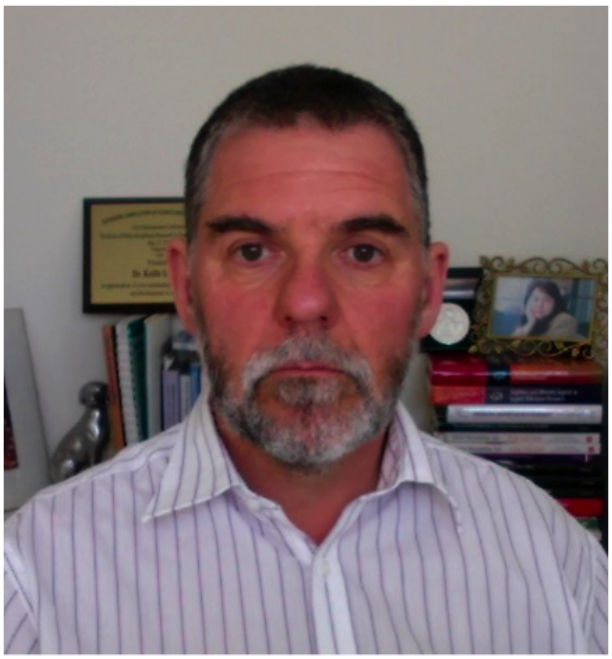
Opinion Page II: *Is Reproducibility a Realistic Norm for Scientific Research into Teaching?*



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 regardless 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 interact-

ively 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 methods 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 multi-layered (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