential threats to the validity of educational experiments, as discussed above, it seems unsurprising that most published studies offer results that are subject to caveats or may offer limited grounds for broad generalisation beyond the original context. Seeing individual studies as part of the incremental build-up of evidence for the general effectiveness of an approach allows users of research to acknowledge the limitations of individual studies, but come to a view based on a wider body of work.

Some of the decision-making required in designing studies is complex and subtle. It is understandable therefore that a reader may conceptualise studies quite differently from their authors, and so may potentially evaluate some of those decisions quite critically. The reader stands outside many practical and contextual considerations that influenced the researchers. Such criticism should therefore be offered with

some humility, and understanding, but may still be important where it has potential to beneficially influence future practice. In this regard, I will here argue that in recent years a particular tradition has developed of experimental studies into aspects of science teaching that are being conceptualised in a way which (a) undermines their potential to contribute to the field, and (b) tends to systematically disadvantage participants assigned to control conditions. I will refer to these as 'rhetorical' experiments (see Figure 8). It is hoped by that by drawing attention to this issue, researchers can be persuaded to shift their conceptualisation of these studies, and will modify their design (as recommended below) when planning future research.

[Figure 8 about here]

Figure 8: Rhetorical experiments are intended to demonstrate that a well-tested teaching approach works in a very specific context

Rhetorical experiments

The labelling of these studies as 'rhetorical experiments' can be understood by analogy with many of the 'experiments' that school children carry out in school science - those laboratory practical activities labelled 'experiments' that are actually demonstrations of well-characterised effects clearly described in the students' textbooks - as part of learning a "rhetoric of conclusions" (Schwab, 1958). These would be genuine experiments for the students if they had no strong expectations of the outcomes in advance, but often the practicals are undertaken after the relevant theory has been taught, rather than in advance to provide 'epistemic relevance' to motivate learning the scientific ideas (Taber, 2015), and the practical may even be entitled 'an experiment to show …'.

I am suggesting that some of the experimental studies reported in the literature are rhetorical in the parallel sense that the researchers clearly expect to demonstrate a well-established effect, albeit in a specific context where it has not previously been demonstrated. The general form of the question 'will this much-tested teaching approach also work here' is clearly set up expecting the answer 'yes'. Indeed, control condition may be chosen to give the experiment the best possible chance of producing a positive outcome for the experimental treatment. Clearly all studies have unique elements, but Figure 8 represents the general logic of many of these rhetorical experiments.

In terms of the analysis offered earlier in this article, such studies are replications, but often made without any strong grounds for suspecting that the context chosen for the study provides a real test for the

teaching innovation. That is, although the particular innovation may not have been tested in that specific context, given the range of prior studies showing it to be widely effective there is no strong reason to suspect that this particular context is sufficiently different from those where the effectiveness has already been demonstrated to motivate reasonable doubts about the outcome of the new study. This may be clear from the published reports themselves.

Some examples of rhetorical studies of this kind are presented in Table 5. What is noteworthy is that as part of the conceptual framework justifying the research readers are told fairly unequivocally that the teaching approach to be tested has already been shown to be clearly superior to (what is sometimes termed) 'traditional' teaching, yet the researchers then seek to test this in a specific context where they set up a control treatment that reflects the very traditional conditions that they have already told readers are ineffective for achieving learning objectives.

[Table 5 about here]

Table 5: Some research studies including control conditions that the researchers claim are already known to be ineffective teaching treatments

Avoiding detrimental control conditions

This raises an ethical issue in such studies that, given the current state of knowledge prior to the research, the researchers employ a control treatment that is considered to be of limited educational value. Students in the control condition are *expected* to be disadvantaged compared to those in the experimental condition. Authors often justify this by reporting that the suboptimal conditions set up for the control are just what these students would experience anyway, and so they are not disadvantaged *compared to not being in the study*. That is only so if authors are correct that 'traditional' teaching, with no elements of more 'progressive' approaches, is endemic in the local context. Whilst studies may present traditional and progressive teaching as being a dichotomy, actual observations of teachers' classroom practice suggest practice is more nuanced and reflects a blend of these two extremes (Bektas & Taber, 2009). These rhetorical studies nominally have level 2 controls (see Table 2) but if a teacher of a control class is asked to "transmit information to students, who receive and memorise it", with "no consideration of the students' existing conceptions", and where learners are 'passive' (see Table 5 for examples) then this may actively prevent teachers engaging any progressive elements that might be part of their normal teaching repertoires. So, these experiments may in practice be better designed as having 'level 2- (two minus)' controls (cf. Table 3).

Quite a few studies of this kind have been reported from Turkey (perhaps unsurprising as it is now one of the most active nations in science education research) where 'reform' teaching along constructivist lines has

been recommended for many years now (Gözütok, 2013). These recommendations have been supported by government policy, changes in teacher education, and a great many studies demonstrating how reform-based teaching can improve learning outcomes. Despite this, study authors often argue that this has not widely impacted teaching practice, and so employing 'traditional' teaching as a control treatment is not detrimental to study participants compared with not taking part in the research. If this is indeed so, then it seems unlikely that one more study demonstrating the greater effectiveness of some progressive teaching approach will persuade teachers in that context to change teaching practices. If researchers are planning studies of this type because they hope to act as catalysts for change, then this strategy is not working.

Good practice in selecting productive control treatments

The framework for thinking about experimental studies into teaching developed in this article suggests a different approach is indicated. Even if it is accepted that control conditions used in rhetorical experiments of this kind do not offer any less educational value than the teaching the particular learners would experience normally, educational researchers who wish to influence teaching practice should decline to adopt such conditions in their studies. In these rhetorical experiments, teachers assigned the experimental classes are prepared to teach using research-informed approaches aligned with reform policies ('are prepared to' both as in 'are trained up to', and as in 'are willing to'), so researchers are certainly able to demonstrate their success in showing individual teachers both that they can teach in these ways, and that such approaches can be effective with their classes.

Acar and Tarhan (2007) comment on the teacher in their study that "because she was experienced on active learning, she adapted the study easily" and as part of her preparation for working with the intervention group, she "was informed about the misconceptions related to electrochemistry and told about which activities had been developed to prevent which misconceptions" (p.353). Yet she was asked to teach the parallel control class "without consideration for student misconceptions". So, a teacher experienced in reform teaching approaches was asked to restrict her professional practice to the detriment of her students, so as to artificially produce a control condition where learning was likely to be limited.

Researchers in these educational contexts should therefore seriously consider looking to abandon testing well-established innovations in new contexts by using nominally level 2 (and perhaps actually level 2-) control conditions, and to instead plan studies with level 3 control conditions (see Table 4). If researchers are working in a context where teachers are expected to adopt 'reform' teaching approaches, then researchers should not undermine this by accepting teaching treatments in control conditions that clearly do not meet the expected educational standards (and so simply demonstrate, once again, the substandard nature of such teaching). Rather, educational researchers should act as change agents, training-up teachers to

offer a range of well-tested teaching approaches in their classes, and then seeking to compare between these to explore which of these superior approaches works best in teaching particular groups of students specific aspects of the curriculum.

Conclusions

This article has reviewed some key issues in designing and interpreting experimental studies intended to test different teaching innovations. Experimental research employing statistical tools is often seen as being more objective than studies based on interpretation of qualitative data, and findings quantified in terms of effect sizes and p values seem to offer definitive results. Yet, all research choices (e.g., how to implement an intervention, how to operationalise a variable, which instruments to use to collect data) involve interpretations, and most studies in education involve some compromises on ideal research designs. Few experiments in education offer large randomly selected or truly representative samples from clearly defined and identified populations, and even such ideal cases can be subject to some potential threats to validity that randomisation cannot overcome.

This certainly does not imply that experiments are not useful, but they are best seen as most informative alongside other types of studies that that have complementary strengths and weaknesses (Taber, 2009) - for example studies that collect detailed data exploring classroom processes. Experimental research of the kind reviewed in this article tests a specific hypothesis about the potential effect of some specific treatment (such as a particular pedagogy or teaching resource). The hypothesis will be based on some theoretical model of how some variable has a causal influence on outcomes of interest (e.g., how pedagogy influences learning). Even when a hypothesis is supported by statistical analysis, that analysis offers no direct support for concluding that the conjectured causal mechanism explains the outcome.

Teaching and learning are complex phenomena. As an example, it may be conjectured that implementing a form of problem-based learning could lead to increases in school test scores because students show greater engagement in classes due to higher motivation, or because it allows a level of peer interaction providing scaffolding of learning, or because it involves high-level thinking skills, or because the group work involved facilitates a more productive kind of discourse, or … A simple experimental study comparing teaching treatments and test scores and finding the problem-based learning condition resulted in significantly better outcomes could not distinguish which mechanism was at work. It is possible several such mechanisms are operating, perhaps synergistically: if students are more motivated and better engaged then they are more open to working outside their existing areas of competence where scaffolding may be effective, and may be more open to productive exploratory discourse - and so forth. Studies that collect data on a wide range of process variables can be used to construct mathematical models using techniques such as structural

equation modelling which offer insights into such complex situations (Schreiber, Nora, Stage, Barlow, & King, 2006), but these studies require more extensive quantitive data (as well as expertise in the methods) than simple experiments, and still require advanced knowledge of the variables that will be measured and included in a model

Processes can also be investigated by 'qualitative' studies using more interpretivist modes of enquiry. Studies that observe teaching, collect classroom talk, and interview teachers and students, can offer valuable indications of productive educational processes (Duit, Roth, Komorek, & Wilbers, 1998; Petri & Niedderer, 1998). These studies may suffer a complementary weakness to experimental studies: so factors identified as salient in qualitative data may not always have a substantive influence on educational outcomes (that needs to be tested); just as showing a specific educational treatment is effective does not imply understanding the causal mechanism at work (an unidentified, confounding, factor could be the cause). Exploratory interpretive studies can be open to considering multiple explanations and to adopting a range of theoretical perspectives to support data analysis (Taber, 2008). Progressing a research programme may then be supported by complementing experimental studies with more interpretive work that can both suggest hypotheses to text experimentally and also question whether the assumed mechanisms underpinning experimental hypotheses seem feasible in terms of what is actually observed in different treatment conditions.

For readers to fully evaluate the implications of experimental studies it is important that authors offer clarity about the units of analysis, the population sampled, what (if anything) has been assigned randomly, and the method used to achieve any randomisation, as well as detailed accounts of the different treatments. As small-scale studies undertaken in particular contexts offer limited inherent generalisability, these should be planned with careful consideration of how they will add to the body of studies testing that particular type of innovation and so contribute to a better understanding of its range of effectiveness. That decision requires a careful examination of both the outcomes and contexts of existing studies to determine what, if any, patterns can be identified for the range of application of the innovation. When researchers report such studies, they should explain the choice of research site and classroom context to help readers appreciate how the new study adds substantially to those previously reported. Context-directed research carried out by teachers in their own classrooms can be justified by the general research question 'will this widely-tested innovation be effective in this particular very specific context where I teach' (Taber, 2013a), but in published research authors should also explain why the particular context has been chosen to be of theoretical interest.

A particular issue arising from the studies reviewed is the choice of control conditions. Comparing an innovation against standard practice is appropriate when the likely effectiveness of the innovation is genuinely uncertain, but when researchers test an approach that has already been widely demonstrated as effective across a broad range of contexts then it is usually more informative to compare it with a

treatment already recognised as good practice. The use of control conditions that reflect teaching that the researchers themselves believe is ineffective, or which is incompatible with local educational policies, should be avoided. Given the current state of knowledge about teaching and learning (Bransford, Brown, & Cocking, 2000; NGSS Lead States, 2013), it seems unlikely that many teachers have classroom practice which fully matches the caricature of 'traditional', 'teacher-centred' practice. Therefore, asking teachers to teach control groups this way (often whilst simultaneously demonstrating competence in much more progressive practice in teaching an intervention group) is difficult to justify ethically or logically.

It is hoped that that this review will provide a framework for reading reports for teachers who may wish to draw upon the research literature to identify innovations that they might consider adopting or testing in their own classrooms, as well as raising some issues that researchers themselves may usefully reflect upon when deciding when to employ an experimental design, or planning an experimental study.

References

Abdi, A. (2014). The Effect of Inquiry-based Learning Method on Students' Academic Achievement in Science Course. *Universal Journal of Educational Research, 2*, 37-41. doi:10.13189/ujer.2014.020104.

Acar, B., & Tarhan, L. (2007). Effect of Cooperative Learning Strategies on Students' Understanding of Concepts in Electrochemistry. *International Journal of Science and Mathematics Education, 5*(2), 349-373. doi:10.1007/s10763-006-9046-7

Acar, B., & Tarhan, L. (2008). Effects of Cooperative Learning on Students' Understanding of Metallic Bonding. *Research in Science Education, 38*(4), 401–420. doi:DOI 10.1007/s11165-007-9054-9

Adey, P. (1999). The Science of Thinking, and Science For Thinking: a description of Cognitive Acceleration through Science Education (CASE). Geneva: International Bureau of Education (UNESCO).

Adey, P., & Shayer, M. (2002). An exploration of long-term far-transfer effects following an extended intervention program in the high school science curriculum. In C. Desforges & R. Fox (Eds.), *Teaching and Learning: The Essential Readings* (pp. 173-209). Oxford: Blackwell Publishing.

Al-Rawahi, N. M., & Al-Balushi, S. M. (2015). The Effect of Reflective Science Journal Writing on Students' Self-Regulated Learning Strategies. *International Journal of Environmental & Science Education, 10*(3), 367-379.

Ausubel, D. P. (1978). In Defense of Advance Organizers: A Reply to the Critics. *Review of Educational Research, 48*(2), 251-257.

Barab, S. A., & Luehmann, A. L. (2003). Building sustainable science curriculum: Acknowledging and accommodating local adaptation. *Science Education, 87*(4), 454-467. doi:doi:10.1002/sce.10083

Bektas, O., & Taber, K. S. (2009). Can science pedagogy in English schools inform educational reform in Turkey? Exploring the extent of constructivist teaching in a curriculum context informed by constructivist principles. *Journal of Turkish Science Education, 6*(3), 66-80.

Berger, R., & Hänze, M. (2015). Impact of Expert Teaching Quality on Novice Academic Performance in the Jigsaw Cooperative Learning Method. *International Journal of Science Education, 37*(2), 294-320. doi:10.1080/09500693.2014.985757

Bramwell-Lalor, S., & Rainford, M. (2013). The Effects of Using Concept Mapping for Improving Advanced Level Biology Students' Lower- and Higher-Order Cognitive Skills. *International Journal of Science Education, 36*(5), 839-864. doi:10.1080/09500693.2013.829255

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How People Learn: Brain, mind, experience & school* (Expanded ed.). Washington D C: National Academy Press.

British Educational Research Association. (2018). *Ethical Guidelines for Educational Research* (4th ed.). London: British Educational Research Association.

Bunterm, T., Lee, K., Ng Lan Kong, J., Srikoon, S., Vangpoomyai, P., Rattanavongsa, J., & Rachahoon, G. (2014). Do Different Levels of Inquiry Lead to Different Learning Outcomes? A comparison between guided and structured inquiry. *International Journal of Science Education*, 1-23. doi: 10.1080/09500693.2014.886347

Çam, A., & Geban, Ö. (2011). Effectiveness of Case-Based Learning Instruction on Epistemological Beliefs and Attitudes Toward Chemistry. *Journal of Science Education and Technology, 20*(1), 26-32. doi:10.1007/s10956-010-9231-x

Chen, S., Chang, W.-H., Lai, C.-H., & Tsai, C.-Y. (2014). A Comparison of Students' Approaches to Inquiry, Conceptual Learning, and Attitudes in Simulation-Based and Microcomputer-Based Laboratories. *Science Education, 98*(5), 905-935. doi:10.1002/sce.21126

Collins, H. (1992). Changing order: Replication and induction in scientific practice: University of Chicago Press.

Dorman, J. P. (2012). The impact of student clustering on the rsults of statistical tests. In B. J. Fraser, K. G. Tobin, & C. J. McRobbie (Eds.), *Second International Handbook of Science Education* (Vol. 2, pp. 1333-1348). Dordrecht: Springer.

Dudai, Y., & Eisenberg, M. (2004). Rites of Passage of the Engram: Reconsolidation and the Lingering Consolidation Hypothesis. *Neuron, 44*(1), 93-100. doi:10.1016/j.neuron.2004.09.003

Duit, R., Roth, W.-M., Komorek, M., & Wilbers, J. (1998). Conceptual change cum discourse analysis to understand cognition in a unit on chaotic systems: towards an integrative perspective on learning in science. *International Journal of Science Education, 20*(9), 1059-1073.

Fabro, S., Smith, R. L., & Williams, R. T. (1967). Toxicity and Teratogenicity of Optical Isomers of Thalidomide. *Nature, 215*, 296. doi:10.1038/215296a0

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505. doi:10.1126/science.1255484

Gauld, C. (1989). A study of pupils' responses to empirical evidence. In R. Millar (Ed.), *Doing Science: images of science in science education* (pp. 62-82). London: The Falmer Press.

Geertz, C. (1973). Thick Description: Toward an Interpretive Theory of Culture *The Interpretation of Cultures: Selected Essays* (pp. 3-30). New York: Basic Books.

Gidena, A., & Gebeyehu, D. (2017). The effectiveness of advance organiser model on students' academic achievement in learning work and energy. *International Journal of Science Education, 39*(16), 2226-2242. doi:10.1080/09500693.2017.1369600

Goldacre, B. (2013). *Building Evidence into Education*. Retrieved from London: http://media.education.gov.uk/assets/files/pdf/b/ben%20goldacre%20paper.pdf

Goswami, U. (2008). Cognitive Development: The Learning Brain. Hove, East Sussex: Psychology Press.

Gözütok, F. D. (2013). Curriculum Studies in Turkey Since 2000. In W. F. Pinar (Ed.), *International Handbook of Curriculum Research*: Routledge.

Grooms, J., Sampson, V., & Golden, B. (2014). Comparing the Effectiveness of Verification and Inquiry Laboratories in Supporting Undergraduate Science Students in Constructing Arguments Around Socioscientific Issues. *International Journal of Science Education, 36*(9), 1412-1433. doi:10.1080/09500693.2014.891160

Günter, T., & Alpat, S. K. (2017). What is the Effect of Case-Based Learning on the Academic Achievement of Students on the Topic of "Biochemical Oxygen Demand?". *Research in Science Education*. doi:10.1007/s11165-017-9672-9

Guthrie, J. T. (1977). Research Views: Follow through: A Compensatory Education Experiment. *The Reading Teacher, 31*(2), 240-244.

Hammersley, M. (2004). Action research: a contradiction in terms? *Oxford Review of Education, 30*(2), 165-181.

Hong, Z.-R., Lin, H.-s., Chen, H.-T., Wang, H.-H., & Lin, C.-J. (2013). The Effects of Aesthetic Science Activities on Improving At-Risk Families Children's Anxiety About Learning Science and Positive Thinking. *International Journal of Science Education, 36*(2), 216-243. doi:10.1080/09500693.2012.758394

Howe, C., Ilie, S., Guardia, P., Hofmann, R., Mercer, N., & Riga, F. (2014). Principled Improvement in Science: Forces and proportional relations in early secondary-school teaching. *International Journal of Science Education, 37*(1), 162-184. doi:10.1080/09500693.2014.975168

Knorr Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, Massachusetts: Harvard University Press.

Koksal, E. A., & Berberoglu, G. (2014). The Effect of Guided-Inquiry Instruction on 6th Grade Turkish Students' Achievement, Science Process Skills, and Attitudes Toward Science. *International Journal of Science Education, 36*(1), 66-78. doi:10.1080/09500693.2012.721942

Kvale, S. (1996). InterViews: An introduction to qualitative research interviewing. Thousand Oaks, California: Sage Publications.

LaPorte, J. (2004). *Natural Kinds and Conceptual Change*. Cambridge: Cambridge University Press.

Leuchter, M., Saalbach, H., & Hardy, I. (2014). Designing Science Learning in the First Years of Schooling. An intervention study with sequenced learning material on the topic of 'floating and sinking'. *International Journal of Science Education, 36*(10), 1751-1771. doi:10.1080/09500693.2013.878482

Millikan, R. G. (1999). Historical Kinds and the "Special Sciences". *Philosophical Studies, 95*(1), 45-65. doi:10.1023/a:1004532016219

Moore, L., Graham, A., & Diamond, I. (2003). On the Feasibility of Conducting Randomised Trials in Education: Case Study of a Sex Education Intervention. *British Educational Research Journal, 29*(5), 673-689. doi:10.2307/1502117

Mortimer, E. F., & Scott, P. H. (2003). Meaning Making in Secondary Science Classrooms. Maidenhead: Open University Press.

National Research Council Committee on Scientific Principles for Educational Research. (2002). *Scientific Research in Education*. Washington DC: National Academies Press.

Next Generation Science Standards: For States, By States. (2013). The National Academies Press.

NGSS Lead States. (2013). Next Generation Science Standards: For States, By States: The National Academies Press.

O'Donnell, C. L. (2008). Defining, Conceptualizing, and Measuring Fidelity of Implementation and Its Relationship to Outcomes in K–12 Curriculum Intervention Research. *Review of Educational Research, 78*(1), 33-84. doi:10.3102/0034654307313793

Park, S., & Oliver, J. S. (2008). Revisiting the Conceptualisation of Pedagogical Content Knowledge (PCK): PCK as a Conceptual Tool to Understand Teachers as Professionals. *Research in Science Education, 38*(3), 261-284. doi:10.1007/s11165-007-9049-6

Petri, J., & Niedderer, H. (1998). A learning pathway in high-school level quantum atomic physics. *International Journal of Science Education, 20*(9), 1075-1088.

Piaget, J. (1970/1972). *The Principles of Genetic Epistemology* (W. Mays, Trans.). London: Routledge & Kegan Paul.

Popper, K. R. (1989). Conjectures and Refutations: The Growth of Scientific Knowledge, (5th ed.). London: Routledge.

Pring, R. (2000). Philosophy of Educational Research. London: Continuum.

Rosenthal, R. (2003). Covert Communication in Laboratories, Classrooms, and the Truly Real World. *Current Directions in Psychological Science, 12*(5), 151-154. doi: 10.1111/1467-8721.t01-1-01250

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom: teacher expectation and pupils' intellectual development. New York: Holt, Rinehart & Winston.

Rosenthal, R., & Jacobson, L. (1970). Teacher's expectations. In L. Hudson (Ed.), *The Ecology of Human Intelligence* (pp. 177-181). Harmondsworth: Penguin.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: the first 345 studies. *Behavioral and Brain Sciences, 1*, 377-386. doi:10.1017/S0140525X00075506

Ruthven, K., Mercer, N., Taber, K. S., Guardia, P., Hofmann, R., Ilie, S., . . . Riga, F. (2016). A research-informed dialogic-teaching approach to early secondary-school mathematics and science: the pedagogical design and field trial of the epiSTEMe intervention. *Research Papers in Education, 32*(1), 18-40. doi:10.1080/02671522.2015.1129642

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research, 99*(6), 323-338. doi:10.3200/JOER.99.6.323-338

Schwab, J. J. (1958). The Teaching of Science as Inquiry. *Bulletin of the Atomic Scientists, 14*(9), 374-379. doi:10.1080/00963402.1958.11453895

Sesen, B. A., & Tarhan, L. (2011). Active-learning versus teacher-centered instruction for learning acids and bases. *Research in Science & Technological Education, 29*(2), 205-226. doi: 10.1080/02635143.2011.581630

Shapin, S., & Schaffer, S. (2011). Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life: Princeton University Press.

Siegler, R. S. (2004). U-Shaped Interest in U-Shaped Development-and What It Means. *Journal of Cognition and Development, 5*(1), 1-10. doi:10.1207/s15327647jcd0501_1

Stake, R. E. (1995). *The Art of Case Study Research*. Thousand Oaks, California: Sage.

Stake, R. E. (2006). *Multiple Case Study Analysis*. New York: The Guilford Press.

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education, 4*(3), 279-282. doi:10.4300/JGME-D-12-00156.1

Taber, K. S. (2000). Case studies and generalisability - grounded theory and research in science education. *International Journal of Science Education, 22*(5), 469-487.

Taber, K. S. (2003). Lost without trace or not brought to mind? - a case study of remembering and forgetting of college science. *Chemistry Education: Research and Practice, 4*(3), 249-277.

Taber, K. S. (2008). Of Models, Mermaids and Methods: The role of analytical pluralism in understanding student learning in science. In I.V. Eriksson (Ed.), *Science Education in the 21st Century* (pp. 69-106). Hauppauge, New York: Nova Science Publishers.

Taber, K. S. (2009). Progressing Science Education: Constructing the scientific research programme into the contingent nature of learning science. Dordrecht: Springer.

Taber, K. S. (2012). Vive la différence? Comparing 'like with like' in studies of learners' ideas in diverse educational contexts. *Educational Research International, 2012*(Article 168741), 1-12. Retrieved from http://www.hindawi.com/journals/edu/2012/168741/ doi: 10.1155/2012/168741

Taber, K. S. (2013a). Classroom-based Research and Evidence-based Practice: An introduction (2nd ed.). London: Sage.

Taber, K. S. (2013b). Modelling Learners and Learning in Science Education: Developing representations of concepts, conceptual structure and conceptual change to inform teaching and research. Dordrecht: Springer.

Taber, K. S. (2013c). Non-random thoughts about research. *Chemistry Education Research and Practice, 14*(4), 359-362. doi:10.1039/c3rp90009f

Taber, K. S. (2014a). Ethical considerations of chemistry education research involving "human subjects". *Chemistry Education Research and Practice, 15*(2), 109-113. doi:10.1039/C4RP90003K

Taber, K. S. (2014b). Methodological issues in science education research: a perspective from the philosophy of science. In M. R. Matthews (Ed.), *International Handbook of Research in History, Philosophy and Science Teaching* (Vol. 3, pp. 1839-1893). Dordrecht: Springer Netherlands.

Taber, K. S. (2015). Epistemic relevance and learning chemistry in an academic context. In I. Eilks & A. Hofstein (Eds.), *Relevant Chemistry Education: From Theory to Practice* (pp. 79-100). Rotterdam: Sense Publishers.

Taber, K. S. (2018). Scaffolding learning: principles for effective teaching and the design of classroom resources. In M. Abend (Ed.), *Effective Teaching and Learning: Perspectives, strategies and implementation* (pp. 1-43). New York: Nova Science Publishers.

Taber, K. S., Ruthven, K., Howe, C., Mercer, N., Riga, F., Hofmann, R., & Luthman, S. (2015). Developing a research-informed teaching module for learning about electrical circuits at lower secondary school level: supporting personal learning about science and the nature of science. In Information Resources Management Association (Ed.), *K-12 STEM Education: Breakthroughs in Research and Practice* (Vol. 1, pp. 1-28). Hershey, Pennsylvania: IGI Global.

Taber, K. S., Ruthven, K., Mercer, N., Riga, F., Luthman, S., & Hofmann, R. (2016). Developing teaching with an explicit focus on scientific thinking. *School Science Review, 97*(361), 75-84.

Taşlidere, E. (2013). The Effect of Concept Cartoon Worksheets on Students' Conceptual Understandings of Geometrical Optics. *Education & Science/Egitim ve Bilim, 38*(167).

Tüysüz, C. (2010). The Effect of the Virtual Laboratory on Students' Achievement and Attitude in Chemistry. *International Online Journal of Educational Sciences, 2*(1), 37-53.

van Driel, J. H., Beijaard, D., & Verloop, N. (2001). Professional development and reform in science education: The role of teachers' practical knowledge. *Journal of Research in Science Teaching, 38*(2), 137-158. doi:doi:10.1002/1098-2736(200102)38:2<137::AID-TEA1001>3.0.CO;2-U

Vygotsky, L. S. (1934/1986). *Thought and Language* (A. Kozulin, E. Hanfmann, & G. Vakar, Trans.). London: MIT Press.

Vygotsky, L. S. (1978). Mind in Society: The development of higher psychological processes. Cambridge, Massachusetts: Harvard University Press.

Wood, D. (1988). How Children Think and Learn: the social contexts of cognitive development. Oxford: Blackwell.

Yin, Y., Tomita, M. K., & Shavelson, R. J. (2013). Using Formal Embedded Formative Assessments Aligned with a Short-Term Learning Progression to Promote Conceptual Change and Achievement in Science. *International Journal of Science Education, 36*(4), 531-552. doi: 10.1080/09500693.2013.787556

Endnote:

1 In this article, the convention in British English spellings (preferred in *Studies in Science Education)* to use 'enquiry' as the normal spelling for the general process of investigating (as the term 'inquiry' is usually reserved for formal proceedings) is followed. This usage is different to the convention with American English spellings. Where works cited use the alternative American spelling, 'inquiry', this has been retained in direct quotations.

Experimental research into teaching innovations: responding to methodological and ethical challenges

Tables and figures

Table 1: A sample of published experimental studies testing teaching innovations

Citation	Independent variable	Dependent variable(s)	Sample	Randomisation
Adbi, 2014	Inquiry-based learning	Students' academic achievement in science	40 5th-grade students from one primary school in Kermanshah, Iran	Two intact classes (n=20, 20) assigned to conditions (same teacher)
Acar & Tarhan, 2007.	Cooperative learning	Understanding of concepts in electrochemistry	41 11th-grade students from two science classes in a high school in Izmir, in Turkey	Two intact classes (n=20, 21) assigned to two conditions (same teacher)
Acar & Tarhan, 2008	Cooperative learning	Students' understanding of metallic bonding	57 9th-grade science students from two science classes in a high school in Izmir, in Turkey	Two intact classes (n = 28, 29) assigned to two conditions (same teacher)
Al-Rawahi & Al-Balushi, 2015	Reflective science journal writing	Self-regulated learning strategies	62 10th-grade students from a public female school in the Ad Dakhiliyah region in Oman	Two intact classes (n = 32, 30) assigned to two conditions (same teacher)
Berger, R., & Hänze, M. (2015).	Expert teaching quality (jigsaw teaching)	Novice academic performance	129 12th-grade students in Nine physics classes from 7 schools in Germany	Students assigned to groups - students acted as both novices and experts during project
Bramwell-Lalor & Rainford, 2013	Concept mapping as a formative assessment tool	Advanced level biology students' cognitive skills	156 A level biology students from (three or more *) schools in Jamaica * details only provided for experimental group	None reported. (Intact classes. Three teachers and 90 students in experimental group; Five other teachers and 66 students in control condition.)
Bunterm, et al. 2014	Form of guidance provided 5E learning cycle model	Science content and process skills	183 10th-grade and 56 7th-grade students from three schools North-Eastern Thailand	Two intact classes assigned in each school (n=42, 44; 49, 48; 27, 29) Within each school, one teacher taught both classes
Çam & Geban, 2011	Effectiveness of case-based learning instruction	Epistemological beliefs and attitudes toward chemistry	63 11th-grade students from two classes of an urban high school in Turkey	Two intact classes (n=28, 35) assigned to conditions (same teacher).
Chen, Chang, Lai & Tsai, 2014	Form of instructional materials	Physics learning, enquiry behaviours, student enjoyment and engagement	68 11th-grade students in two physics classes at an urban high school in Taipei, Taiwan	Two intact classes (n=32, 36) randomly assigned to conditions (same teacher)
Gidena & Gebeyehu, 2017	Effectiveness of the advance organiser model	Students' academic achievement in learning about work and energy	139 11th-grade natural science students from a preparatory school, in Northern- West zone of Tigray region, Ethiopia	Two intact classes (n=46, 46) assigned to conditions (same teacher)
Grooms, Sampson & Golden, 2014	Enquiry-based undergraduate laboratories in relation to	Students' abilities to construct arguments relating to socioscientific issues	73 chemistry undergraduates from a two-year community college; and 79 chemistry undergraduates from a four-year university; in the same City in the Southeastern USA.	None. (College students made up intervention; and university students the comparison condition.)
Günter & Alpat, 2017	Case-based learning	Academic achievement of students on the topic of biochemical oxygen demand	18 4th or 5th year undergraduates attending the chemistry teaching programme in a university in Izmir, Turkey	Students randomly assigned to conditions (n = 10, 8)
Hong, Lin, Chen, Wang & Lin, 2013	Aesthetic science activities	At-risk families children's anxiety about learning science and positive thinking	133 4th-grade school children from two elementary schools in the Chi-Jin district of Kaohsiung city in Taiwan	36 children volunteered for the intervention; "97 typical 4th graders were randomly selected as the comparison group" (p. 222)

Leuchter, Saalbach, & Hardy, 2014	Structured learning materials	Understanding of floating and sinking	15 classes (244 children) age 4-9 years plus 2 classes (22 children) as a control group in Central Switzerland	No randomisation reported.
Moore, Graham and Diamond, 2003)	Teacher-led intervention	Teenagers' knowledge of emergency contraception	24 schools in Avon, South-West England who responded to a invitation to all 49 eligible schools partake in the study	12 schools assigned to each condition
Ruthven et al, 2016	Design of teaching units	Learning and attitudes	11-12 year old pupils in 70 intact classes in schools in Eastern England	25 schools schools assigned to two conditions
Sesen & Tarhan, 2011	Active-learning versus teacher-centered instruction	Learning acids and bases	45 [sic] high-school students (average age 17 years) from two different classes in a high school in Turkey.	Two intact classes (n=21, 25) assigned to conditions (same teacher)
Taşlidere, 2013	Concept cartoon worksheets	Students' conceptual understanding of geometrical optics	121 pre-service science teachers, sophomores (2nd year undergraduates), taking General Physics-III at a state university in Turkey	Two intact classes (n=63 58) were assigned to each condition (same lecturer)
Tüysüz, 2010	Virtual laboratory	Students' achievement and attitude in chemistry	341 9th-grade high school students in Turkey	Students divided into two groups (n=174, 167)
Yin, Tomita & Shavelson, 2013	Learning progression-aligned formal embedded formative assessment	Conceptual change and achievement in middle-school science	52 6th-graders from a university laboratory school in Honolulu, Hawaii	Students assigned to conditions (n=26, 26)

Table 2: Distinct levels of control in experimental designs according to the nature of the educational 'treatment' experience by the control or comparison group

Type	Experimental group	Control/comparison group	Purpose
Level 1: treatment vs.no treatment	A treatment is applied which is hypothesised to have an educational effect	Outcomes for the experimental group are compared with outcomes for a matched group not receiving any relevant educational treatment	To test whether a particular form of treatment leads to educationally desirable outcomes
Level 2: innovation vs. standard treatment	An innovative treatment is applied which is hypothesised to have a greater educational effect than the standard treatment	Outcomes for the group subject to the innovation are compared with outcomes for a matched group subject to the relevant standard educational input	To test whether an innovative form of treatment leads to greater educational outcomes than current practice
Level 3: innovation vs. enhanced treatment	An innovative treatment of unknown efficacy is applied	Outcomes for the group subject to the innovation are compared with outcomes for a matched group subject to a treatment recognised as good practice	A treatment is tested to see how effective it is compared to another treatment previously shown to be effective

Table 3: Examples of different 'levels' of control condition

Citation	Focus	Experimental treatment	Comparison condition	Level characterisation
Moore, Graham and Diamond, 2003	An intervention to improve teenagers' knowledge of emergency contraception	An extra lesson to be delivered to 14-15 year-old students in addition to existing normal sex education	No supplement to existing sex education provision	Level 1
Hong, Lin, Chen, Wang and Lin, 2013	Intervention programme of inquiry-based aesthetic science activities	12 weeks programme of extra-curricular activities: "hands-on pedagogical strategy", "inquiry teaching theory" and "aesthetic understanding teaching method"; and including "introductory hands-on activities, displays, team competitions, peer tutoring, small group discussions, demonstrations of scientific myths, and aesthetic science activities" (p.222)	No relevant extra-curricular provision	Level 1
Leuchter, Saalbach and Hardy, 2014	Curriculum intervention in the topic of floating and sinking	"An instructional design with sequenced and problem-based tasks which are supposed to stimulate conceptual change in the area of 'floating and sinking' in children in the first years of schooling…[enacted through] a structured and problem-based learning environment…[during] a 4-week experiment-based instruction" (p.1757)	"Usual curriculum"to *exclude* "any curriculum on floating and sinking between pre- and posttests" (p.1762)	Level 1
Grooms, Sampson and Golden, 2014	Construct arguments relating to socio-scientific issues	"A series of [six] argument-based lab activities" alongside 5 of "more 'cookbook' style" "a chemistry laboratory course aligned with the argument-driven inquiry" (p.1412) that emphasised "scientific argumentation, group collaboration, and peer review" (p.1417)	All eleven laboratory activities followed the "more traditional laboratory approach" p.1412 "instruction followed a more 'cookbook' style, where the students were provided the steps needed to complete each investigation and typically worked as individuals" (p.1417)	Level 2
Bramwell-Lalor and Rainford, 2013	Concept mapping as a formative assessment tool in developing students' higher level cognitive skills	Concept mapping added to the teaching of topics by "lectures, discussion and practical work."	"The same biology curriculum during the period under study. The topics that they were taught was done over the same time period as the treatment groups … [through] lectures, discussion and practical work" (pp.850-851)	Level 2
Yin, Tomita and Shavelson, 2013	"Learning progression-aligned formal embedded formative assessment on conceptual change and achievement in middle-school science"	Formal formative assessments added to teaching provision	Equal amount of time on the same day gathering additional data and discussing patterns found in their experiment	Level 2+

Bunterm, et al., 2014	The 5E Learning Cycle Model	Enquiry learning following lesson plans adapted to support guided enquiry,	Enquiry learning following lesson plans adapted to support structured enquiry,	Level 3
Chen, Chang, Lai, & Tsai, 2014	Using a pre-test to diagnose student's misconceptions relating to diodes	Responding to diagnosed alternative conceptions by engaging in the P-O-E (Predict-Observe-Explain) sequence	Responding to diagnosed alternative conceptions by providing students with remedial input	level 3

Table 4: Guidance on the logic of selecting control conditions

Context of study	Type of control condition
There are question about whether the teaching innovation can lead to learning gains in the context (e.g., students may be too young to benefit)	Level 1 - comparison with learners not receiving any teaching
It is unclear if it would be beneficial to provide some supplementary input in addition to current standard provision	Level 1 - comparison with learners not receiving any supplement to standard teaching
There is genuine uncertainty about the potential of the teaching intervention to lead to learning outcomes as positive as those obtained by current practice (i.e., the innovation has yet to be tested in any reasonably comparable context)	Level 2 - comparison with learners receiving standard teaching
An innovation is suspected to offer potential advantages over current practice, and there are no other alternatives already demonstrated to be effective that could feasibly substitute for current practice	Level 2 - comparison with learners receiving standard teaching
An innovation is suspected to offer potential advantages over current practice, where there are other alternatives already demonstrated to be effective that could feasibly substitute for current practice	Level 3 - comparison with learners receiving an alternative teaching treatment already demonstrated to be effective

Table 5: Some research studies including control conditions that the researchers claim are already known to be ineffective teaching treatments

Citation	Intervention condition	Background assumptions	Control condition
Abdi, 2014	"Student [sic, students] in the experimental group were instructed with inquiry-based instruction supported 5E learning cycle. In the instruction based on 5E learning cycle method, teaching and learning activities and lesson plans were designed to maximize students active involvement in the learning process." (p. 39)	"The inquiry-based teaching approach is supported on knowledge about the learning process that has emerged from research…. In inquiry-based science education, children become engaged in many of the activities and thinking processes that scientists use to produce new knowledge. (p.37) "the traditional classroom often looks like a one-person show with a largely uninvolved learner. Traditional classes are usually dominated by direct and unilateral instruction. Students are expected to blindly accept the information they are given without questioning the instructor… Traditional approach followers assume that there is a fixed body of knowledge that the student must come to know. … The teacher seeks to transfer thoughts and meanings to the passive student leaving little room for student-initiated questions, independent thought or interaction between students" (p.37)	"In the control group, a teacher directed strategy representing the traditional approach was used… where students are completely passive…" (p.39) "The teacher used direct teaching and question and answer methods … In this group, the teacher provided instruction through lecture and discussion methods to teach the concepts. The teacher … wrote notes on the chalkboard about the definition of concepts, and passed out worksheets for students to complete. The primary underlying principle was that knowledge takes the form of information that is transmitted to students. …" (p.39)
Acar & Tarhan, 2007	"…cooperative learning instruction based on a constructivist approach" (p.353)	"Construction of the knowledge occurs best in an active learning environment. Active learning methods such as cooperative learning encourages students to be active participants in the construction of their own knowledge during the learning process…The benefits of cooperative learning for students' social and academic skills have been well documented by researchers...Based on the literature it can be said that cooperative learning based on the constructivist approach is effective for remediation of misconceptions" (pp. 351-352).	"The control group was taught [by the same teacher] with a teacher-centered traditional didactic lecture format. Teaching strategies were dependent on teacher expression without consideration for student misconceptions. …students were required to use their textbooks; students were passive participants and rarely asked questions; they did not benefit from the library or internet sources; activities such as computer animations or brainstorming were not used; generally the teacher wrote the concepts on the board and then explained them; students listened and took notes as the teacher lectured on the content." (p.358)
Acar & Tarhan, 2008	"…newly developed material based on cooperative learning instruction was used in the experimental group" (p.407) "The teacher required students to actively participate in the learning process… asking some key questions such as "What are you doing?" "Why are you doing it?" "How will it help you understanding the subject?" "Why are you researching it?" (p.407) "At the beginning of the instruction, students' groups were required to activate their prior knowledge" p.408	"…the most important factor that affects learning is the student's existing conceptions" (p.401) "The benefits of cooperative learning on students' academic and social skills have been well-documented…" (p.404)	"…the control group was taught [by the same teacher] …using teacher-centred traditional didactic lecture format. Teaching strategies were dependent on teacher expression. The students were required to use their textbooks…there are not any student centred active activities [that] depend on constructivism. Students were passive participants during the lessons and they only listened and took notes as the teacher lectured on the content" (pp. 408-409).

Çam & Geban, 2011	"The EG [experimental group] was treated with case-based learning instruction by small group format … The instruction was student-centered rather than teacher centered education. … Teacher is a facilitator who assists small groups of self-directed students as they work through a case. She kept the groups on track and stimulated the functioning of the groups. She were [sic, did] not lecture or directly teach the students. She taught students to find answers to their own questions and provided students with feedback. (p.29)	"…people construct their knowledge by actively creating their own understanding rather [than] receiving knowledge from others" (p.26) "Case based learning instruction … promotes students' active participation and students could construct their own learning." (p.26)	"Students in CG [control group] were instructed by lecturing method, discussion and sometimes students performed the laboratory activities in that students were passive listeners and teacher's role was to transmit the facts and concepts to the students.…Teacher did not give emphasis on students' misconceptions. Students were passive listeners and they were taking notes. In the laboratory activity section, students were required to do experiment by using the handout.…like "cookbook", described the all steps of the experiment (p.29).
Sesen & Tarhan, 2011	"…a variety of specific student-centered instructional strategies […including] experimental activities, brain-storming, video presentations, demonstrations, computer animations, and learning together activities that engage active participation of students in the learning process" (p.209).	"In an active-learning environment, in contrast to teacher-centered instruction, a teacher acts as a facilitator, engages active participation of students in the learning process, and puts less emphasis on memorizing information and more emphasis on inquiry through which students develop a deeper knowledge and appreciation of the nature of science … when students are actively involved in the learning task, they learn more than when they are passive recipients of instruction" (p.208).	"…teacher-centered instruction, [where] learning focuses on the mastery of content, with little development of the skills and attitudes necessary for scientific inquiry. The teacher transmits information to students, who receive and memorize it. …The curriculum is loaded with many facts and a large number of vocabulary words, which encourages a lecture format of teaching. (p.216) "…the control group were instructed via teacher-centered didactic lecture format…The students were instructed with regular chemistry textbooks. They listened to the teacher carefully, took notes and solved algorithmic problems" (p.216).
Gidena & Gebeyehu, 2017	"The lesson plan for the experimental group was prepared using the AOM. … This lesson was prepared in such a way that those students actively participated with guidance of the teacher in the starter activity, main activity, and concluding activity of the lesson." p.2233	"AOM [Advance organiser model] provides support for effective teaching and learning process …provides a framework to enable students to learn new ideas or information by meaningfully linking these ideas to the existing knowledge." (p.2227) "…theories, concepts, and techniques are better understood when lectures are accompanied with demonstration, hands-on experiments through self-discovery, and questions that require students to ponder what will happen in an experiment and why" (p.2227).	"…was taught using the lesson plan based on the conventional teaching method" (p.2226) "…the conventional teaching method, which was commonly practised in that school…in which the teacher dominants [sic], whereas the learners remain passive" (p.2233)
Taşlidere, 2013	"For the three-week treatment period, the experimental group was instructed the application of concept cartoon worksheets" p.148	"…it is reported that traditional physics instruction is ineffective in helping students develop a scientific view and their conceptual understandings … In general, the approaches encouraging active participation of learners in learning environment are thought to help students construct knowledge meaningfully" (p. 145)	"traditional instruction which relied on instructors' explanations with no consideration of the students' misconceptions. The instructor used overhead projector to show the definitions of concepts, explained the facts, solved the questions, meanwhile students took notes through the lessons" (p.154)
Tüysüz, 2010	"…taught by a constructivist based instructional approach which was enriched by computer animations at the computer laboratory" (p.43)	"As accepted throughout the world the idea of using student centred constructivist based instructional methods is widely accepted, since teacher centred, traditional instructional methods has given insufficient opportunities for student to construct their own learning. Eliciting students' individual capabilities, intelligence and creative thinking can only be achieved through student centered instructional methods" (p.37)	"…using chalk and talk method as commonly known name, the traditional method" (p.43)

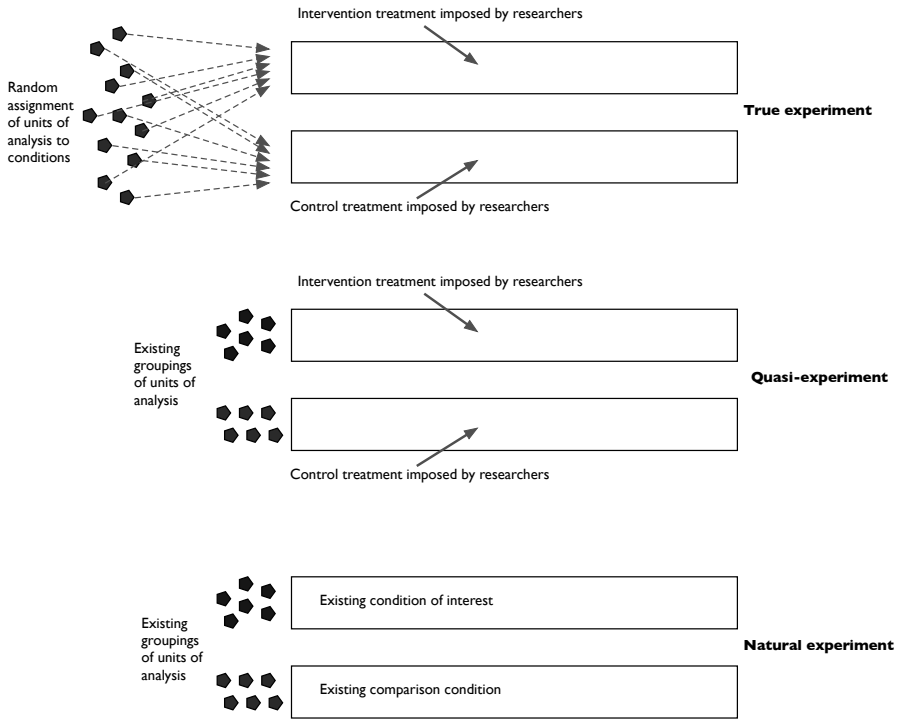

Figure 1: Experimental designs may be categorised as true experiments, quasi-experiments and natural experiments

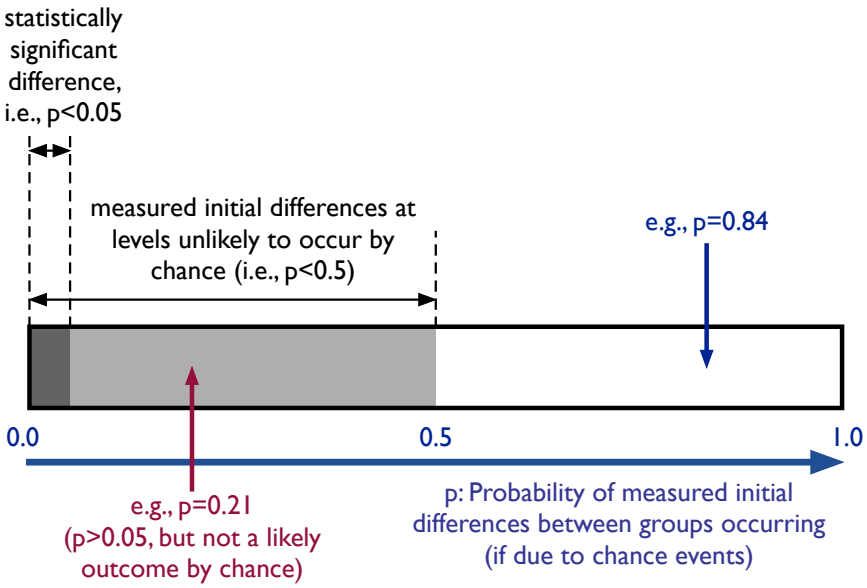

Figure 2: Evaluations of equivalence between different groups should be more rigorous than simply excluding differences reaching statistical significance

		Topic 1 instruction			Topic 2 instruction	
Group 1	Topic 1 pre-test	Intervention (innovative treatment)	Topic 1 post-test	Topic 2 pre-test	Comparison (customary treatment)	Topic 2 post-test
Group 2	Topic 1 pre-test	Comparison (customary treatment)	Topic 1 post-test	Topic 2 pre-test	Intervention (innovative treatment)	Topic 2 post-test

Figure 3 - A compensatory research design where both groups experience the innovation

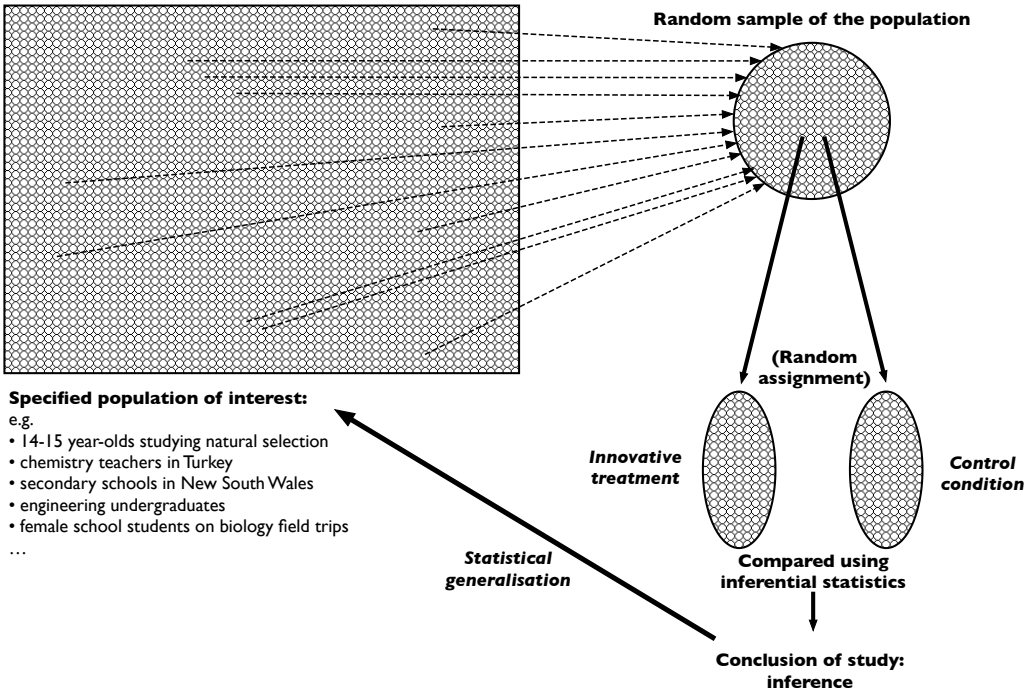

Figure 4: When an experiment tests a sample drawn at random from a wider population, then the findings of the experiment can be assumed to apply (on average) to the population

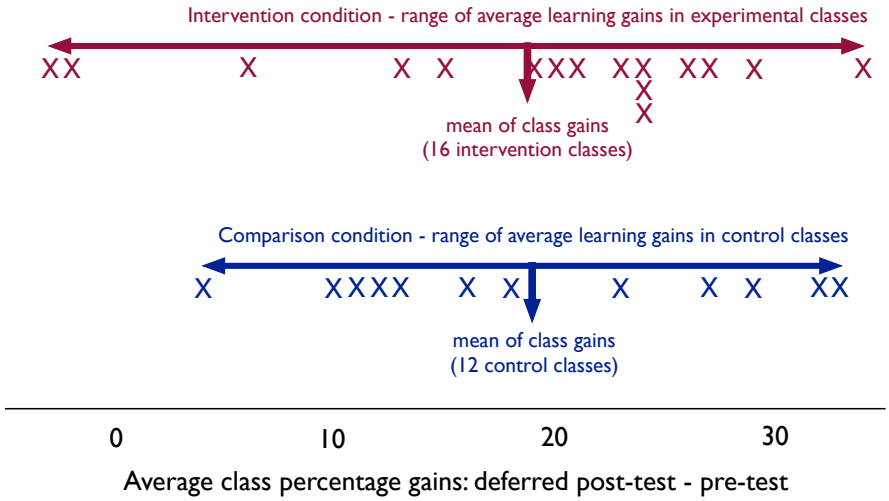

Figure 5: Results from a randomised trial showing the range of within-condition outcomes (Taber et al., 2016)

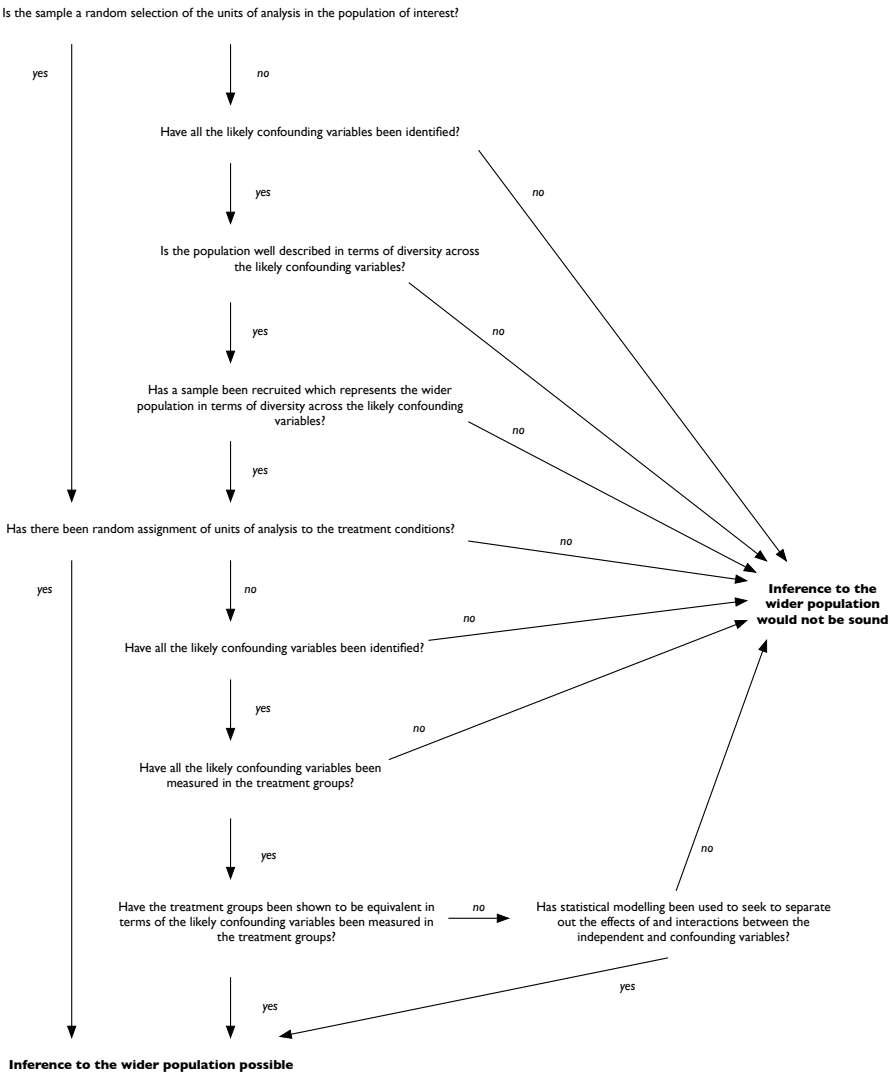

Figure 6: Many educational experiments do not meet the conditions that allow statistical generalisation to a wider population

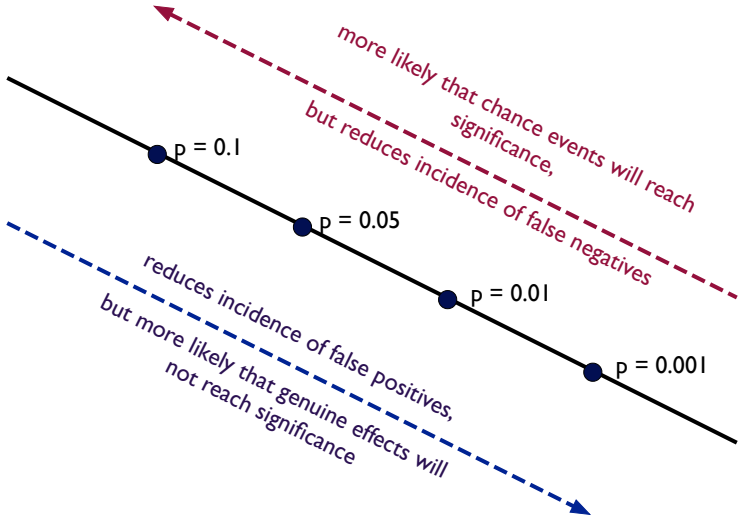

Figure 7: Choice of confidence level reflects a balance between admitting false positives (due to chance events) and false negatives (where real effects are not distinguished from chance events)

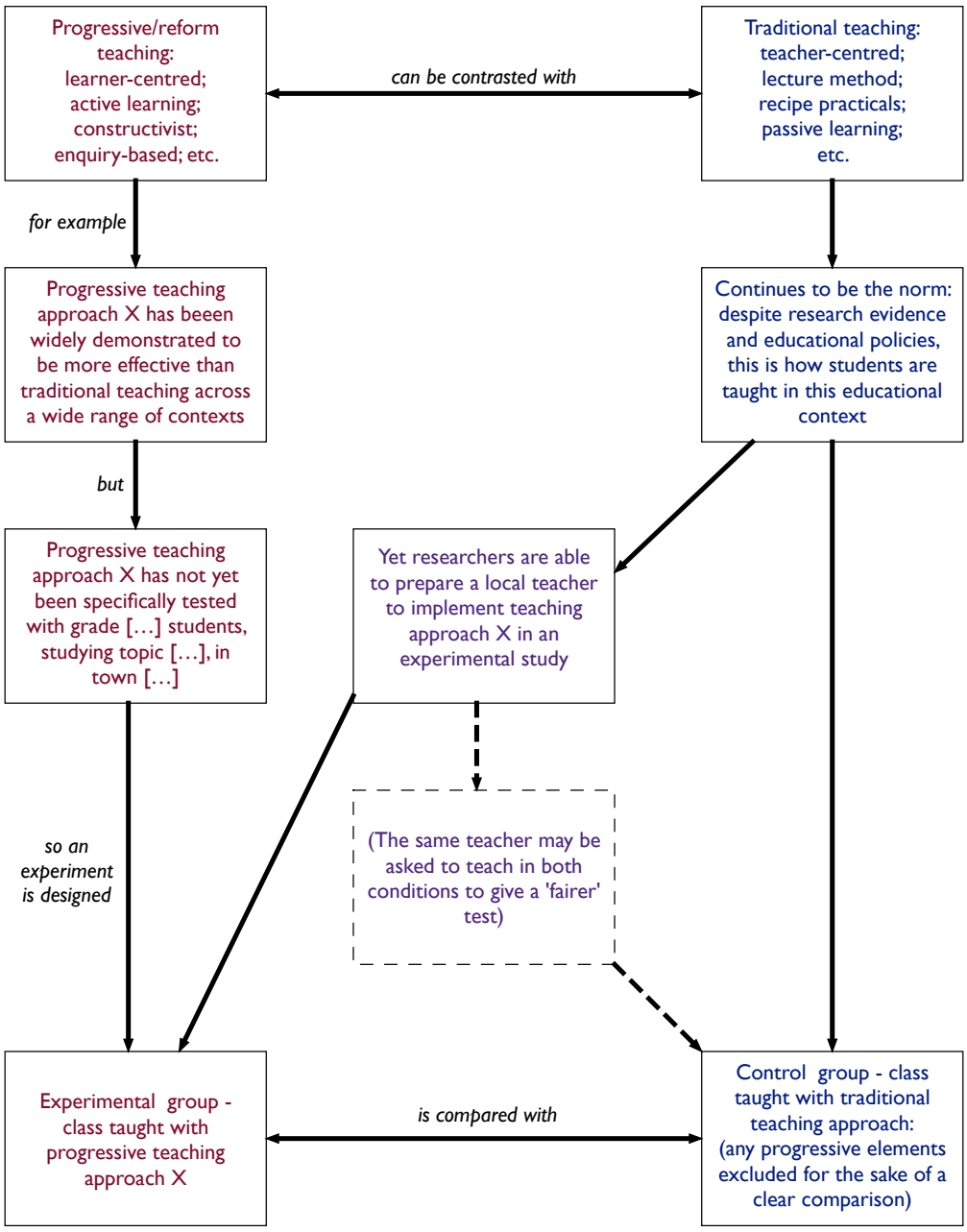

Figure 8: Rhetorical experiments are intended to demonstrate that a well-tested teaching approach works in a very specific context