Opinion Page

Opinion Page I: *Replicable, Reproducible, and Generalisable: Implications of Scientific Hallmarks for Research in Education*

Norm Lederman
Illinois Institute of Technology Chicago
Illinois
USA

Introduction

As a not so innocent undergraduate student sitting in an animal physiology class (approximately 50 years ago), my professor said prior to a laboratory investigation, 'No matter how controlled the situation is with respect to sampling, research procedures, ambient temperature, ambient lighting, organism satiety, and all other relevant variables *the rat will do whatever he damn well pleases*'. It was a humorous comment to which my classmates ascribed no significant meaning. I was reminded of my professor's comment when reading Alger's (2020) commentary on the importance of reproducibility in science in the February issue of the HPS&ST NEWSLETTER and his brief discussion of the influence of the personalities of rats and mice on experimental investigations. Indeed, my professor's comment was more poignant than I thought at the time. Additionally, Alger's discussion is also reminiscent of the 2002 debates about "scientific research in education (NRC, 2002) and the formation of the What Works Clearinghouse (WWC) by the U.S. Department of Education.

The discussion of the hallmarks of scientific research conveniently begins with the idea of 'ways of knowing'. As humans we have various approaches to knowing. Each of the academic disciplines we encounter are characterised by different ways of knowing. That is, each develops knowledge in similar, but different, ways. Some of the more common ways of knowing are appeals to authority, *a priori* knowledge, and scientific (also known as the scholarly approach). Appeals to authority are exemplified by the numerous marketing advertisements that state things such as 'nine out of 10 doctors say…' or 'Linus Pauling says that vitamin C cures the common cold' or 'John Dewey says…'

Alternatively, sometimes knowledge is established through logic. For example, 'since vitamins are good, the more the better' or 'the more subject matter knowledge a teacher has, the better they will be as a teacher'. These are examples of *a priori* knowledge and are based on logical connections to assertions that we have always known to be true. Finally, there is the category of scientific knowledge, which is a systematic, empirically based,

and self-correcting. Furthermore, scientific research generally follows three general forms: descriptive, correlational, and experimental. Education, and the social sciences, value and try to adhere to the scientific approach to knowing. Hence the knowledge established by educational research seemingly should be replicable, reproducible and generalisable. But is this the case or is it even a reasonable expectation?

What is Meant by 'Replicable' and Why is it Important?

In general, replicable refers to the ability of a researcher to exactly, as much as possible, follow the same procedures (e.g., data collection approach, instrumentation, data analysis, etc.) in an investigation to see if the same results are obtained. This does not mean that the specific statistical values are the same, but rather that the overall findings of rejecting or failure to reject hypotheses are achieved. For example, do humans with higher levels of cholesterol consistently exhibit higher levels of heart disease? If the results of an investigation can be consistently repeated, researchers have more confidence in the findings.

At the most mundane level, statistical analysis has a stated level of significance (e.g., .05 or lower), which means that any probability of .05 or lower is considered as evidence that the achieved results are the result of something other than chance. Alternatively, it can also mean that the significant results also have a .05 level of probability of occurring by chance. By convention researchers have agreed that a .05 probability of results is rare enough that something other than chance is at play. It also means, unfortunately, that if the same study is performed enough times, significant results will be achieved most of the time, both significant and non-significant results will be achieved. The researcher does not know from a single investigation whether the results achieved are part of the chance occurrence group or the rare and significant group. Consequently, it is essential that attempts to replicate findings are necessary.

If you are attempting to see if you are holding a normal two-sided coin, you could flip the coin 100 times. If the coin is "normal" you would expect the number of times it lands on one side versus the other side to be approximately equal. If we get one side appearing 95 times or higher, we would conclude that something other than chance has occurred. However, we could be wrong because there is a chance, albeit small, that the coin may land on one side 95 or more times due to chance. Hence, the importance of have replications of science investigations.

This point was stressed by Alger (2020) and we all stress this to our students in science classes as well as in our education research courses. So, it would be appropriate to conclude that replication is important in science research as well as in educational research. The one glitch is that in educational research we are usually studying animate objects as opposed to inanimate objects and "the rat will do whatever he damn well pleases." This same point is made eloquently by Cziko (1989) in his discussion of free will. It is certainly true that in many areas of science animate subjects are used. But most of the time the target of interest are molecules and their interactions. These molecules do not make mindful decisions as to how they will respond to a certain condition.

What is Meant by 'Reproducible' and Why is it Important?

Reproducibility is the main focus of Alger's (2020) commentary and he discusses the topic in much more detail than will be found here. The concept is closely related to replication, but slightly different, for practical reasons. Let us return to the previous example relating cholesterol to heart disease. Two different studies may be attempting to answer the same question. However, it is inevitable that different samples will be used and there also could be different approaches to the measurement of cholesterol levels as well as the form of the cholesterol when it was taken into the human system. For example, was the cholesterol in the same foods eaten or was there a variety of foods that lead to the ingestion of cholesterol? So, in such cases, the studies are not replications of each other. Additionally, the two studies may have handled 'outliers' differently in their data analyses. There are inevitable differences, but the results may be equivalent and lead to the same conclusion about the relationship of heart disease to cholesterol.

Consistency across these studies, as well as others, is important and leads to building confidence among researchers, even though the studies are not technically replications. The same factors that can impact conclusions reached in replication studies hold in studies that document reproducibility. As with replication, reproducing findings is valued in education research, but with important qualifications, that will be discussed later. In the end 'the rat will still do whatever he damn well pleases'.

What is meant by 'Generalisable' and Why is it Important?

The third hallmark of scientific research is generalisability. Quite simply it pertains to how well the research findings from an investigation apply to situations extending beyond the current investigation and its context. For example, do the findings of research on cholesterol and heart disease in Massachusetts apply to samples of individuals from other locations in the U.S. or other countries around the world? This is difficult to establish because of a multitude of extraneous variables (i.e., variables that affect the results of an investigation that were ignored or were unknown). Certainly, replication and reproducibility studies enhance the ability to generalise findings, but they do not guarantee it. Generalisability is a cornerstone of progress in science. It is this hallmark that lends the most confidence to research findings and marks the progress of science with respect to one or more aspects of the natural world. However, it is difficult to achieve.

Generalisability is more easily attained with inanimate objects. Most would agree that gravity functions in the same manner globally. Although the effects on the earth's surface differ because of the earth's oblate shape, the law of gravitation does not take a different form in South Africa, Israel, and in Japan. Animate objects provide another perspective, and this is especially true in educational research. Students vary with respect to gender, culture, identity, religion, cognitive ability, among others. Each of these characteristics individually, or in tandem, impact the success of a particular curriculum. This is a perennial problem for education researchers. Generalisability is generally valued, but quite difficult to achieve. Additionally, it is important to note that some qualitative

researchers make no claim that their findings are generalisable, nor do they think that it should be a goal of their research.

What is the Applicability of Scientific Hallmarks to Educational Research?

This commentary has provided a general overview of the concepts of replicable, reproducible, and generalisable in science. They are related hallmarks of scientific knowledge and goals of science as a way of knowing. However, Alger (2020) provides a much more tempered view of their importance in science, as well as their necessity. His view is refreshing.

Since educational research attempts to develop knowledge in a scientific way it would seem that the aforementioned hallmarks are important to research in education. Their importance in education must be qualified even further. The concepts are largely positivistic and posit a view of nature of science and scientific inquiry that is reminiscent of the 2002 debates about "scientific research in education (NRC, 2002) and the formation of the *What Works Clearinghouse* (WWC) by the U.S. Department of Education. These debates focused on what constitutes scientific evidence and the relative value of qualitative and quantitative research paradigms.

The film and book titled *Never Cry Wolf* chronicles an actual account of a scientific study designed to document that the decrease in the population of caribou in the arctic was due to wolves preying on the caribou. Scientific data to document this assertion did not exist and the work of Tyler (i.e, the scientist) was to provide these data and lead to a plan that would intervene by reducing the wolf population. Upon arrival in the arctic location, it became clear to Tyler that much of the supplies were not useable in the frigid north. Nevertheless, the research plan was to kill several wolves and then investigate the contents of their stomachs. Tyler did not see any wolves at first so he began to observe their scat for remnants of caribou. No remnants were found. When he finally did observe some wolves, he noted that they were not getting their daily sustenance from consuming caribou. It was clear that the original premise of the research was not adequate to solve the original problem and so he drastically altered the design of his investigation. He proceeded to observe all aspects of the wolves' behaviour; he tried to integrate himself as much as possible into the wolves' daily lives. He chose to perceive with '*what worked*'.

In the end, Tyler found that the wolves were living primarily on rodents. The times that they did prey on caribou was on the weakest and most diseased members of the heard. In short, the wolves were actually enhancing the genetic pool of the heard and helping the future survival of caribou as opposed to being a menace to their ultimate survival. The scientist's change of plans worked. He was able to answer his research question by finding 'what worked'.

What this book illustrated was that scientists pursue the answers to their questions in various ways. These approaches differ within the various sciences and vary even more across the different sciences. There is no single set or sequence of steps that scientists always follow. There is no *single scientific method* or any single approach that can be used to characterise all of science. The questions that scientists have guide their research approaches/design and scientists, within certain limits, do 'what works'. Much of the tempered view of science taken by Alger (2020) is echoed in this book.

Never Cry Wolf and Alger's (2020) commentary remind me of the policy debates about scientifically-based educational research in 2002-2003 (Berliner, 2002; Eisenhart & Towne, 2003; Erikson & Gutierez, 2002; National Research Council, 2002; Slavin, 2002; St. Pierre, 2002). These debates lead to the development of the WWC by the Department of Education, whose mantra was '*what works*'. The WWC advocated that educational research needed to be more scientific and that the reason we are in a quandary about what constitutes good practice in educational settings. It was clear that varying methods were appropriate in educational research, designs that established cause were strongly preferred. Underlying the debates was a misrepresentation of 'what is science?' This is quite similar to the position of Alger (2020) that certain attitudes about reproducibility are less than accurate.

Historically, during the latter part of the 20th century, a systematic and concerted effort to study teaching and learning was undertaken. Researchers borrowed from the same models of agricultural designs used in mainstream science. Perhaps the primary reason for this decision was the cultural status possessed by science and/or the reigning popularity of positivist thinking. Although this approach to research was virtually the same as that advocated by the WWC, it provided important foundational knowledge about teaching and learning, but it was clearly limited. In depth understandings of teachers' thought processes and how students mediate instructional experiences were not accessible through such means. Educators realised that many questions remained, and new questions had arisen. In their search to find out 'what works', they needed to consider alternative research approaches that worked. The situation was really no different than what confronted Tyler and his wolves.

Educational researchers in all areas began to view classrooms as systems and cultures. They began to see the importance of the dynamic interactions among participants (i.e., teachers and students), as groups and as individuals. Borrowing from anthropology and sociology, educational researchers began to research instruction from a totally different perspective than what was afforded by the 'scientific' agricultural designs (Howe, & Eisenhart,1990). The situation is not much different than the shift from reductionist to systems thinking in environmental studies. As a consequence, there are few today who do not realise the difficulty in generalising educational research from one class to another (with the same teacher), let alone generalising across teachers, schools, states, and countries.

Interestingly, classroom teachers have known this all along. Most teachers' complaints about research findings that failed to resonate with their local situations were in response to rigorously quantitative studies that over generalised in deference to sampling theory. The difficulty, or the impossibility, of generalising as is commonly conceived is aligned with Alger's (2020) as well as within current thinking by educational researchers.

Misconceptions some have about the existence of a single scientific method aside, there are other problems with the application of classical experimental scientific research designs to classroom situations. It is absolutely critical, if one wants to imply cause, to carefully control or account for extraneous variables in research. There are problems, however, when you are dealing with situations involving living organisms that exhibit voluntary or free will and individuals that react differently, for a variety of reasons, to the same environmental conditions/stimuli. Remember my

physiology professor's rat and how it behaved.

There has been a history of attempts to conduct carefully controlled experiments in classroom settings. However, the situation becomes so contrived that little external validity can be ascribed to the investigation. Quite simply, the situation is so deviant from general classroom life across settings that attempts to generalise to other situations have become futile at best. Much of the research conducted in 'laboratory schools' suffered from this problem. The research, in and of itself, was fine for the specific situation, but generalising to other populations was difficult. Nevertheless, the WWC would like to pursue this path again, to the degree that they place little value on designs that do not attempt to make definitive causal claims.

The WWC, and those promoting that educational research become more scientific, claim to be moving educational practices toward a 'medical model'. The medical clinical trial has been posited as the 'gold standard'. That is, educational practitioners are asked to seek the results of 'scientifically controlled studies (like clinical trials)' to make instructional decisions. To be clear, most medical research, for ethical reasons, does not follow experimental research models. It is simply not acceptable to randomly solicit participants for an investigation and then randomly assign them to treatments, one of which has potential harm. In most cases, medical research involves ex-post facto designs (e.g., heart disease studies, smoking/cancer studies), which are correlational by nature.

Surely, many readers will now want to direct my attention to the plethora of experimental studies in medicine that involve human models (substitutes for humans in terms of physiology). Surely, they would say, all of the research done on various drugs and medicines began with experimental studies on rats or other mammals with the only inference being the similarities between the physiology of the human and the physiology of the animal being used as a model. In this case, my detractors would be absolutely correct.

However, there is a vast difference between generalising results of experimental medical studies using human models and generalising experimental studies in education. The studies with drugs, medicine, etc., involve inanimate effects in the sense that what is involved is the interaction of various molecules within the physiological systems of the human or human model. In education we transcend the organic level and have to grapple with motivation, free will, emotions, attitudes, etc.

Certainly, inanimate factors influence all of these human characteristics, but virtually everyone interested in learning beyond a passing curiosity knows learning to be far more complex. I don't know anyone who would currently assume that using a particular teaching approach with birds would generalise to human learning. Wasn't this the problem that we all had with operant conditioning and the work of behaviourists? When it comes to complex thinking, human behaviour is just not that simple. Or it is not simple enough to allow the high levels of predictability and generalisability.

It is interesting to note that there was a period of time in recent history when experimental studies in learning involving human models was in vogue. Do you remember the investigations in which worms or rodents were taught certain skills and then were sacrificed and fed to other animals of the same type? This approach was entrenched in the belief that learning was organic and learning could be transferred through the transfer of

organic material. Again, the medical model of experimental research only holds for investigations involving the inanimate, not such things as complex learning in humans or other mammals.

Conclusions

By now you may have concluded that educational researchers must be total relativists and would not admit to any progress in our knowledge of teaching or learning. Nothing could be further from the truth. We are strong supporters of the value of both quantitative and qualitative research. My colleagues and I also believe that studies with small sample sizes can be as valuable as studies with large sample sizes. The most critical issue is the relationship among the research questions, research design, and the nature of the data collected.

Research questions should guide design and data choice. Researchers should pursue "what works," and this depends on the question being asked, not some idealised scientific method that is incorrectly purported to be the only method to produce scientific evidence. In addition, it is critically important that all researchers remain intimately aware of the assumptions embedded in their research questions, designs, and analyses and the implications these assumptions have for results being replicable, reproducible, and generalisable to the rest of the world. Such a perspective is consistent with the views of Alger (2020) about the expectations for scientific knowledge.

Overall, although the intentions of the WWC and those advocating that educational research become more scientific are admirable, the advocacy is flawed for at least two critical reasons. There is a clear underlying (and sometimes not so subtle) belief that scientific evidence can only be provided by causal research designs (aka The Scientific Method) and that research findings from studies of teaching and learning can be reproducible, replicable, and generalised freely across contexts and situations if derived from rigorously controlled studies scientific studies. In our attempts to enhance teaching and learning from systematically collected empirical evidence let us never lose sight of the unpredictability and indeterminate nature of human behaviour (Cziko, 1989).

References

Alger, B.E. (2020). Is non-reproducibility a crisis for science? *HPS&ST NEWSLETTER*, February, 2020, pp. 9-17.

Berliner, D.C. (2002). Educational research: The hardest science of all. *Educational Researcher*, 31(8), pp. 18-20.

Cziko, G.A. (1989). Unpredictability and indeterminism in human behavior: Arguments and implications for educational research. *Educational Researcher*, 18 (3), pp. 17-25.

Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on "scientifically based" educational research. *Educational Researcher*, 32 (31), pp. 31-38.

Erickson, F., & Gutierrez, K. (2002). Culture, rigor, and science in educational research. *Educational Researcher*, 31 (8), pp. 21-24.

Howe, K., & Eisenhart, M. (1990). Standards for qualitative (and quantitative) research: A prolegomenon. *Educational Researcher*, 19 (4), pp. 2-9.

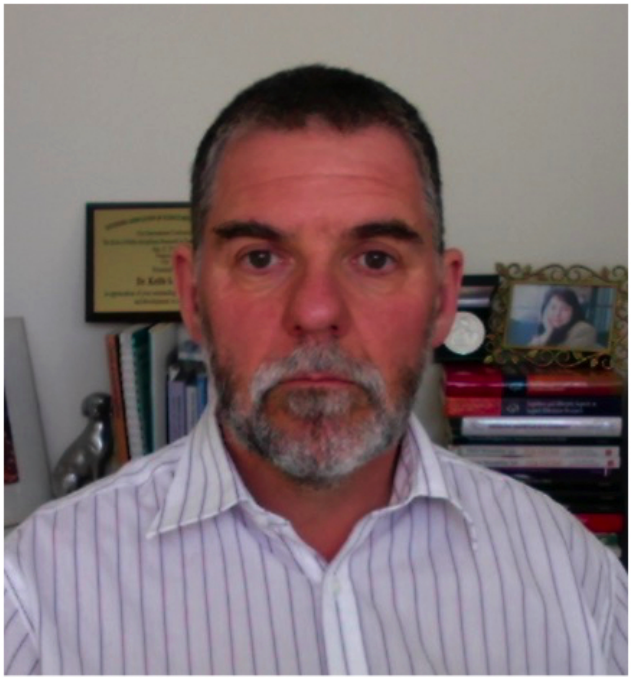
National Research Council. (2002). *Scientific research in education*. R.J. Shavelson & L. Towne

(Eds.). Committee on Scientific Principles for Education Research. Washington, DC: National Academy Press.

Slavin, R.E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31 (7), pp. 15-21.

St. Pierre, E.A. (2002). "Science" rejects postmodernism. *Educational Researcher*, 31 (8), pp. 25- 27.

Opinion Page II: *Is Reproducibility a Realistic Norm for Scientific Research into Teaching?*



Keith S. Taber
Professor of Science Education, University of Cambridge
UK

In a recent Opinion piece in the HPS&ST NEWSLETTER, Bradley Alger (February 2020) asked if we should be concerned about non-reproducibility in science. It could be argued that if much scientific work is not (and, perhaps, cannot be) replicated, then this might present some kind of existential crisis for science.

At the risk of reducing a nuanced argument to a few bullet points, Alger was suggesting that reproducibility is seen as norm in science, perhaps sometimes even a criterion for work to be seen as scientific, and that recent discussions of the extent to which published studies may resist attempts at replication presented a challenge for the scientific community. If a substantial proportion of the studies in the scientific literature cannot be replicated, then we may seem to be faced with a choice between downgrading reproducibility as a criterion for science, or, alternatively, accepting that there is something very rotten in the state of the scientific literature.

Alger explored different understandings of reproducibility, and highlighted the implications of using statistical significance as a basis for claiming that an experiment gives a positive result. Given that the conclusions of many studies are based on inferential statistics, there is an inherent tolerance (in the technical sense) for a level of non-reproducibility that is to be expected across published studies, such that the scientific community should show a level of tolerance (in the psychological sense) of non-reproducibility of published results.

Alger's essay concerned scientific studies; that is, research in what are commonly termed the natural sciences. In this essay I wish to *complement* Auger's discussion by focusing on educational research. Education is commonly seen as falling within the social, rather than the natural, sciences. However, the breadth of work actually undertaken

in education is very wide – from research in cognitive science that employs laboratory conditions and adopts strict experimental paradigms; to studies that deconstruct texts through literary analyses, or are purely philosophical in nature, and which are better considered part of the humanities. This is too great a spectrum to readily consider as being of a kind (even a 'social kind' as discussed below), and so my remarks here relate to a subcategory within empirical educational research investigating classroom teaching. What I have to say here applies across that category, but the focus will be taken to be studies on science teaching.

An impression given by the research literature is that many of my fellow science education researchers, having initially trained in the natural sciences, consider that the methods of those sciences can be relatively unproblematically applied to educational research. Consideration of the different nature of natural and social phenomena suggests that this is not so.

I will argue that research into education cannot be approached like research in physics laboratories because educational research concerns social kinds (such as 'teachers', 'classes', 'lessons', etc.) which do not support the assumptions made about natural kinds underpinning much work in the natural sciences. It might also be provocatively suggested that to the extent that research into teaching is sometimes like some studies in natural science, this is because some scientific studies fall subject to the kinds of complexities inherent to social science research (e.g., inability to identify all relevant variables; inability to isolate the phenomenon of interest from interactions with its context), as in the examples of 'unforeseen impediments to reproducibility' cited by Alger.

A key purpose, indeed rationale, of educational research is to inform teaching. There is a substantive literature on effective pedagogy in science teaching, including many empirical studies claiming to test the effectiveness of different teaching approaches, or reform curricula, or new learning resources. Many of these studies adopt an experimental approach, or, at least, a quasi-experimental approach, and then draw conclusions on the relative effectiveness of some pedagogy or resource or curriculum innovation, usually based on measures of student learning gains or changes in attitudes. These studies draw upon the experimental method used in the natural sciences to compare outcomes in two or more distinct conditions that differ in terms of some independent variable. So, perhaps, in one condition a group of 13-14 year old students study the topic of acids and alkalis through co-operative group-work, whilst in the 'control' condition a similar group of students is taught the same topic through what might be (and often is) termed 'traditional' instruction. If statistical analysis suggests that the students in the co-operative group-work condition show significantly greater learning gains (or shifts in attitude, etc.) than those in the traditional condition then researchers conclude that the co-operative group work condition is superior.

I have suggested these studies *draw upon*, rather than *adopt*, the experimental method used in the natural sciences, as, although educational experiments are superficially like experiments in the physical sciences, an engagement with the details of research designs often raises serious concerns. There are certainly substantive issues in relation to generalisability and control of variables.

I have recently undertaken a review of work of this kind, exploring the methodological and ethical challenges of experimental work intended to test teaching innovations (Taber, 2019a). The key

ethical issue raised is how learners in 'control' or 'comparisons' are treated instrumentally in many studies that pass journal peer review: it may be perfectly acceptable to set up a control condition expected to be deleterious when the experimental subject is a copper wire, or a crystal, or a rivet, or even a mustard seedling (or indeed, to many people's thinking, a rat), but it is a different matter to set up teaching conditions which can reasonably be expected to disadvantage the learning of whole classes of students attending public schools. It is not only that many studies actually do this: it is often made quite explicit in the published reports that this is a deliberate aspect of the research design.

Here, however, I want to consider some of the *methodological* challenges of such work, and how the nature of what is being studied undermines notions of replication or reproducibility of studies. Space here only allows a limited discussion, and interested readers are referred to the full review article (Taber, 2019a) for further detail.

Natural kinds and social kinds

One very relevant concern that is often ignored in educational research (and indeed from my reading, often also in psychological research) is the distinction between natural kinds (Dupré, 1981) and what have been called social kinds. The notion of natural kinds is based on an ontological assumption: that nature offers us certain regularities of experience that justify classifying recurrent features of our experience as having inherent, essential, properties. We can class this as gold (which will therefore have *these* properties) and that as phosphorus (which will therefore have *those* properties); we recognise *this*, but not *that*, as an instance of torque. *These* are dogs, and *those* are cats. Of course, there are limits to this. There is probably no such thing as an absolutely pure sample of gold, but we can decide on a level of impurity low enough to be negligible. There is a diversity of varieties of dogs, and, since Darwin, we no longer think that there are absolute boundaries between species or indeed any taxonomic categories. We may be fairly clear what extant specimens are, or are not, mammals; but go back into the fossil record, and there will a point where it becomes a matter of debate, and indeed, potentially, scientific controversy.

We can identify a particular strain of microorganism that we know has certain genetic characteristics that give it particular properties in certain environments. At the type of writing the world is suffering from, and responding to, a global pandemic. The disease concerned is known as COVID-19, and it is considered to be caused by infection with a transmissible virus known as SARS-CoV-2 (severe acute respiratory syndrome coronavirus-2). Not only the scientific community, but the population generally, accepts both that the infectious agent is a virus, and, moreover, that it is a particular type of virus, indeed a particular type of coronavirus (SARS-CoV-2); *and* also that one specimen of SARS-CoV-2 is much like another in that any specimen has the inherent property of being able to infect human beings leading to the 'same' disease.

Even if the layperson has never heard the term 'natural kinds', and indeed has no technical knowledge about such matters as the means by which viruses cause disease, they will still have an implicit notion of the natural world such that they have no difficulty accepting that billions of specimens of virus particles around the world are causing the same disease because they are of one kind.

Generally, the subjects of scientific studies are samples/specimens of natural kinds where it can either be assumed (i) that swapping the particular specimen would not change the results, or (ii) that there is some relevant variation between specimens such that we should work with a sample and draw conclusions statistically, so that drawing a new sample from the same population should give substantially the same results. If we find a pure copper rod is a good conductor, then this applies to all pure copper rods and not just the one(s) we decided to test. If we find a sample of a variety of wheat plants grow taller on average when we provide a phosphorus-based fertiliser then we assume this will generally apply to other samples of wheat of that kind.

Often, in science education, we find that learners have intuitions about the natural world that are contrary to canonical science and can impede school learning (Taber, 2009). So, many students have intuitions contrary to Newton's first law for example, and so expect all motion to naturally come to a stop; or assume the force acting on an orbiting body acts tangentially to, rather than perpendicular to, its instantaneous direction of motion. Many such 'intuitive theories' or 'preconceptions', as they have been called, make learning canonical science more challenging.

However, an informal commitment to objects being specimens of natural kinds is a common intuition which works to the advantage of teachers of physics and chemistry. So, learners usually readily accept, for example, that all protons have the same amount of positive charge, all samples of copper wire will conduct electricity, and that all samples of potassium dichromate will act as oxidising agents. It is quite common for school laboratory work to be used to generalise results from a single run on one sample of materials (usually without any explicit attention to the grounds for, or validity of, making such generalisations), before moving on in the next lab. session to a completely different practical demonstrating some other principle or concept. This approach tends to persuade most students – if at the same time providing an unauthentic representation of how science is actually done.

This same intuition *often* helps biology teachers, too, as students do not question that they are shown models of 'the' human skeleton or asked to label diagrams showing the parts of 'the' digestive system. If they learn the function of 'the' kidneys then they do not need to ask who's kidneys in particular they are discussing, and what alternative functions someone else's kidneys might have. However, this common ontological intuition may work against learning about arguably the most important organising idea in biology – evolution by natural selection, which shows that species are not completely discrete, but blend into one another. That is, to see species as natural kinds is only approximately or contextually true (for example, not when considering geological timescales). The assumption that different kinds of animals and plants (and fungi, of course) are separate fixed types generally works well in most everyday contexts, but is counter to the insight underpinning much of modern biology (Taber, 2017).

In the social sciences we are not dealing with natural kinds at all, but what are sometimes called social kinds. Science teachers, classes of learners, schools, and the like do not have the degree of essential qualities we expect of natural kinds. What science teachers have in common *qua* science teachers is largely contingent – science teachers are developed ('trained') and not born.

Genuine natural kinds retain their properties re-

gardless of human culture (even if what humans *know of* their properties can clearly change). Arguably, a much-used category like acid (or oxidising agent) does not strictly label a natural kind in the way potassium does (Taber, 2019b). The potassium concept has changed over time, but the natural kind, potassium, itself has not. Yet the acid concept – if indeed there is a single canonical concept, which is moot – has changed its defining properties in ways such that membership of the category has changed over time. (That is, not just the range of acids we know of has changed, but so has which substances *should* be considered acids according to different historical scientific accounts.) This is not a matter of better understanding the qualities of a natural kind, acids, but of chemists redefining the acids concept to be more convenient – and so shifting the demarcation between acid and not acid. However, those (more genuine) natural kinds subsumed under the broader acid concept (sulphuric acid, ethanoic acid, etc.) can be considered to have their own essential properties.

Social kinds are quite different. What actually counts as a school is a matter of social convention and can change over time. The same point can be made of effective teaching. A quiet classroom where all the students sit at their desks with their eyes on their textbooks or writing under the watchful gaze of a teacher would have been seen as a positive indicator in some cultural contexts, and a busy, noisy, classroom with students moving about and interacting in groups while one of their classmates actively disputes their teacher's presented account of some subject matter would have been seen as unacceptable. This has shifted over time, but not to the same extent in all national contexts.

Is replication overplayed in science?

We all learn that reproducibility is important in science, and this is indeed so. When it was claimed that power could be generated by 'cold fusion', scientists did not simply accept this, but went about trying it for themselves (Close, 1990). Over a period of time a (near) consensus developed that, when sufficient precautions were made to measure energy inputs and outputs accurately, there was no basis for considering a new revolutionary means of power generation had been discovered. That this process took some time reflects something bench scientists will know, but which does not fit the popular image of science. It has long been recognised that there is a tacit dimension to scientific work (Polanyi, 1962), and the formal published technical account of a novel experiment is often insufficient by itself to allow scientists to replicate each other's work (Collins, 1992). Indeed, it has been claimed more generally:

In the normal way, scientific phenomena are not reproducible with great reliability, but this is usually explained as being a consequence of scientists' mistakes, or 'anomalies', or some anodyne formulation such as 'gremlins' or the 'fifth law of thermodynamics'. (Collins & Pinch, 1982/2009, p. 159)

However, it has also been argued that when historical cases of scientific replications are studied, it is found that, generally, scientists do not spend a great deal of time trying to precisely reproduce the published studies of others (at least, not unless they have reasons to suspect flawed work), but actually usually set out to deliberately undertake a related, but modified, experiment (Shapin & Schaffer, 2011).

In part, this may relate to the widely discussed

belief that getting published in the most prestigious journals is unlikely when your paper reports that you did exactly what was reported in a previously published study and found entirely comparable results. Even if scientists value replication as a principle, the community awards novelty. Nobel laureates are not normally cited for their careful replication studies and contributions to the reproducibility of someone else's novel findings. However, this is also related to that assumption about natural kinds: if one person has carefully obtained a result with a sample or specimen of some natural kind, then, as long as they have worked carefully using appropriate, well-maintained, and calibrated apparatus, the reasonable default assumption is that others should get similar findings when working with another sample or specimen of the same kind (Millikan, 1999). Precise replications are therefore more likely to be attempted to challenge, rather than support, published results.

We can (or, rather, should) seldom make such an assumption in educational research. *This* class of 14 years old students learning physics cannot be assumed to respond to our interventions the same way as *that* class of 14 years old students learning physics; *this* chemistry teacher cannot be assumed to be able to master a new pedagogy as well as *that* chemistry teacher; *this* biology undergraduate cannot be assumed to have the same intuitions about the natural world as *that* biology undergraduate.

To a lesser extent, the life sciences face similar issues: even genetically identical individuals can vary considerably (Vogt et al., 2008) which is why biologists, where practicable, commonly use large sample sizes and statistical methods rather than compare one mouse in condition A with one mouse in condition B. However, nearly always biologists are only having to deal with physiological variation – and do not have to consider cultural issues, such as social class, cultural norms, language of instruction, local national curriculum, school ethos, and so forth.

The ideal of random control trials

Social scientists know how to respond to such a challenge in principle. Perhaps you want to know whether having students work in pairs will better support learning about forces than individual working of 11-12 year old students. To set up a study all you need to do is:

1. Define your population of interest – so perhaps you do not claim your research is about 11-12 year olds *per se*, but rather about 11-12 year olds in England (or perhaps 11-12 year olds attending state schools in England, or perhaps 11-12 year olds attending non-selective state schools in England, or perhaps 11-12 year olds attending mixed-gender, non-selective state schools in England, or…).

2. Then you identify the members of that population – so, all the 11-12 year olds in England (or all those attending state schools, or…).

3. Then you select a random sample of the population, large enough for the statistical tests you intend to apply to be able to *potentially* offer positive outcomes, and randomly assign the sample to the two conditions.

Even readers with no experience of educational research are likely to appreciate that this never happens. Steps 2 and 3 are clearly non-feasible when dealing wither large populations of this kind. Even the issue of the unit of analysis is problematic: unless one has the resources to set up experimental

classes in laboratory conditions, one usually relies on intact classes in schools being assigned to conditions.

Not only do studies rarely draw upon a national population, but many published studies in decent journals are based only upon one class being assigned to each of two conditions. This usually means that results can only be obtained by considering the individual learners as the units of analysis (even though it is well known that there are interactions within classes such that the learner variables cannot be assumed to be changing independently).

Despite few, if any, studies approaching procedures including steps 2 and 3 above, it is notable that both the titles and conclusions of so many educational studies offer universally generalised findings about such social kinds as '14 year old students learning physics' or 'chemistry teachers' or 'biology undergraduates'. Just as the results of physics experiments carried out on Earth are assumed to also apply in the vicinity of Alpha Centauri, many educational studies are reported as though their findings about classes of 14 years old students learning physics, or chemistry teachers, or biology undergraduates, would be just as applicable whether these students or teachers were based in Oxford, Tehran, St. Helena, or, indeed, on a planet somewhere near Alpha Centauri.

There are many other complications: such as choosing between having different teachers in the different conditions, or assuming that employing the same teacher for both classes controls for the teacher effect – as if any teacher is just as effective in any teaching condition or working with any class. (Again, the reader is referred to Taber, 2019a for further discussion). Arguably, being a teacher is a social role, and is enacted interactively with a particular class: most teachers will acknowledge that there is a sense in which they have not been the same teacher across all their classes. (Just as mischievous schoolchildren tend to be naughty with some, but not all, their teachers.) This is before it is considered that, unlike the experimental subjects manipulated in the physical sciences, teachers and school children's behaviour and teaching/learning can be strongly affected by their expectations about the research they are part of (Rosenthal & Rubin, 1978). Teachers are regularly reminded that it is important to have high expectations of their students as this can make a substantial difference to classroom outcomes – yet this factor is seldom mentioned as a caveat in the published reports of experimental studies in education.

Experimental work in the science laboratory can be useful because it allows identification, control, and measurement of variables. Educational experiments seldom identify all relevant variables (as phenomena such as classroom teaching and learning are very complex, and embedded in diverse and very particular contexts), let alone control or measure them all. That does not make an experimental study invalid *in its own context* – but it raises very substantial barriers to generalisation from the context to some wider population.

What makes educational work scientific?

This leads me to the question, hinted at near the outset of this essay, of whether we make our educational studies more scientific by aping scientific research. I think that depends what is taken as the model. A good deal of scientific work is experimental in nature: yet, certainly not all. When experimental methods are inappropriate or not feasible, then more naturalistic, observational meth-

ods are the more 'scientific' approach.

This has been recognised in education as well (National Research Council Committee on Scientific Principles for Educational Research, 2002). Unfortunately, some national governments and funding agencies are seduced by the perceived gold standard of the randomised control trial (Phillips, 2005), rather than recognising that when the conditions needed for rigorous experiments are not possible, it is better to choose what is viable in the actual fieldwork circumstances that researchers face, rather than look to an ideal that needs to be so compromised that studies cannot possibly be judged robust. In educational contexts, this will often (certainly not always) mean that more is learnt from an in-depth case study of an authentic episode of teaching-learning in a well characterised and described particular context, than attempts to use small, non-random, unrepresentative samples of populations to attempt to draw general conclusions about 'what [universally] works'.

Unfortunately, although these debates are widely rehearsed in educational research circles, science education is disproportionally staffed by scientists! Unlike, for example, history teachers or literature teachers, science teachers (and so, often, science teacher educators, and science education researchers) come to educational work with a background in the natural sciences where working with natural kinds, and the *implicit* assumptions that such experimental subjects allow one to make in undertaking and reporting research, colours how they think about educational studies.

Seeking incremental generalisation, rather than reproducibility, in educational research

Inevitably, the evidence for the effectiveness of most pedagogic, curriculum, or resource innovations is not based on random control trials undertaken with representative samples of the populations that results are claimed to apply to. For any particular innovation, it is likely the positive evidence comes from a handful of studies, perhaps scattered across different types of schools, different grade levels, different languages of instruction, and carried out with somewhat arbitrary (rather than random) teachers and classes where researchers could negotiate access and persuade teachers to implement something novel in that context.

Perhaps, where these scattered studies do report positive results from a wide range of teaching and learning contexts we might be encouraged: something seems to work both in elite schools and in comprehensives; when taught in Spanish, and in Chinese, language contexts; in single-sex Catholic schools, and in mixed-gender community schools serving multi-cultural communities; etcetera. However, we then run into the problem of publication bias (Franco, Malhotra, & Simonovits, 2014): the likelihood that the literature is systematically biased to report studies that found significant differences, over those that failed to obtain 'positive' results. With so many variables at work, we cannot be confident that there are not just as many unpublished studies, from a similarly diverse set of unique teaching-learning contexts, where the innovation did not seem to offer any improvement in desired educational outcomes.

This is without considering the contribution of studies that report those 'rhetorical' experiments I referred to earlier, where the researchers ensure

that the comparison condition by which the innovation is judged is a teaching condition widely recognised as ineffective. This almost guarantees that the experimental conditions, where the teacher is given special training and the class have a learning experience notably different from the norm, will be more effective than the deliberately humdrum instruction in the control condition, almost regardless of the actual innovation supposedly being tested.

Despite being quite critical of the state of experimental research in education, I do not think the situation is hopeless, as long as the community can become more scientific by better following *the logic behind* experimental work, rather than simply trying to transfer the appearance of controlled laboratory studies into messy social contexts where meaningful control is never going to be possible.

One of the arguments made in my review (Taber, 2019a) is that even if strict replication is never going to be feasible in educational contexts, there is still much value in seeing whether what worked in one context will also work in another. It is never possible to entirely characterise something as complex as a teaching-learning episode embedded in, and entangled with, its particular multilayered (classroom, plus institutional, plus curriculum, plus cultural) context – or even to specify the relevant characteristics of different classroom teachers observed in different studies (what night make a difference: age? gender? years of experience? teacher preparation regime? degree specialism? relationship with own parents?…).

There are going to be some teaching or curricular innovations that will be generally effective when implemented by enthusiastic and well-prepared teachers. However, others will quite reasonably only tend to be found useful with, say, older students, or with higher achieving students, or with students in countries with a strong Confucian tradition, or in contexts where students have already mastered the basic skills needed for productive classroom dialogue, or perhaps only in a particular educational context that is found on that planet somewhere near Alpha Centauri.

It is therefore very important to move away from treating social kinds as if they are natural kinds, and so expecting that pedagogic or other innovations either 'work' or 'do not work' and so be universally worth (or not worth) implementing. It is possible, however, to make some judgements about *where and when* particular innovations are worth recommending and expending resources implementing, if instead of focusing on replication *per se* we put the emphasis on profiling generalisation in terms of the range of effective application.

This will only be possible when researchers (and journal editors) recognise the importance of characterising the study context as well as they can for readers of the research, rather than just reporting along the lines that the work was undertaken with 15-16 year old students from an urban school in Melbourne. Efficient *programmes* of research of this kind require those planning individual studies to be able to gauge the variation across previously published studies. If the literature suggests mixed outcomes from previous testing, then what is indicated are further tests which can help determine the kinds of conditions that (do and do not) favour the effectiveness of the innovation *from within the broad range of populations that have given inconsistent outcomes*. If, however, the literature suggests something is very widely effective, then further tests will be most useful in situations *outside the scope of existing studies* (has it yet been

tested with very young learners, with very disengaged learners, with the gifted, with traumatised students in migrant camps, with visually impaired students…?)

Over time, then, such programmes of 'replications' offer an opportunity to build up an account of the (multi-dimensional) ranges of effectiveness of different teaching approaches/curricula/resources. This does rely on 'negative' results being published as well as 'positive' results. Knowing the characteristics of contexts where some innovation does not seem to be effective avoids wasting the expenditure of precious teacher time and other resources implementing something when the available evidence suggests (we can never be sure of course) it is unlikely to offer an educational return in a particular teaching context. Indeed, it is not appropriate to think of study outcomes as *positive* or *negative* replications, but contributions to building up a *profile* of the pedagogic effectiveness of some innovation. In this context, reporting a poor educational outcome is as valuable as reporting a good outcome – as long as we ignore our intuitions about research studying natural kinds, and sufficiently characterise *the particular* class of 14 years old students learning physics, or chemistry teacher, or biology undergraduate, that the study focuses on.

References

Alger, B.E. (2020). Opinion: Is Reproducibility a Crisis for Science? *HPS&ST NEWSLETTER* (February), 9-18.

Close, F. (1990). *Too Hot to Handle: the story of the race for cold fusion*. London: Allen and Unwin.

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

Collins, H., & Pinch, T. J. (1982/2009). *Frames of meaning: The social construction of extraordinary science*. Abingdon, Oxon.: Routledge.

Dupré, J. (1981). Natural Kinds and Biological Taxa. *The Philosophical Review*, 90(1), 66-90. doi:10.2307/2184373

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

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

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

Phillips, D.C. (2005). The contested nature of empirical educational research (and why philosophy of education offers little help). *Journal of Philosophy of Education*, 39(4), 577-597.

Polanyi, M. (1962). *Personal Knowledge: Towards a post-critical philosophy* (Corrected version ed.). Chicago: University of Chicago Press.

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

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

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

Taber, K.S. (2017). Representing evolution in science education: The challenge of teaching about natural selection. In B. Akpan (Ed.), *Science Education: A Global Perspective* (pp. 71-96). Switzerland: Springer International Publishing.

Taber, K.S. (2019a). Experimental research into teaching innovations: responding to methodological and ethical challenges. *Studies in Science Education*, 55(1), 69-119. doi:10.1080/03057267.2019.1658058

Taber, K.S. (2019b). *The Nature of the Chemical Concept: Constructing chemical knowledge in teaching and learning*. Cambridge: Royal Society of Chemistry.

Vogt, G., Huber, M., Thiemann, M., van den Boogaart, G., Schmitz, O. J., & Schubart, C. D. (2008). Production of different phenotypes from the same genotype in the same environment by developmental variation. *Journal of Experimental Biology*, 211(4), 510-523. doi:10.1242/jeb.008755