Taber, Keith S. (2023) **Science, superstition, or confidence trick. *Do science educators have too much faith in the experiment?*** Seminar paper presented to the *Science Education Group* at the University College London's *Institute of Education*: 3rd May 2023 (on-line via Zoom).

The presentation can be viewed at
https://science-education-research.com/publications/miscellaneous/science-superstition-or-confidence-trick/

Science, superstition, or confidence trick:

Do science educators have too much faith in the experiment?

Abstract

A rigorous experiment is rightly considered an especially informative research tool. But doing rigorous experiments in education is very challenging. A poorly designed experiment may tell us very little. Yet the literature includes a vast number of experimental studies in science education. Here I make an argument that:

- *Often these are very small scale studies with unrepresentative populations;*
- *Often the extent of control of variables would not pass for 'fair testing' in a school laboratory exercise;*
- *There is a sleight of hand commonly used to (mis)apply statistical tests to treat samples as much larger than they are;*
- *Sometimes authors draw conclusions contrary to their results; and*
- *Sometimes inappropriate (unethical) control conditions are imposed on learners for the sake of research.*

I will pose, and reflect on, the question of why so many of these flawed studies get undertaken in science education and published in peer-reviewed journals

The thesis

The thesis I am advancing is that there is a phenomenon to be explained. That phenomenon is that the science education research literature includes a very large number of studies that are experimental in nature, but which do not meet the criteria for being valid experiments. Many of these are published in regional or national journals, but there are plenty in more prestigious journals. Perhaps everybody involved - authors, peer reviewers, editors - recognise the problem with these studies, and see their findings as purely indicative - but often that is not the impression given by their framing and phrasing.

I argue that

- Experimental method is often the appropriate approach in seeking new knowledge in the natural sciences;
- Experimental method is much more challenging in the social sciences;

- Valid experiments are (at the least) very rare in small-scale studies of teaching and learning;
- Yet, the science education literature includes a vast number of such studies, which do not support robust conclusions.

This needs explaining! My focus here is on experiments that involve a small number of teachers and classes…

Indeed, the prototype of this type of study involves

- one class in the experimental treatment condition
- another class being the comparison class

Sometimes both classes are taught by the same teacher, sometime not. Sometimes the classes are not even from the same school. Other studies may have just two or three classes in each condition. A good many published studies are of this kind.

Key issues

There is a host of difficulties in doing experimental studies into classroom teaching, and I have published an account of a number of these in a review for *Studies in Science Education*. Here I wish to focus on, and illustrate, a few themes

- Control of variables
- Representativeness and generalisability
- Identifying independent units of analysis
- Inadequate tests of equivalence
- Rhetorical experiments and unethical controls
- Falsifying conclusions

Control of variables

Now a good experiment involves three classes of variable:

- the thing we are deliberately changing,
- the thing we allow to vary to see if, and if so how, it changes, and
- everything else which could have an effect, and, so, is not allowed to change.

To do an experiment could be seen as in large part controlling all the other things that could effect our results and confuse the possible dependency of the dependent variable upon our planned intervention. A poor experiment has a fourth class of variable. These are all the things that should be in the controlled category, but which we do not control. Confounding variables confound our study because we logically need to caveat the conclusions with an 'unless this was due to something else'.

Now, even in the natural sciences, there can be confounding variables, because it is never possible to control everything that we might imagine could be considered a variable, so, on theoretical grounds, we dismiss such possibilities as the relevance of a researcher's hair colour or whether they use their left or right eye to look into the microscope. We use existing theoretical knowledge to judge what we can ignore. Even here, it sometimes transpires there was a pertinent variable that influences results, but which had been assumed to be irrelevant, or was even unknown to the researchers.

In the social sciences, control of variables is very much more challenging. For one thing, we usually work in naturally occurring social contexts, rather than isolate systems for laboratory manipulation. Another major difference is that physicists and chemists do not have to consider what their research materials think about them, or expect to happen in their experiments. A piece of alloy, or a solution of an oxidising agent, has no preconceptions about the outcome of the experiment, and does not have an attitude to the researcher. You do not need to develop rapport with your test tubes.

Indeed, if you have an active imagination, you can likely think of feasible scenarios when any number of things could effect the outcomes of an educational experiment such as student learning. I do not think it is an exaggeration to argue that in the case of an educational experiment:

- we often *cannot identify* all the variables which might have an effect;
- even when we can identify them, we often *do not know how to meaningfully measure* them;
- even if we can measure them, we may *not have a way of holding them at a constant value*.

That need not prevent experimental studies where you have large enough samples that are representative of a population to be able to assume that statistics can tell you whether any outcomes are unlikely to to be due to such chance factors. This is a point I will return to. It is a problem, though, in any studies with small, unrepresentative, samples. And most of the published experiments in the educational literature have small, unrepresentative, samples.

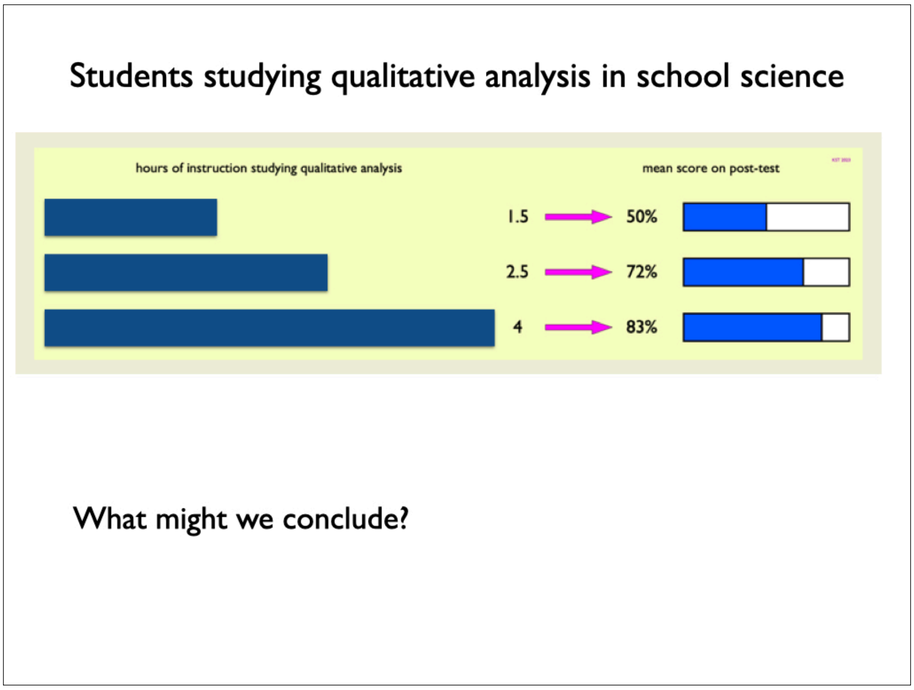
As one example of the kinds of issues that can arise, I was impressed to read this comment in a research paper.

> "The ambiguous results of research comparing IBSE [enquiry-based science education] with other teaching methods may result from the fact that often teaching methods used in the control groups have not been clearly defined, merely referred to as 'traditional teaching methods' with no further specification, or there has been no control group at all."

The authors had noticed that in many experimental studies the experimental treatment is well defined, but the control condition is anything but controlled. It is actually laissez-faire, anything goes, as long as the teacher avoids the approach being taken in the experimental condition. This seemed a fair point, so I read on to see how they managed this issue in their own study:

> "The teaching method as an independent variable was manipulated to identify its effect on the dependent variable (in this case, knowledge and skills)...
>
> In the control group, teachers revised the topic using methods of their choice, e.g. questions & answers, oral and written revision, textbook studying, demonstration experiments, laboratory work."

It seems that raising an issue which undermines the ability to draw clear conclusions from a study does not impose a requirement to address the issue in your own study!



Students studying qualitative analysis in school science

What might we conclude?

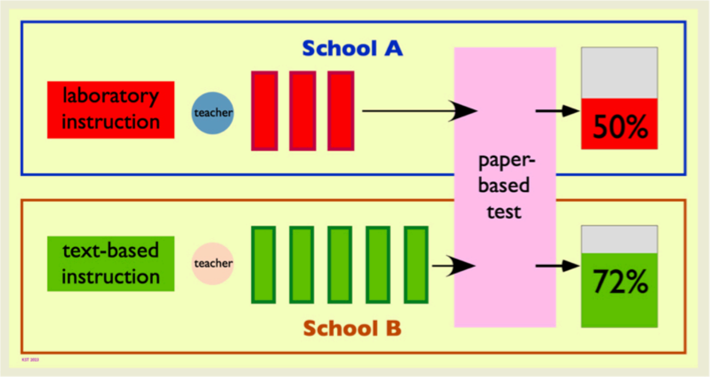
Now here I present the results I found in a published research study. As you see, three classes were taught the topic of chemical qualitative analysis for different lengths of time. Afterwards, the average performance of the students in these classes was found to vary. I wonder what you think we might be able to conclude from this study?

Now, as you may have guessed, I was not giving you all the details. Here I reveal more. One class had three lessons in the lab. But that class was outperformed by a class that had five lessons of paper-based learning activities. The third class, spent three lessons in the lab. and also had the five lessons of text-based activities.

It perhaps seems reasonable to conclude that the class that had both kinds of activity learnt the most. But what about comparing the first two classes? I would mischievously suggest five lessons can lead to more learning than three, but the authors thought this was evidence of the superiority of the text-based learning approach. Presumably, the peer-reviewers and editor were sufficiently convinced.

But I was not convinced.

Students studying qualitative analysis in school science



What might we [reasonably] conclude?

https://science-education-research.com/shock-result-more-study-time-leads-to-higher-test-scores

The two classes were taught by different teachers. The two classes were drawn from different schools. They were considered to be similar schools, but even so. For that matter, the assessment tool was a paper-based test, so how do we know that if a laboratory-based assessment had been used, the results would not have been very different? After all, to actually do qualitative analysis, you need to work with real samples and reagents in a laboratory.

You might feel that I am only stating the obvious. But if it is so obvious, how does such work get published without strong caveats, and sometimes even in the more prestigious journals. After all, we expect 14 years olds to do better than this. Arguably, science educators are teaching skills they then fail to display in their own work.

One of the variables not controlled in that study, was the teacher. Perhaps the two different teachers were very similar in all relevant characteristics, but that is not very likely. Sometimes the 'teacher variable ' is 'controlled' by asking the same teacher to teach differently in two different conditions. That is, the teacher is asked to use two different teaching approaches in different classes. Jig-saw learning here, but computer simulations there. Or, sadly, more likely, enquiry-based teaching here, and dictation of notes there. More on that choice later. Or, a teacher is asked to trial a new curriculum module whilst teaching a parallel class according to the established scheme of work.

This assumes, at least implicitly, the teacher will have the same competence and confidence when switching to do something different, even when it is novel to them. The same teacher, teaching different classes, is assumed to have controlled that variable, but I do not find that very convincing. One of the issues is teacher beliefs, which much research shows often have an effect on outcomes.

If the teacher is persuaded the experimental treatment is an improvement, or is entirely unconvinced by it, then that may be enough to make a difference. Even if the teacher simply lacks confidence in their competence to teach in a different way, then this may make a difference.

It is issues such as this that have led to medical studies adopting double blind conditions in drug trials, so the neither the patients nor the clinicians administering treatment know whether a tablet or injection actually contains the substance being tested or not. Of course if one takes double blind protocols too far, one might run into ethical issues.

I was astonished to read this description of studies where a novel angina treatment was tested by doing sham surgeries alongside genuine interventions.

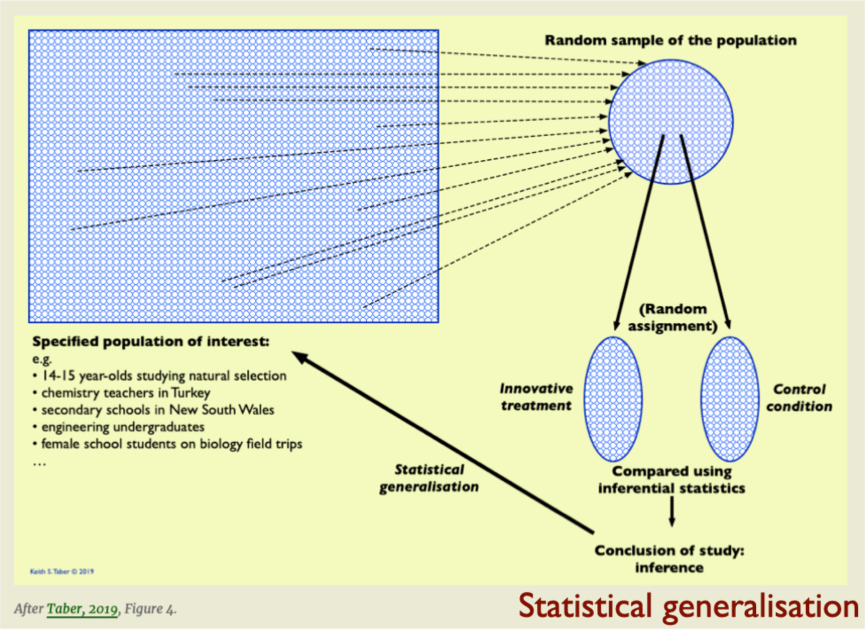
> "In the late 1950s and early 60s two different surgical teams...did double-blind trials of a ligation procedure - the closing of a duct or tube using a clip - for very ill patients suffering from severe angina, a condition in which pain radiates from the chest to the outer extremities as a result of poor blood supply to the heart. The surgeons were not told until they arrived in the operating theatre which patients were to receive a real ligation and which were not. **All the patients, whether or not they were getting the procedure, had their chest cracked open and their heart lifted out.** But only half the patients actually had their arteries rerouted so that their blood could more efficiently bathe its pump ..."

Reports based on the accounts of patients and their doctors had previously claimed that angina symptoms were relieved somewhat by doing some re-routing of blood in the chest though closing off some vessels. However, the experimental studies found that, actually, those given this procedure showed no more improvement than patients wheeled into theatre for a sham procedure.

I was less astonished when I sourced the original research papers, as it seems the sham surgery only involved some superficial incisions made under local anaesthetic. In any case, it is much harder to blind learners and, especially, teachers to the educational treatment they are assigned to.

Generalisation

Social kinds are also different from natural kinds, in that all pure samples of copper or all E. coli specimens have much in common - but not all schools, or all classes, or all teachers, or all lessons, have so much in common. This creates the very big issue in educational research, that what works in one classroom in one school, does not always work in another school, or even when adopted by another teacher in the same school.



Random sample of the population

Specified population of interest:
e.g.
- 14-15 year-olds studying natural selection
- chemistry teachers in Turkey
- secondary schools in New South Wales
- engineering undergraduates
- female school students on biology field trips
- …

(Random assignment)

Innovative treatment

Control condition

Compared using inferential statistics

Statistical generalisation

Conclusion of study: inference

Keith S. Taber © 2019

After Taber, 2019, Figure 4.

Statistical generalisation

https://science-education-research.com/research-methodology/research-design/sampling/

Even experiments that are designed to allow us to generalise to populations only tell us that what was found to be most effective in the research is more likely than not to be effective in the wider population. But there is diversity in those populations. We know from some of the few very large studies undertaking in schooling, that what is most effective overall, is not most effective in every context; what is found to be least effective overall, can have been the most effective approach with some classes.

Knowing what most often works best is still useful. But to do this kind of work well we need not only to work at scale, but to randomise our sample to conditions, and to either use a random sample of the population or be confident that our sample is representative of the diversity of the population.

But what are the populations?

A paper title in the natural sciences might refer to a class of star; a superconductor with a specific composition; a variant of SARS-CoV-2, or some such natural kind. If we look at the titles of papers in science education, we find these papers seem to be about broad groups - sometimes National groups, but sometimes they are apparently about 'children' or 'adolescents' or 'primary school teachers' quite generally!

Of course, the participants in such studies can seldom be considered statistically representative of such broad groups. My point is that we - as authors as well as readers - fall into the trap of generalising, at least implicitly,

> 'We studied what (some) 14 year old Australian students knew about natural selection, so now we know what 14 year old Australian students know about natural selection.'

Again, we would not accept this kind of sloppy thinking from school children.

Units of analysis

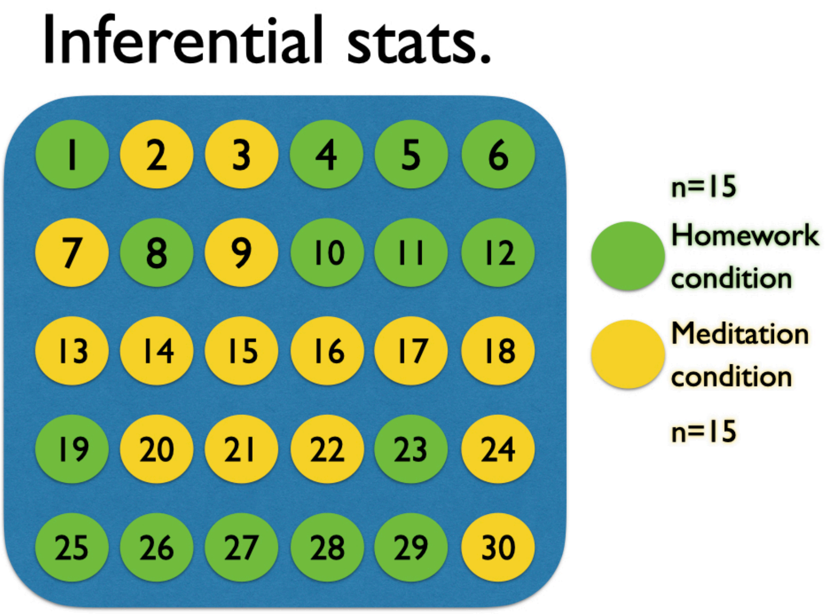
A major issue with many small scale studies is the identification of the unit of analysis to use in statistical testing. In a true experiment we randomise the so-called units of analysis to the treatments, the conditions. Sometimes this is possible in education. Perhaps we enrol fifty schools and assign each randomly to an experimental or control condition. If we can consider the schools to be experiencing the assigned treatment independently of each other, then this seems fair enough.

But many experiments in education are undertaken with learners as the unit of analysis, and often these are not individually assigned to treatments, but are members of pre-established classes. Of course, there are very good reasons why schools would not be happy with researchers coming in and breaking up established classes to randomise students.

Perhaps the logic here should be that as we cannot meet the requirements for an experiment, we should do a different kind of study. Often, instead, the logic adopted is that as we cannot meet the requirements for a valid experiment, we are justified in carrying on regardless and just ignoring that requirement.

If a manufacturer of pickled onions was selling jars of vinegar as pickled onions, it is likely their customers would not be prepared to accept this just because the company was having difficulty sourcing onions. Yet, readers of research papers are assumed to be less demanding of rigour. Science education researchers often sell jars of vinegar labelled as pickled onions.

Now, I do not want to give the impression that I do not think there might be circumstances when treating students within a class as independent units would be appropriate. One has to consider the overall research design and purpose.



Inferential stats.

🟢 1	🟡 2	🟡 3	🟢 4	🟢 5	🟢 6
🟡 7	🟢 8	🟡 9	🟢 10	🟢 11	🟢 12
🟡 13	🟡 14	🟡 15	🟡 16	🟡 17	🟡 18
🟢 19	🟡 20	🟡 21	🟡 22	🟢 23	🟡 24
🟢 25	🟢 26	🟢 27	🟢 28	🟢 29	🟡 30

n=15

🟢 Homework condition

🟡 Meditation condition

n=15

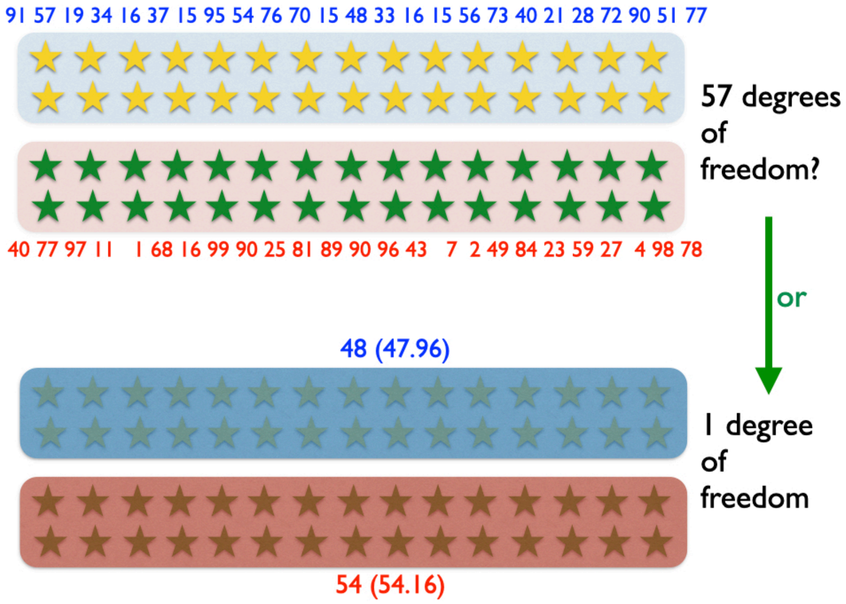
In this hypothetical case, a teacher has read a lot of material about mindfulness, student anxiety, relaxation techniques, and the like; and suspects that asking students to do two half-hours of meditation each week would be just as beneficial for their science learning as asking them to do subject-based homework. Being a science teacher, she tests this idea by randomly assigning students to either a homework condition or a meditation condition, and at the end of term compares test scores.

Randomisation does not ensure matched groups, rather just avoids any systematic bias. So, how does the teacher decide if the difference in profiles of scores in the two conditions is just down to chance effects of who ended-up in each condition? Inferential statistics will allow her to see if any difference in performance is likely to just be down to chance.

Of course, even if there is a very low probability value, and it is concluded there is a significant difference, strictly this only applies to this class with this teacher, and perhaps even this topic [...when taught at this time of year, in this classroom...]? It may be more generally applicable, but we should not assume that.

Here it is reasonable to assume we can treat the learners as independent units of analysis even though they are from the same class. Here, in effect, the class is the population of interest. Yes, the students will influence each other in class, but they are all in the same class so those in both conditions are exposed to the same influences. If the students are off doing homework or meditating individually, and do not collude on the end of term test, it seems reasonable to assume we have fifteen units of analysis in each condition. That is still a small sample size, but at least it is more than unity.

In that circumstance, it is not a problem that all the students are taught together in class and no doubt, at least one might hope, interact during lessons.



But, if we were comparing between two classes, and one class was assigned to the homework condition and another to the meditation condition, then it surely does matter. In this situation, we would need to consider the class as the unit of analysis with one mean outcome score. But, of course, if there are only two classes in the experiment we are not going to find any difference in outcomes that can be statistically significant. The only way we can get positive results here, is by pretending that we have a lot of independent outcomes scores in each condition.

But, that seems like cheating. "*Can't you see the onions in the jar?*" Perhaps you feel I am wrong. I suspect that, like me, you may have taught a range of different classes over the years.

- My own experience is that a class has its own character that is emergent, and is not just an aggregate of the characters in the class.
- My own experience is that parallel classes, or successive cohorts, that are nominally equivalent can actually be very different.
- My own experience is that the same class can be experienced by different teachers as quite different.

When working in school I sometimes found that one or two students in a class could have a disproportional influence on the class environment and progress - and that could be either for better or worse.

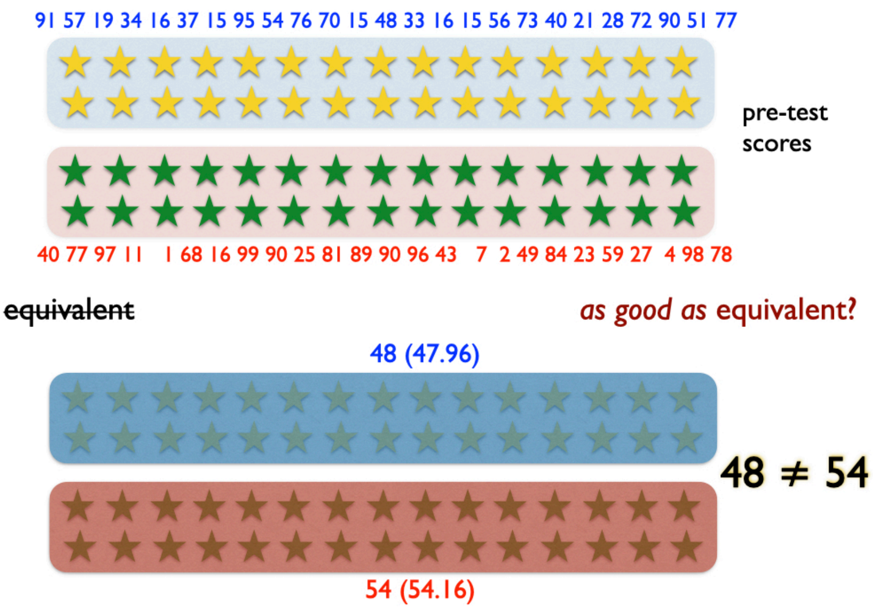
But, perhaps your experiences have been different. *If you can honestly say that you feel that the attitudes; progress; learning, of students in classes is not influenced by the rest of the class, and occurs completely independently of the others in the lessons, then, yes, it is fair to treat the learner as the unit of analysis.*

I thought an analogy might be testing a drug that helped blood supply to the extremities where blood circulation was measured for each digit. That seems sensible if the scores are to be aggregated to give an overall score for the patient. But It would not make sense to consider blood supply to different digits to be independent when it is all part of the same circulatory system, being pumped around by the same heart!

Initial equivalence

A very common procedure used in experimental studies is testing for initial equivalence between groups. This is especially important when there is no random assignment as if there are systematic differences between groups, then any difference in final outcomes may just reflect differences at the start.

Even if researchers showed there was no difference between groups at the outset, this does not negate concerns about students in different classes being influenced by the class context and not learning independently. But leaving that aside, there is a conceptual problem with testing for initial equivalence, because if researchers were really checking for equivalence between groups at the start of start of an experiment then they would very rarely find complete equivalence.

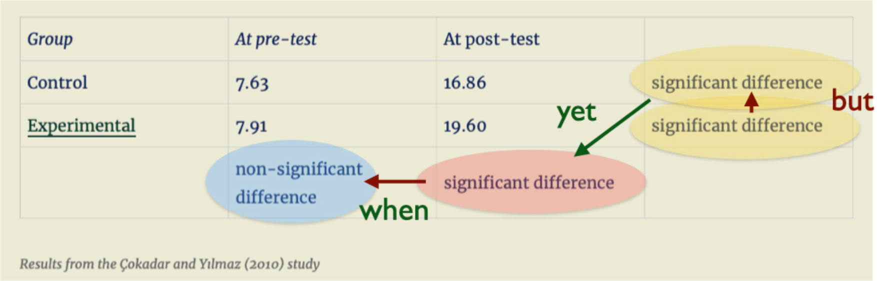


So, imagine two classes we have given what we consider the most relevant pre-test in relation to the study outcomes to be measured at post-test. Actually these are randomly generated numbers, so there is no systematic bias in the scores assigned to the students in the conditions. The average score in one class is about 48, but in the other it is about 54. 48 is not strictly equivalent to 54. If you were appointed to a job on an annual salary of £54 000, but were only paid £48 000, you would probably *not* accept the argument that these two figures are equivalent. But one is seldom going to get precisely the same scores on a pre-test (even with random numbers, as we see here!) So, the question becomes how close is close enough to be seen as as good as equivalent. And this is where I think the most common approach is seriously flawed.

Here is a real example.

initial equivalence

"the effect of creative drama-based instruction on seventh graders' science achievements in the ecology and matter cycles unit …"



Group	*At pre-test*	At post-test
Control	7.63	16.86
Experimental	7.91	19.60

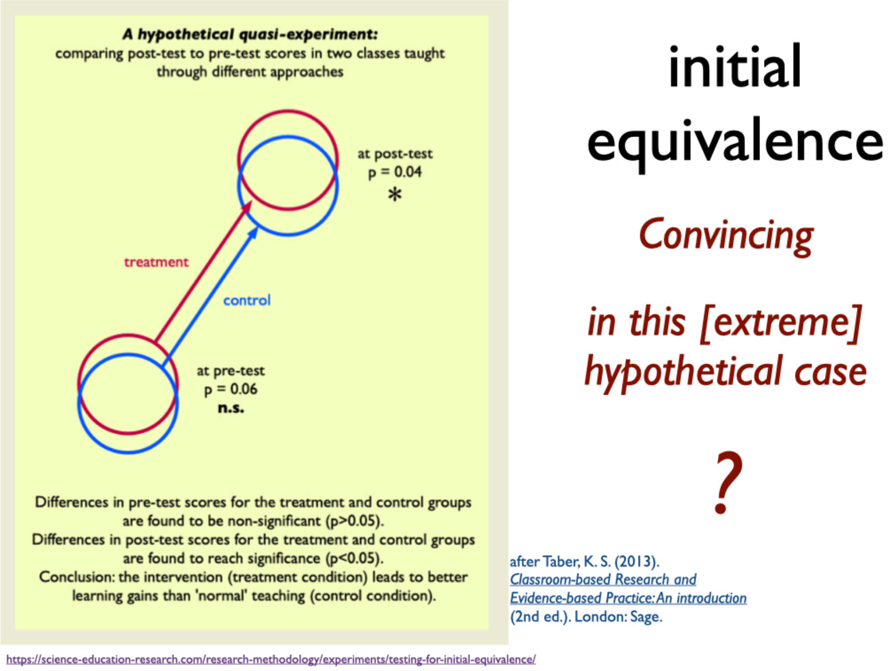
significant difference
but
significant difference
yet
significant difference
when
non-significant difference

Results from the Çokadar and Yılmaz (2010) study

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

In this study the experimental group out-performed the control group at the end of the experiment. And pre-tests were used, so we can compare between pre- and post- intervention, as well as between the two conditions. Statistical tests tell us that the experimental group did significantly better on the post-test than it did on the pre-test. But by itself that's not very informative, as even fairly ineffective teaching is likely to produce some learning. And indeed the control group also showed significant increases between the two tests, so both conditions seem to bring about learning. However, the statistics also tell us that the experimental group out-performed the control group at post-test by a difference that was statistically significant. That means this difference was unlikely to be due to chance effects. But what if the two groups were starting from different bases? This is discounted because there is a significant difference between groups at the end of the experiments, when there was no such difference at the outset.

This seems logical, and perhaps the numbers here look quite convincing. But I think, in general, there is a logical problem here. Let's consider a hypothetical marginal case.



A hypothetical quasi-experiment:
comparing post-test to pre-test scores in two classes taught through different approaches

Differences in pre-test scores for the treatment and control groups are found to be non-significant (p>0.05).
Differences in post-test scores for the treatment and control groups are found to reach significance (p<0.05).
Conclusion: the intervention (treatment condition) leads to better learning gains than 'normal' teaching (control condition).

initial equivalence

Convincing

in this [extreme] hypothetical case

?

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

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

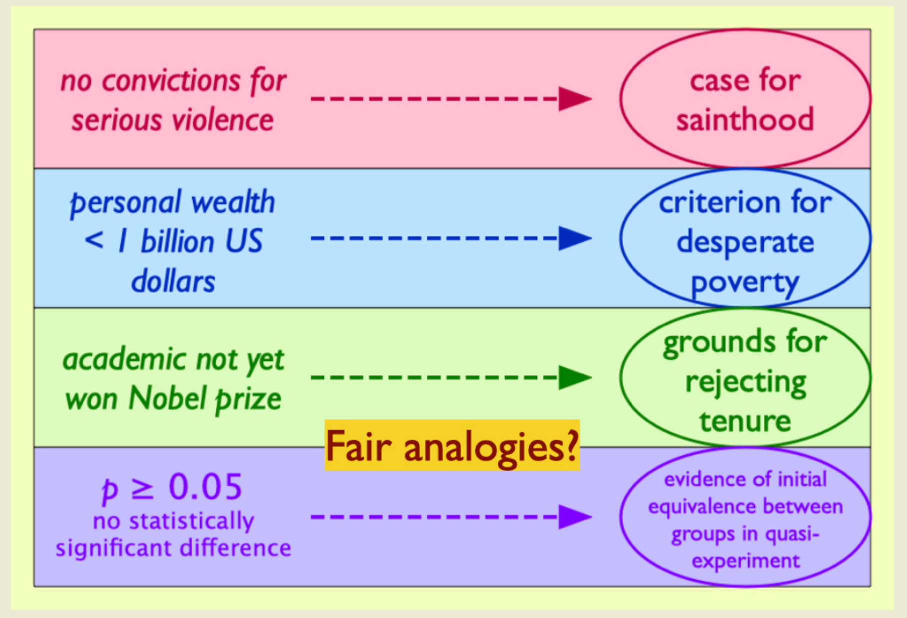
Here there is a small measured difference before the experiment, and also a small measured difference after the experiment.

- The difference at pre-test just failed to reach significance.
- The difference at post test just reached significance.
- Conclusion: the experimental intervention made a difference.

But, surely, initial differences can be magnified in subsequent teaching. It is a common phenomena that differences between learners tend to increase over time. In this hypothetical case, can we really be confident that initial differences were not a factor? Of course, we might argue that there are more sophisticated statistical approaches which look at how factors co-vary in a study. Indeed, there are, but my target here is the simple test for equivalence that is commonly used in published studies to supposedly establish a level playing field.

This common approach is to test to see if there is a very unlikely difference between the pre-test measures in the different conditions. This means that differences which are unlikely to be due to chance effects, but which are not so unlikely to get a p value below 0.05 are found 'equivalent'. I do not think this is a sensible test for equivalence. It is a weak test, and, indeed, inadequate.

- Imagine you were looking for a test to decide if someone should be considered a Saint. How high would you set the bar?
- What if you wanted to identify those who should be considered in poverty. Certainly, not having a great deal of money would be a relevant criterion, but perhaps not quite exclusive enough.
- Maybe you were on a committee that was considering rejecting the permanent re-appointment of a colleague on probation as their research record was not strong enough. But you should not have unreasonable expectations.


no convictions for serious violence
case for sainthood
personal wealth < 1 billion US dollars
criterion for desperate poverty
academic not yet won Nobel prize
grounds for rejecting tenure
Fair analogies?
$p \geq 0.05$ no statistically significant difference
evidence of initial equivalence between groups in quasi-experiment

Some situations where stronger evidence would be useful

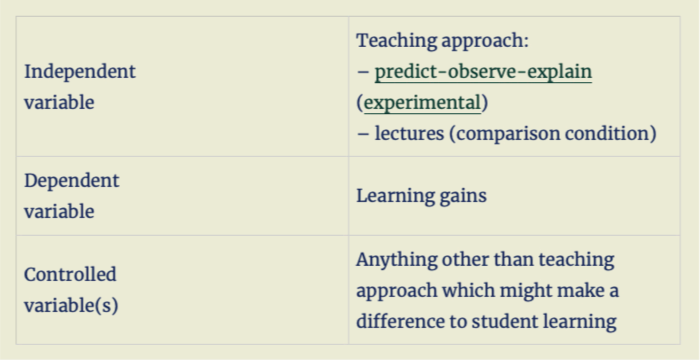
https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

I think these are similar situations, in the sense that the criteria being adopted are certainly relevant and perhaps necessarily apply, but are by no means sufficient. In a sense, we are looking at the wrong end of the distribution. We should be asking how probable we need measured differences to be, not simply excluding the most improbable.

Here is another real example.

The effect of predict-observe-explain strategy…

"Forty nine (49) learners (31 males and 18 females) were from school A and acted as the experimental group (EG) whereas the control group (CG) consisted of 44 learners (18 males and 26 females) from school B."



Independent variable	Teaching approach: – predict-observe-explain (experimental) – lectures (comparison condition)
Dependent variable	Learning gains
Controlled variable(s)	Anything other than teaching approach which might make a difference to student learning

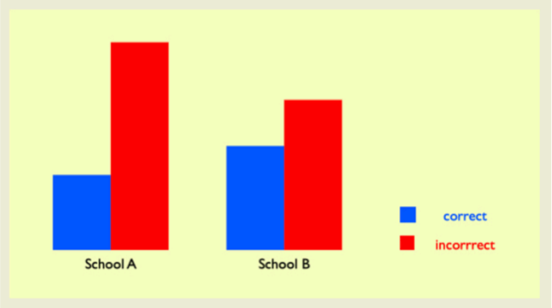
https://science-education-research.com/quasi-experiment-or-crazy-experiment/

The study looked to compare between two teaching conditions. I struggle to understand why researchers think it is acceptable to require teachers to lecture school children, but I will come back to that later. One thing we might notice is the very different gender composition of classes. Is that relevant? Perhaps it should not be. In some cultural contexts though, it might be. The two groups are in different schools, which I think it really troubling in these kinds of studies, as schools vary so much, and in so many ways. But a pre-test was used, and the researchers claimed that on none of the items did differences between groups reach significance. This was seen to assure equivalence. From the data given, I prepared this chart showing the performance on one of the items at pre-test.

The effect of predict-observe-explain strategy…

"The results reveal that there was **no significant difference** between the pre-test achievement scores of the CG [control group] and EG [experimental group] for questions.

*The **p value for these questions was greater than 0.05**.*"



https://science-education-research.com/quasi-experiment-or-crazy-experiment/

We are told that there was 'no significant difference'. Certainly, in both groups most students got the question wrong. But if this is an equivalent performance in the two classes, then the word '*equivalent*' means something very different to its normal sense. Surely, despite the lack of statistical significance, one of these classes is better placed to build on their level of prior learning than the other? This does not persuade me of equivalence. This is not an isolated case. This technique is very widely used.

Ethics of control conditions

Another major concern I have with some published studies, is that they seem to be examples of rhetorical research. That is the study is done to demonstrate what the researchers already expect, indeed believe, and not in a spirit of open-ended enquiry. I have noticed that many school science practicals undertaken to demonstrate well-established scientific findings are often incorrectly referred to as 'experiments', but surely professional science educators know that genuine experiments have uncertain outcomes?

Of course, if one includes more dubious journals in one's purview, one can find extreme examples where no doubt the researcher was entirely convinced by their research, but it seems unlikely there was ever any serious peer review or editorial evaluation beyond checking the publication fee had been submitted. Sadly, there are now a large number of predatory journals out there.

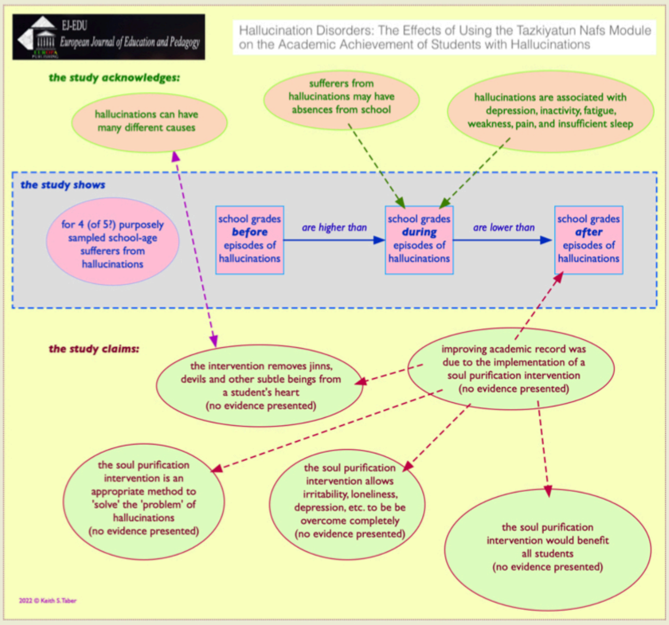
I've detailed a number of examples of both honest and dishonest nonsense published in such journals on my website, by which I mean both work which has been submitted in good faith, but which should never have been published; as well as things that presumably even the authors knew were complete nonsense.

An example of the former is an alternative version of periodic table including a whole raft of new elements that had previously been missed. An example of the latter was a paper which suggested that deforestation of the Amazon was a good idea. After some digging around I found the entire paper was a collage of translations of paragraphs form other published works. Poor translations. 'Deforestation' had been 'poisoning' in the source. One paper I found basically plagiarised a published study, but simply replaced key words to make the paper about something else entirely. [A strong clue to the quality of the work was how the topic of the paper seemed unrelated to its title, and neither seemed related to the remit of the publishing journal. Actually, it seems the author first stole someone else's work; and then plagiarised his own plagiarism by making a modified copy of the same paper with a different focus].

Improving school grades by purifying souls to remove devils:

The researcher seemed convinced…

[We learn about the researcher's beliefs from the study]


EJ-EDU European Journal of Education and Pedagogy

Hallucination Disorders: The Effects of Using the Tazkiyatun Nafs Module on the Academic Achievement of Students with Hallucinations

the study acknowledges:

hallucinations can have many different causes

sufferers from hallucinations may have absences from school

hallucinations are associated with depression, inactivity, fatigue, weakness, pain, and insufficient sleep

the study shows

for 4 (of 5?) purposely sampled school-age sufferers from hallucinations

school grades **before** episodes of hallucinations — are higher than → school grades **during** episodes of hallucinations — are lower than → school grades **after** episodes of hallucinations

the study claims:

the intervention removes jinns, devils and other subtle beings from a student's heart (no evidence presented)

improving academic record was due to the implementation of a soul purification intervention (no evidence presented)

the soul purification intervention is an appropriate method to 'solve' the 'problem' of hallucinations (no evidence presented)

the soul purification intervention allows irritability, loneliness, depression, etc. to be be overcome completely (no evidence presented)

the soul purification intervention would benefit all students (no evidence presented)

2022 © Keith S. Taber

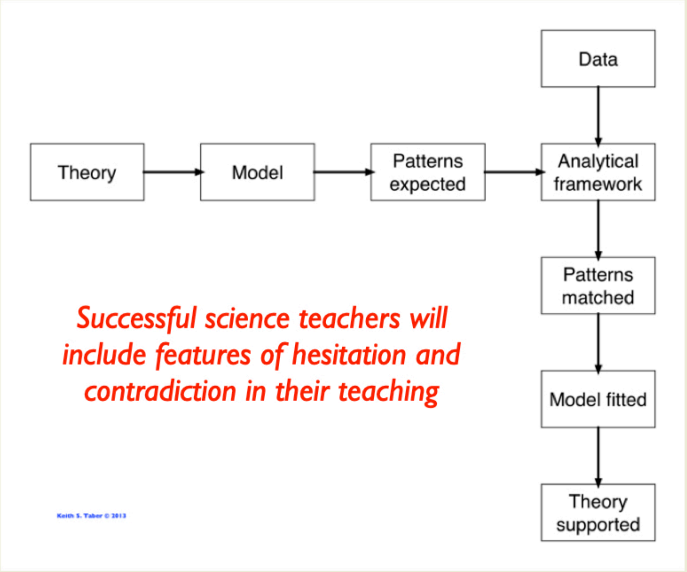
A 'peer-reviewed' study claims to improve academic performance by purifying the souls of students suffering from hallucinations

https://science-education-research.com/delusions-of-educational-impact/

I can only assume that the *European Journal of Education and Pedagogy* does not use serious peer review as the paper I have summarised here does not logically show that school grades can be improved by exorcism, though I strongly suspect the author was genuine enough about the work. That is one of the key points for rigorous peer review - we can have blind spots in our own work.

It is very easy for qualitative [i.e., interpretivist] studies to become rhetorical. We look for something. We find evidence. If we believe that enquiry and argumentation are essential to good science teaching; and so devise an observation schedule based on indicators of enquiry and argumentation; and then observe the lessons of good science teachers, we are likely to find indicators of enquiry and argumentation. The question is, so what?

Danger of 'qualitative" rhetorical research



Successful science teachers will include features of hesitation and contradiction in their teaching

Keith S. Taber © 2013

Rhetorical research: Simply finding some positive examples of instances in data that match codes deriving from a preconceived analytical framework is a weak form of analysis

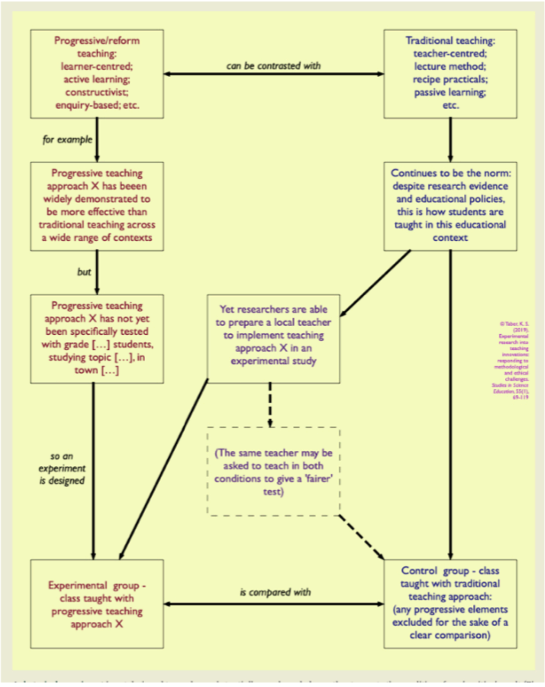
https://science-education-research.com/research-methodology/analysis/coding/coding-in-confirmatory-research/

If instead, we believed that hesitation and contradiction were essential to good science teaching; and so devised an observation schedule to spot incidents of hesitation and contradiction; and then observed the lessons of good science teachers, I suspect we could also find evidence of some level of hesitation and contradiction. So, what? That's one reason why quantitative, experimental studies have more kudos. Simple forms of qualitative studies are good at identifying potentially interesting points, but not suitable for testing substantive hypotheses.

But there is a genre of rhetorical experiments as well.

Rhetorical experiments

(a genre?)


Progressive/reform teaching: learner-centred; active learning; constructivist; enquiry-based; etc.
can be contrasted with
Traditional teaching: teacher-centred; lecture method; recipe practicals; passive learning; etc.
for example
Progressive teaching approach X has beeen widely demonstrated to be more effective than traditional teaching across a wide range of contexts
Continues to be the norm: despite research evidence and educational policies, this is how students are taught in this educational context
but
Progressive teaching approach X has not yet been specifically tested with grade […] students, studying topic […], in town […]
Yet researchers are able to prepare a local teacher to implement teaching approach X in an experimental study
so an experiment is designed
(The same teacher may be asked to teach in both conditions to give a 'fairer' test)
Experimental group - class taught with progressive teaching approach X
is compared with
Control group - class taught with traditional teaching approach: (any progressive elements excluded for the sake of a clear comparison)

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

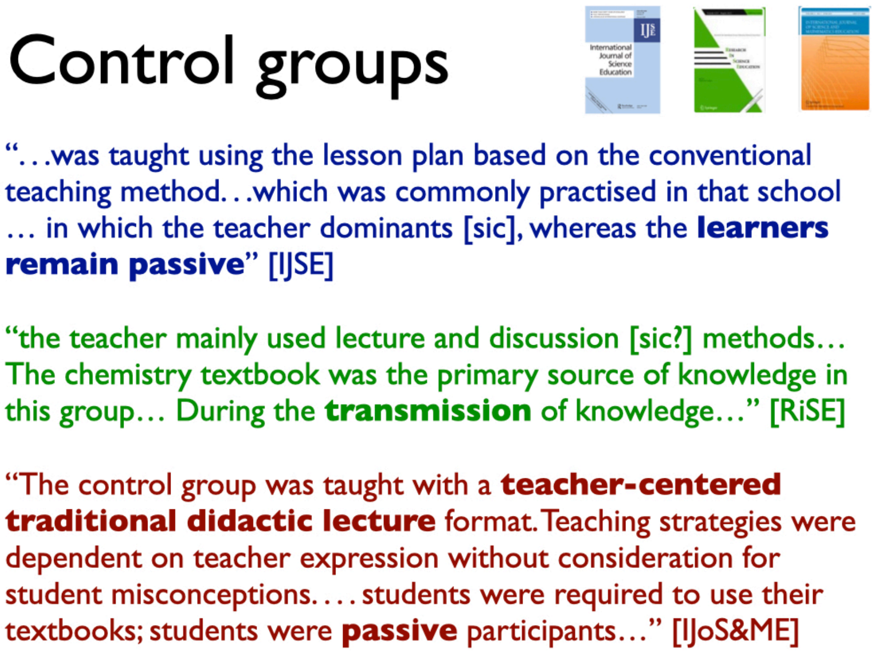
These papers usually start by making very strong claims about the merits of constructivism and some particular pedagogy associated with it. These papers will offer strong, convincing, theoretical arguments why such teaching is so much more effective than what may be labelled 'traditional' teaching (though by now I would have hoped constructivist-minded teaching would be traditional, but anyway). They then report a whole raft of prior studies which have shown the superiority of the focal pedagogy in a wide range of contexts - across geographical locations, age ranges, topics, languages of instruction, school cultures, etc.

But, they tell us, no one has yet tested the effectiveness of this pedagogy with, say, 13-14 year-olds studying the specific topic of the periodic table, in rural schools in South Cambridgeshire. Of course, from everything that has been reported, there is every reason to think this pedagogy will be effective in such a context. No sense of Karl Popper's notion of scientists testing bold conjectures, here: we only test dead-cert hypotheses.

But to make absolutely sure that the experiment works, we compare our preferred pedagogy with a teacher lecturing a class. There may be some 'discussion' in the sense of the teacher answering questions, but we do not allow group-work or any real dialogue, practical work, or access to digital technologies; and we restrict written work to exercises at the lower end of Bloom's taxonomy: recall, comprehension, and some application. The teacher teaches as if any educational thinking post about 1870 never happened. That is the comparison condition for our theoretically-sound, widely-proven, pedagogy.

I find this utterly unethical. It is unethical in at least two senses.

- Any research which inconveniences people for no good reason, is pointless and so a waste of resource. A demonstration posing as an experiment falls into this category.
- Asking control teachers to deliberately avoid good practice in their classroom, so to ensure a positive study outcome, is also clearly wrong.


IJSE
International Journal of Science Education
Research in Science Education

Control groups

"…was taught using the lesson plan based on the conventional teaching method…which was commonly practised in that school … in which the teacher dominants [sic], whereas the **learners remain passive**" [IJSE]

"the teacher mainly used lecture and discussion [sic?] methods… The chemistry textbook was the primary source of knowledge in this group… During the **transmission** of knowledge…" [RiSE]

"The control group was taught with a **teacher-centered traditional didactic lecture** format. Teaching strategies were dependent on teacher expression without consideration for student misconceptions.… students were required to use their textbooks; students were **passive** participants…" [IJoS&ME]

And here I have seen quite a few of these studies, including some published in good journals. How come reviewers do not spot that this is not genuine enquiry, and it is not an acceptable way to treat study participants? I can only assume we are all blinded by the assumed superiority of the experiment in science.

Falsifying conclusions

My final issue also reflects this idea, in that sometimes authors are so convinced about their expectations that they fail to observe the logic of their experiment.

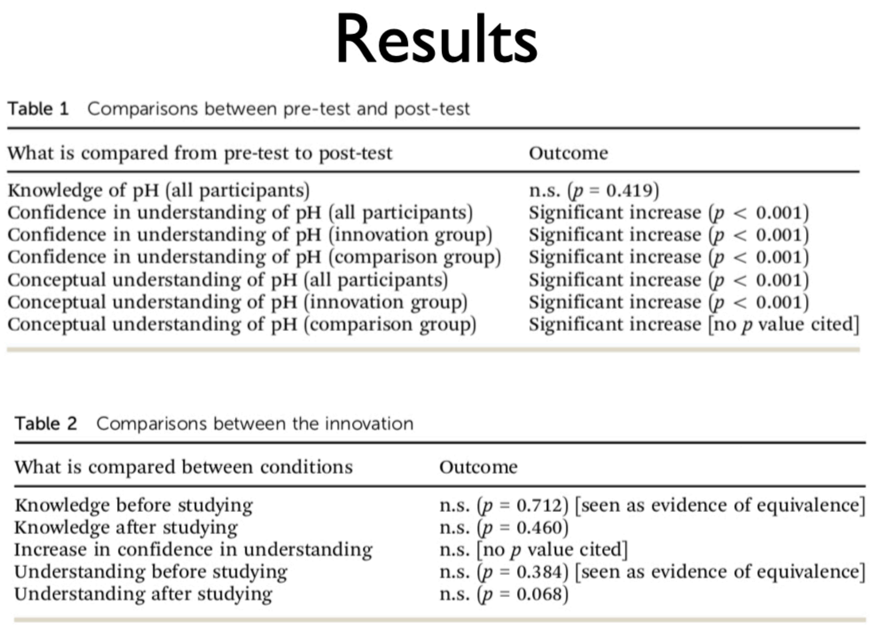
The confidence level chosen to determine whether differences between outcomes are statistically significant is arbitrary. Nearly always, we use a cut-off of 0.05: but there is nothing magical about that value. I suspect aliens with different numbers of fingers to us may have chosen a different critical value.

Whatever critical value we choose, the outcome is only an indicator. We are always subject to false positives where results are found significant, but unbeknown to us are actually due to chance effects; and false negatives, where results are found to be non-significant despite there being some kind of causal effect too small to reach significance in small samples.

But, if we are going to use inferential stats, then the conclusions of our experiments, certainly subject to appropriate caveats, should be determined by the outcomes of the analysis undertaken.

I am going to briefly refer to two papers, one published in each of the two most important chemistry education journals

A paper in *Chemistry Education Research and Practice*, reported positive outcomes from a study into a learning approach, and made a point of suggesting this refuted a previous study which had not found a significant effect. The whole question of replication is worthy of a talk of its own, but I am merely going to comment here that it was questionable whether the two studies were similar enough for one to be able to refute the other.



Results

Table 1 Comparisons between pre-test and post-test

What is compared from pre-test to post-test	Outcome
Knowledge of pH (all participants)	n.s. (p = 0.419)
Confidence in understanding of pH (all participants)	Significant increase (p < 0.001)
Confidence in understanding of pH (innovation group)	Significant increase (p < 0.001)
Confidence in understanding of pH (comparison group)	Significant increase (p < 0.001)
Conceptual understanding of pH (all participants)	Significant increase (p < 0.001)
Conceptual understanding of pH (innovation group)	Significant increase (p < 0.001)
Conceptual understanding of pH (comparison group)	Significant increase [no p value cited]

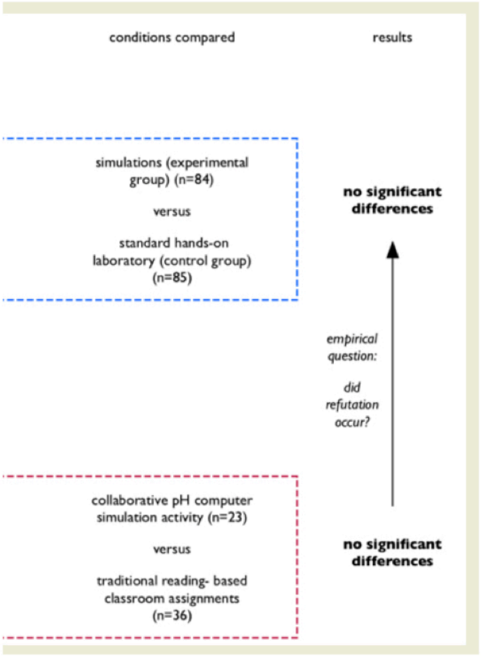
Table 2 Comparisons between the innovation

What is compared between conditions	Outcome
Knowledge before studying	n.s. (p = 0.712) [seen as evidence of equivalence]
Knowledge after studying	n.s. (p = 0.460)
Increase in confidence in understanding	n.s. [no p value cited]
Understanding before studying	n.s. (p = 0.384) [seen as evidence of equivalence]
Understanding after studying	n.s. (p = 0.068)

The study presented a range of results, comparing pre-test to post-test, and between the experimental and comparison groups. The two groups were considered equivalent at pretest, for the familiar reason that the differences were not so different as to reach statistical significance. My interpretation of a p value of 0.384 is that the differences between the two groups probably were NOT due to chance effects, but I think I have already said enough about that.

But surely that is not really important *in this study*, as there was not a significant difference between the two groups *after* the intervention. Any differences in scores between the two groups were not sufficiently unlikely so as be considered statistically significant.

So, in terms of the experimental design set out before collecting data, this is a negative result.


conditions compared
results
simulations (experimental group) (n=84)
versus
standard hands-on laboratory (control group) (n=85)
no significant differences
empirical question:
did refutation occur?
collaborative pH computer simulation activity (n=23)
versus
traditional reading- based classroom assignments (n=36)
no significant differences

"This research study found that collaborative computer sim group members experienced higher mean scores regarding pH knowledge and conceptual understanding, and indicated higher levels of pH-related confidence from the beginning to the end of the semester when compared to the traditional group members"

There is certainly nothing wrong with reporting a negative result, and indeed there is strong belief that research literature is distorted by a bias towards authors submitting, and journals preferentially publishing, positive outcomes. But, these authors are claiming to refute a study that did not find significant differences by having carried out another study that ALSO did not find significant differences. Here the negative result is imply ignored when presenting the study conclusions as being positive outcomes.

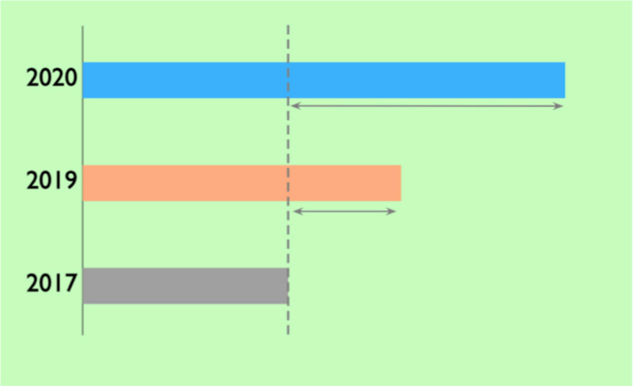
Sadly, having noticed this, I felt it necessary to write up a comment - which to be fair to the journal, was published. Again, we would not accept this kind of sloppy work from school children.

For my final example I turn to the other top chemistry education specific journal, the Journal of Chemical Education published by the American Chemical Society, whose journals are, according to the American Chemical Society itself, at least, 'most trusted'.

This is based on a figure showing results reported in a study in the Journal of Chemical Education. After 2017, the researchers modified their medicinal chemistry course, by implementing what they called 'Student-Centred Team-Based Learning Teaching Method', and, as you can see, student results improved thereafter.

A paper in *J.Chem.Ed.* reported

"…our results suggest that the SCTBL method is an effective way to improve teaching quality and student achievement."

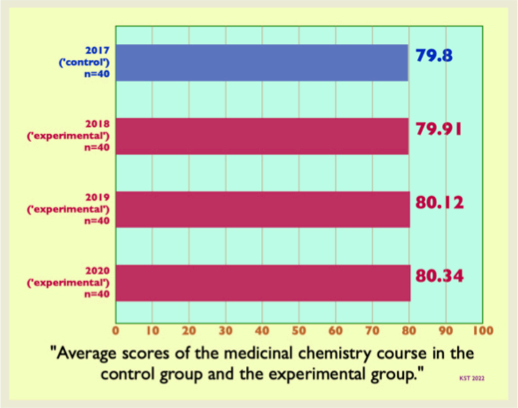


redrawn from Li, Ouyang, Xu & Zhang, 2022 in *Journal of Chemical Education*

You will notice I have failed to include any numbers on this figure, which is disingenuous of me, because the original did include the numbers. And with the numbers we see that we are viewing a graph with a truncated axis. That is a perfectly valid technique used to emphasise differences, but I wondered if here they might have over-emphasised the differences? They are just focusing on a small range of values. And if I present the whole graph as it might have been drawn, the change seems less impressive.

…*and compared to the full range*

"…our results suggest that the SCTBL method is an effective way to improve teaching quality and student achievement."



"Average scores of the medicinal chemistry course in the control group and the experimental group."

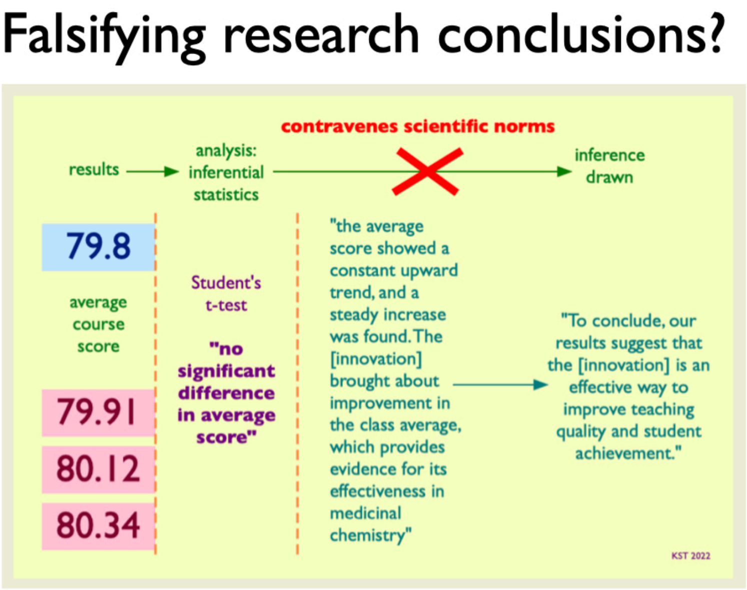
"the average score showed a constant upward trend, and a steady increase was found"

https://science-education-research.com/falsifying-research-conclusions/

Perhaps some of you are used to marking student work, and perhaps you feel your marking is entirely objective and very precise. But, given how cohorts shift from year to year, I really do not think average course scores to two places of decimals are meaningful. We would not accept this from school children. So, I have reworked their results:

cohort	*mean assessment score (2 s.f.)*
2017	80
2018	80
2019	80
2020	80

But even if I am being too fussy, and you think this level of precision can be justified, we might wonder how such small differences led to a statistically significant outcome. Of course, they did not. The authors did the analysis, and reported there was no significant difference between the scores before and after the new approach was implemented. Yet, these authors felt they could ignore this analysis and reach a positive conclusion.



Falsifying research conclusions?

contravenes scientific norms

results

analysis: inferential statistics

inference drawn

79.8

average course score

79.91

80.12

80.34

Student's t-test

"no significant difference in average score"

"the average score showed a constant upward trend, and a steady increase was found. The [innovation] brought about improvement in the class average, which provides evidence for its effectiveness in medicinal chemistry"

"To conclude, our results suggest that the [innovation] is an effective way to improve teaching quality and student achievement."

KST 2022

https://science-education-research.com/falsifying-research-conclusions/

Presumably the peer reviewers, and the editor, thought this was fine. I think this goes completely against proper scientific practice.

This is leaving aside such a comparison only makes sense if we think subsequent cohorts can be considered equivalent. The authors did not even use a weak test of equivalence of the kind I have discussed earlier, because they did not use any test of equivalence. That just assumed that the 40 students admitted to the course each year could be treated as equivalent. It was also clear from details in the paper that there had been some necessary modification in the assessment process in moving to the new teaching approach. This may have only effected a minor component of the final score, but that seems relevant when they think they are measuring to a hundredth of a percentage point. I just cannot see how peer reviewers thought this was okay.

So, I wrote to the editor about it. I was told my comments could be considered for publication if I submitted them as a formal submission. So, I did. Apparently, the editor initially asked three reviewers to read my submission. Two thought it should be published. One thought some changes were needed before it could be published. Now, I have been a journal editor, so I am pretty clear how I would respond to that review profile as an editor. But this editor was unsure. The editor then asked a fourth person, who thought the comment should be rejected. And so it was. I was told I could prepare a resubmission, but that if I did so it would be better to focus on general issues and not the specific paper I wanted to critique. That, of course, would have been a completely different article. I must admit to having been shocked by this outcome.

In conclusion

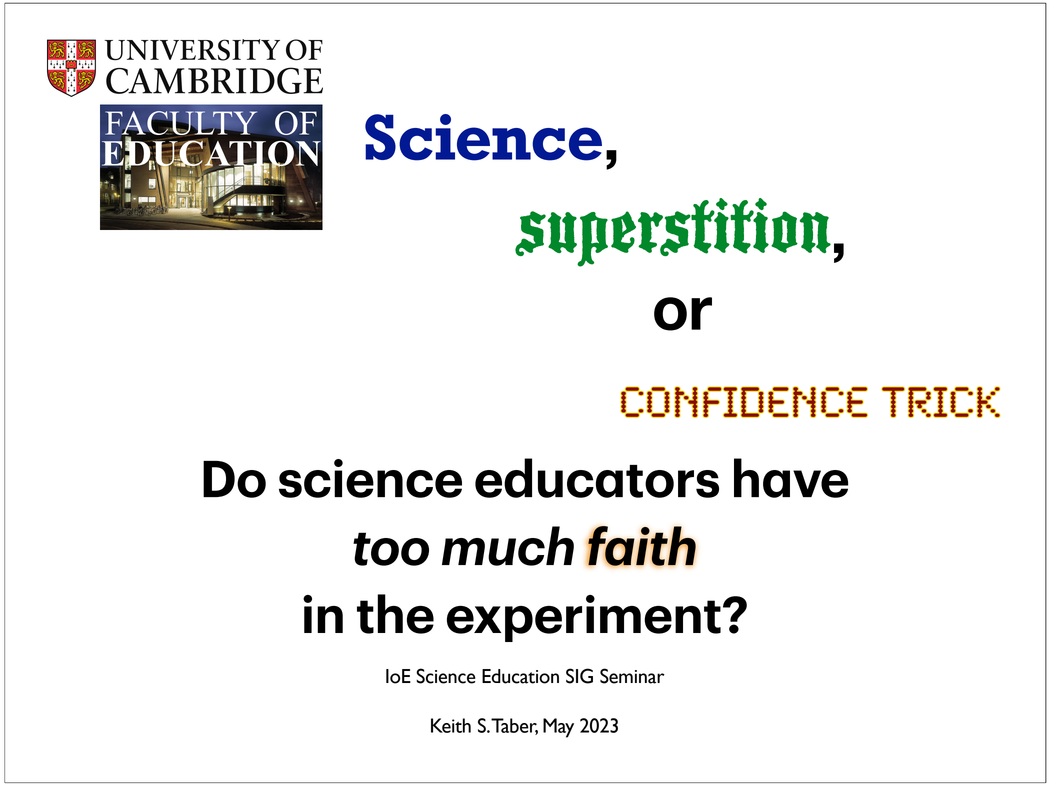
We are scientists, or we like to think we are, and scientists do experiments. I know from my time working in science teacher preparation that, generally, science graduates tend to think experiment is the method of choice when we ask them to undertake small-scale enquiry into their teaching.

- *Perhaps* our scientific training so promotes the merits of experimental procedures that science educators have an implicit bias that is strong enough to overcome any concerns about features that invalidate so many educational experiments of this kind.
- *Perhaps* our scientific education leads us to think that the world can be organised into natural kinds such that one copper wire of a certain gauge is assumed to be able to stand for any other copper wire of the same dimensions; and so one teacher, or one class of fifteen-year-olds, can stand for any other?
- *Perhaps* the scientific mindset that objectifies the natural world is so strong that we see people as experimental subjects that respond to our treatments without regard to the inter-subjective nature of our interactions with them and what they might think of us and our experiments?
- *Perhaps* this is also why we tend to see classes as arrays of individuals and forget that people in groups interact and influence each other.

I would suggest that when a valid experiment is possible, it is usually to be preferred. But if we cannot do a valid experiment, we should not do an experiment at all.

- An invalid experiment is not scientific, and is a waste of valuable resource - including researcher time and participant goodwill.
- An invalid experiment carried out by a researcher undermines their claim to be a competent scientist.

So, *what does the science education community's propensity for publishing invalid experiments in its journals say about us collectively?*



UNIVERSITY OF CAMBRIDGE
FACULTY OF EDUCATION

Science, superstition, or CONFIDENCE TRICK

Do science educators have *too much faith* in the experiment?

IoE Science Education SIG Seminar

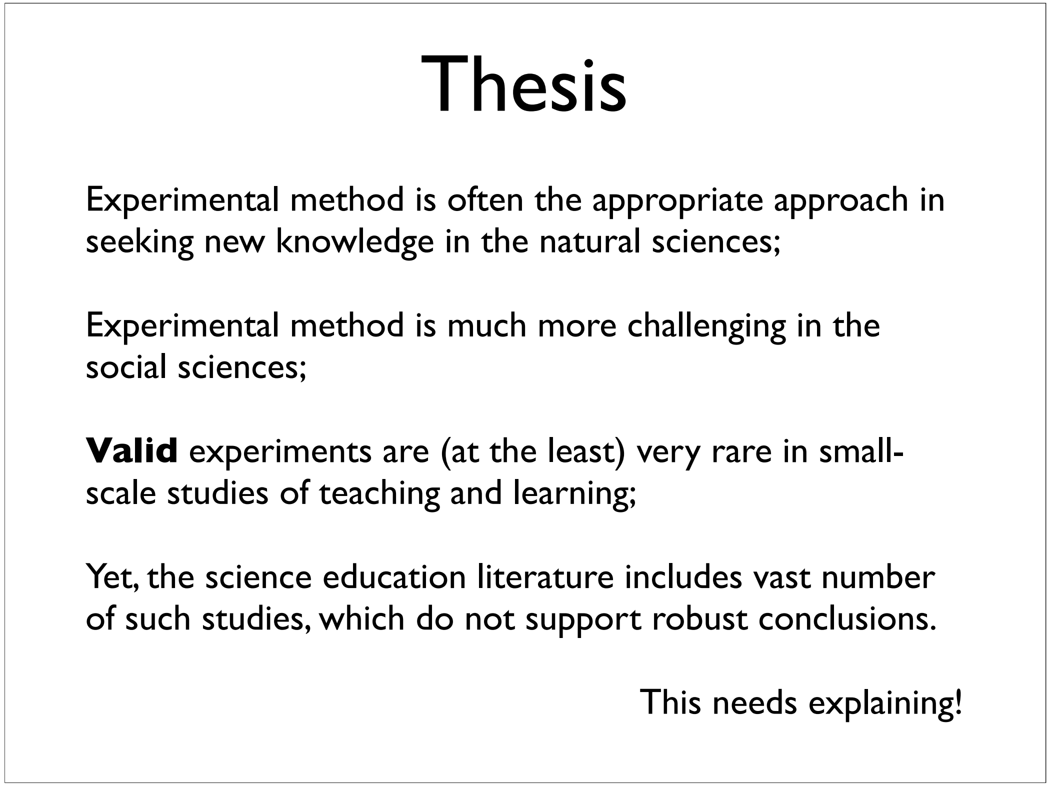
Keith S. Taber, May 2023

Abstract:

A rigorous experiment is rightly considered an especially informative research tool. But doing rigorous experiments in education is very challenging. A poorly designed experiment may tell us very little. Yet the literature includes a vast number of experimental studies in science education.

Here I make an argument that:
* Often these are very small scale studies with unrepresentative populations;
* Often the extent of control of variables would not pass for 'fair testing' in a school laboratory exercise;
* There is a sleight of hand commonly used to (mis)apply statistical tests to treat samples as much larger than they are;
* Sometimes authors draw conclusions contrary to their results; and
* Sometimes inappropriate (unethical) control conditions are imposed on learners for the sake of research.

I will pose, and reflect on, the question of why so many of these flawed studies get undertaken in science education and published in peer-reviewed journals.



Thesis

Experimental method is often the appropriate approach in seeking new knowledge in the natural sciences;

Experimental method is much more challenging in the social sciences;

Valid experiments are (at the least) very rare in small-scale studies of teaching and learning;

Yet, the science education literature includes vast number of such studies, which do not support robust conclusions.

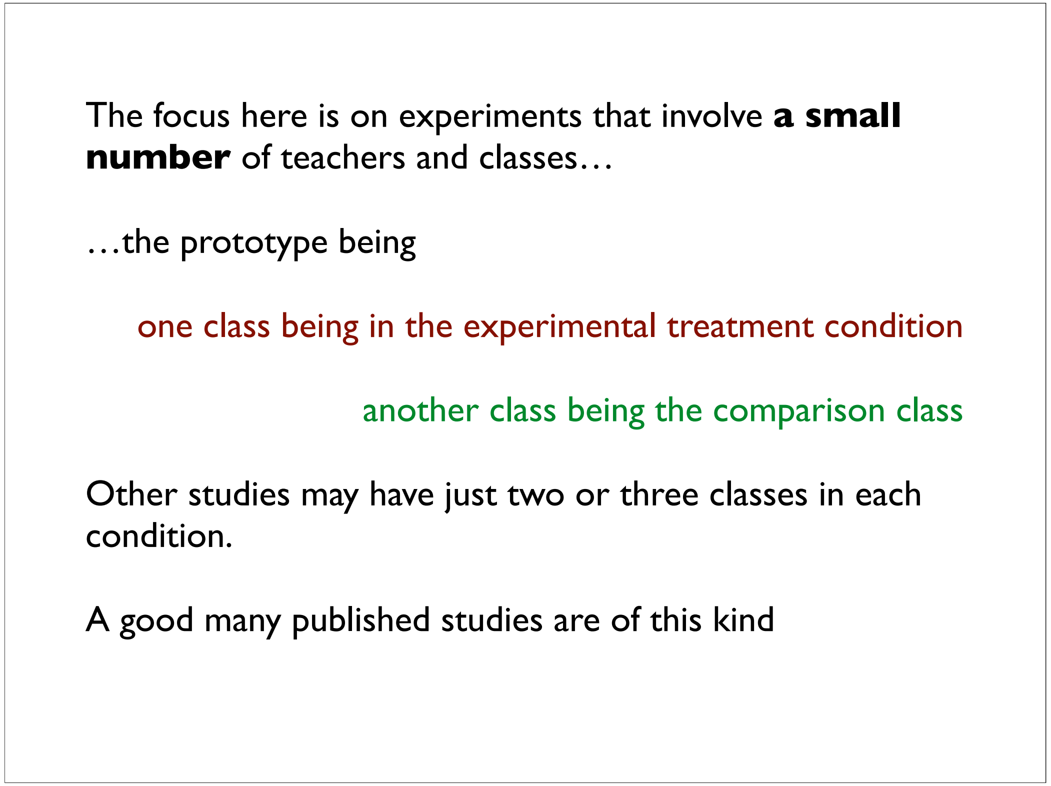
This needs explaining!

The thesis I am advancing is that there is a phenomenon to be explained. That phenomenon is that the science education research literature includes a very large number of studies that are experimental in nature, but which do not meet the criteria for being valid experiments. Many of these are published in regional or national journals, but there are plenty in more prestigious journals. Perhaps everybody involved - authors, peer reviewers, editors - recognise the problem with these studies, and see their findings as purely indicative - but often that is not the impression given by their framing and phrasing.

I argue that

- Experimental method is often the appropriate approach in seeking new knowledge in the natural sciences;
- Experimental method is much more challenging in the social sciences;
- Valid experiments are (at the least) very rare in small-scale studies of teaching and learning;
- Yet, the science education literature includes a vast number of such studies, which do not support robust conclusions.

This needs explaining!



The focus here is on experiments that involve **a small number** of teachers and classes…

…the prototype being

one class being in the experimental treatment condition

another class being the comparison class

Other studies may have just two or three classes in each condition.

A good many published studies are of this kind

My focus here is on experiments that involve a small number of teachers and classes…

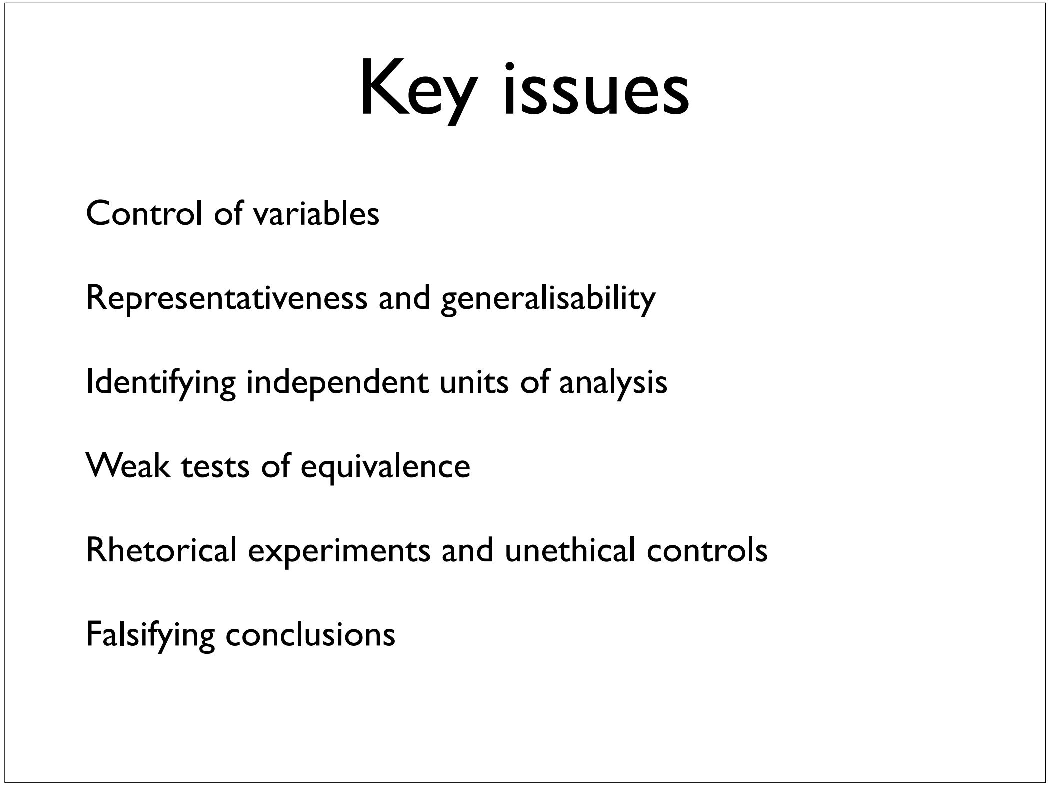
Indeed, the prototype of this type of study involves

- one class in the experimental treatment condition
- another class being the comparison class

Sometimes both classes are taught by the same teacher, sometime not. Sometimes the classes are not even from the same school.

Other studies may have just two or three classes in each condition.

A good many published studies are of this kind.



Key issues

Control of variables

Representativeness and generalisability

Identifying independent units of analysis

Weak tests of equivalence

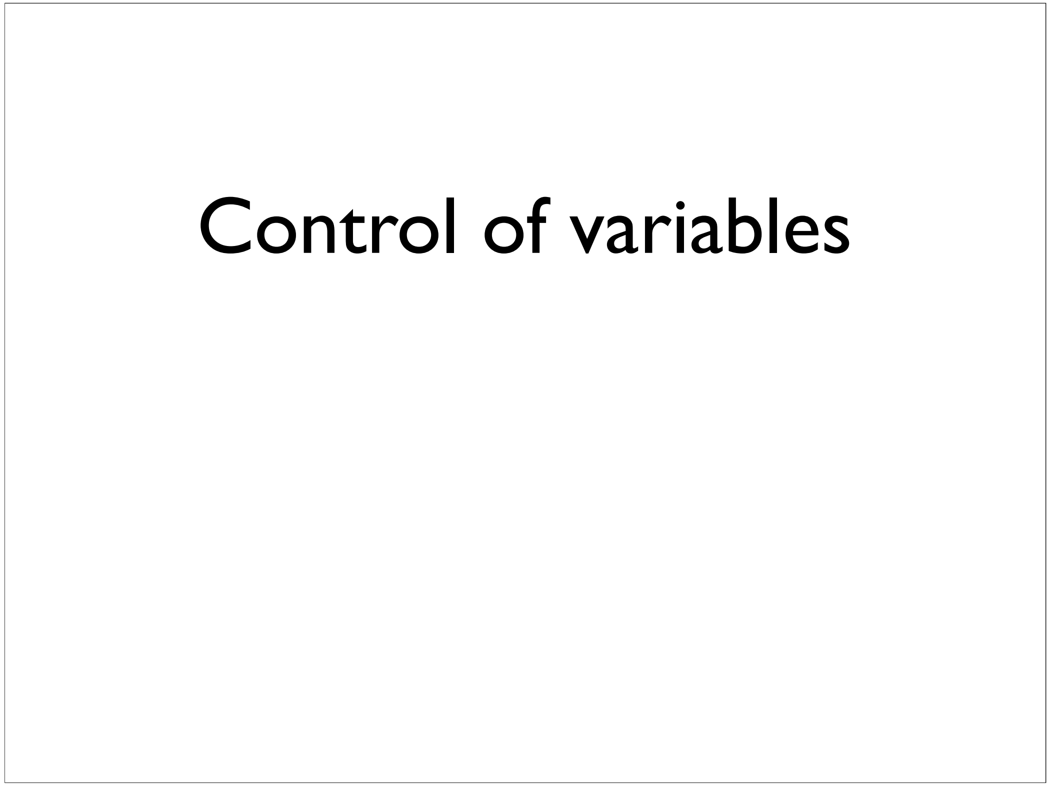
Rhetorical experiments and unethical controls

Falsifying conclusions

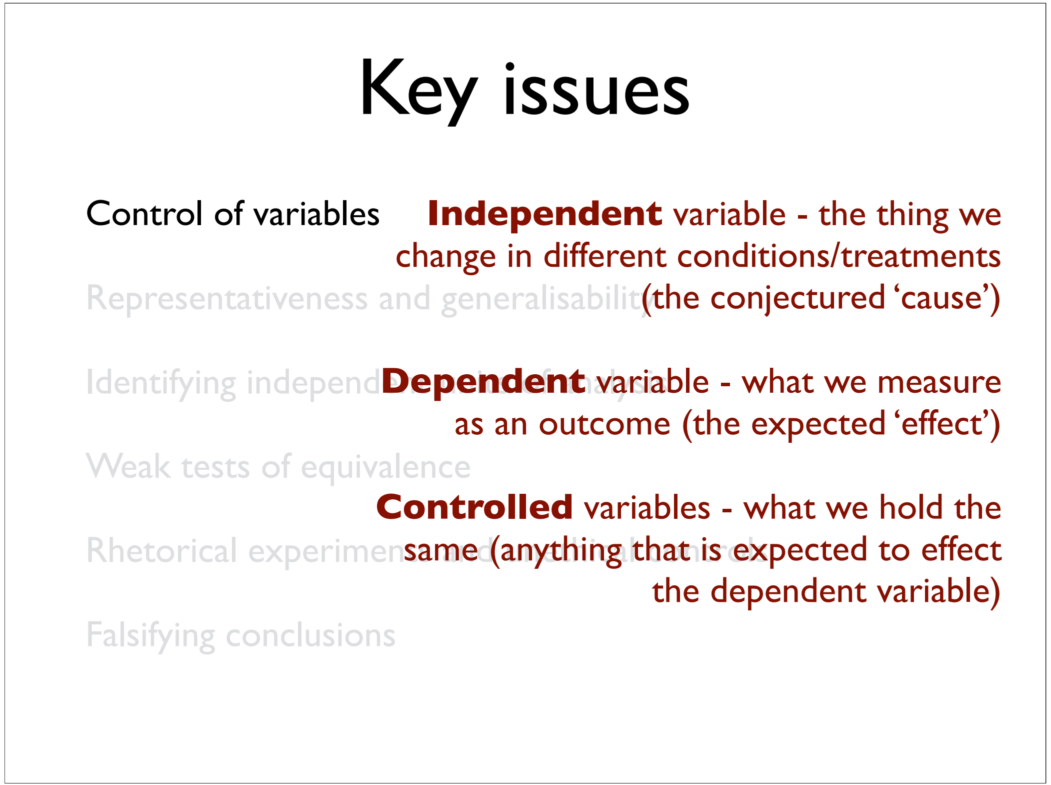
There is a host of difficulties in doing experimental studies into classroom teaching, and I have published an account of a number of these in a review for *Studies in Science Education* [*]. Here I wish to focus on, and illustrate, a few themes

- Control of variables
- Representativeness and generalisability
- Identifying independent units of analysis
- Inadequate tests of equivalence
- Rhetorical experiments and unethical controls
- Falsifying conclusions

[* Taber, K. S. (2019). 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 https://science-education-research.com/publications/papers/experimental-research-into-teaching-innovations/]



Control of variables



Key issues

Control of variables

Representativeness and generalisability

Identifying independent units of analysis

Weak tests of equivalence

Rhetorical experiments and medical controls

Falsifying conclusions

Independent variable - the thing we change in different conditions/treatments (the conjectured 'cause')

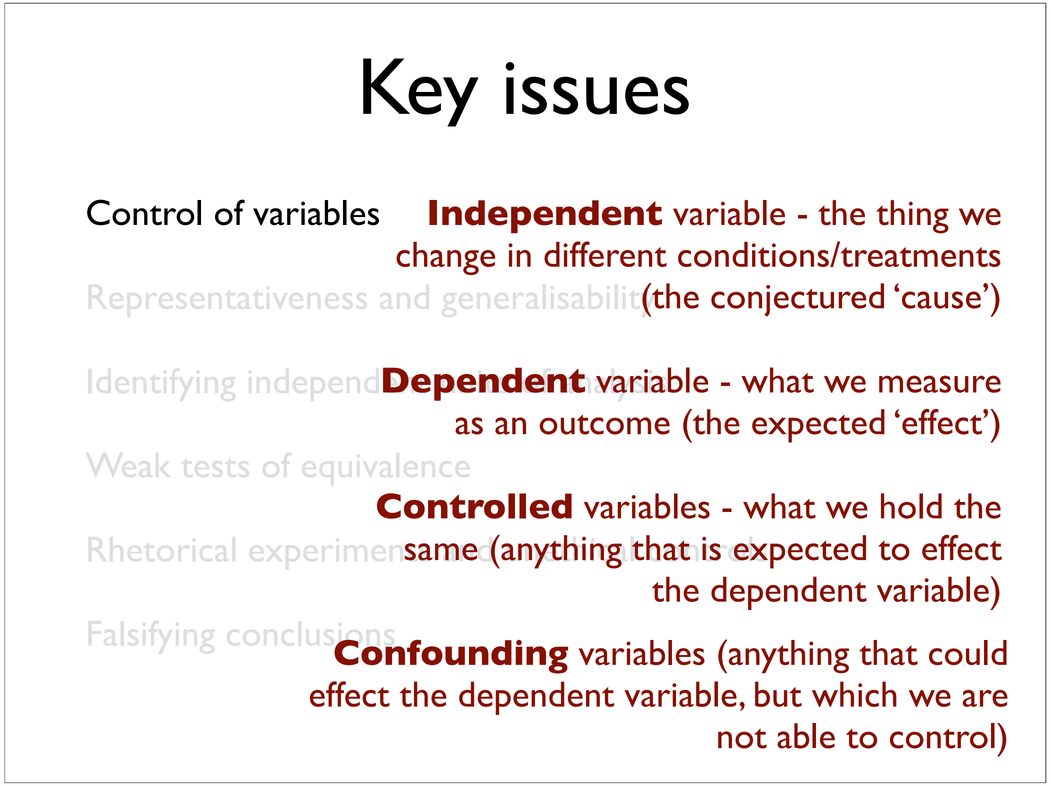
Dependent variable - what we measure as an outcome (the expected 'effect')

Controlled variables - what we hold the same (anything that is expected to effect the dependent variable)

Now a good experiment involves three classes of variable:

- the thing we are deliberately changing,
- the thing we allow to vary to see if, and if so how, it changes, and
- everything else which could have an effect, and, so, is not allowed to change.

To do an experiment could be seen as in large part controlling all the other things that could effect our results and confuse the possible dependency of the dependent variable upon our planned intervention.



Key issues

Control of variables

Representativeness and generalisability

Identifying independent units of analysis

Weak tests of equivalence

Rhetorical experiments and unethical controls

Falsifying conclusions

Independent variable - the thing we change in different conditions/treatments (the conjectured ‘cause’)

Dependent variable - what we measure as an outcome (the expected ‘effect’)

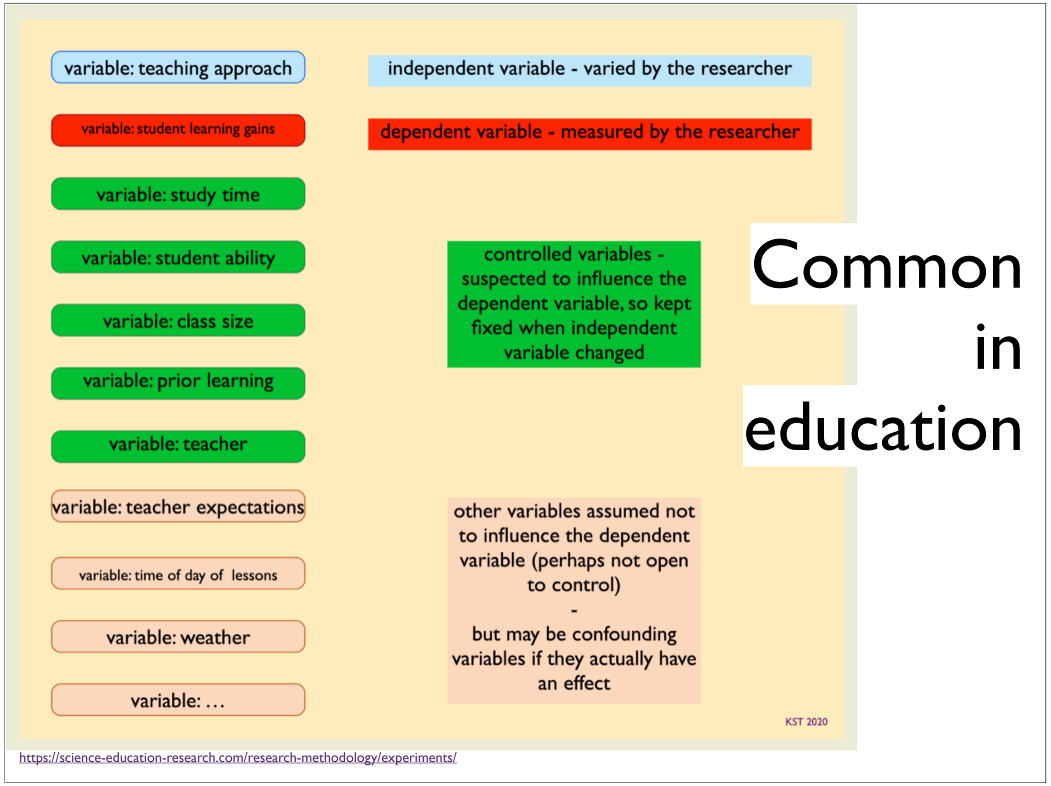
Controlled variables - what we hold the same (anything that is expected to effect the dependent variable)

Confounding variables (anything that could effect the dependent variable, but which we are not able to control)

A poor experiment has a fourth class of variable. These are all the things that should be in the controlled category, but which we do not control.

Confounding variables confound our study because we logically need to caveat the conclusions with an ‘unless this was due to something else’.

Now, even in the natural sciences, there can be confounding variables, because it is never possible to control everything that we might imagine could be considered a variable, so, on theoretical grounds, we dismiss such possibilities as the relevance of a researcher’s hair colour or whether they use their left or right eye to look into the microscope. We use existing theoretical knowledge to judge what we can ignore. Even here, it sometimes transpires there was a pertinent variable that influences results, but which had been assumed to be irrelevant, or was even unknown to the researchers.



Common in education

variable: teaching approach

variable: student learning gains

variable: study time

variable: student ability

variable: class size

variable: prior learning

variable: teacher

variable: teacher expectations

variable: time of day of lessons

variable: weather

variable: …

independent variable - varied by the researcher

dependent variable - measured by the researcher

controlled variables - suspected to influence the dependent variable, so kept fixed when independent variable changed

other variables assumed not to influence the dependent variable (perhaps not open to control)
-
but may be confounding variables if they actually have an effect

KST 2020

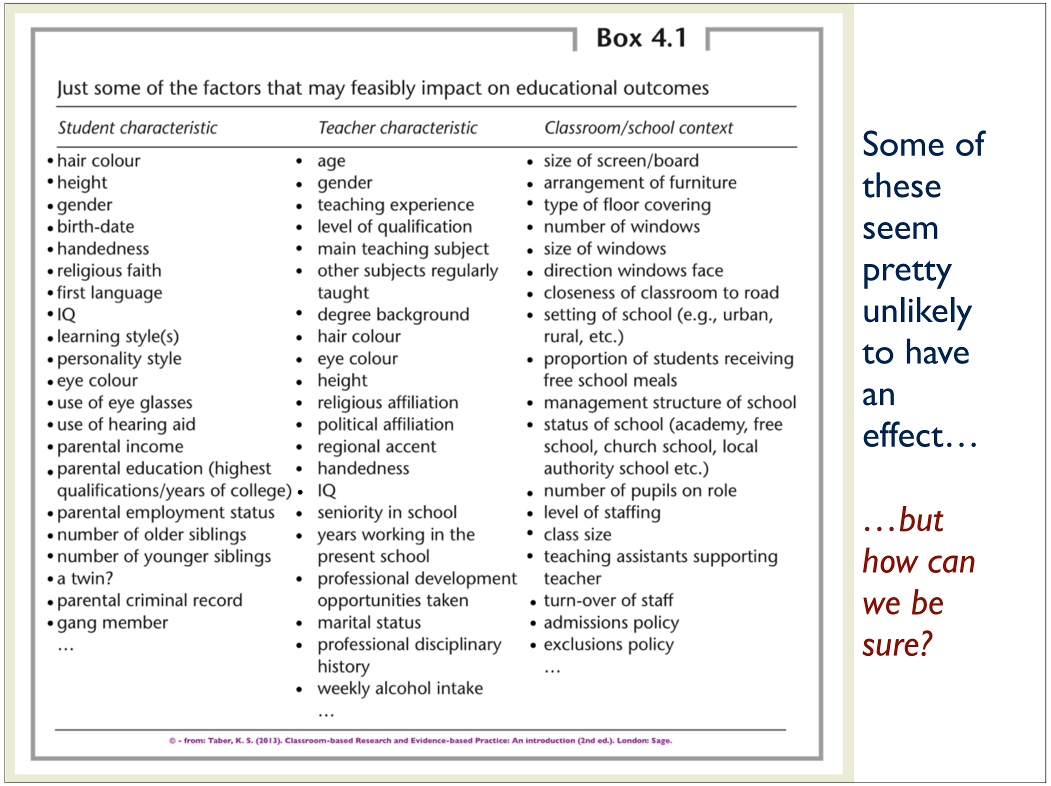
https://science-education-research.com/research-methodology/experiments/

In the social sciences, control of variables is very much more challenging.

For one thing, we usually work in naturally occurring social contexts, rather than isolate systems for laboratory manipulation.

Another major difference is that physicists and chemists do not have to consider what their research materials think about them, or expect to happen in their experiments. A piece of alloy, or a solution of an oxidising agent, has no preconceptions about the outcome of the experiment, and does not have an attitude to the researcher. You do not need to develop rapport with your test tubes.

https://science-education-research.com/research-methodology/experiments/



Box 4.1

Just some of the factors that may feasibly impact on educational outcomes

Student characteristic

- hair colour
- height
- gender
- birth-date
- handedness
- religious faith
- first language
- IQ
- learning style(s)
- personality style
- eye colour
- use of eye glasses
- use of hearing aid
- parental income
- parental education (highest qualifications/years of college)
- parental employment status
- number of older siblings
- number of younger siblings
- a twin?
- parental criminal record
- gang member

…

Teacher characteristic

- age
- gender
- teaching experience
- level of qualification
- main teaching subject
- other subjects regularly taught
- degree background
- hair colour
- eye colour
- height
- religious affiliation
- political affiliation
- regional accent
- handedness
- IQ
- seniority in school
- years working in the present school
- professional development opportunities taken
- marital status
- professional disciplinary history
- weekly alcohol intake

…

Classroom/school context

- size of screen/board
- arrangement of furniture
- type of floor covering
- number of windows
- size of windows
- direction windows face
- closeness of classroom to road
- setting of school (e.g., urban, rural, etc.)
- proportion of students receiving free school meals
- management structure of school
- status of school (academy, free school, church school, local authority school etc.)
- number of pupils on role
- level of staffing
- class size
- teaching assistants supporting teacher
- turn-over of staff
- admissions policy
- exclusions policy

…

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

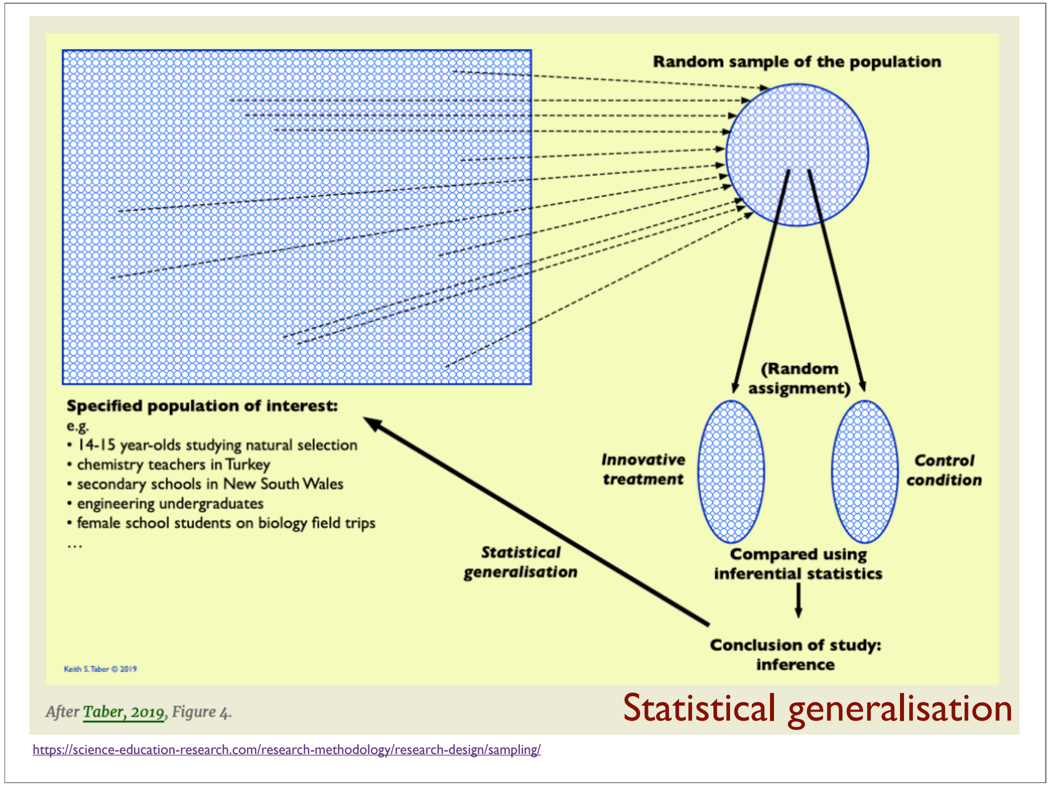
Some of these seem pretty unlikely to have an effect…

…but how can we be sure?

Indeed, if you have an active imagination, you can likely think of feasible scenarios when any number of things could effect the outcomes of an educational experiment such as student learning.

I do not think it is an exaggeration to argue that in the case of an educational experiment:

- we often cannot identify all the variables which might have an effect;
- even when we can identify them, we often do not know how to meaningfully measure them;
- even if we can measure them, we may not have a way of holding them at a constant value.



Random sample of the population

(Random assignment)

Innovative treatment

Control condition

Compared using inferential statistics

Conclusion of study: inference

Statistical generalisation

Specified population of interest:
e.g.
- 14-15 year-olds studying natural selection
- chemistry teachers in Turkey
- secondary schools in New South Wales
- engineering undergraduates
- female school students on biology field trips

…

Keith S.Taber © 2019

After Taber, 2019, Figure 4.

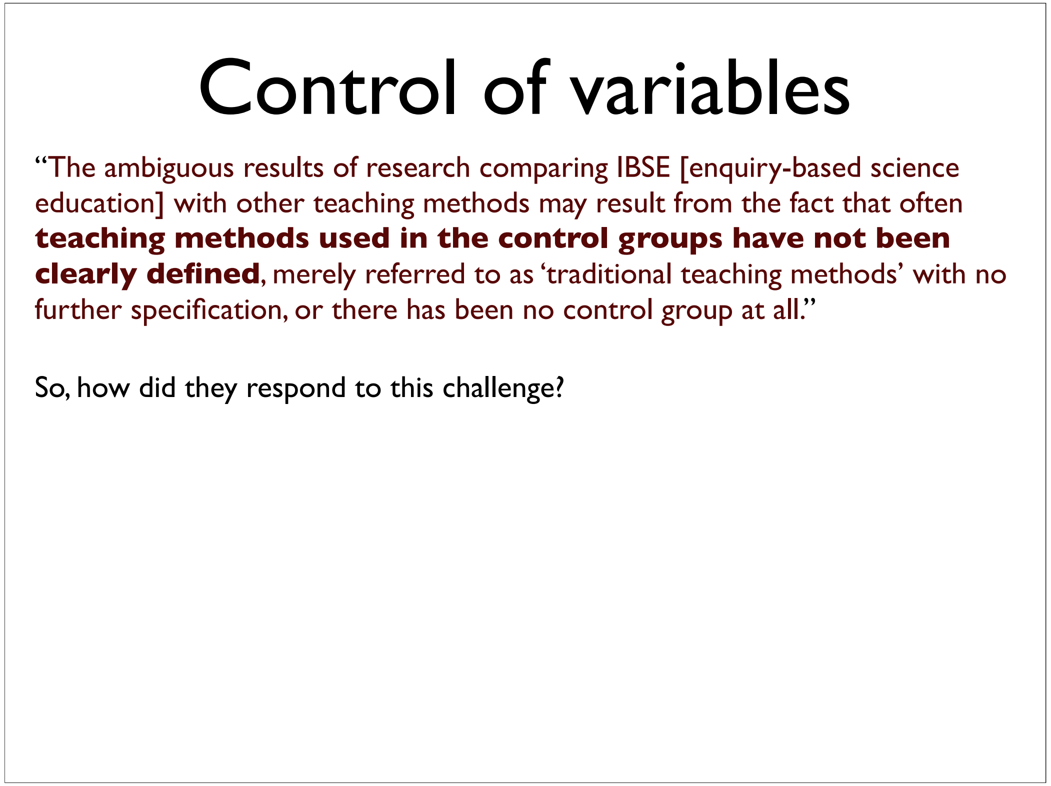
Statistical generalisation

https://science-education-research.com/research-methodology/research-design/sampling/

That need not prevent experimental studies where you have large enough samples that are representative of a population to be able to assume that statistics can tell you whether any outcomes are unlikely to be due to such chance factors. This is a point I will return to.

It is a problem, though, in any studies with small, unrepresentative, samples. And most of the published experiments in the educational literature have small, unrepresentative, samples.

https://science-education-research.com/research-methodology/research-design/sampling/



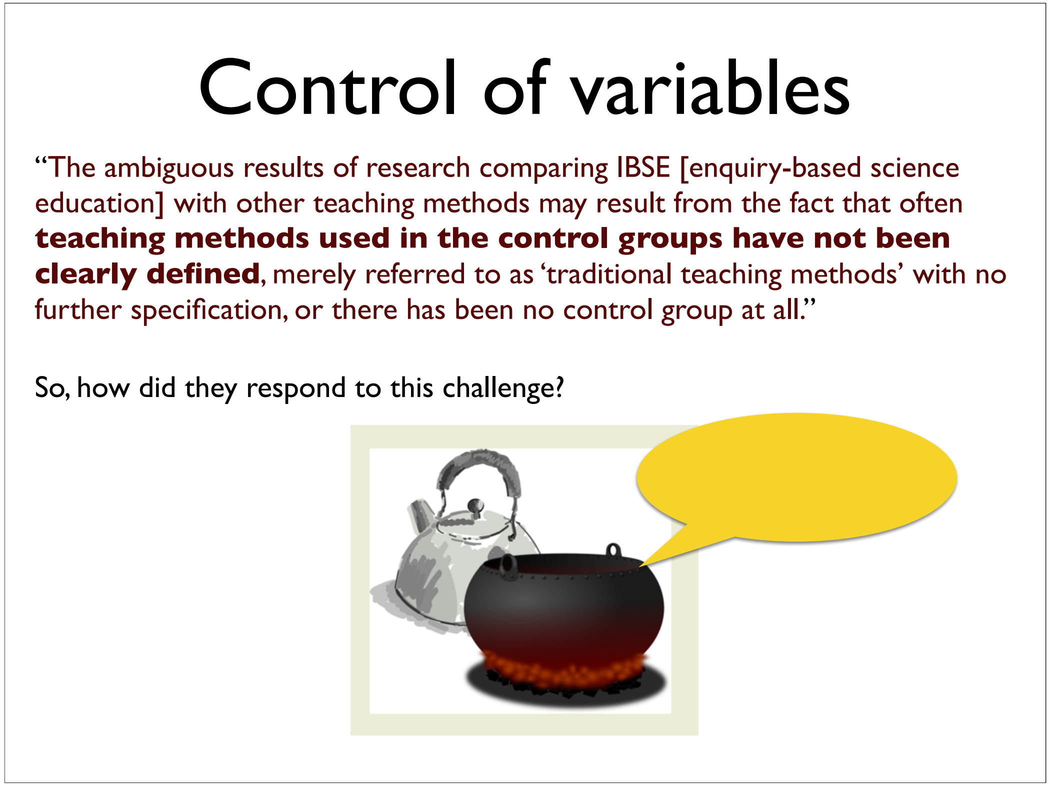
Control of variables

"The ambiguous results of research comparing IBSE [enquiry-based science education] with other teaching methods may result from the fact that often **teaching methods used in the control groups have not been clearly defined**, merely referred to as 'traditional teaching methods' with no further specification, or there has been no control group at all."

So, how did they respond to this challenge?

As one example of the kinds of issues that can arise, I was impressed to read this comment in a research paper.

The authors had noticed that in many experimental studies the experimental treatment is well defined, but the 'control' condition is anything but controlled. It is actually laissez-faire, anything goes, as long as the teacher avoids the approach being taken in the experimental condition.

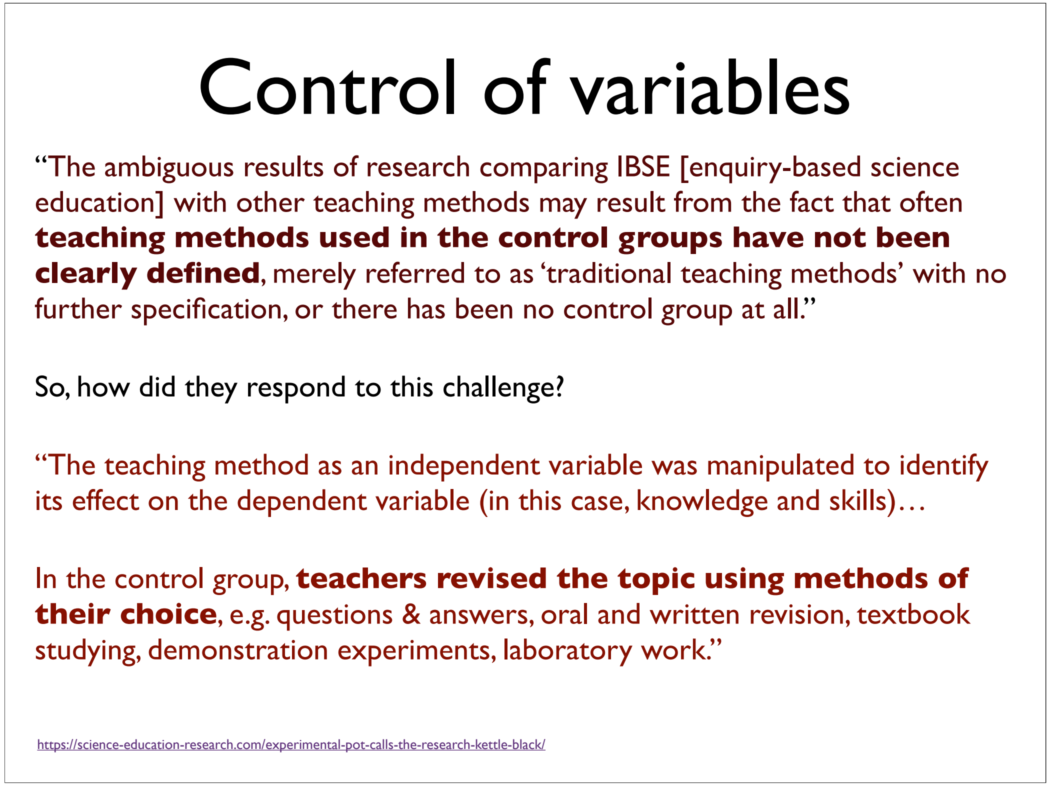


Control of variables

“The ambiguous results of research comparing IBSE [enquiry-based science education] with other teaching methods may result from the fact that often **teaching methods used in the control groups have not been clearly defined**, merely referred to as ‘traditional teaching methods’ with no further specification, or there has been no control group at all.”

So, how did they respond to this challenge?

This seemed a fair point, so I read on to see how they managed this issue in their own study.



Control of variables

"The ambiguous results of research comparing IBSE [enquiry-based science education] with other teaching methods may result from the fact that often **teaching methods used in the control groups have not been clearly defined**, merely referred to as 'traditional teaching methods' with no further specification, or there has been no control group at all."

So, how did they respond to this challenge?

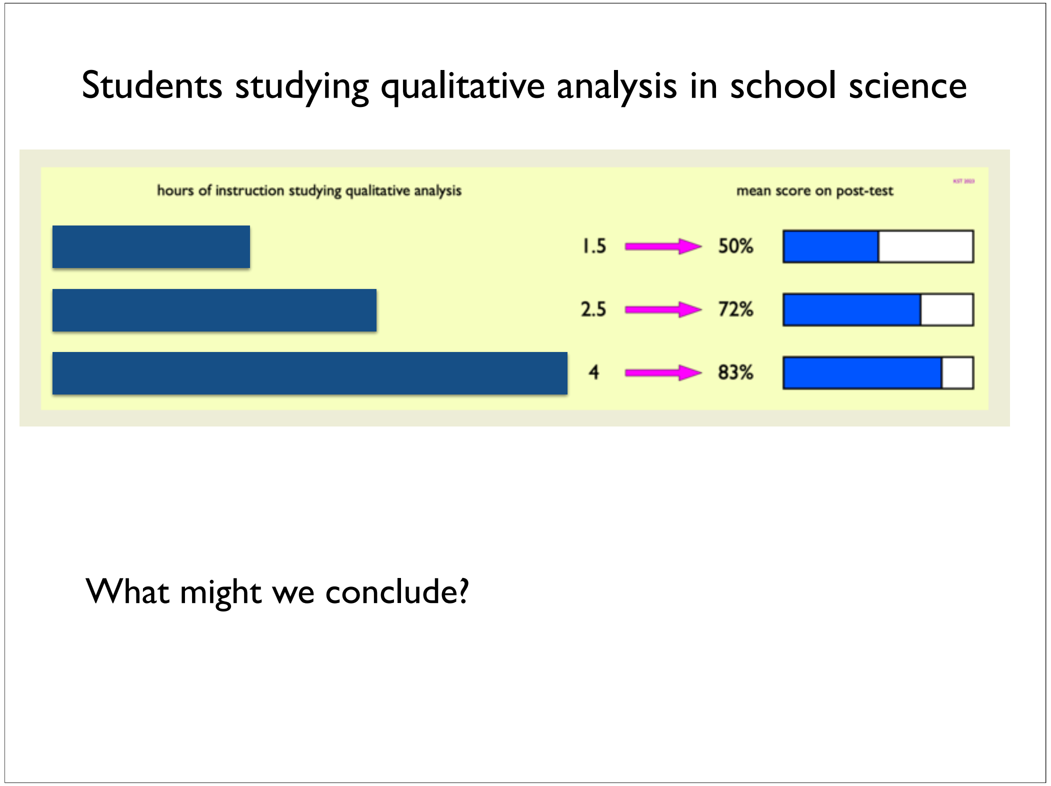
"The teaching method as an independent variable was manipulated to identify its effect on the dependent variable (in this case, knowledge and skills)…

In the control group, **teachers revised the topic using methods of their choice**, e.g. questions & answers, oral and written revision, textbook studying, demonstration experiments, laboratory work."

https://science-education-research.com/experimental-pot-calls-the-research-kettle-black/

It seems that raising an issue which undermines the ability to draw clear conclusions from a study does not impose a requirement to address the issue in your own study!

https://science-education-research.com/experimental-pot-calls-the-research-kettle-black/



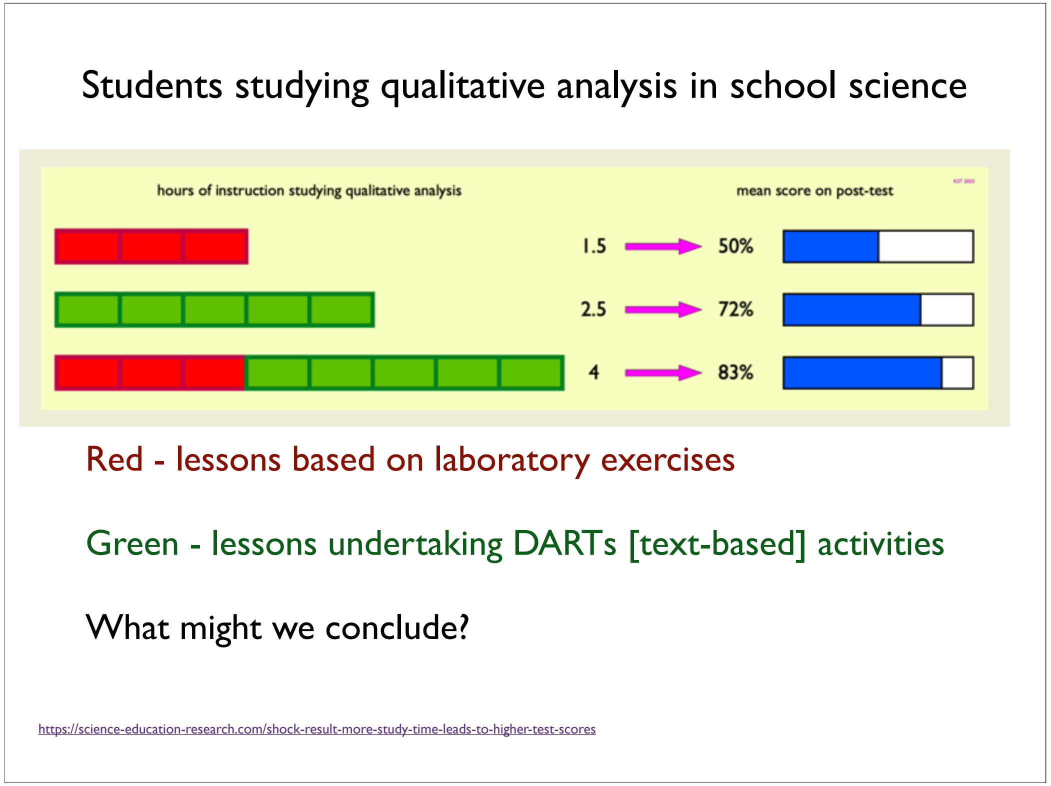
Students studying qualitative analysis in school science

What might we conclude?

Now here I present the results I found in a published research study.

As you see, three classes were taught the topic of chemical qualitative analysis for different lengths of time. Afterwards, the average performance of the students in these classes was found to vary.

I wonder what you think we might be able to conclude from this study?



Students studying qualitative analysis in school science

Red - lessons based on laboratory exercises

Green - lessons undertaking DARTs [text-based] activities

What might we conclude?

https://science-education-research.com/shock-result-more-study-time-leads-to-higher-test-scores

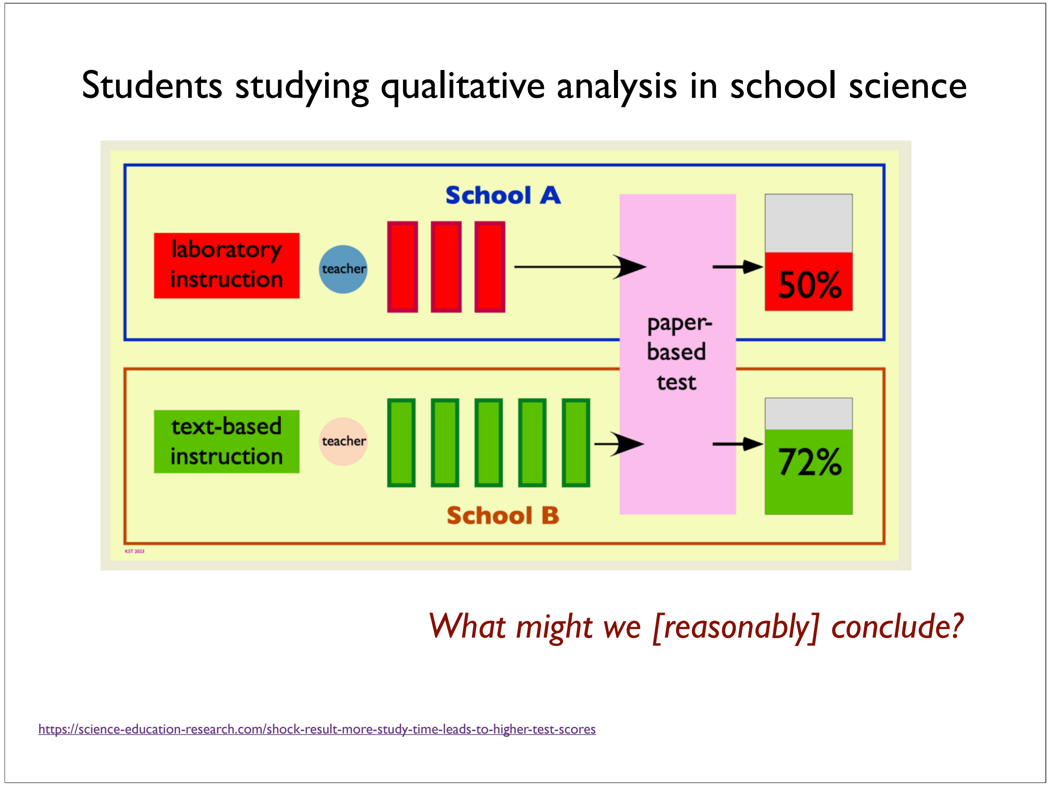
Now, as you may have guessed, I was not giving you all the details. Here I reveal more.

One class had three lessons in the lab. But that class was outperformed by a class that had five lessons of paper-based learning activities. The third class, spent three lessons in the lab. and also had the five lessons of text-based activities.

It perhaps seems reasonable to conclude that the class that had both kinds of activity learnt the most. But what about comparing the first two classes?

I would mischievously suggest five lessons can lead to more learning than three, but the authors thought this was evidence of the superiority of the text-based learning approach. Presumably, the peer-reviewers and editor were sufficiently convinced.

https://science-education-research.com/shock-result-more-study-time-leads-to-higher-test-scores/



Students studying qualitative analysis in school science

What might we [reasonably] conclude?

https://science-education-research.com/shock-result-more-study-time-leads-to-higher-test-scores

But I was not convinced.

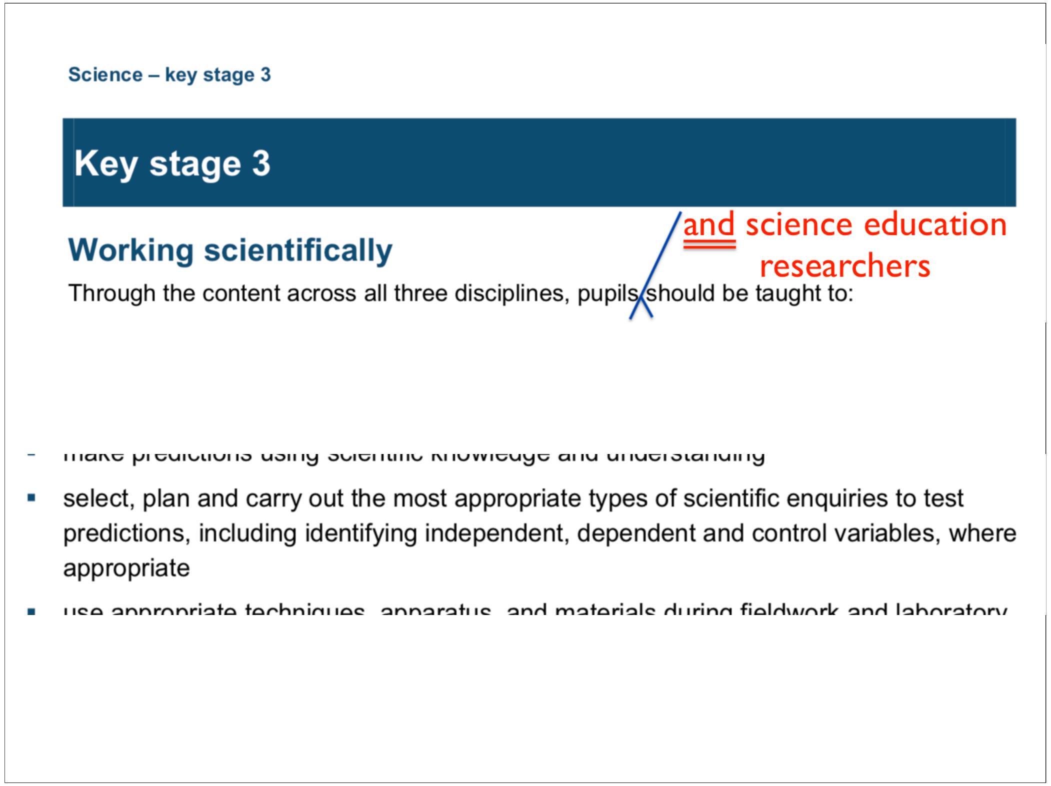
The two classes were taught by different teachers.

The two classes were drawn from different schools. They were considered to be similar schools, but even so.

For that matter, the assessment tool was a paper-based test, so how do we know that if a laboratory-based assessment had been used, the results would not have been very different? After all, to actually do qualitative analysis, you need to work with real samples and reagents in a laboratory.

You might feel that I am only stating the obvious. But if it is so obvious, how does such work get published without strong caveats, and sometimes even in the more prestigious journals.

https://science-education-research.com/shock-result-more-study-time-leads-to-higher-test-scores



Science – key stage 3

Key stage 3

Working scientifically

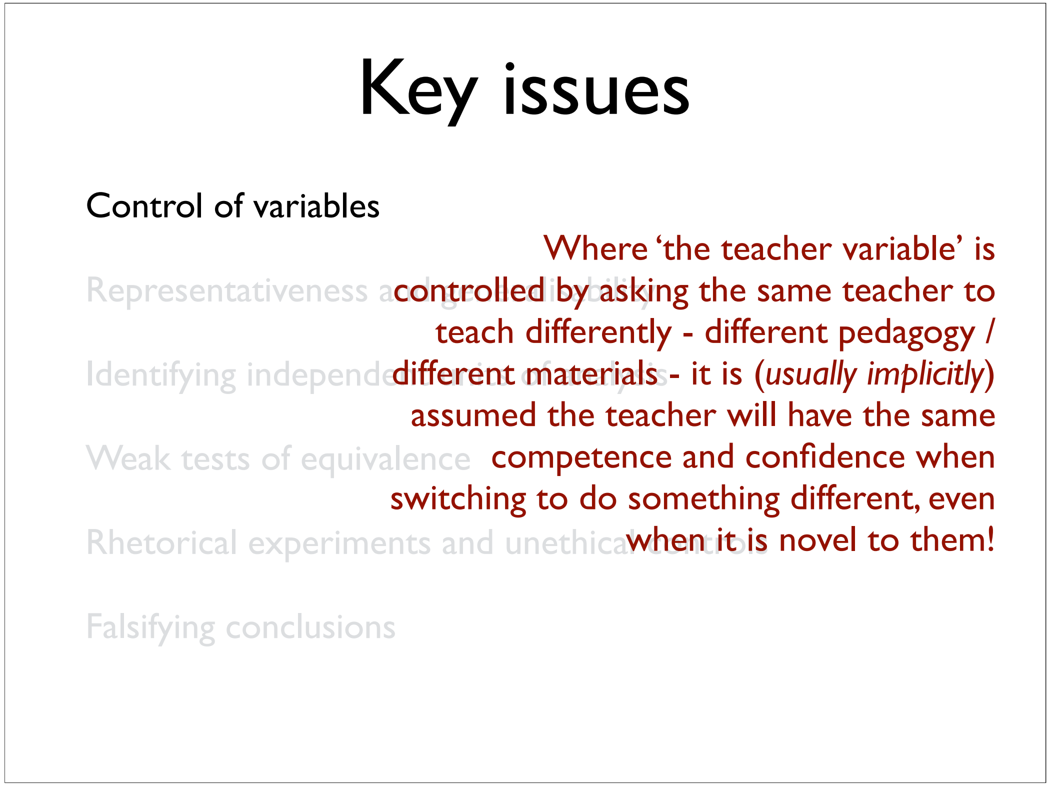
and science education researchers

Through the content across all three disciplines, pupils should be taught to:

- make predictions using scientific knowledge and understanding
- select, plan and carry out the most appropriate types of scientific enquiries to test predictions, including identifying independent, dependent and control variables, where appropriate
- use appropriate techniques, apparatus, and materials during fieldwork and laboratory

After all, we expect 14 years olds to do better than this.

Arguably, science educators are teaching skills they then fail to display in their own work.



Key issues

Control of variables

Where ‘the teacher variable’ is controlled by asking the same teacher to teach differently - different pedagogy / different materials - it is (*usually implicitly*) assumed the teacher will have the same competence and confidence when switching to do something different, even when it is novel to them!

Representativeness and generalisability

Identifying independent units of analysis

Weak tests of equivalence

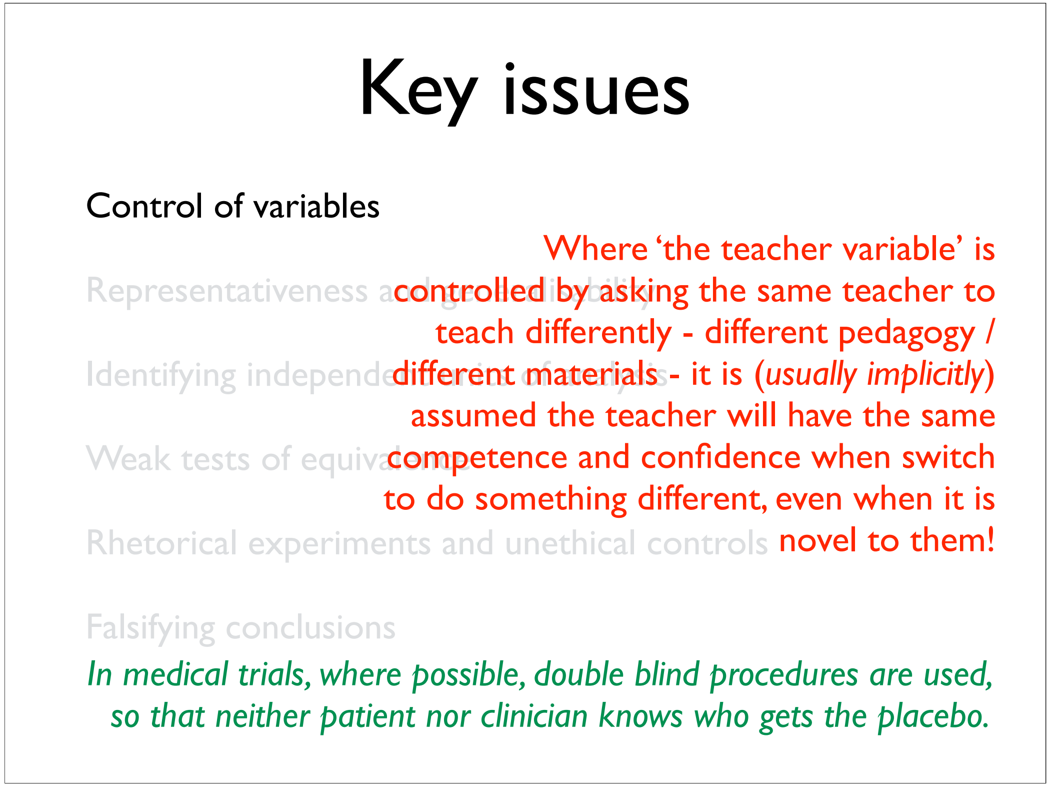
Rhetorical experiments and unethical controls

Falsifying conclusions

One of the variables not controlled in that study, was the teacher. Perhaps the two different teachers were very similar in all relevant characteristics, but that is not very likely.

Sometimes the ‘teacher variable ’ is ‘controlled’ by asking the same teacher to teach differently in two different conditions. That is, the teacher is asked to use two different teaching approaches in different classes. Jig-saw learning here, but computer simulations there. Or, sadly, more likely, enquiry-based teaching here, and dictation of notes there. More on that choice later. Or, a teacher is asked to trial a new curriculum module whilst teaching a parallel class according to the established scheme of work.

This assumes, at least implicitly, the teacher will have the same competence and confidence when switching to do something different, even when it is novel to them. The same teacher, teaching different classes, is assumed to have controlled that variable, but I do not find that very convincing.



Key issues

Control of variables

Representativeness and generalisability

Identifying independent units of analysis

Weak tests of equivalence

Rhetorical experiments and unethical controls

Falsifying conclusions

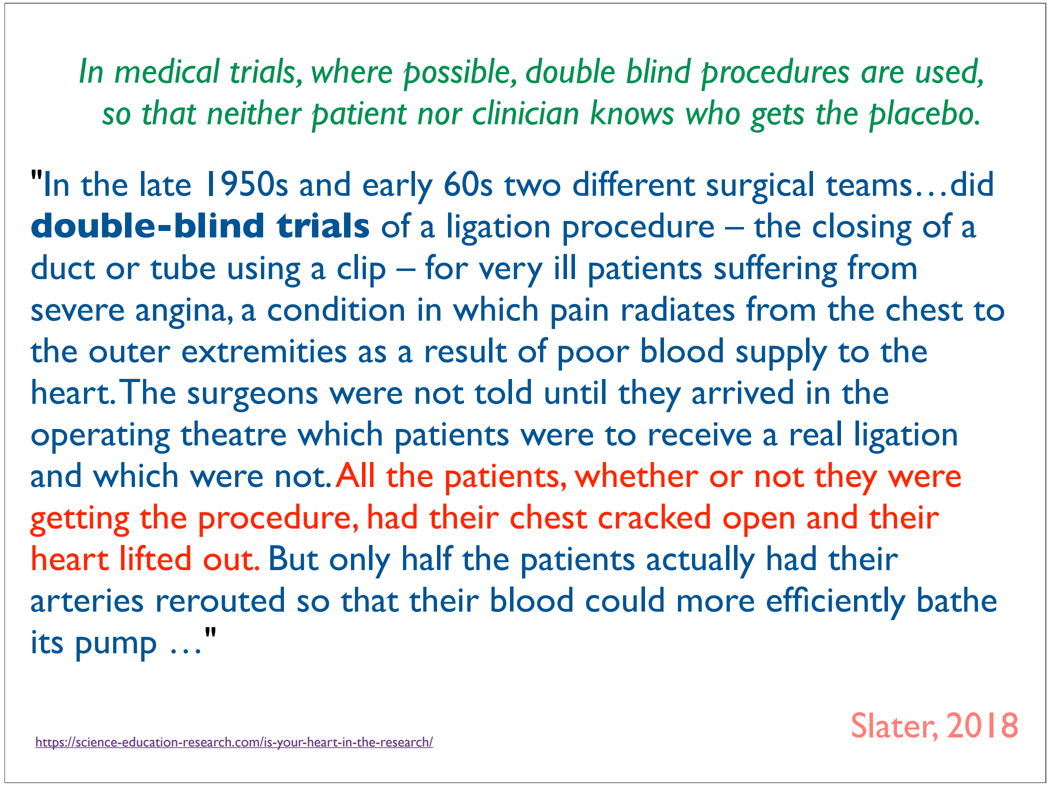
Where ‘the teacher variable’ is controlled by asking the same teacher to teach differently - different pedagogy / different materials - it is (*usually implicitly*) assumed the teacher will have the same competence and confidence when switch to do something different, even when it is novel to them!

In medical trials, where possible, double blind procedures are used, so that neither patient nor clinician knows who gets the placebo.

One of the issues is teacher beliefs, which much research shows often have an effect on outcomes.

If the teacher is persuaded the experimental treatment is an improvement, or is entirely unconvinced by it, then that may be enough to make a difference. Even if the teacher simply lacks confidence in their competence to teach in a different way, then this may make a difference.

It is issues such as this that have led to medical studies adopting double blind conditions in drug trials, so the neither the patients nor the clinicians administering treatment know whether a tablet or injection actually contains the substance being tested or not.



In medical trials, where possible, double blind procedures are used, so that neither patient nor clinician knows who gets the placebo.

"In the late 1950s and early 60s two different surgical teams…did **double-blind trials** of a ligation procedure – the closing of a duct or tube using a clip – for very ill patients suffering from severe angina, a condition in which pain radiates from the chest to the outer extremities as a result of poor blood supply to the heart. The surgeons were not told until they arrived in the operating theatre which patients were to receive a real ligation and which were not. All the patients, whether or not they were getting the procedure, had their chest cracked open and their heart lifted out. But only half the patients actually had their arteries rerouted so that their blood could more efficiently bathe its pump …"

https://science-education-research.com/is-your-heart-in-the-research/

Slater, 2018

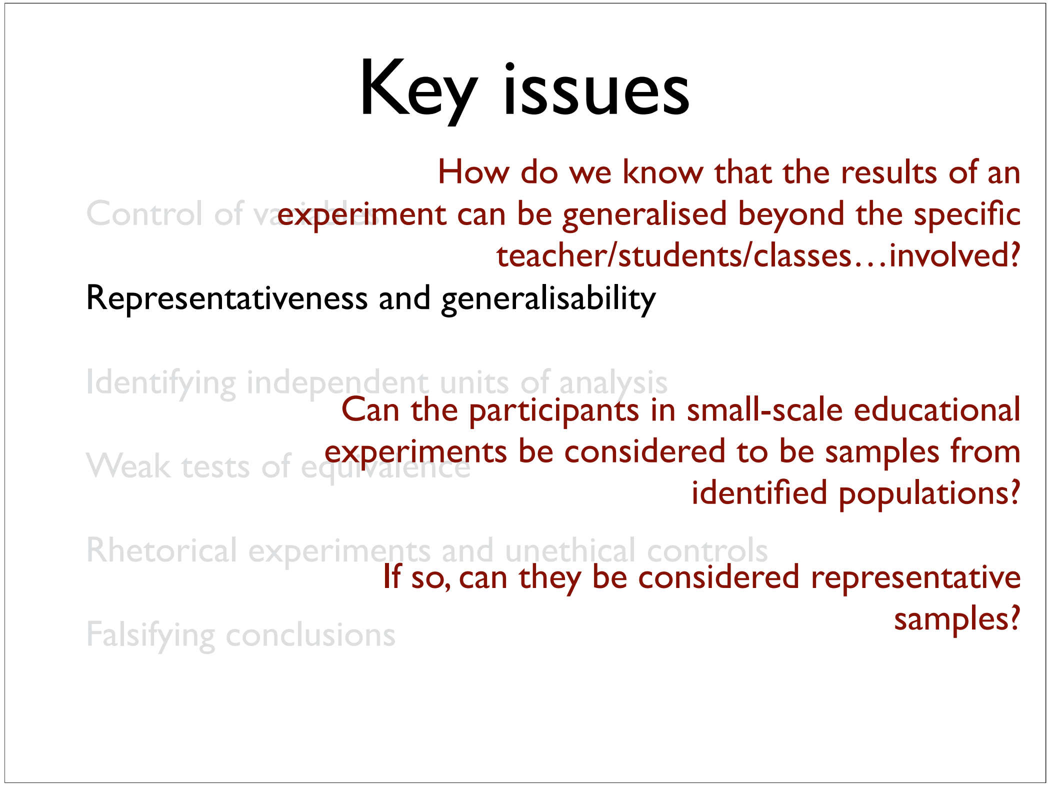
Of course if one takes double blind protocols too far, one might run into ethical issues. I was astonished to read this description of studies where a novel angina treatment was tested by doing sham surgeries alongside genuine interventions.

Reports based on the accounts of patients and their doctors had previously claimed that angina symptoms were relieved somewhat by doing some re-routing of blood in the chest though closing off some vessels. However, the experimental studies found that actually those given this procedure showed no more improvement than patients wheeled into theatre for a sham procedure.

I was less astonished when I sourced the original research papers, as it seems the sham surgery only involved some superficial incisions made under local anaesthetic. In any case, it is much harder to blind learners and, especially, teachers to the educational treatment they are assigned to.

https://science-education-research.com/is-your-heart-in-the-research/

Generalisation



Key issues

How do we know that the results of an experiment can be generalised beyond the specific teacher/students/classes…involved?

Control of variables

Representativeness and generalisability

Identifying independent units of analysis

Can the participants in small-scale educational experiments be considered to be samples from identified populations?

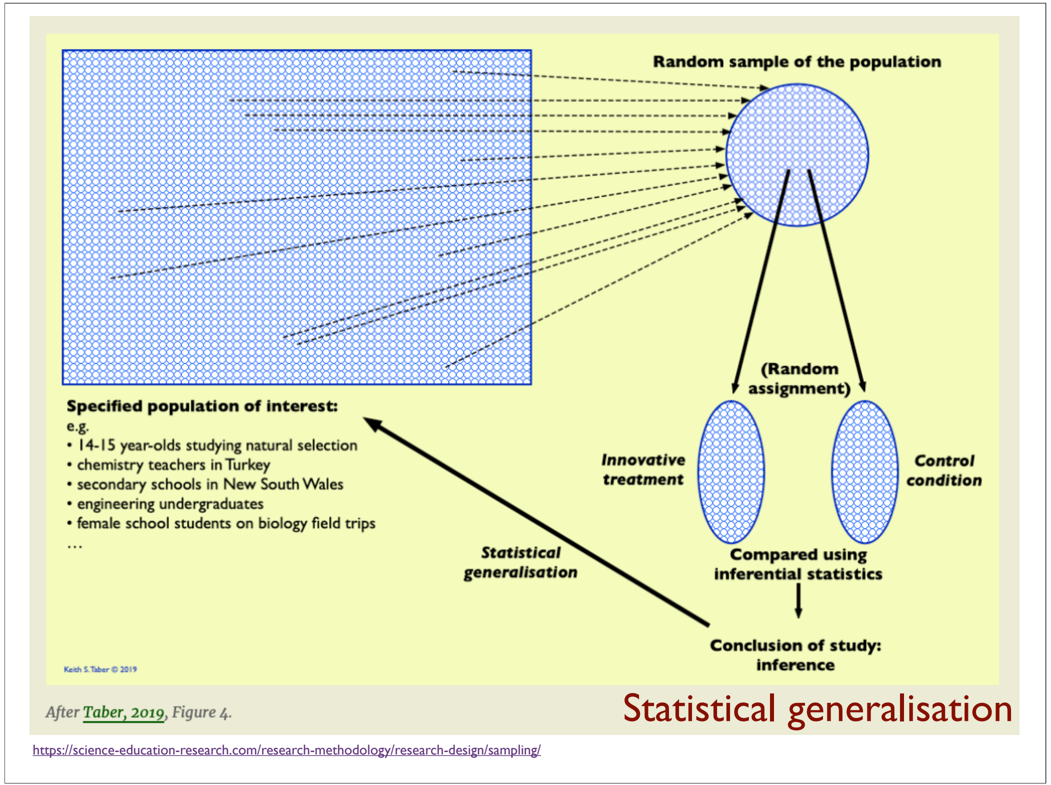
Weak tests of equivalence

Rhetorical experiments and unethical controls

If so, can they be considered representative samples?

Falsifying conclusions

Social kinds are also different from natural kinds, in that all pure samples of copper or all E. coli specimens have much in common - but not all schools, or all classes, or all teachers, or all lessons, have so much in common. This creates the very big issue in educational research, that what works in one classroom in one school, does not always work in another school, or even when adopted by another teacher in the same school.



Random sample of the population

(Random assignment)

Innovative treatment

Control condition

Compared using inferential statistics

Conclusion of study: inference

Statistical generalisation

Specified population of interest:
e.g.
- 14-15 year-olds studying natural selection
- chemistry teachers in Turkey
- secondary schools in New South Wales
- engineering undergraduates
- female school students on biology field trips

…

Keith S Taber © 2019

After Taber, 2019, Figure 4.

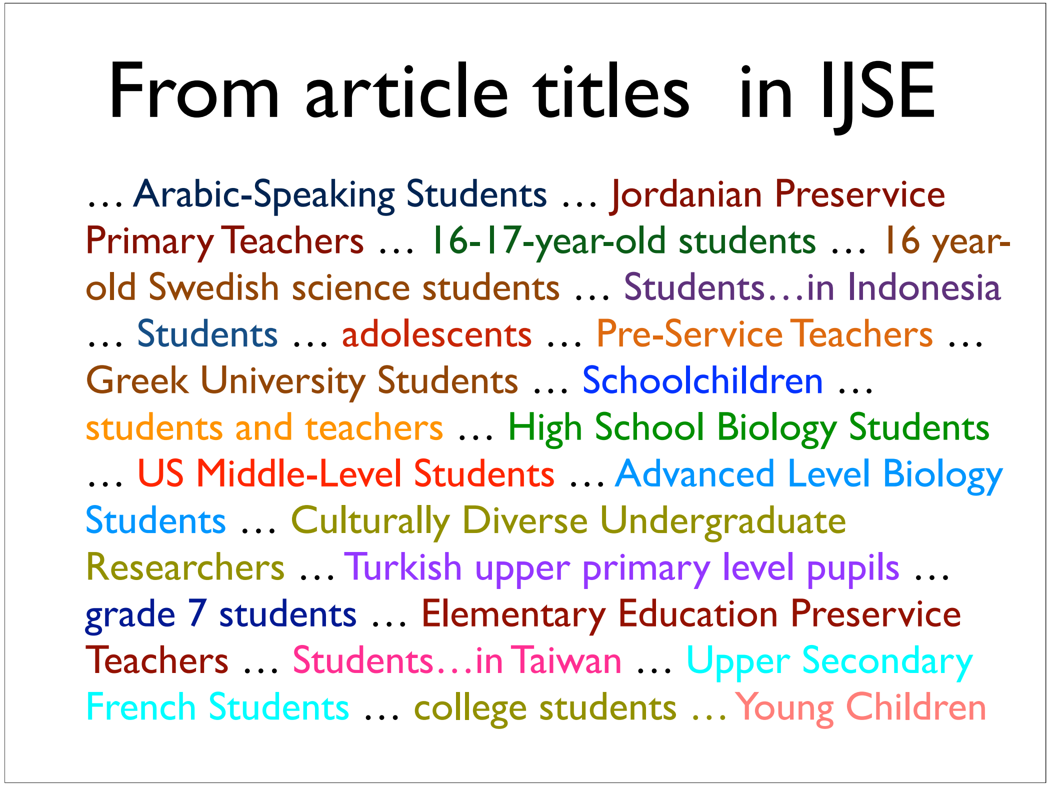
Statistical generalisation

https://science-education-research.com/research-methodology/research-design/sampling/

Even experiments that are designed to allow us to generalise to populations only tell us that what was found to be most effective in the research is more likely than not to be effective in the wider population. But there is diversity in those populations. We know from some of the few very large studies undertaking in schooling, that what is most effective overall, is not most effective in every context; what is found to be least effective overall, can have been the most effective approach with some classes.

Knowing what most often works best is still useful. But to do this kind of work well we need not only to work at scale, but to randomise our sample to conditions, and to either use a random sample of the population or be confident that our sample is representative of the diversity of the population.

https://science-education-research.com/research-methodology/research-design/sampling/



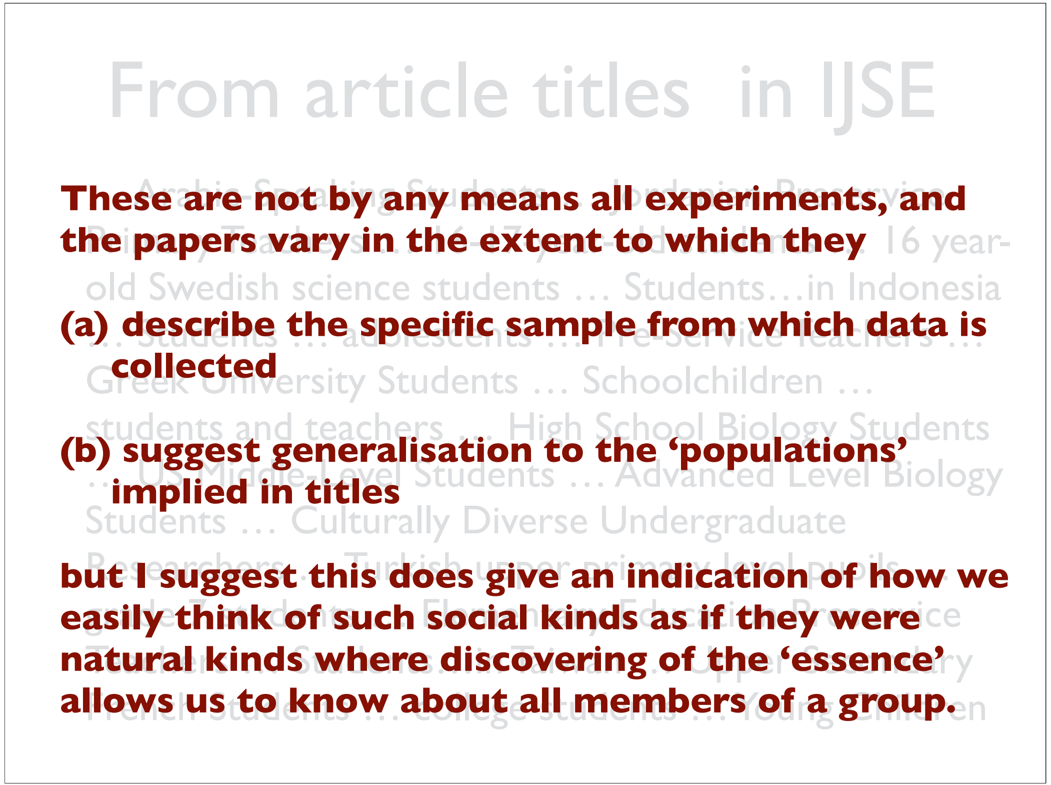
From article titles in IJSE

… Arabic-Speaking Students … Jordanian Preservice Primary Teachers … 16-17-year-old students … 16 year-old Swedish science students … Students…in Indonesia … Students … adolescents … Pre-Service Teachers … Greek University Students … Schoolchildren … students and teachers … High School Biology Students … US Middle-Level Students … Advanced Level Biology Students … Culturally Diverse Undergraduate Researchers … Turkish upper primary level pupils … grade 7 students … Elementary Education Preservice Teachers … Students…in Taiwan … Upper Secondary French Students … college students … Young Children

But what are the populations? A paper title in the natural sciences might refer to a class of star; a superconductor with a specific composition; a variant of SARS-CoV-2, or some such natural kind.

If we look at the titles of papers in science education, we find these papers seem to be about broad groups - sometimes National groups, but sometimes they are apparently about ‘children’ or ‘adolescents’ or ‘primary school teachers’ quite generally!

Of course, the participants in such studies can seldom be considered statistically representative of such broad groups.



From article titles in IJSE

Arabic-Speaking Students ... Jordanian Preservice Primary Teachers ... 16-17 year-old students ... 16 year-old Swedish science students ... Students…in Indonesia ... Students ... adolescents ... Pre-Service Teachers ... Greek University Students ... Schoolchildren ... students and teachers ... High School Biology Students ... US Middle-Level Students ... Advanced Level Biology Students ... Culturally Diverse Undergraduate Researchers ... Turkish upper primary level pupils ... grade 7 students ... Elementary Education Preservice Teachers ... Students…in Taiwan ... Upper Secondary French students ... college students ... Young Children

These are not by any means all experiments, and the papers vary in the extent to which they

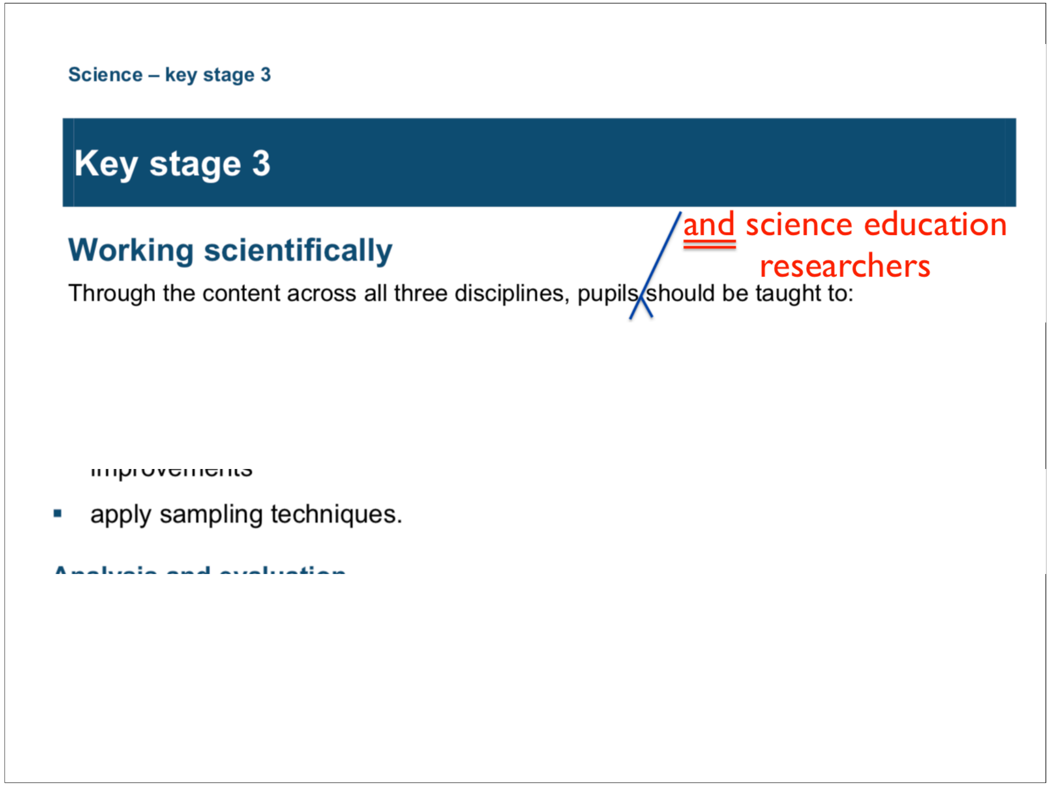
(a) describe the specific sample from which data is collected

(b) suggest generalisation to the 'populations' implied in titles

but I suggest this does give an indication of how we easily think of such social kinds as if they were natural kinds where discovering of the 'essence' allows us to know about all members of a group.

My point is that we - as authors as well as readers - fall into the trap of generalising, at least implicitly.

'We studied what (some) 14 year old Australian students knew about natural selection, so now we know what 14 year old Australian students know about natural selection.'



Science – key stage 3

Key stage 3

Working scientifically

Through the content across all three disciplines, pupils and science education researchers should be taught to:

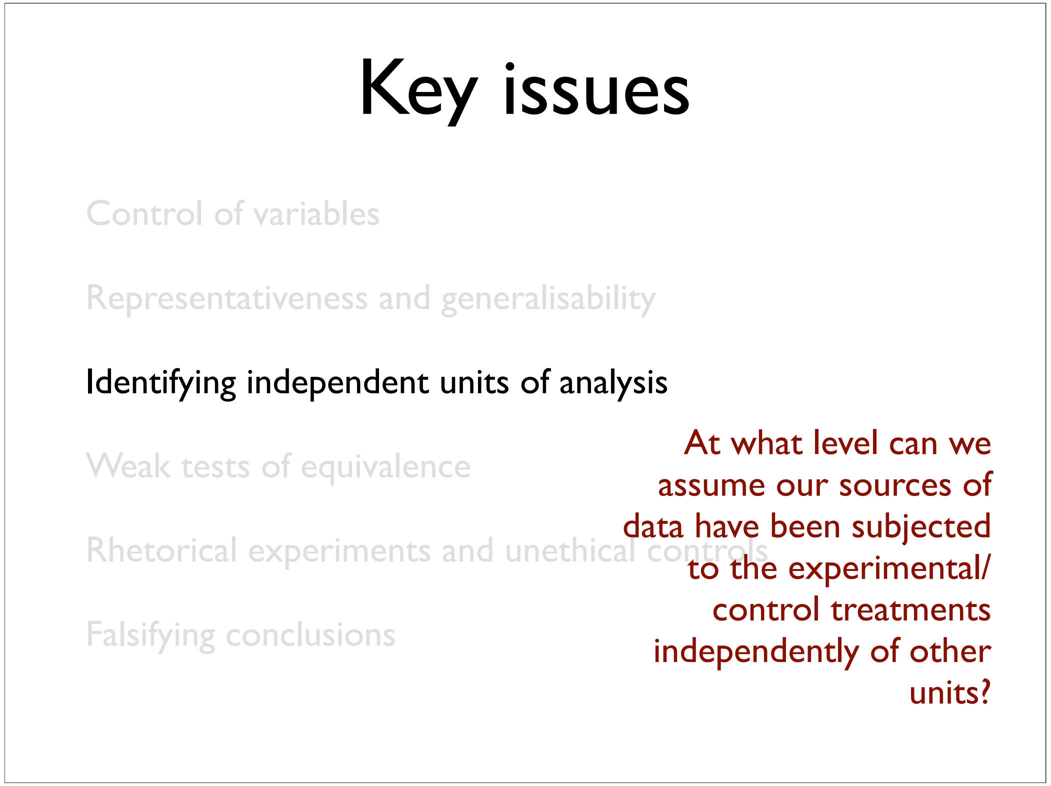
improvements

- apply sampling techniques.

Analysis and evaluation

Again, we would not accept this kind of sloppy thinking from school children.

Units of analysis



Key issues

Control of variables

Representativeness and generalisability

Identifying independent units of analysis

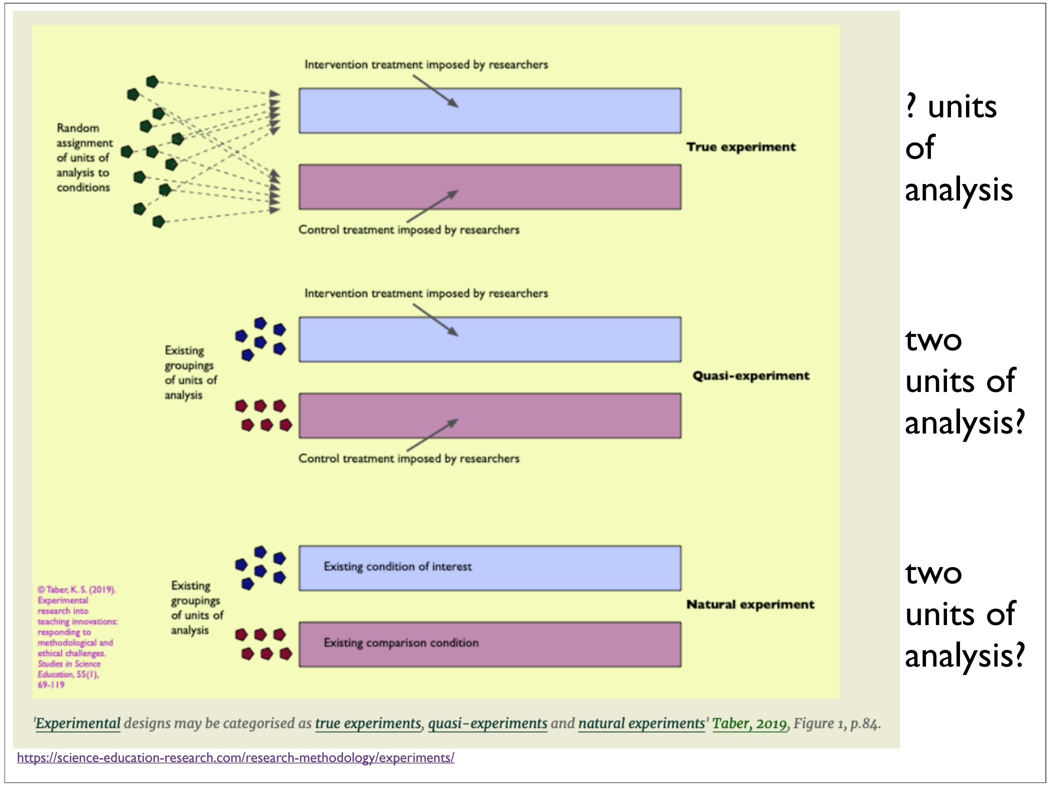
Weak tests of equivalence

Rhetorical experiments and unethical controls

Falsifying conclusions

At what level can we assume our sources of data have been subjected to the experimental/ control treatments independently of other units?

A major issue with many small scale studies is the identification of the unit of analysis to use in statistical testing.


Intervention treatment imposed by researchers
Random assignment of units of analysis to conditions
True experiment
Control treatment imposed by researchers
Intervention treatment imposed by researchers
Existing groupings of units of analysis
Quasi-experiment
Control treatment imposed by researchers
Existing condition of interest
Existing groupings of units of analysis
Natural experiment
Existing comparison condition

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

'Experimental designs may be categorised as true experiments, quasi-experiments and natural experiments' Taber, 2019, Figure 1, p.84.

https://science-education-research.com/research-methodology/experiments/

? units of analysis

two units of analysis?

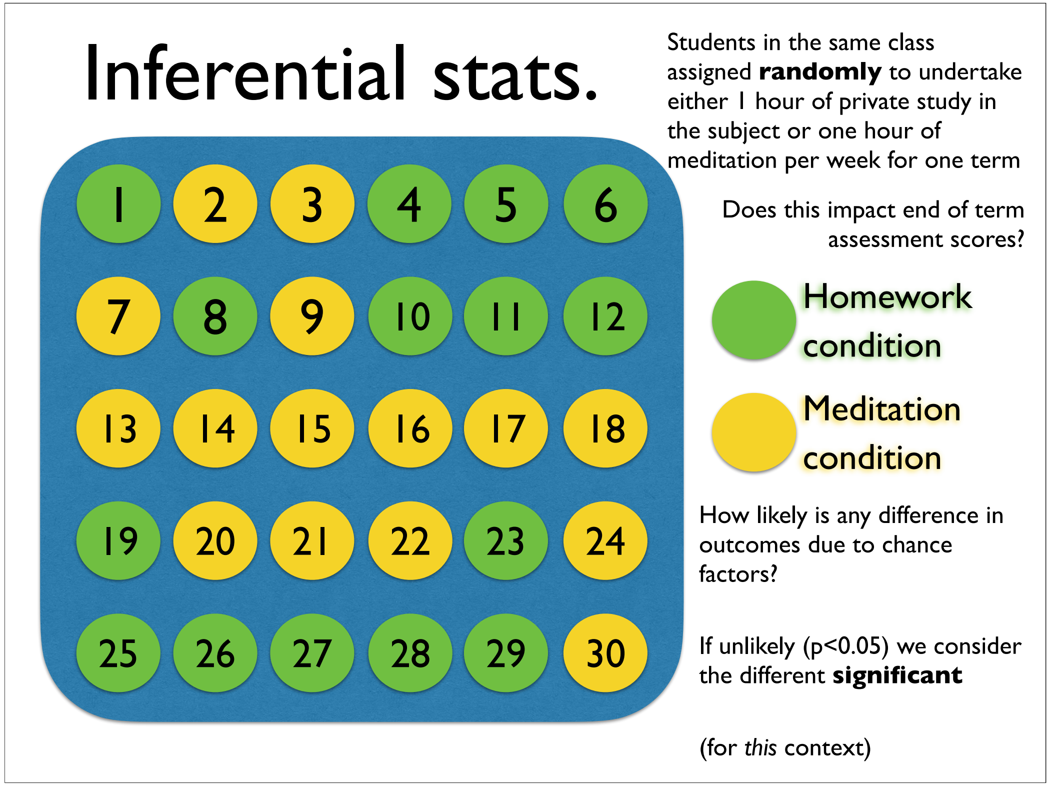
two units of analysis?

In a true experiment we randomise the so-called units of analysis to the treatments, the conditions. Sometimes this is possible in education. Perhaps we enrol fifty schools and assign each randomly to an experimental or control condition. If we can consider the schools to be experiencing the assigned treatment independently of each other, then this seems fair enough.

But many experiments in education are undertaken with learners as the unit of analysis, and often these are not individually assigned to treatments, but are members of pre-established classes. Of course, there are very good reasons why schools would not be happy with researchers coming in and breaking up established classes to randomise students. Perhaps the logic here should be that as we cannot meet the requirements for an experiment, we should do a different kind of study. Often, instead, the logic adopted is that as we cannot meet the requirements for a valid experiment, we are justified in carrying on regardless and just ignoring that requirement.

If a manufacturer of pickled onions was selling jars of vinegar as pickled onions, it is likely their customers would not be prepared to accept this just because the company was having difficulty sourcing onions. Yet, readers of research papers are assumed to be less demanding of rigour. Science education researchers often sell jars of vinegar labelled as pickled onions.

https://science-education-research.com/research-methodology/experiments/



Inferential stats.

Students in the same class assigned **randomly** to undertake either 1 hour of private study in the subject or one hour of meditation per week for one term

Does this impact end of term assessment scores?

🟢 1	🟡 2	🟡 3	🟢 4	🟢 5	🟢 6
🟡 7	🟢 8	🟡 9	🟢 10	🟢 11	🟢 12
🟡 13	🟡 14	🟡 15	🟡 16	🟡 17	🟡 18
🟢 19	🟡 20	🟡 21	🟡 22	🟢 23	🟡 24
🟢 25	🟢 26	🟢 27	🟢 28	🟢 29	🟡 30

🟢 Homework condition

🟡 Meditation condition

How likely is any difference in outcomes due to chance factors?

If unlikely ($p<0.05$) we consider the different **significant**

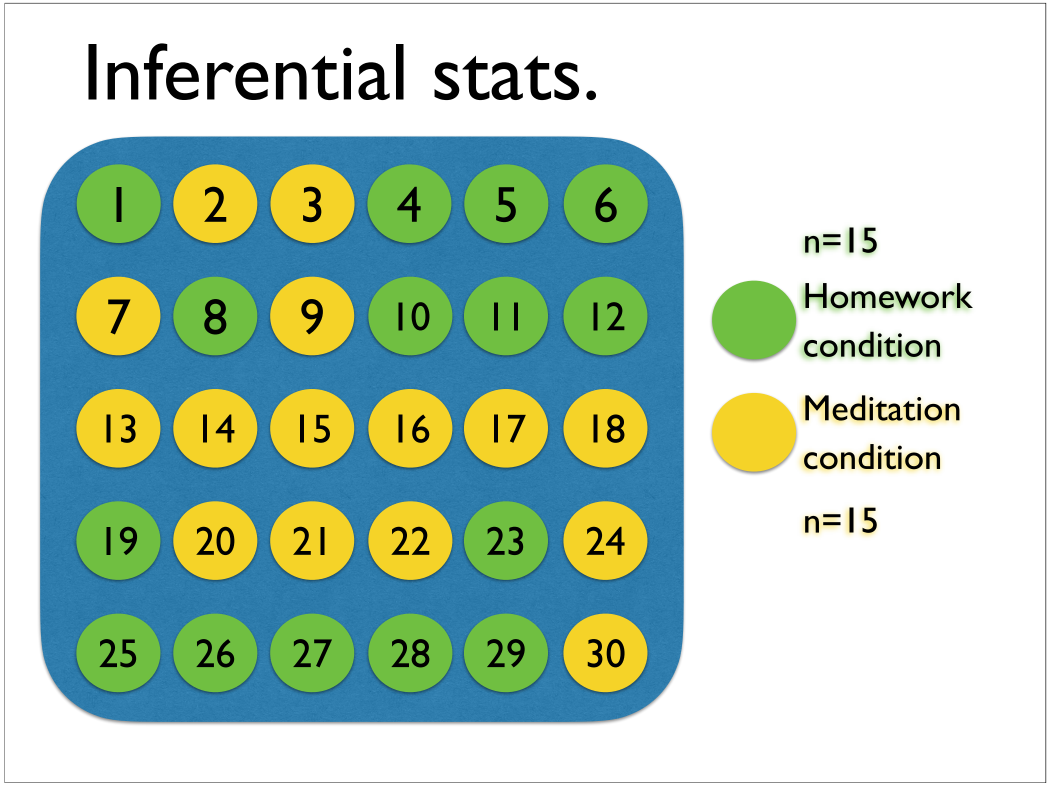
(for *this* context)

Now, I do not want to give the impression that I do not think there might be circumstances when treating students within a class as independent units would be appropriate. One has to consider the overall research design and purpose.

In this hypothetical case, a teacher has read a lot of material about mindfulness, student anxiety, relaxation techniques, and the like; and suspects that asking students to do two half-hours of mediation each week would be just as beneficial for their science learning as asking them to do subject-based homework. Being a science teacher, she tests this idea by randomly assigning students to either a homework condition or a meditation condition, and at the end of term compares test scores.

Randomisation does not ensure matched groups, rather just avoids any systematic bias. So, how does the teacher decide if the difference in profiles of scores in the two conditions is just down to chance effects of who ended-up in each condition? Inferential statistics will allow her to see if any difference in performance is likely to just be down to chance.

Of course, even if there is a very low probability value, and it is concluded there is a significant difference, strictly this only applies to this class with this teacher, and perhaps even this topic? It may be more generally applicable, but we should not assume that.



Inferential stats.

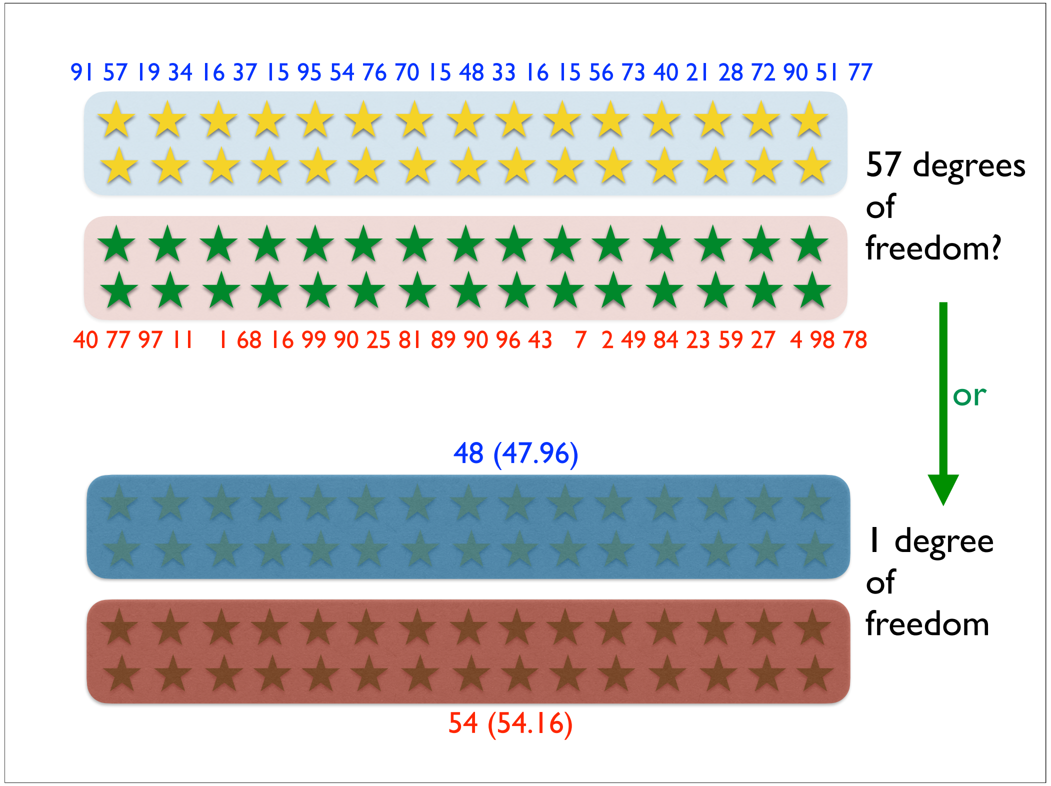
n=15

🟢 Homework condition

🟡 Meditation condition

n=15

Here it is reasonable to assume we can treat the learners as independent units of analysis even though they are from the same class. Here, in effect, the class is the population of interest. Yes, the students will influence each other in class, but they are all in the same class so those in both conditions are exposed to the same influences. If the students are off doing homework or meditating individually, and do not collude on the end of term test, it seems reasonable to assume we have fifteen units of analysis in each condition. That is still a small sample size, but at least it is more than unity.

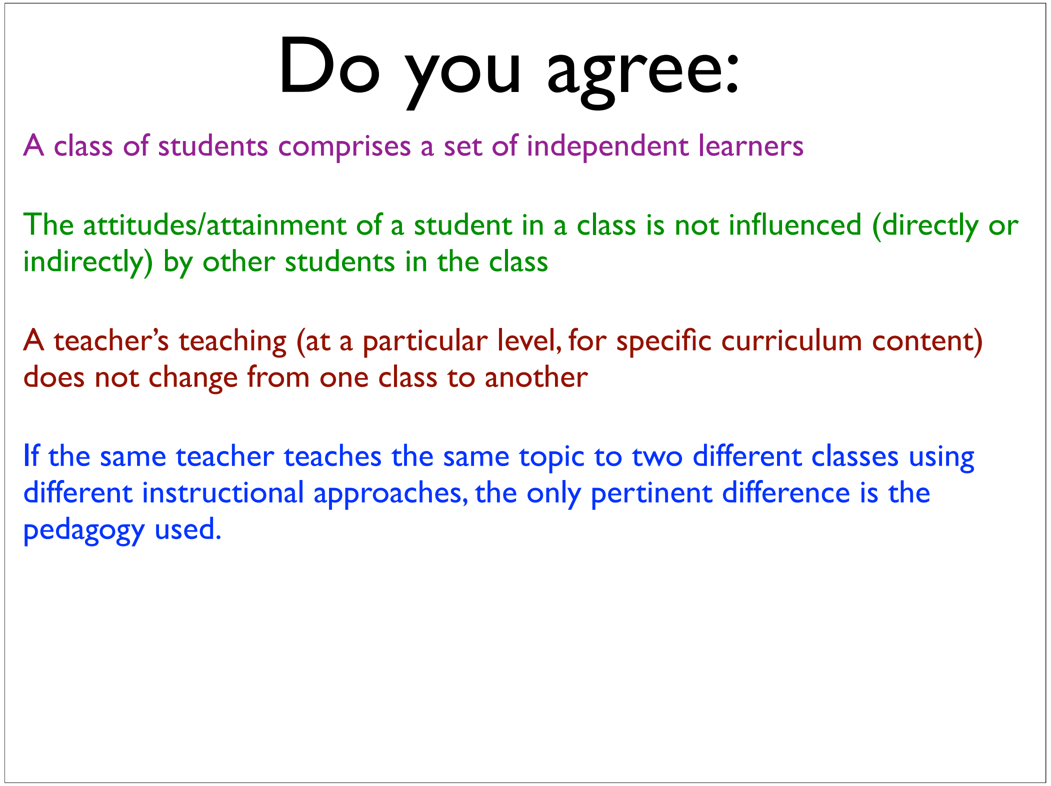


In that circumstance, it is not a problem that all the students are taught together in class and no doubt, at least one might hope, interact during lessons.

But, if we were comparing between two classes, and one class was assigned to the homework condition and another to the meditation condition, then it surely does matter. In this situation, we would need to consider the class as the unit of analysis with one mean outcome score.

But, of course, if there are only two classes in the experiment we are not going to find any difference in outcomes that can be statistically significant. The only way we can get positive results here, is by pretending that we have a lot of independent outcomes scores in each condition.

But, that seems like cheating. "Can't you see the onions in the jar?"



Do you agree:

A class of students comprises a set of independent learners

The attitudes/attainment of a student in a class is not influenced (directly or indirectly) by other students in the class

A teacher’s teaching (at a particular level, for specific curriculum content) does not change from one class to another

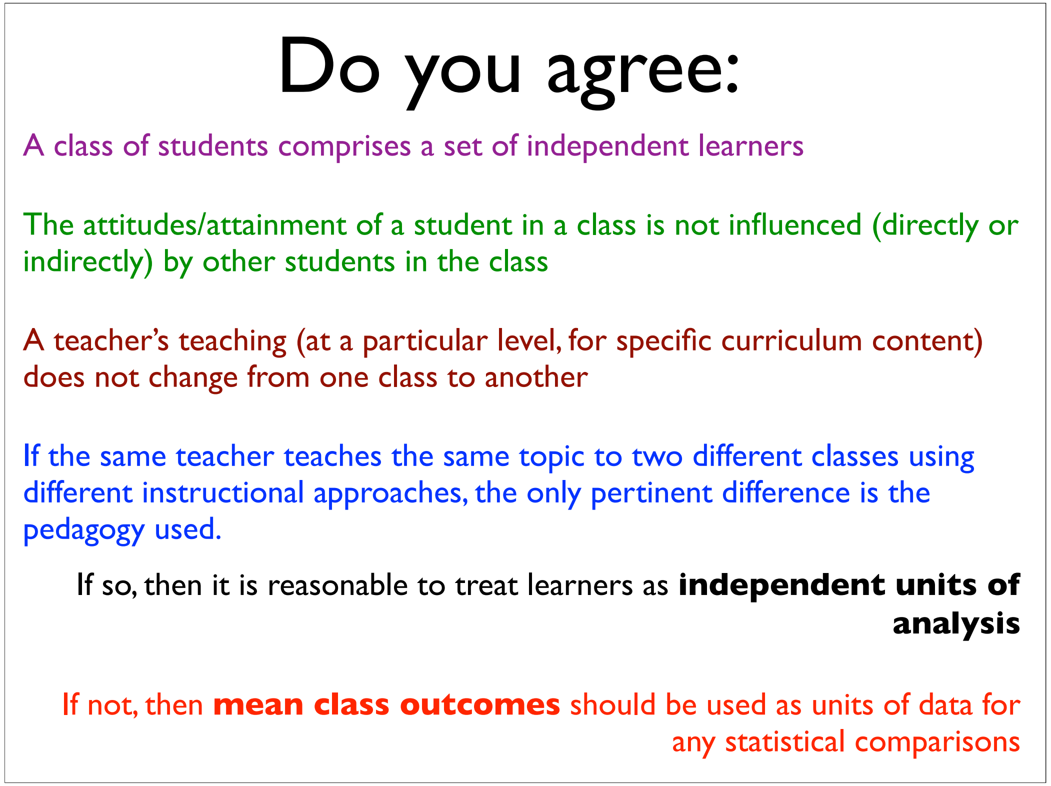
If the same teacher teaches the same topic to two different classes using different instructional approaches, the only pertinent difference is the pedagogy used.

Perhaps you feel I am wrong.

I suspect that like me you my have taught a range of different classes over the years.

My own experience is that a class has its own character that is emergent, and is not just an aggregate of the characters in the class.
My own experience is that parallel classes, or successive cohorts, that are nominally equivalent can actually be very different.
My own experience is that the same class can be experienced by different teachers as quite different.
When working in school I sometimes found that one or two students in a class could have a disproportional influence on the class environment and progress - and that could be either for better or worse.

But, perhaps your experiences have been different.



Do you agree:

A class of students comprises a set of independent learners

The attitudes/attainment of a student in a class is not influenced (directly or indirectly) by other students in the class

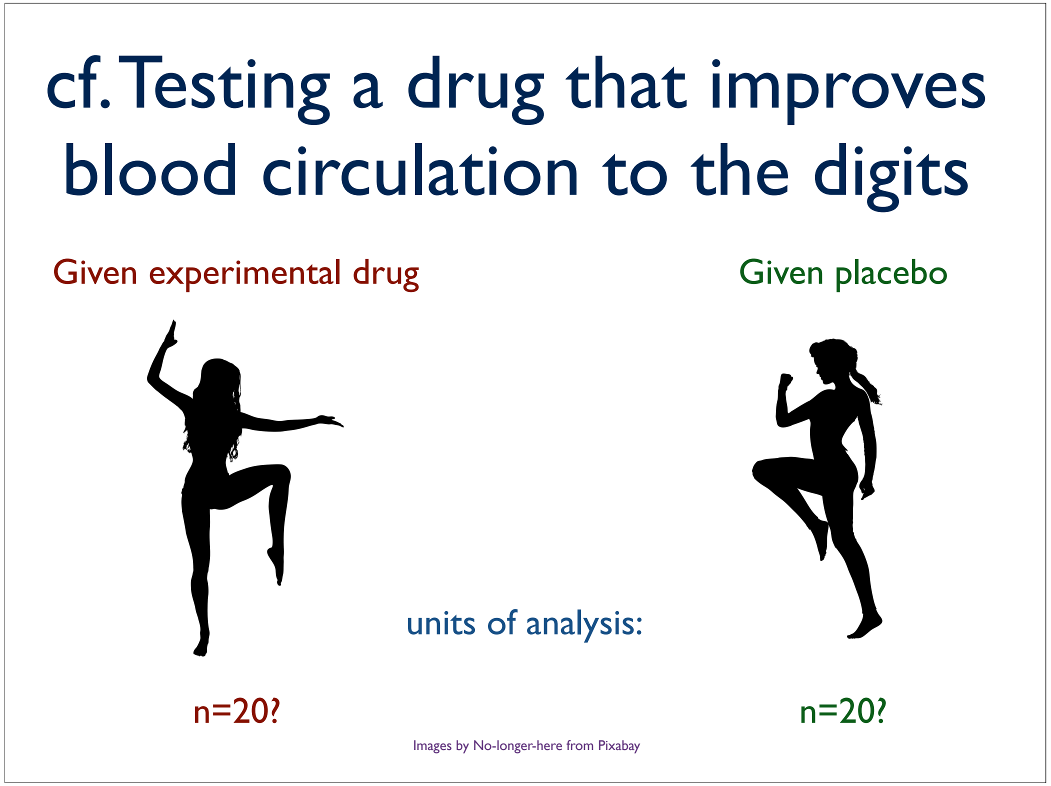
A teacher’s teaching (at a particular level, for specific curriculum content) does not change from one class to another

If the same teacher teaches the same topic to two different classes using different instructional approaches, the only pertinent difference is the pedagogy used.

If so, then it is reasonable to treat learners as **independent units of analysis**

If not, then **mean class outcomes** should be used as units of data for any statistical comparisons

If you can honestly say that you feel that the attitudes; progress; learning, of students in classes is not influenced by the rest of the class, and occurs completely independently of the others in the lessons, then, yes, it is fair to treat the learner as the unit of analysis.



cf. Testing a drug that improves blood circulation to the digits

Given experimental drug

Given placebo

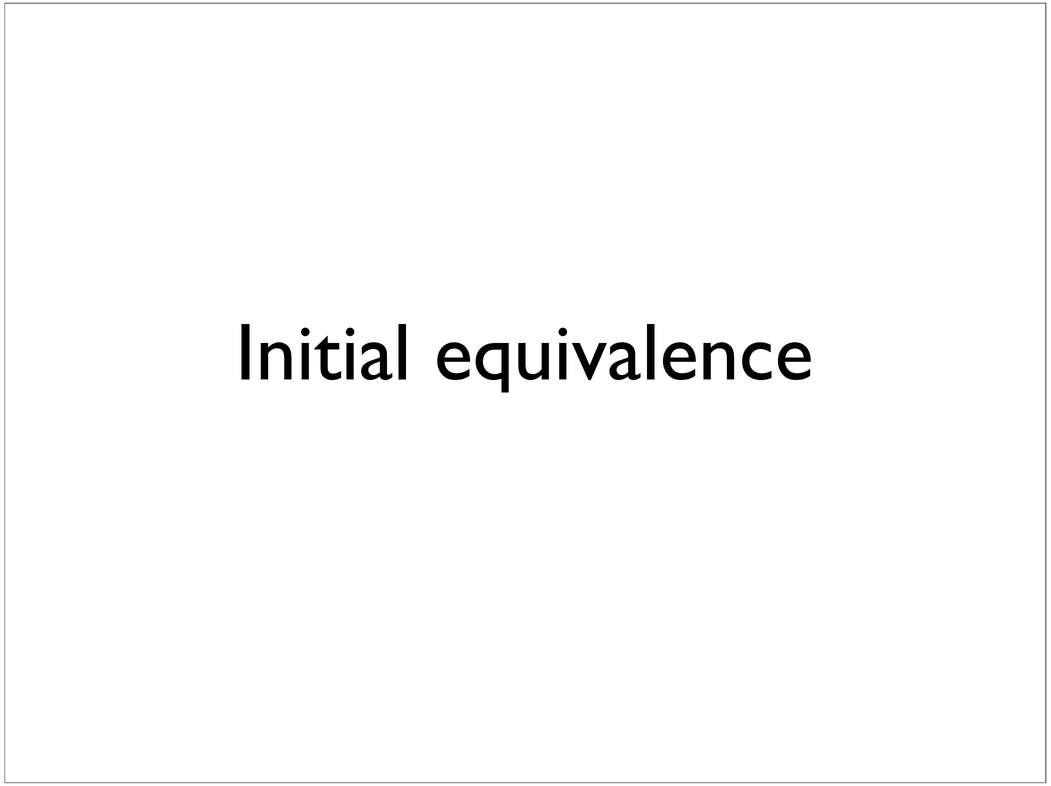
units of analysis:

n=20?

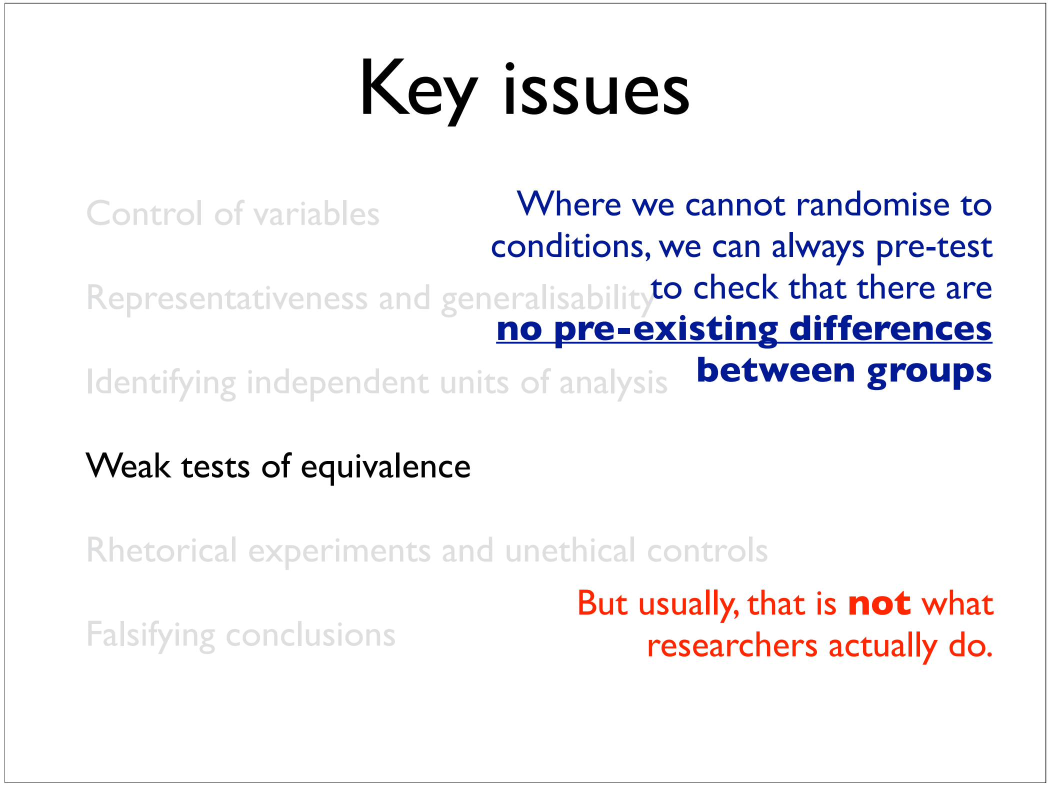
n=20?

Images by No-longer-here from Pixabay

I thought an analogy might be testing a drug that helped blood supply to the extremities where blood circulation was measured for each digit. That seems sensible if the scores are to be aggregated to give an overall score for the patient. But It would not make sense to consider blood supply to different digits to be independent when it is all part of the same circulatory system, being pumped around by the same heart!



Initial equivalence



Key issues

Control of variables

Representativeness and generalisability

Identifying independent units of analysis

Weak tests of equivalence

Rhetorical experiments and unethical controls

Falsifying conclusions

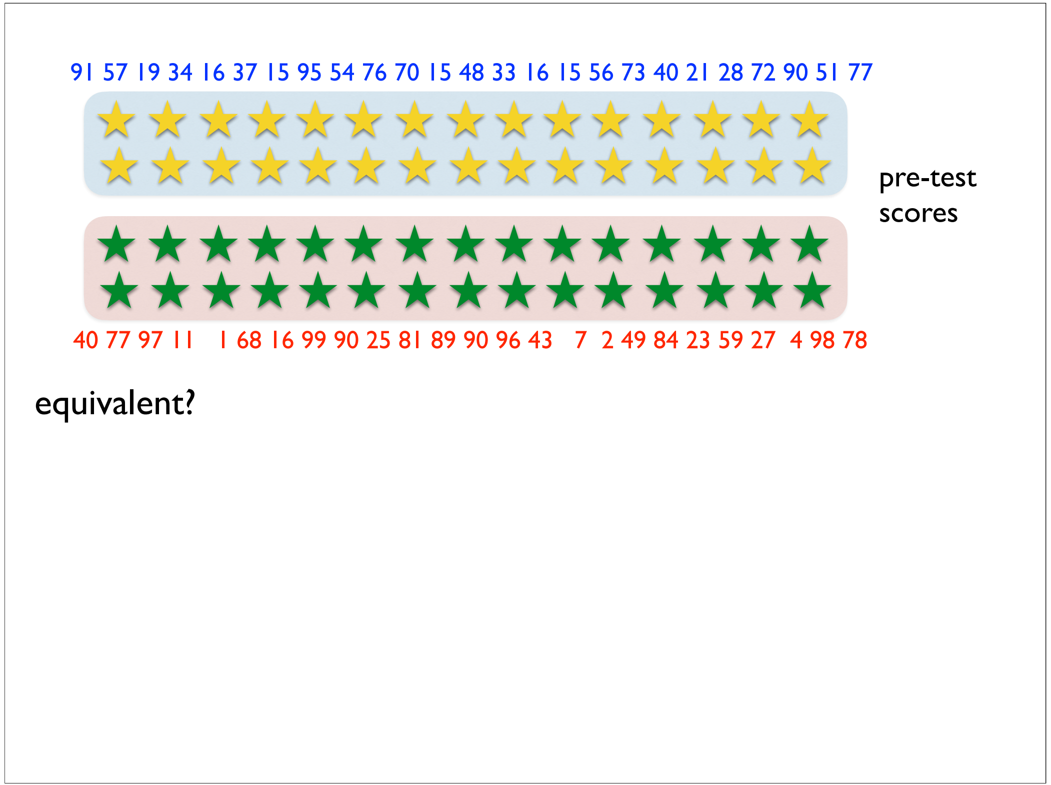
Where we cannot randomise to conditions, we can always pre-test to check that there are **no pre-existing differences between groups**

But usually, that is **not** what researchers actually do.

A very common procedure used in experimental studies is testing for initial equivalence between groups. This is especially important when there is no random assignment as if there are systematic differences between groups, then any difference in final outcomes may just reflect differences at the start.

Even if researchers showed there was no difference between groups at the outset, this does not negate concerns about students in different classes being influenced by the class context and not learning independently.

But leaving that aside, there is a conceptual problem with testing for initial equivalence, because if researchers were really checking for equivalence between groups at the start of start of an experiment then they would very rarely find complete equivalence.



91 57 19 34 16 37 15 95 54 76 70 15 48 33 16 15 56 73 40 21 28 72 90 51 77

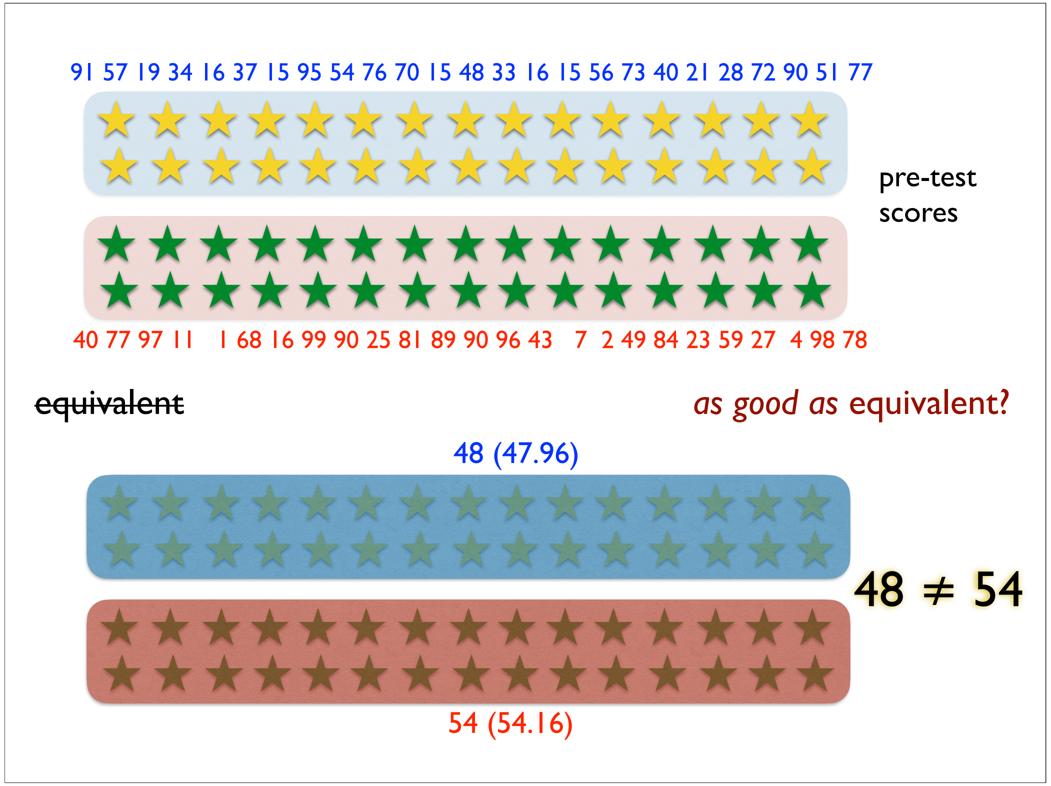
pre-test
scores

40 77 97 11 1 68 16 99 90 25 81 89 90 96 43 7 2 49 84 23 59 27 4 98 78

equivalent?

So, imagine two classes we have given what we consider the most relevant pre-test in relation to the study outcomes to be measured at post-test.

Actually these are randomly generated numbers, so there is no systematic bias in the scores assigned to the students in the conditions.



91 57 19 34 16 37 15 95 54 76 70 15 48 33 16 15 56 73 40 21 28 72 90 51 77

pre-test
scores

40 77 97 11 1 68 16 99 90 25 81 89 90 96 43 7 2 49 84 23 59 27 4 98 78

~~equivalent~~

as good as equivalent?

48 (47.96)

48 ≠ 54

54 (54.16)

The average score in one class is about 48, but in the other it is about 54.

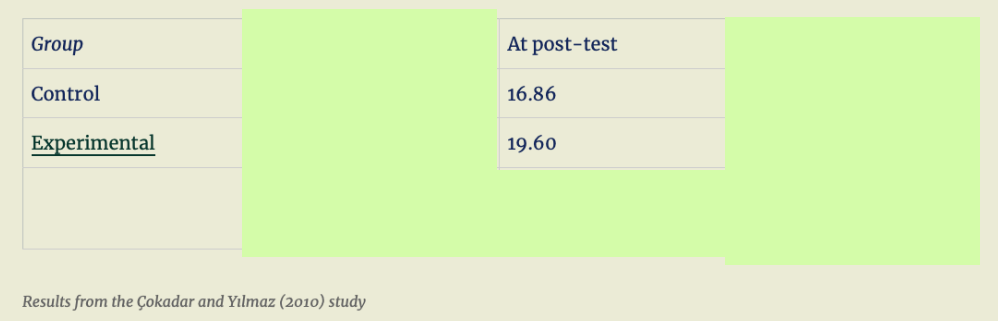
48 is not strictly equivalent to 54.

If you were appointed to a job on an annual salary of 54 000 pounds, but were only paid £48 000, you would probably not accept the argument that these two figures are equivalent.

But one is seldom going to get precisely the same scores on a pre-test (even with random numbers, as we see here!) So, the question becomes *how close* is to be seen as *as good as* equivalent. And this is where I think the most common approach is seriously flawed.

initial equivalence

"the effect of creative drama-based instruction on seventh graders' science achievements in the ecology and matter cycles unit …"



Group	[redacted]	At post-test	[redacted]
Control	[redacted]	16.86	[redacted]
Experimental	[redacted]	19.60	[redacted]
	[redacted]	[redacted]	[redacted]

Results from the Çokadar and Yılmaz (2010) study

So?

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

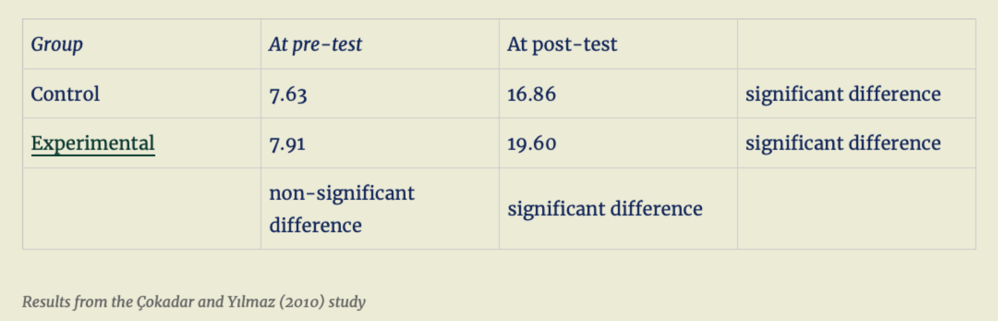
He is a real example to make clear hat these tests are about.

In this study the experimental group out-performed the control group at the end of the experiment.

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

initial equivalence

"the effect of creative drama-based instruction on seventh graders' science achievements in the ecology and matter cycles unit …"



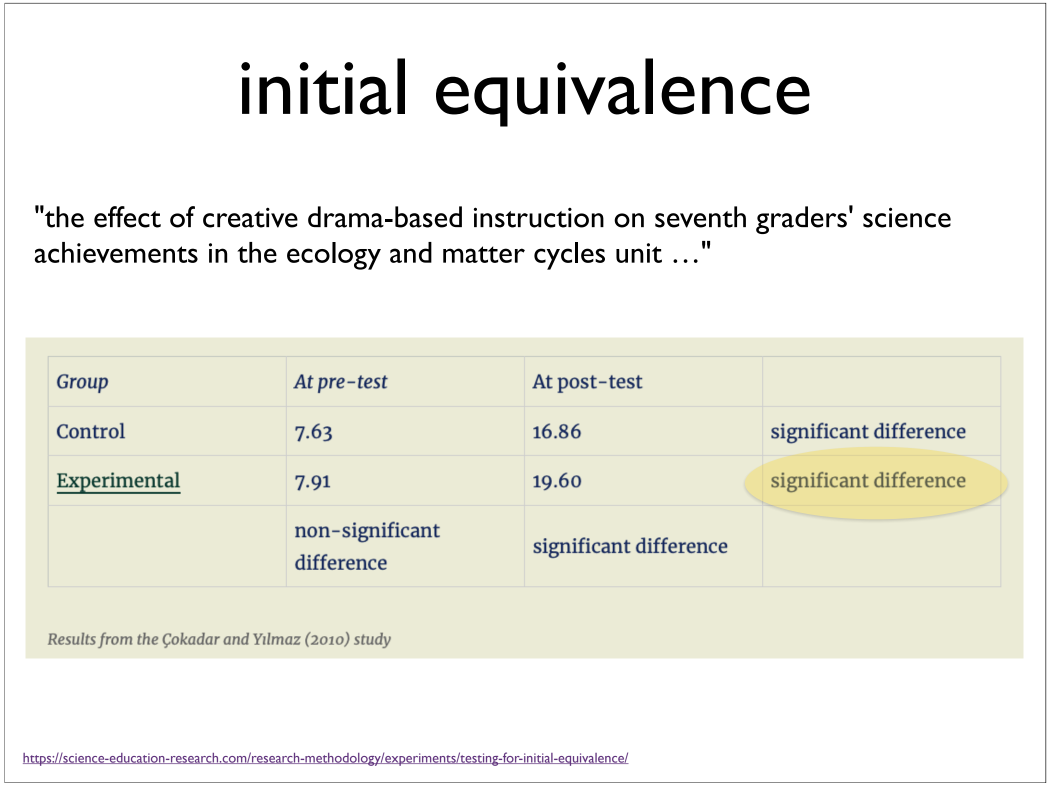
Group	*At pre-test*	At post-test	
Control	7.63	16.86	significant difference
Experimental	7.91	19.60	significant difference
	non-significant difference	significant difference	

Results from the Çokadar and Yılmaz (2010) study

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

And pre-tests were used, so we can compare between pre- and post- intervention, as well as between the two conditions.

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/



initial equivalence

"the effect of creative drama-based instruction on seventh graders' science achievements in the ecology and matter cycles unit …"

Group	*At pre-test*	At post-test	
Control	7.63	16.86	significant difference
Experimental	7.91	19.60	significant difference
	non-significant difference	significant difference	

Results from the Çokadar and Yılmaz (2010) study

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

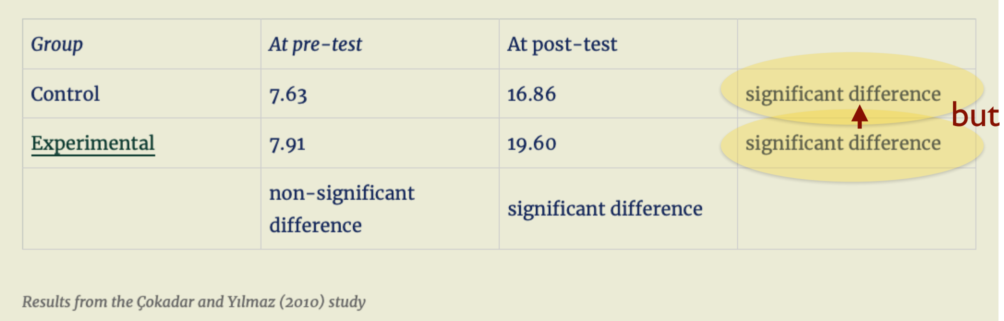
Statistical tests tell us that the experimental group did significantly better on the post-test than it did on the pre-test.

But by itself that's not very informative, as even fairly ineffective teaching is likely to produce some learning.

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

initial equivalence

"the effect of creative drama-based instruction on seventh graders' science achievements in the ecology and matter cycles unit …"



Group	*At pre-test*	At post-test	
Control	7.63	16.86	significant difference
Experimental	7.91	19.60	significant difference
	non-significant difference	significant difference	

Results from the Çokadar and Yılmaz (2010) study

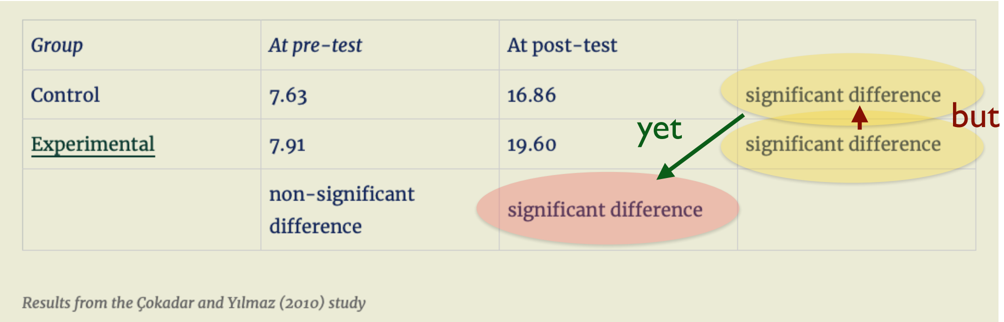
https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

And indeed the control group also showed significant increases between the two tests, so both conditions seem to bring about learning.

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

initial equivalence

"the effect of creative drama-based instruction on seventh graders' science achievements in the ecology and matter cycles unit …"



Group	*At pre-test*	At post-test	
Control	7.63	16.86	significant difference
Experimental	7.91	19.60	significant difference
	non-significant difference	significant difference	

Results from the Çokadar and Yılmaz (2010) study

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

However, the statistics also tell us that the experimental group out-performed the control group at post-test by a difference that was statistically significant.

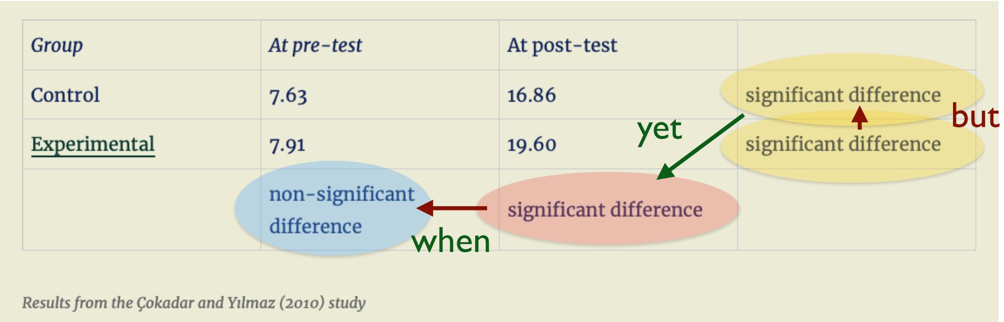
That means this difference was unlikely to be due to chance effects.

But what if the two groups were starting from different bases?

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

initial equivalence

"the effect of creative drama-based instruction on seventh graders' science achievements in the ecology and matter cycles unit …"



Group	*At pre-test*	At post-test	
Control	7.63	16.86	significant difference
Experimental	7.91	19.60	significant difference
	non-significant difference	significant difference	

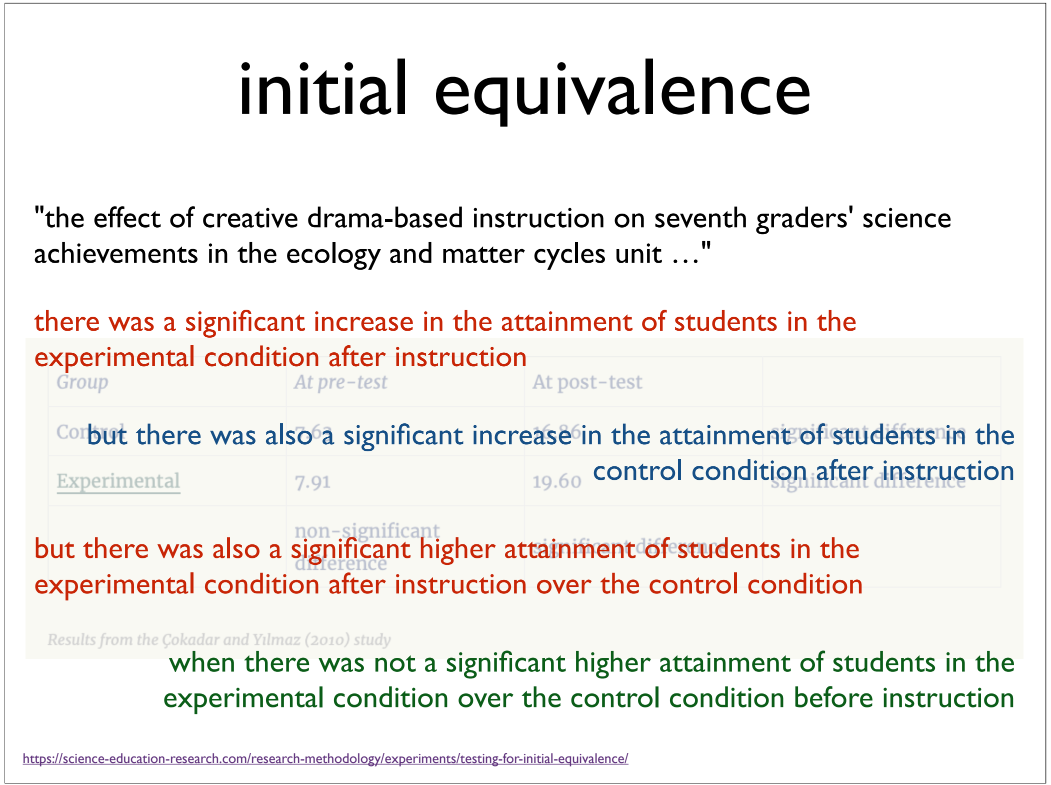
yet
but
when

Results from the Çokadar and Yılmaz (2010) study

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

This is discounted because there is a significant difference between groups at the end of the experiments, when there was no such difference at the outset.

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/



initial equivalence

"the effect of creative drama-based instruction on seventh graders' science achievements in the ecology and matter cycles unit …"

there was a significant increase in the attainment of students in the experimental condition after instruction

Group	At pre-test	At post-test	
Control	7.63	16.86	significant difference
Experimental	7.91	19.60	significant difference
	non-significant difference	significant difference	

Results from the Çokadar and Yılmaz (2010) study

but there was also a significant increase in the attainment of students in the control condition after instruction

but there was also a significant higher attainment of students in the experimental condition after instruction over the control condition

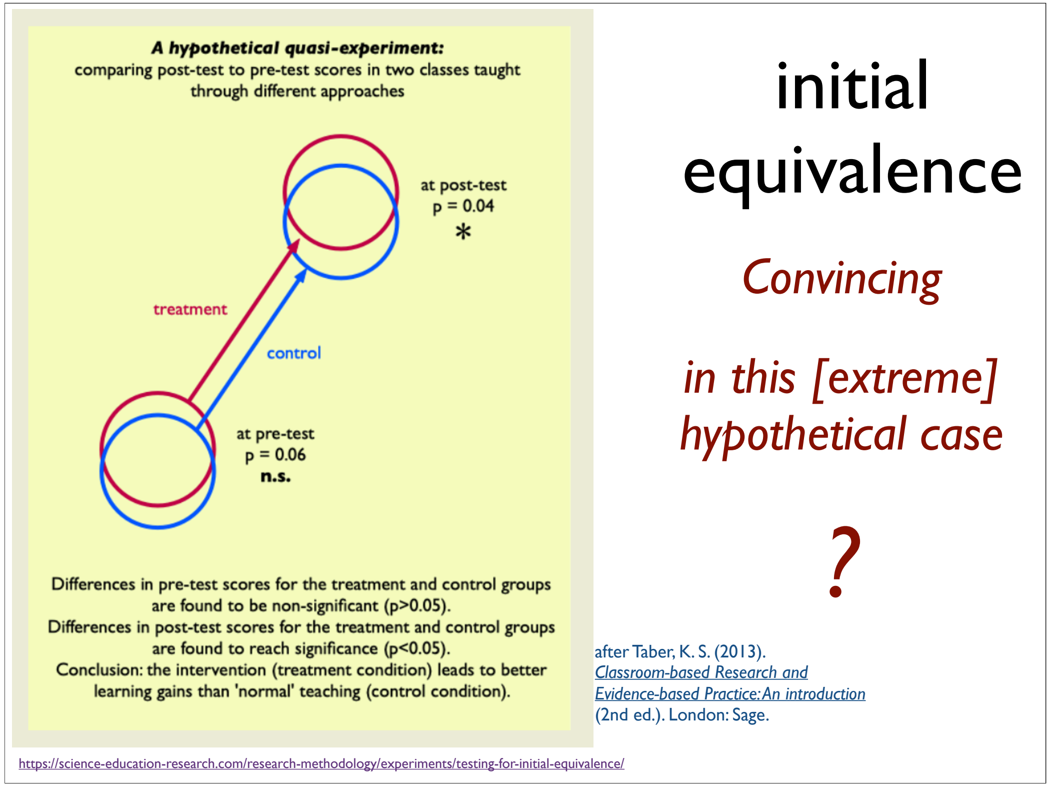
when there was not a significant higher attainment of students in the experimental condition over the control condition before instruction

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

This seems logical, and perhaps the numbers here look quite convincing.

But I think, in general, there is a logical problem here.

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/



A hypothetical quasi-experiment:
comparing post-test to pre-test scores in two classes taught through different approaches

Differences in pre-test scores for the treatment and control groups are found to be non-significant (p>0.05).
Differences in post-test scores for the treatment and control groups are found to reach significance (p<0.05).
Conclusion: the intervention (treatment condition) leads to better learning gains than 'normal' teaching (control condition).

initial equivalence

Convincing

in this [extreme] hypothetical case

?

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

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

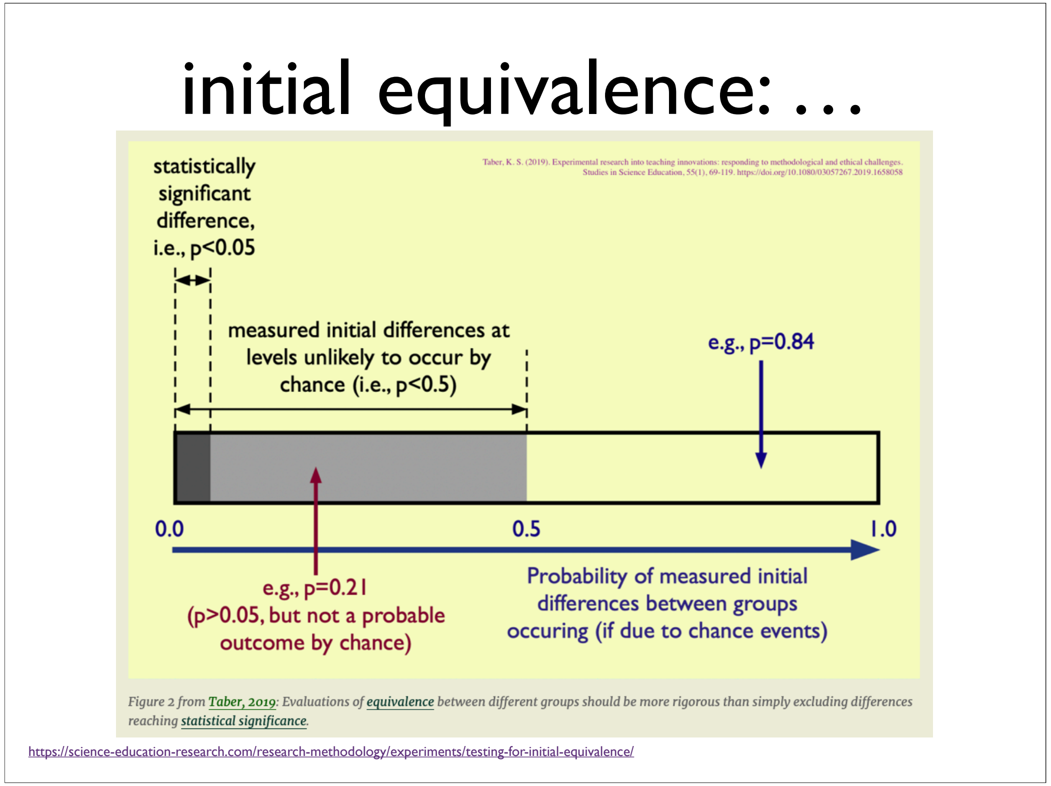
Let's consider a hypothetical marginal case. Here there is a small measured difference before the experiment, and also a small measured difference after the experiment.

> The difference at pre-test just failed to reach significance.
> The difference at post test just reached significance.
> Conclusion: the experimental intervention made a difference.

But, surely, initial differences can be magnified in subsequent teaching. It is a common phenomena that differences between learners tend to increase over time. In this hypothetical case, can we really be confident that initial differences were not a factor?

Of course, we might argue that there are more sophisticated statistical approaches which look at how factors co-vary in a study. Indeed, there are, but my target here is the simple test for equivalence that is commonly used in published studies to supposedly establish a level playing field.

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/



initial equivalence: …

statistically significant difference, i.e., p<0.05
Taber, K. S. (2019). Experimental research into teaching innovations: responding to methodological and ethical challenges. Studies in Science Education, 55(1), 69-119. https://doi.org/10.1080/03057267.2019.1658058
measured initial differences at levels unlikely to occur by chance (i.e., p<0.5)
e.g., p=0.84
0.0
0.5
1.0
e.g., p=0.21 (p>0.05, but not a probable outcome by chance)
Probability of measured initial differences between groups occuring (if due to chance events)

Figure 2 from Taber, 2019: Evaluations of equivalence between different groups should be more rigorous than simply excluding differences reaching statistical significance.

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

This common approach is to test to see if there is a very unlikely difference between the pre-test measures in the different conditions.

This means that differences which are unlikely to be due to chance effects, but which are not so unlikely to get a p value below 0.05 are found ‘equivalent’.

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

Fair analogies?

Some situations where stronger evidence would be useful

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

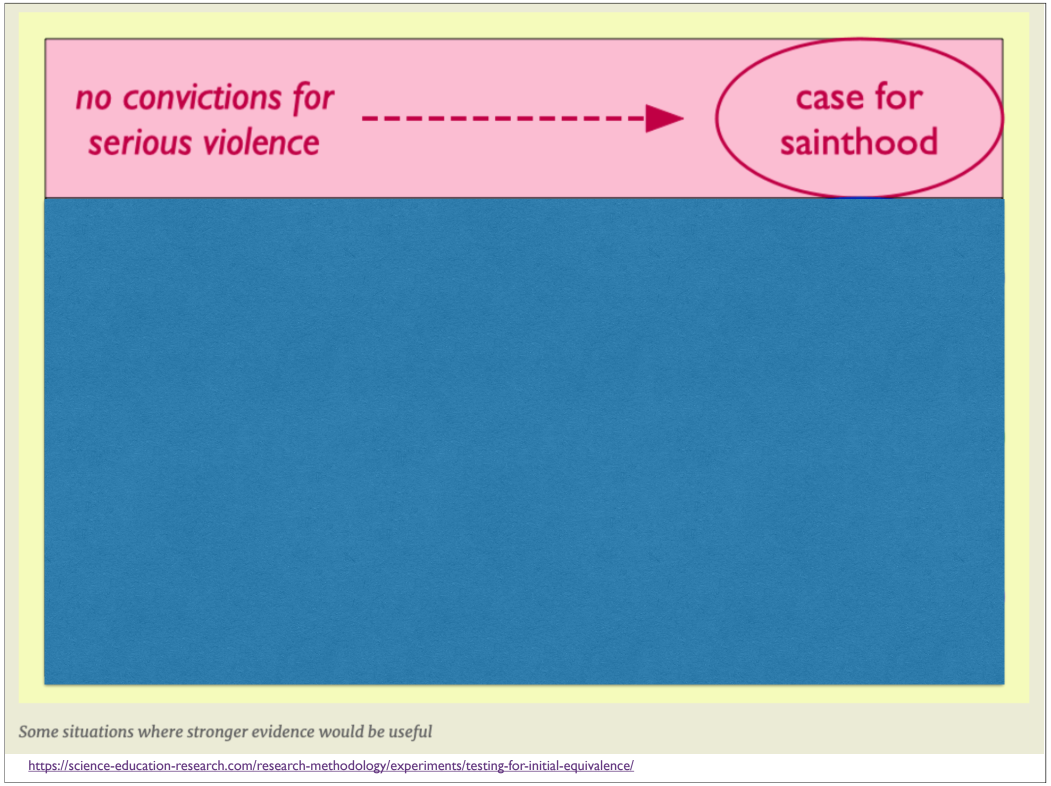
I do not think this is a sensible test for equivalence. It is a weak test, and, indeed, inadequate.

case for sainthood

Some situations where stronger evidence would be useful

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

Imagine you were looking for a test to decide if someone should be considered a Saint.



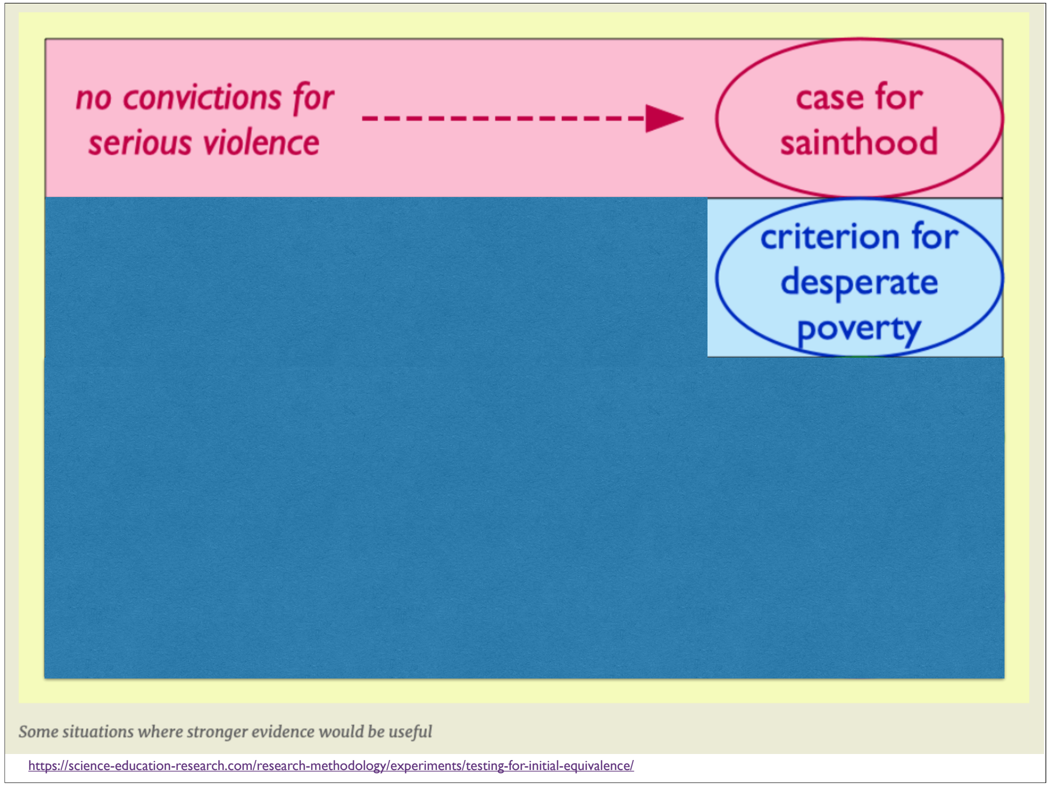
no convictions for serious violence

case for sainthood

Some situations where stronger evidence would be useful

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

How high would you set the bar?

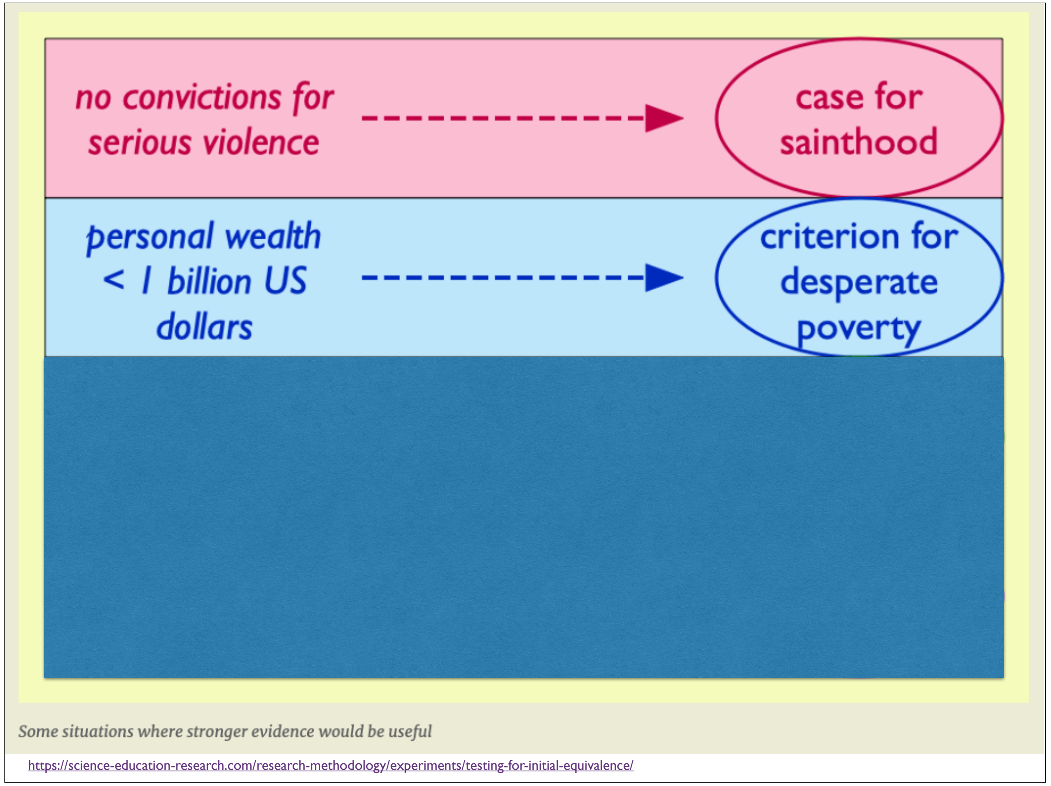


no convictions for serious violence
case for sainthood
criterion for desperate poverty

Some situations where stronger evidence would be useful

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

What if you wanted to identify those who should be considered in poverty.



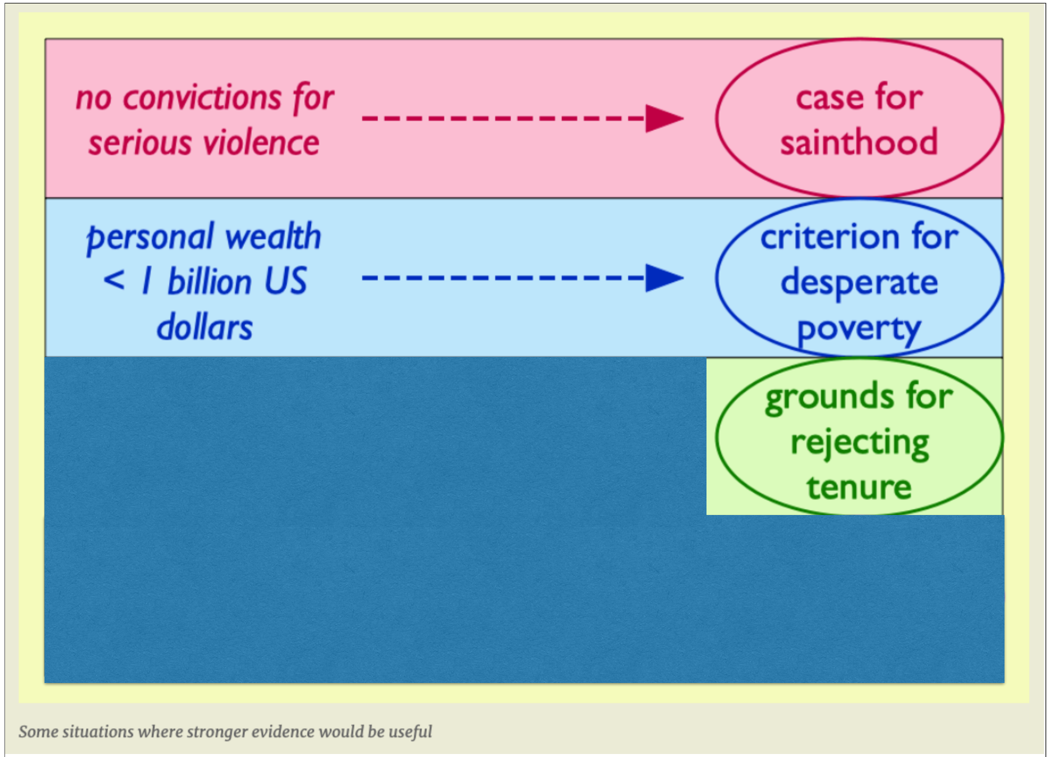
no convictions for serious violence ------------► case for sainthood

personal wealth < 1 billion US dollars ------------► criterion for desperate poverty

Some situations where stronger evidence would be useful

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

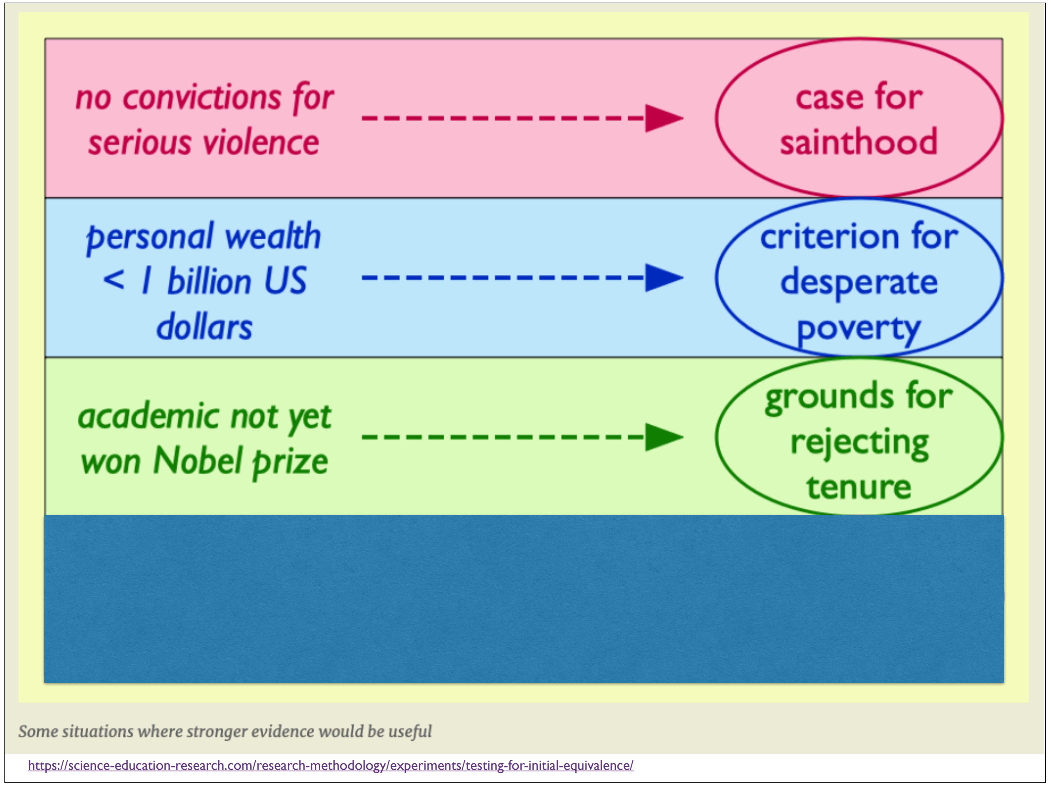
Certainly, not having a great deal of money would be a relevant criterion, but perhaps not quite exclusive enough.


no convictions for serious violence
case for sainthood
personal wealth < 1 billion US dollars
criterion for desperate poverty
grounds for rejecting tenure

Some situations where stronger evidence would be useful

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

Maybe you were on a committee that was considering rejecting the permanent re-appointment of a colleague on probation as their research record was not strong enough.



no convictions for serious violence → case for sainthood

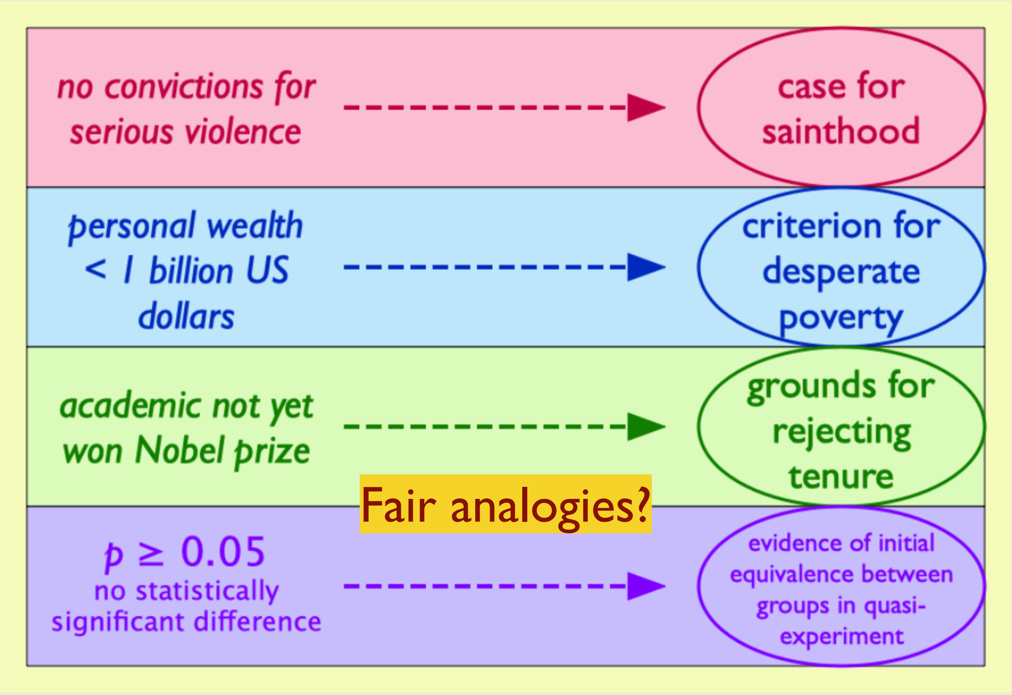
personal wealth < 1 billion US dollars → criterion for desperate poverty

academic not yet won Nobel prize → grounds for rejecting tenure

Some situations where stronger evidence would be useful

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

But you should not have unreasonable expectations.



no convictions for serious violence → case for sainthood

personal wealth < 1 billion US dollars → criterion for desperate poverty

academic not yet won Nobel prize → grounds for rejecting tenure

$p \geq 0.05$ no statistically significant difference → evidence of initial equivalence between groups in quasi-experiment

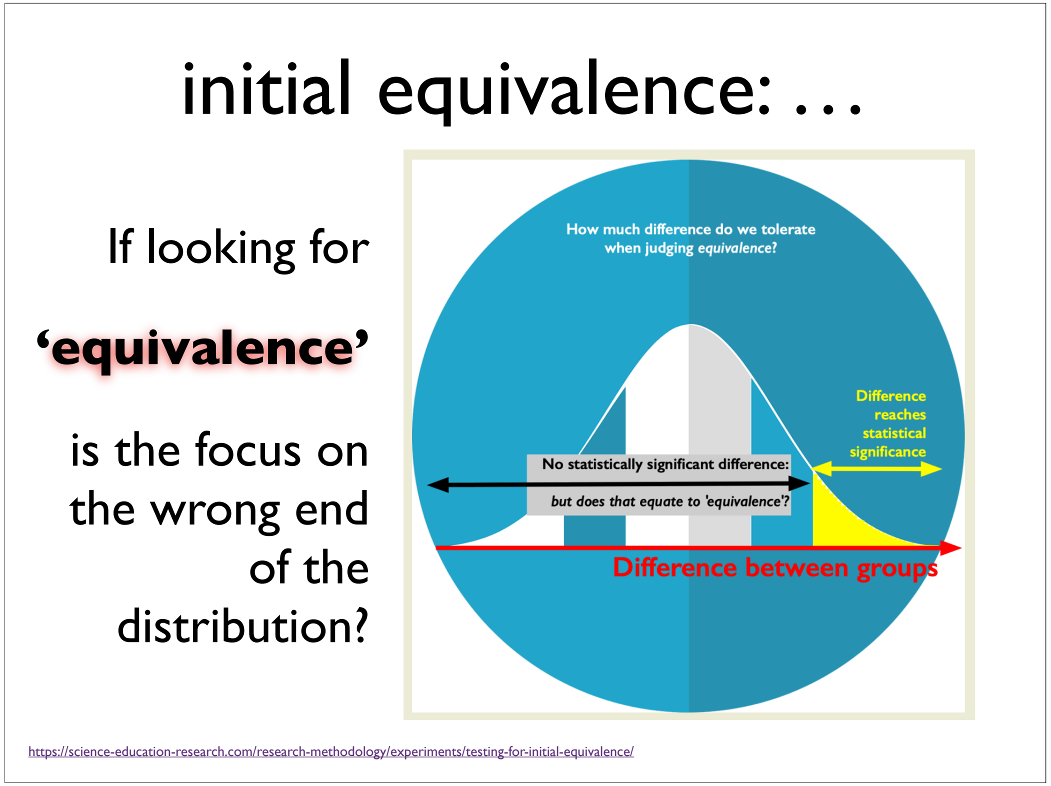
Fair analogies?

Some situations where stronger evidence would be useful

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

I think these are similar situations, in the sense that the criteria being adopted are certainly relevant and perhaps necessarily apply, but are by no means sufficient.

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/



initial equivalence: …

If looking for

‘equivalence’

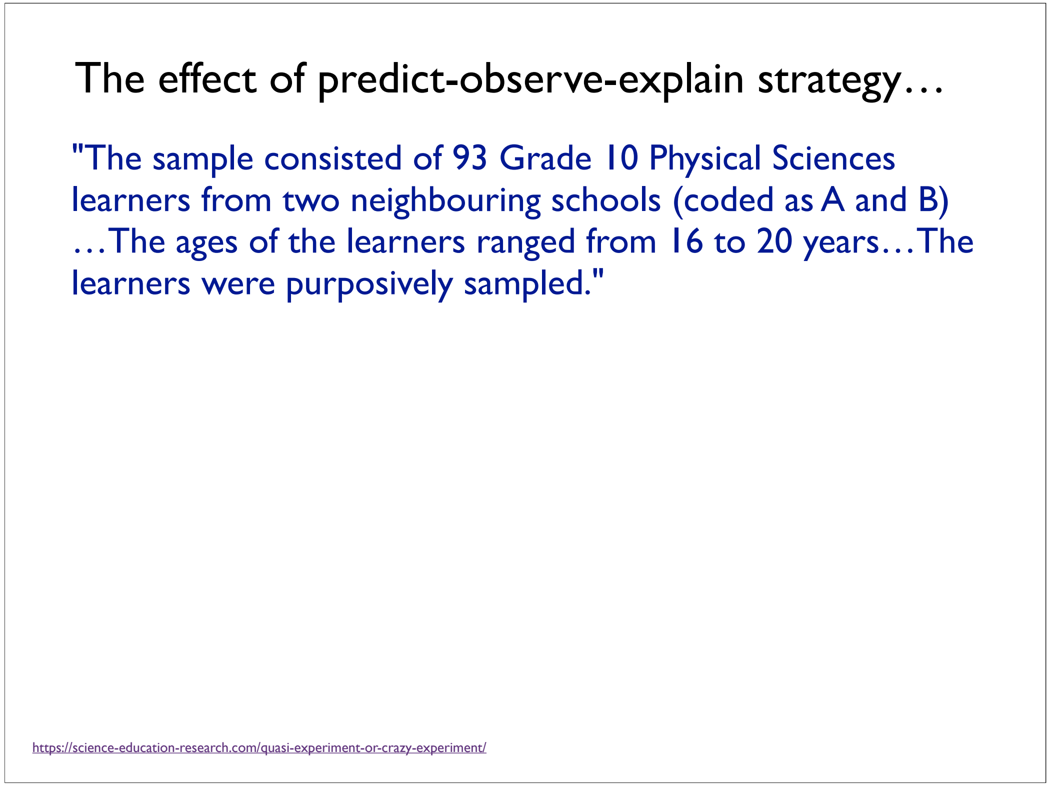
is the focus on the wrong end of the distribution?

How much difference do we tolerate when judging *equivalence*?
Difference reaches statistical significance
No statistically significant difference:
but does that equate to 'equivalence'?
Difference between groups

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/

In a sense, we are looking at the wrong end of the distribution. We should be asking how probable we need measured differences to be, not simply excluding the most improbable.

https://science-education-research.com/research-methodology/experiments/testing-for-initial-equivalence/



The effect of predict-observe-explain strategy…

"The sample consisted of 93 Grade 10 Physical Sciences learners from two neighbouring schools (coded as A and B) …The ages of the learners ranged from 16 to 20 years…The learners were purposively sampled."

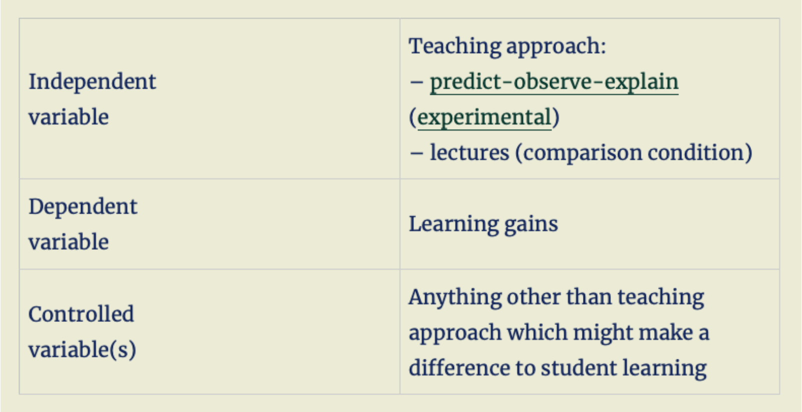
https://science-education-research.com/quasi-experiment-or-crazy-experiment/

Here is another real example.

https://science-education-research.com/quasi-experiment-or-crazy-experiment/

The effect of predict-observe-explain strategy…

"Forty nine (49) learners (31 males and 18 females) were from school A and acted as the experimental group (EG) whereas the control group (CG) consisted of 44 learners (18 males and 26 females) from school B."

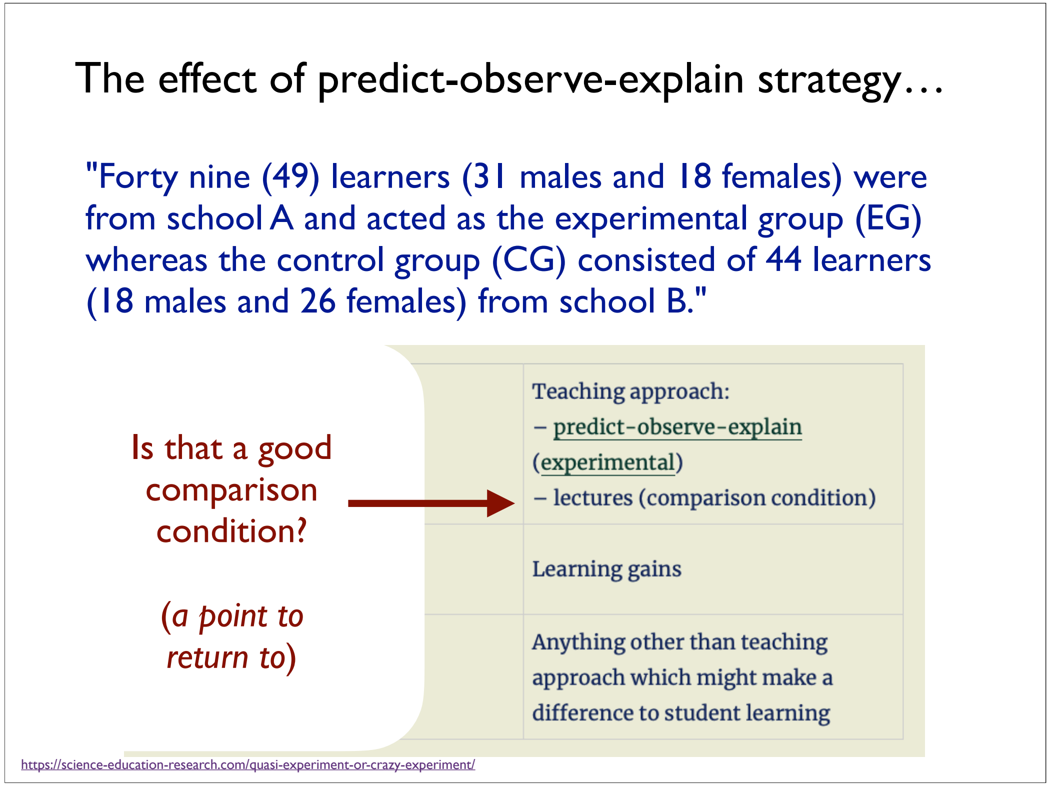


Independent variable	Teaching approach: – predict-observe-explain (experimental) – lectures (comparison condition)
Dependent variable	Learning gains
Controlled variable(s)	Anything other than teaching approach which might make a difference to student learning

https://science-education-research.com/quasi-experiment-or-crazy-experiment/

The study looked to compare between two teaching conditions.

https://science-education-research.com/quasi-experiment-or-crazy-experiment/



The effect of predict-observe-explain strategy…

"Forty nine (49) learners (31 males and 18 females) were from school A and acted as the experimental group (EG) whereas the control group (CG) consisted of 44 learners (18 males and 26 females) from school B."

Is that a good comparison condition?

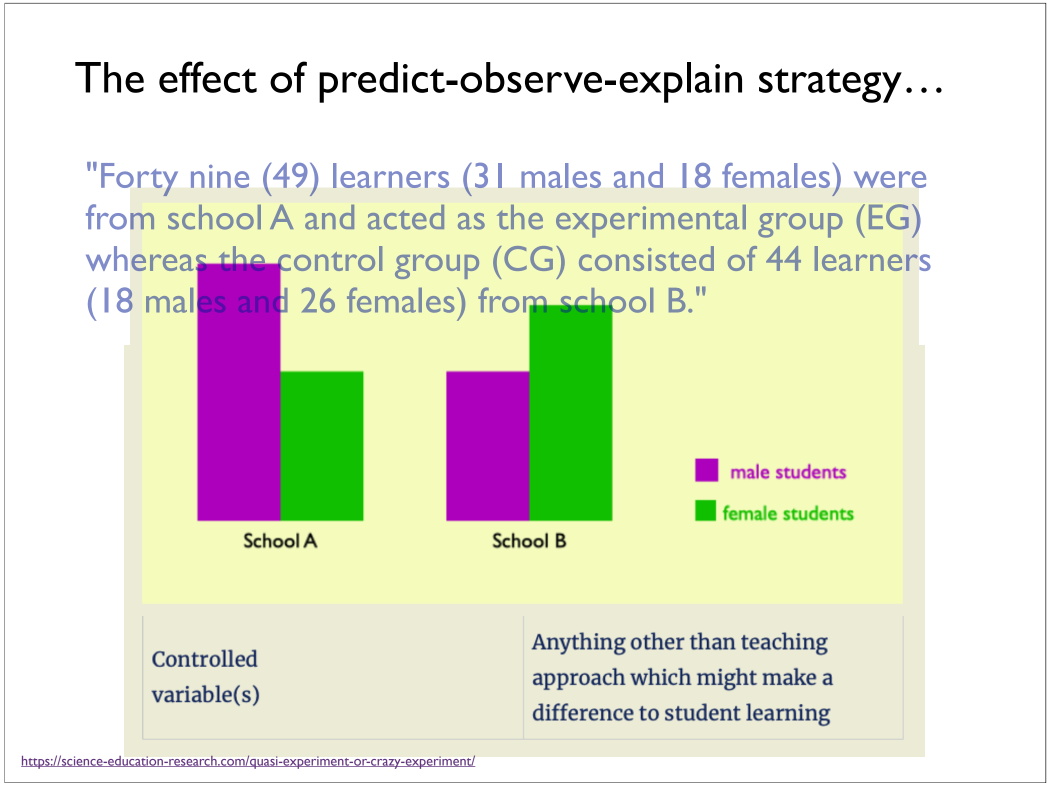
(*a point to return to*)

	Teaching approach: – predict-observe-explain (experimental) – lectures (comparison condition)
	Learning gains
	Anything other than teaching approach which might make a difference to student learning

https://science-education-research.com/quasi-experiment-or-crazy-experiment/

I struggle to understand why researchers think it is acceptable to require teachers to lecture school children, but I'll come back to that later.

https://science-education-research.com/quasi-experiment-or-crazy-experiment/



The effect of predict-observe-explain strategy…

"Forty nine (49) learners (31 males and 18 females) were from school A and acted as the experimental group (EG) whereas the control group (CG) consisted of 44 learners (18 males and 26 females) from school B."

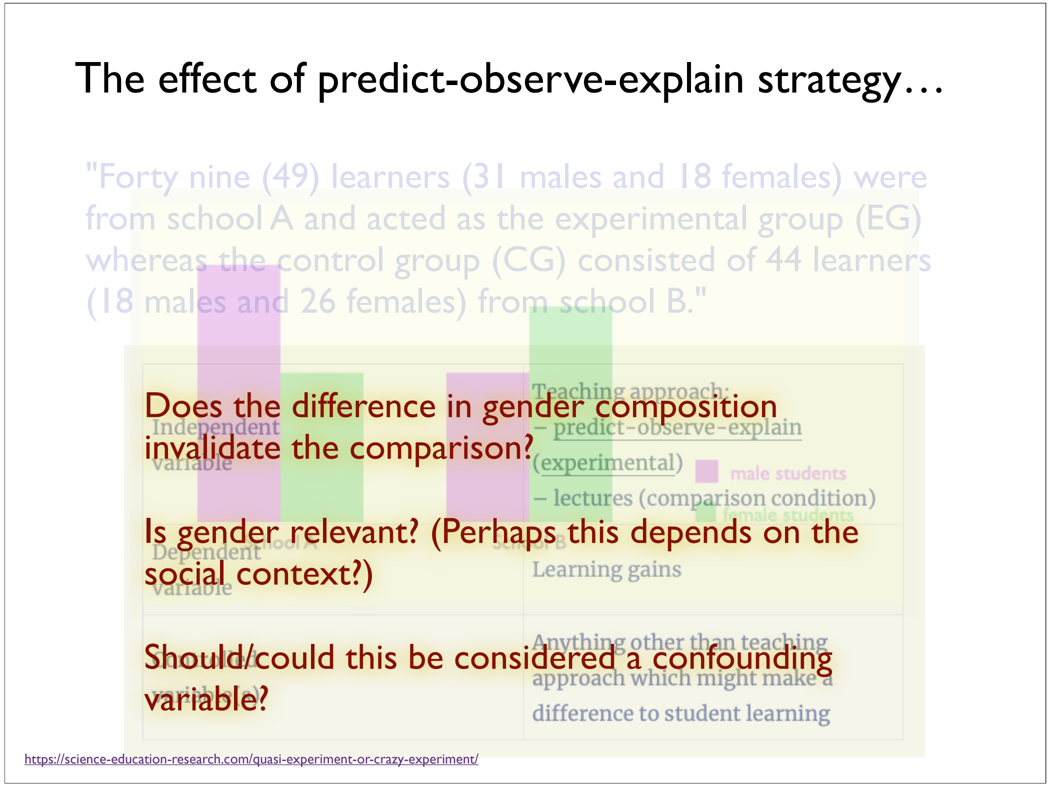
Controlled variable(s)	Anything other than teaching approach which might make a difference to student learning

https://science-education-research.com/quasi-experiment-or-crazy-experiment/

One thing we might notice is the very different gender composition of classes.

Is that relevant? Perhaps it should not be.

https://science-education-research.com/quasi-experiment-or-crazy-experiment/



The effect of predict-observe-explain strategy…

"Forty nine (49) learners (31 males and 18 females) were from school A and acted as the experimental group (EG) whereas the control group (CG) consisted of 44 learners (18 males and 26 females) from school B."

Does the difference in gender composition invalidate the comparison?

Is gender relevant? (Perhaps this depends on the social context?)

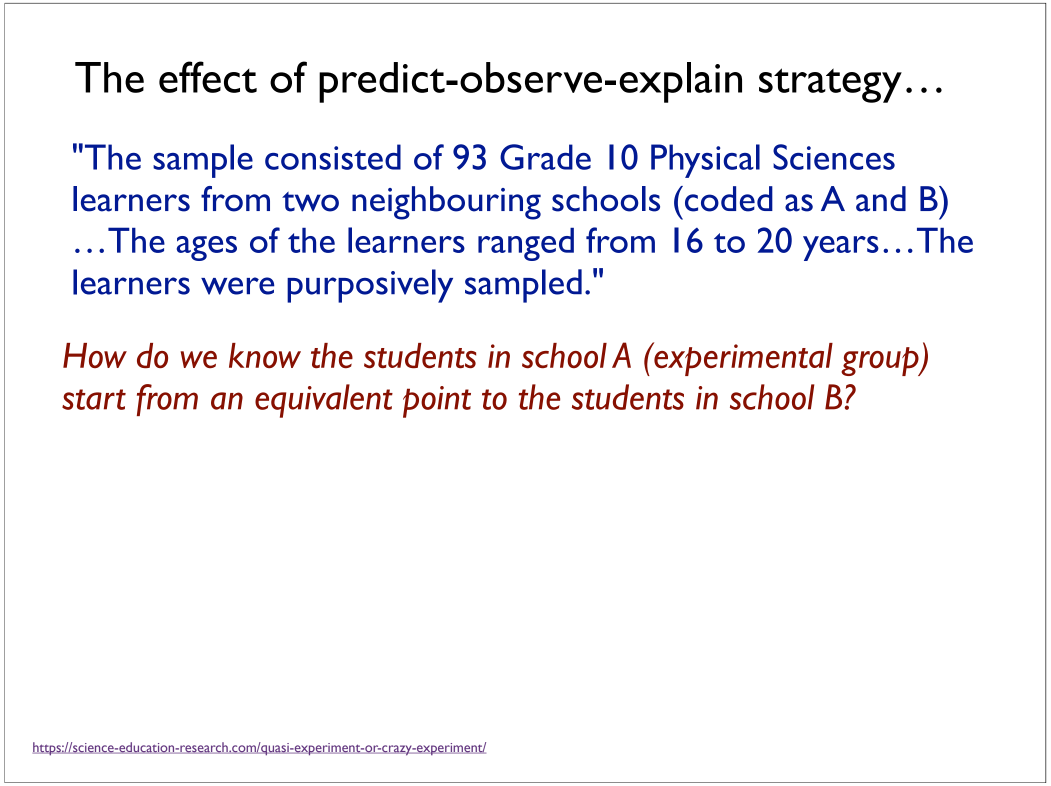
Should/could this be considered a confounding variable?

Independent variable	Teaching approach: – predict-observe-explain (experimental) – lectures (comparison condition)
Dependent variable	Learning gains
Controlled variable(s)	Anything other than teaching approach which might make a difference to student learning

https://science-education-research.com/quasi-experiment-or-crazy-experiment/

In some cultural contexts though, it might be.

https://science-education-research.com/quasi-experiment-or-crazy-experiment/



The effect of predict-observe-explain strategy…

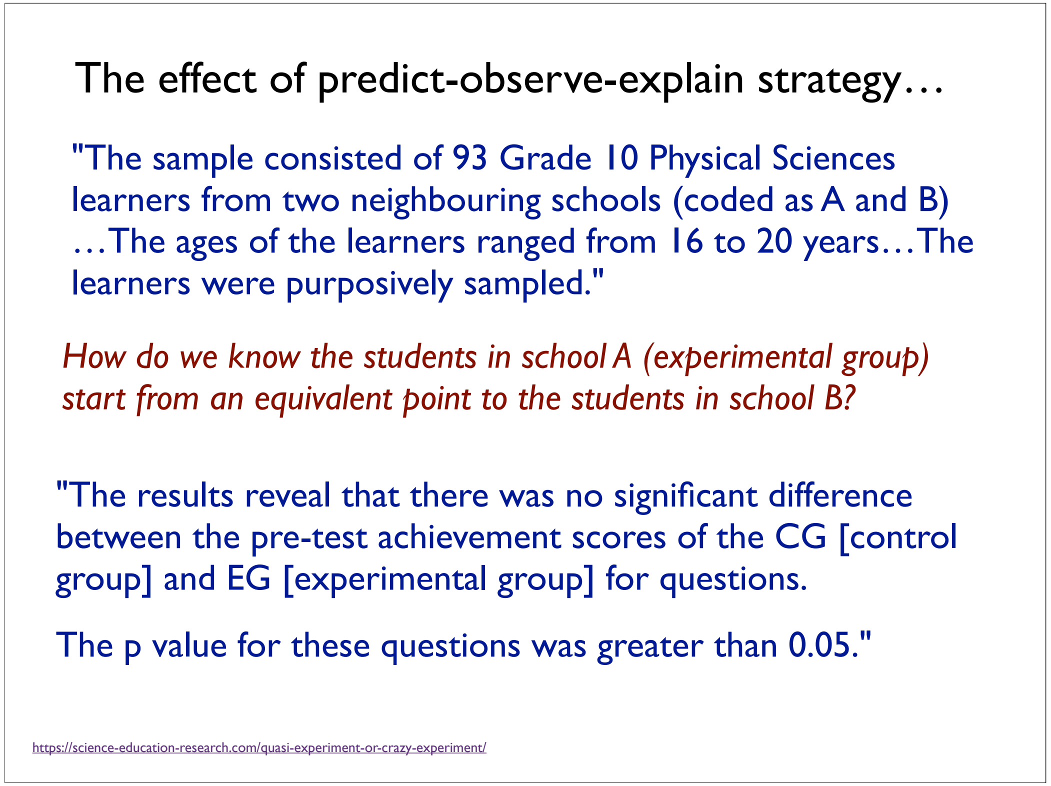
"The sample consisted of 93 Grade 10 Physical Sciences learners from two neighbouring schools (coded as A and B) …The ages of the learners ranged from 16 to 20 years…The learners were purposively sampled."

How do we know the students in school A (experimental group) start from an equivalent point to the students in school B?

https://science-education-research.com/quasi-experiment-or-crazy-experiment/

The two groups are in different schools, which I think it really troubling in these kinds of studies, as schools vary so much, and in so many ways.

https://science-education-research.com/quasi-experiment-or-crazy-experiment/



The effect of predict-observe-explain strategy…

"The sample consisted of 93 Grade 10 Physical Sciences learners from two neighbouring schools (coded as A and B) …The ages of the learners ranged from 16 to 20 years…The learners were purposively sampled."

How do we know the students in school A (experimental group) start from an equivalent point to the students in school B?

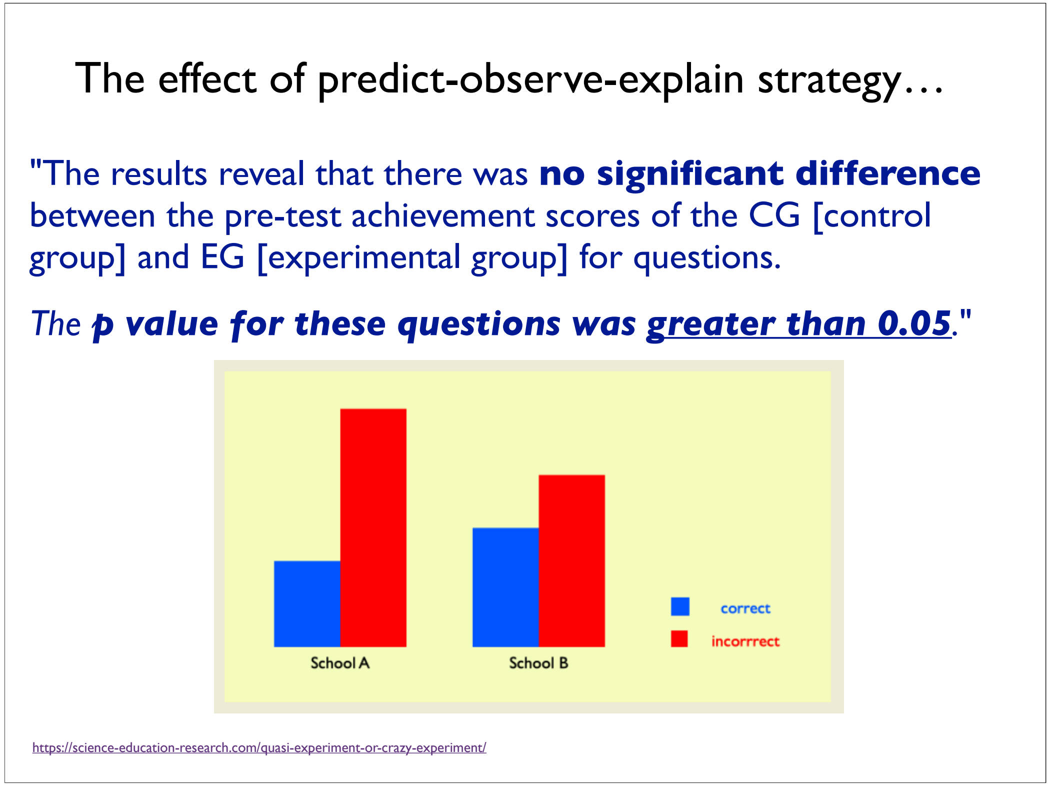
"The results reveal that there was no significant difference between the pre-test achievement scores of the CG [control group] and EG [experimental group] for questions.

The p value for these questions was greater than 0.05."

https://science-education-research.com/quasi-experiment-or-crazy-experiment/

But a pre-test was used, and the researchers claimed that on none of the items did differences between groups reach significance. This was seen to assure equivalence.

https://science-education-research.com/quasi-experiment-or-crazy-experiment/



The effect of predict-observe-explain strategy…

"The results reveal that there was **no significant difference** between the pre-test achievement scores of the CG [control group] and EG [experimental group] for questions.

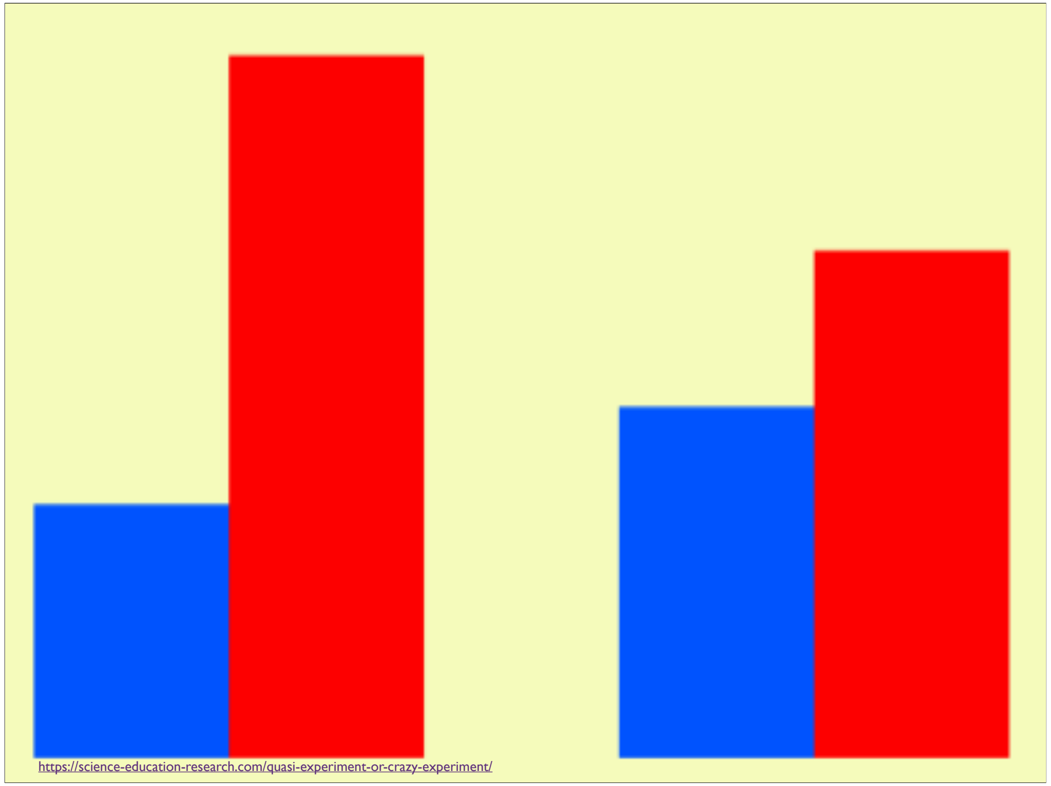
*The **p value for these questions was greater than 0.05**."*

https://science-education-research.com/quasi-experiment-or-crazy-experiment/

From the data given, I prepared this chart showing the performance on one of the items at pre-test.

We are told that there was ‘no significant difference’. Certainly, in both groups most students got the question wrong.

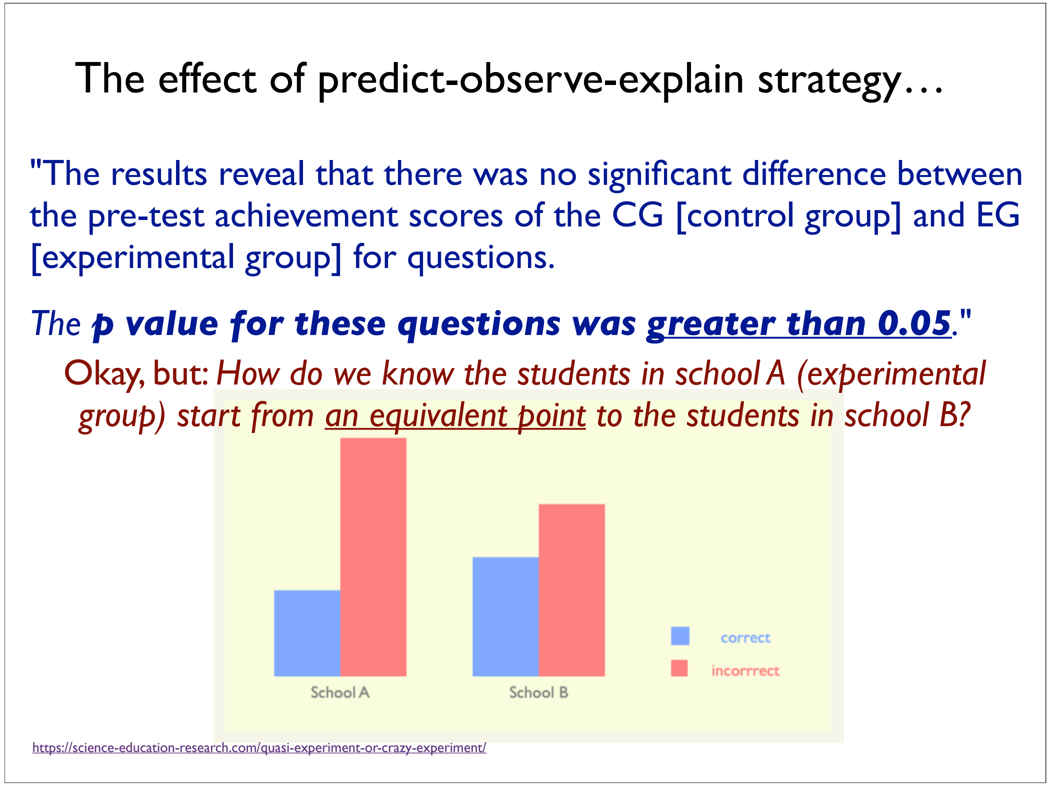
https://science-education-research.com/quasi-experiment-or-crazy-experiment/



https://science-education-research.com/quasi-experiment-or-crazy-experiment/

But if this is an equivalent performance in the two classes, then the word 'equivalent' means something very different to its normal sense.

Surely, despite the lack of statistical significance, one of these classes is better placed to build on their level of prior learning than the other?



The effect of predict-observe-explain strategy…

"The results reveal that there was no significant difference between the pre-test achievement scores of the CG [control group] and EG [experimental group] for questions.

*The **p value for these questions was <u>greater than 0.05</u>**.*"

Okay, but: *How do we know the students in school A (experimental group) start from <u>an equivalent point</u> to the students in school B?*

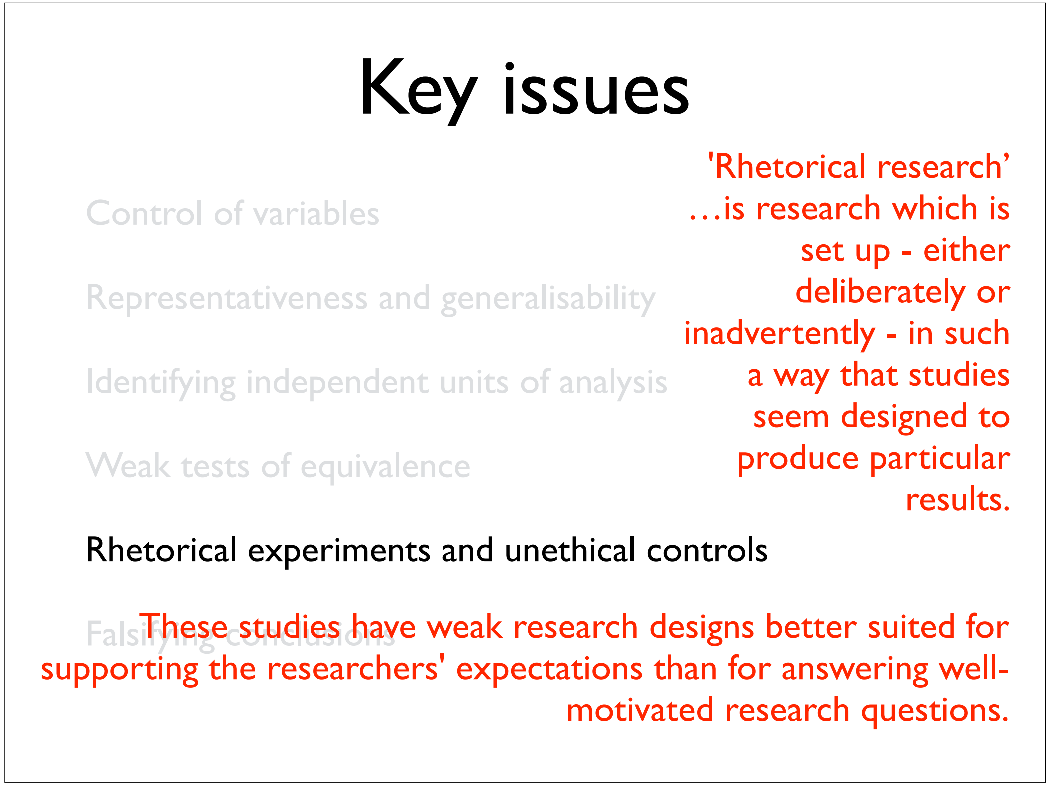
https://science-education-research.com/quasi-experiment-or-crazy-experiment/

This does not persuade me of equivalence.

This is not an isolated case. This technique is very widely used.

https://science-education-research.com/quasi-experiment-or-crazy-experiment/

ethics of control conditions



Key issues

'Rhetorical research'
…is research which is set up - either deliberately or inadvertently - in such a way that studies seem designed to produce particular results.

Control of variables

Representativeness and generalisability

Identifying independent units of analysis

Weak tests of equivalence

Rhetorical experiments and unethical controls

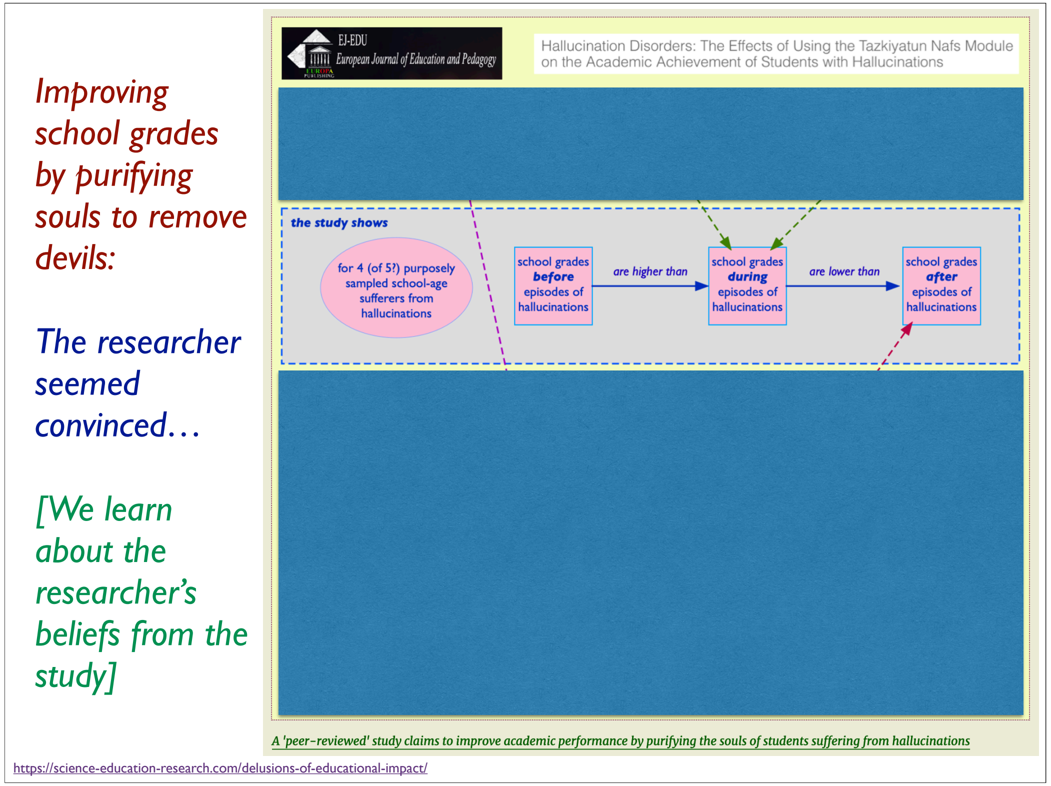
Falsifying conclusions

These studies have weak research designs better suited for supporting the researchers' expectations than for answering well-motivated research questions.

Another major concern I have with some published studies, is that they seem to be examples of rhetorical research.

That is the study is done to demonstrate what they researchers already expect, indeed believe, and not in a spirit of open-ended enquiry.

I have noticed that many school science practicals undertaken to demonstrate well-established scientific findings are often incorrectly referred to as 'experiments', but surely professional science educators know that genuine experiments have uncertain outcomes?



Improving school grades by purifying souls to remove devils:

The researcher seemed convinced…

[We learn about the researcher's beliefs from the study]

EJ-EDU European Journal of Education and Pedagogy

Hallucination Disorders: The Effects of Using the Tazkiyatun Nafs Module on the Academic Achievement of Students with Hallucinations

the study shows

for 4 (of 5?) purposely sampled school-age sufferers from hallucinations

school grades **before** episodes of hallucinations *are higher than* school grades **during** episodes of hallucinations *are lower than* school grades **after** episodes of hallucinations

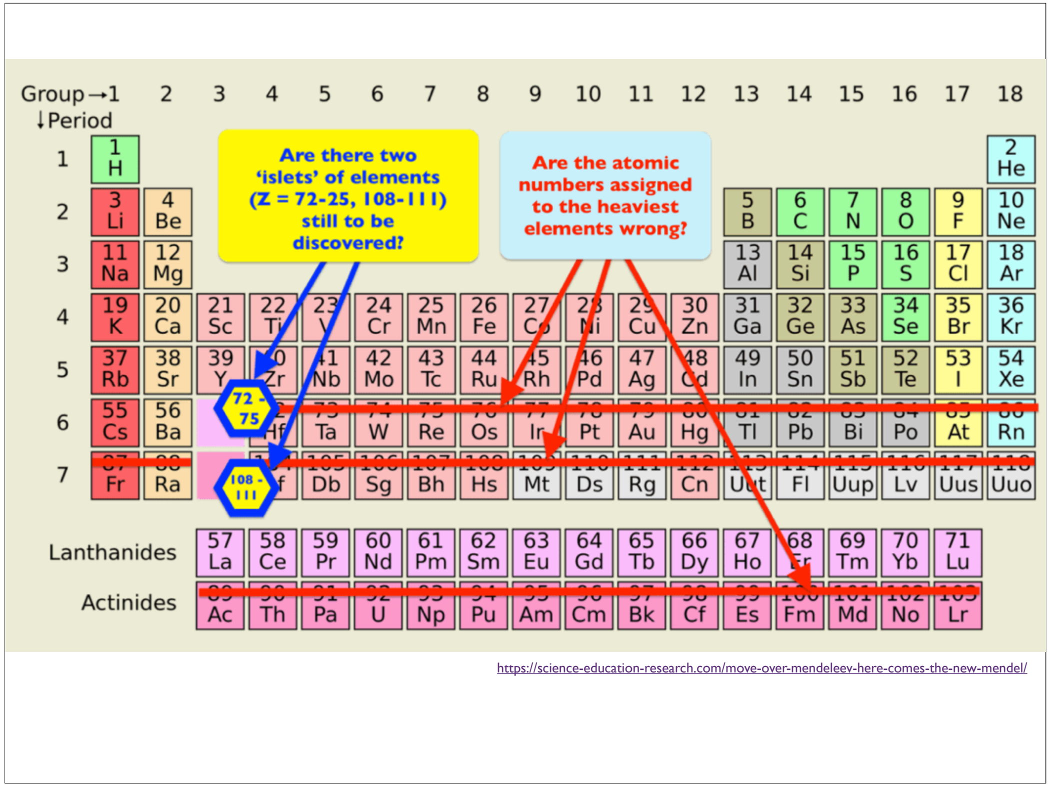
A 'peer-reviewed' study claims to improve academic performance by purifying the souls of students suffering from hallucinations

https://science-education-research.com/delusions-of-educational-impact/

Of course, if one includes more dubious journals in one's purview, one can find extreme examples where no doubt the researcher was entirely convinced by their research, but it seems unlikely there was ever any serious peer review or editorial evaluation beyond checking the publication fee had been submitted. Sadly, there are now a large number of predatory journals out there.

I've detailed a number of examples of both honest and dishonest nonsense published in such journals on my website, by which I mean both work which has been submitted in good faith, but which should never have been published; as well as things that presumably even the authors knew were complete nonsense.

https://science-education-research.com/delusions-of-educational-impact/



Group→ ↓Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba		72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra		104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Uut	114 Fl	115 Uup	116 Lv	117 Uus	118 Uuo
Lanthanides			57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu	
Actinides			89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr	

Are there two 'islets' of elements (Z = 72-25, 108-111) still to be discovered?

Are the atomic numbers assigned to the heaviest elements wrong?

https://science-education-research.com/move-over-mendeleev-here-comes-the-new-mendel/

An example of the former is an alternative version of periodic table including a whole raft of new elements that had previously been missed.

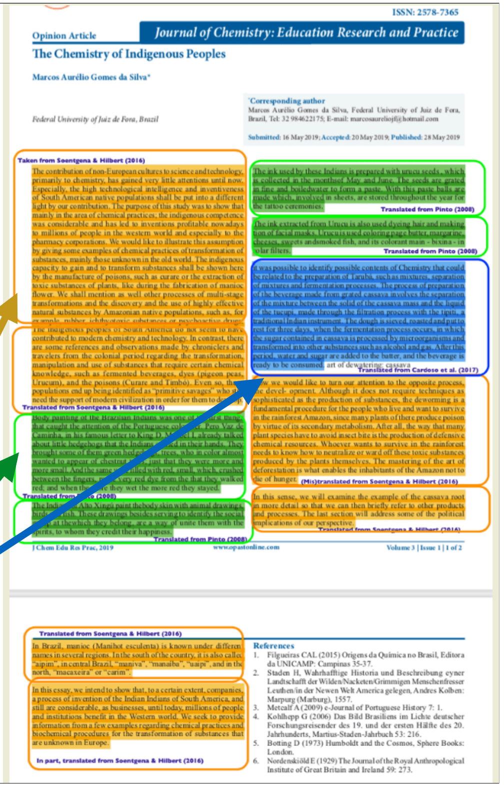
https://science-education-research.com/move-over-mendeleev-here-comes-the-new-mendel/

“The mastering of the art of deforestation is what enables the inhabitants of the Amazon not to die of hunger.”

Marcos Aurélio Gomes da Silva,
Federal University of Juiz de Fora, Brazil

highlighting showing paragraphs from different published sources

https://science-education-research.com/can-deforestation-stop-indigenous-groups-starving/



ISSN: 2578-7365

Opinion Article

Journal of Chemistry: Education Research and Practice

The Chemistry of Indigenous Peoples

Marcos Aurélio Gomes da Silva*

Federal University of Juiz de Fora, Brazil

*Corresponding author
Marcos Aurélio Gomes da Silva, Federal University of Juiz de Fora, Brazil, Tel: 32 984622175; E-mail: marcosaureliojf@hotmail.com

Submitted: 16 May 2019; Accepted: 20 May 2019; Published: 28 May 2019

Taken from Soentgena & Hilbert (2016)

The contribution of non-European cultures to science and technology, primarily to chemistry, has gained very little attentions until now. Especially, the high technological intelligence and inventiveness of South American native populations shall be put into a different light by our contribution. The purpose of this study was to show that mainly in the area of chemical practices; the indigenous competence was considerable and has led to inventions profitable nowadays to millions of people in the western world and especially to the pharmacy corporations. We would like to illustrate this assumption by giving some examples of chemical practices of transformation of substances, mainly those unknown in the old world. The indigenous capacity to gain and to transform substances shall be shown here by the manufacture of poisons, such as curare or the extraction of toxic substances of plants, like during the fabrication of manioc flower. We shall mention as well other processes of multi-stage transformations and the discovery and the use of highly effective natural substances by Amazonian native populations, such as, for example, rubber, ichthyotoxic substances or psychoactive drugs.

The indigenous peoples of South America do not seem to have contributed to modern chemistry and technology. In contrast, there are some references and observations made by chroniclers and travelers from the colonial period regarding the transformation, manipulation and use of substances that require certain chemical knowledge, such as fermented beverages, dyes (pigeon peas, Urucum), and the poisons (Curare and Timbó). Even so, these populations end up being identified as "primitive savages" who still need the support of modern civilization in order for them to develop.

Translated from Soentgena & Hilbert (2016)

Body painting of the Brazilian Indians was one of the first things that caught the attention of the Portuguese colonizer. Pero Vaz de Caminha, in his famous letter to King D. Manuel I, already talked about little hedgehogs that the Indians carried in their hands. They brought some of them green hedgehogs, trees, who in color almost wanted to appear of chestnut trees; just that they were more and more small. And the same were filled with red, small, which, crushed between the fingers, made very red dye from the that they walked red; and when the more they wet the more red they stayed.

Translated from Pinto (2008)

The Indians of Alto Xingú paint thebody skin with animal drawings, birds and fish. These drawings besides serving to identify the social group at thewhich they belong, are a way of unite them with the spirits, to whom they credit their happiness.

Translated from Pinto (2008)

The ink used by these Indians is prepared with urucu seeds , which is collected in the monthsof May and June. The seeds are grated in fine and boiledwater to form a paste. With this paste balls are made which, involved in sheets, are stored throughout the year for the tattoo ceremonies.

Translated from Pinto (2008)

The ink extracted from Urucu is also used dyeing hair and making tion of facial masks. Urucu is used coloring page butter, margarine, cheeses, sweets andsmoked fish, and its colorant main - bixina - in solar filters.

Translated from Pinto (2008)

it was possible to identify possible contents of Chemistry that could be related to the preparation of Tarubá, such as mixtures, separation of mixtures and fermentation processes. The process of preparation of the beverage made from grated cassava involves the separation of the mixture between the solid of the cassava mass and the liquid of the tucupi, made through the filtration process with the tipiti, a traditional Indian instrument. The dough is sieved, roasted and put to rest for three days, when the fermentation process occurs, in which the sugar contained in cassava is processed by microorganisms and transformed into other substances such as alcohol and gas. After this period, water and sugar are added to the batter, and the beverage is ready to be consumed. art of dewatering: cassava

Translated from Cardoso et al. (2017)

Now we would like to turn our attention to the opposite process, the devel- opment. Although it does not require techniques as sophisticated as the production of substances, the deworming is a fundamental procedure for the people who live and want to survive in the rainforest Amazon, since many plants of there produce poison by virtue of its secondary metabolism. After all, the way that many plant species have to avoid insect bite is the production of defensive chemical resources. Whoever wants to survive in the rainforest needs to know how to neutralize or ward off these toxic substances produced by the plants themselves. The mastering of the art of deforestation is what enables the inhabitants of the Amazon not to die of hunger.

(Mis)translated from Soentgena & Hilbert (2016)

In this sense, we will examine the example of the cassava root in more detail so that we can then briefly refer to other products and processes. The last section will address some of the political implications of our perspective.

Translated from Soentgena & Hilbert (2016)

J Chem Edu Res Prac, 2019 www.opastonline.com Volume 3 | Issue 1 | 1 of 2

Translated from Soentgena & Hilbert (2016)

In Brazil, manioc (Manihot esculenta) is known under different names in several regions. In the south of the country, it is also called "aipim", in central Brazil, "maniva", "manaíba", "uaipi", and in the north, "macaxeira" or "carim".

In this essay, we intend to show that, to a certain extent, companies, a process of invention of the Indian Indians of South America, and still are considerable, as businesses, until today, millions of people and institutions benefit in the Western world. We seek to provide information from a few examples regarding chemical practices and biochemical procedures for the transformation of substances that are unknown in Europe.

In part, translated from Soentgena & Hilbert (2016)

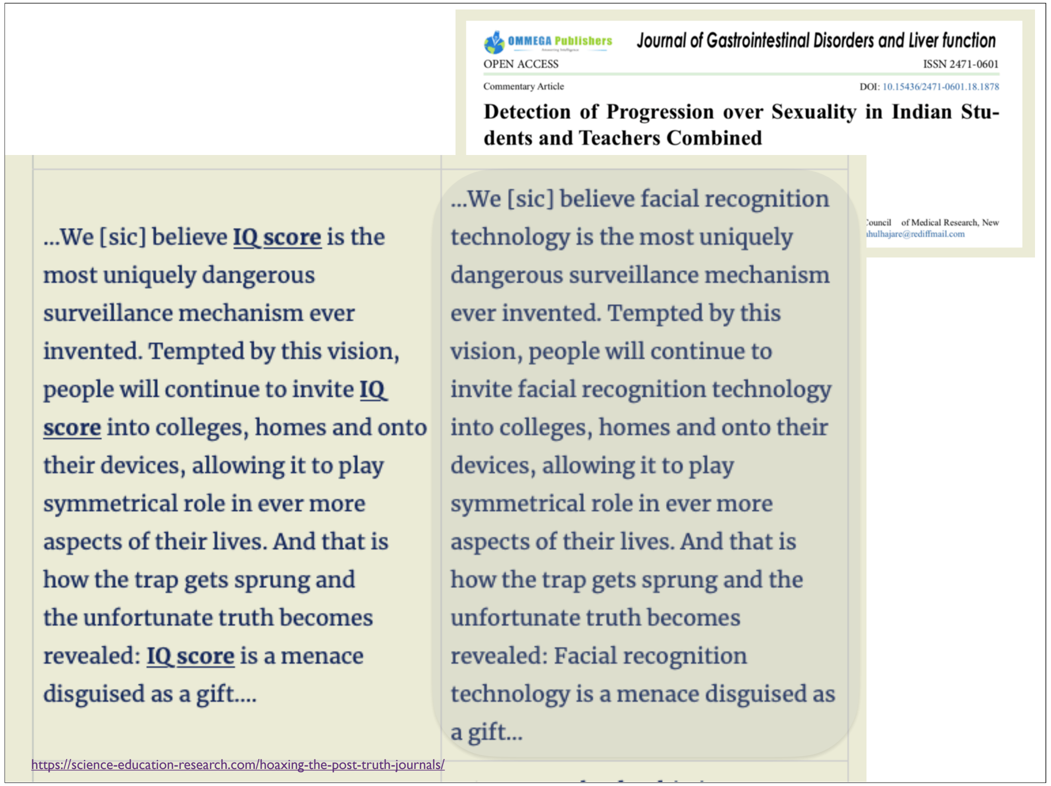
References

1. Filgueiras CAL (2015) Origens da Química no Brasil, Editora da UNICAMP: Campinas 35-37.
2. Staden H, Wahrhafftige Historia und Beschreibung eyner Landtschafft der Wilden/Nacketen/Grimmigen Menschenfresser Leuthen/in der Newen Welt America gelegen, Andres Kolben: Marpurg (Marburg), 1557.
3. Metcalf A (2009) e-Journal of Portuguese History 7: 1.
4. Kohlhepp G (2006) Das Bild Brasiliens im Lichte deutscher Forschungsreisender des 19. und der ersten Hälfte des 20. Jahrhunderts, Martius-Staden-Jahrbuch 53: 216.
5. Botting D (1973) Humboldt and the Cosmos, Sphere Books: London.
6. Nordenskiöld E (1929) The Journal of the Royal Anthropological Institute of Great Britain and Ireland 59: 273.

An example of the latter was a paper which suggested that deforestation of the Amazon was a good idea.

After some digging around I found the entire paper was a collage of translations of paragraphs form other published works.
Poor translations. ‘Deforestation’ had been ‘poisoning’ in the source.

https://science-education-research.com/can-deforestation-stop-indigenous-groups-starving/



OMMEGA Publishers Journal of Gastrointestinal Disorders and Liver function

OPEN ACCESS ISSN 2471-0601

Commentary Article DOI: 10.15436/2471-0601.18.1878

Detection of Progression over Sexuality in Indian Students and Teachers Combined

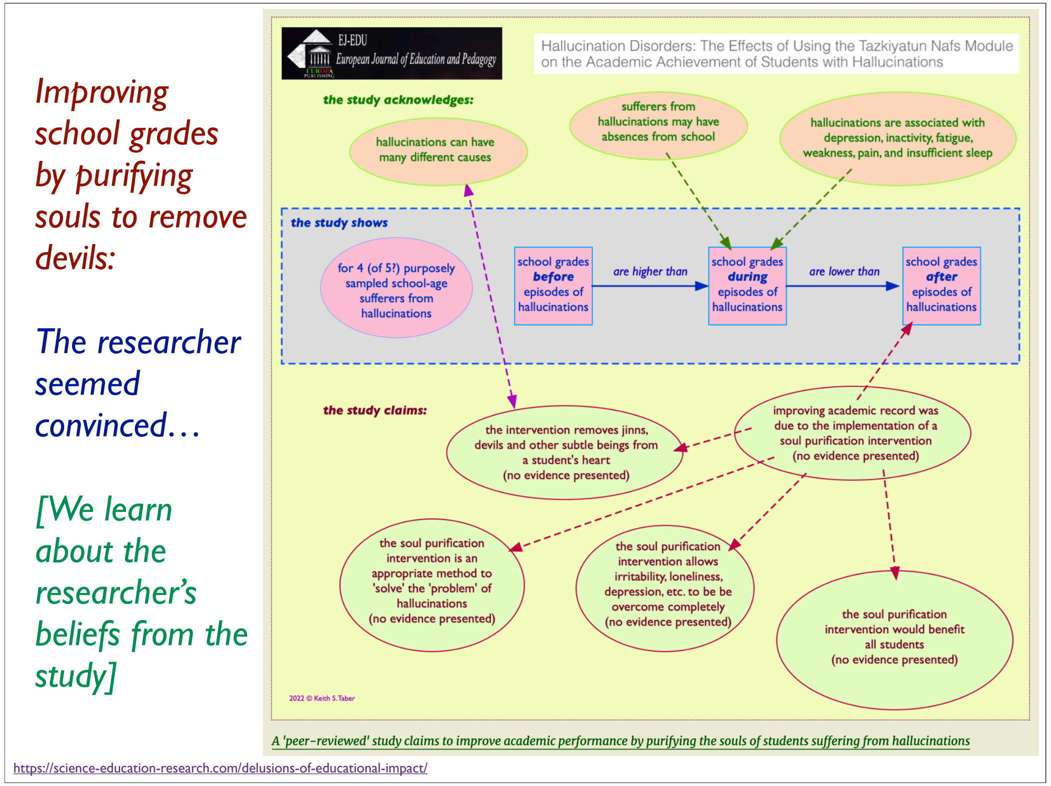
ouncil of Medical Research, New
hulhajare@rediffmail.com

...We [sic] believe **IQ score** is the most uniquely dangerous surveillance mechanism ever invented. Tempted by this vision, people will continue to invite **IQ score** into colleges, homes and onto their devices, allowing it to play symmetrical role in ever more aspects of their lives. And that is how the trap gets sprung and the unfortunate truth becomes revealed: **IQ score** is a menace disguised as a gift....	...We [sic] believe facial recognition technology is the most uniquely dangerous surveillance mechanism ever invented. Tempted by this vision, people will continue to invite facial recognition technology into colleges, homes and onto their devices, allowing it to play symmetrical role in ever more aspects of their lives. And that is how the trap gets sprung and the unfortunate truth becomes revealed: Facial recognition technology is a menace disguised as a gift...

https://science-education-research.com/hoaxing-the-post-truth-journals/

One paper I found basically plagiarised a published study, but simply replaced key words to make the paper about something else entirely.

https://science-education-research.com/hoaxing-the-post-truth-journals/



Improving school grades by purifying souls to remove devils:

The researcher seemed convinced…

[We learn about the researcher's beliefs from the study]

Hallucination Disorders: The Effects of Using the Tazkiyatun Nafs Module on the Academic Achievement of Students with Hallucinations

the study acknowledges:

hallucinations can have many different causes

sufferers from hallucinations may have absences from school

hallucinations are associated with depression, inactivity, fatigue, weakness, pain, and insufficient sleep

the study shows

for 4 (of 5?) purposely sampled school-age sufferers from hallucinations

school grades **before** episodes of hallucinations — *are higher than* → school grades **during** episodes of hallucinations — *are lower than* → school grades **after** episodes of hallucinations

the study claims:

the intervention removes jinns, devils and other subtle beings from a student's heart (no evidence presented)

improving academic record was due to the implementation of a soul purification intervention (no evidence presented)

the soul purification intervention is an appropriate method to 'solve' the 'problem' of hallucinations (no evidence presented)

the soul purification intervention allows irritability, loneliness, depression, etc. to be be overcome completely (no evidence presented)

the soul purification intervention would benefit all students (no evidence presented)

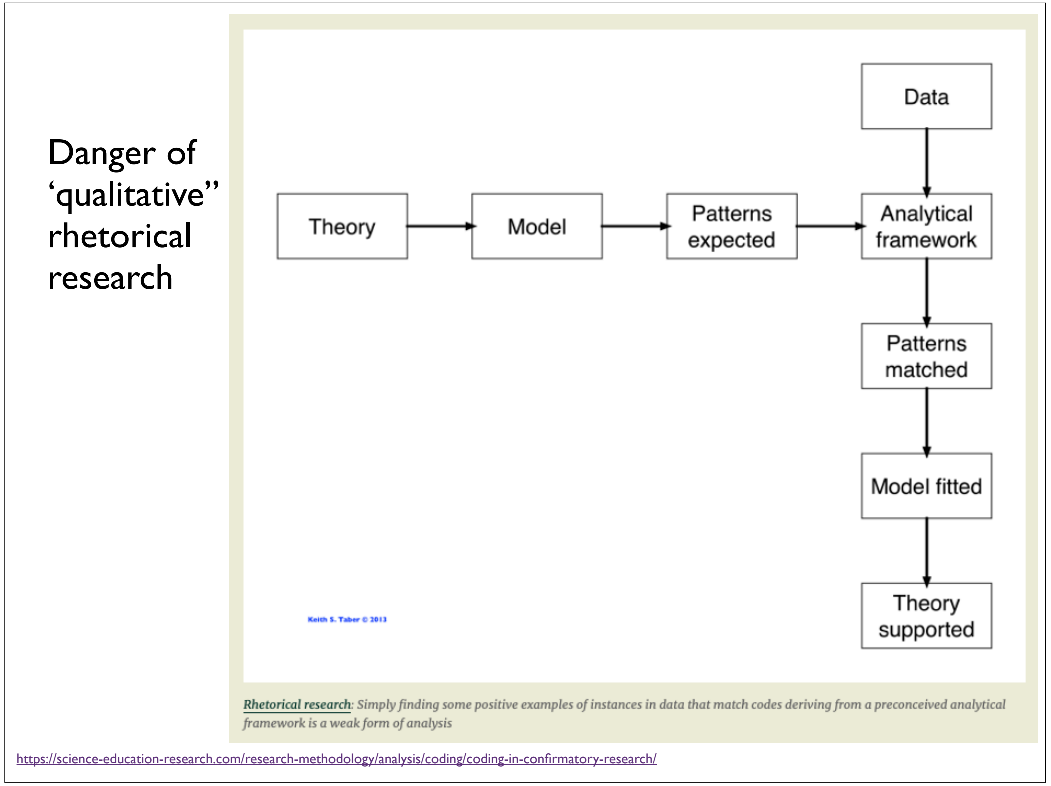
2022 © Keith S.Taber

A 'peer-reviewed' study claims to improve academic performance by purifying the souls of students suffering from hallucinations

https://science-education-research.com/delusions-of-educational-impact/

I can only assume that the European Journal of Education and Pedagogy does not use serious peer review as the paper I have summarised here does not logically show that school grades can be improved by exorcism, though I strongly suspect the author was genuine enough about the work. That is one of the key points for rigorous peer review - we can have blind spots in our own work.

https://science-education-research.com/delusions-of-educational-impact/



Danger of ‘qualitative” rhetorical research

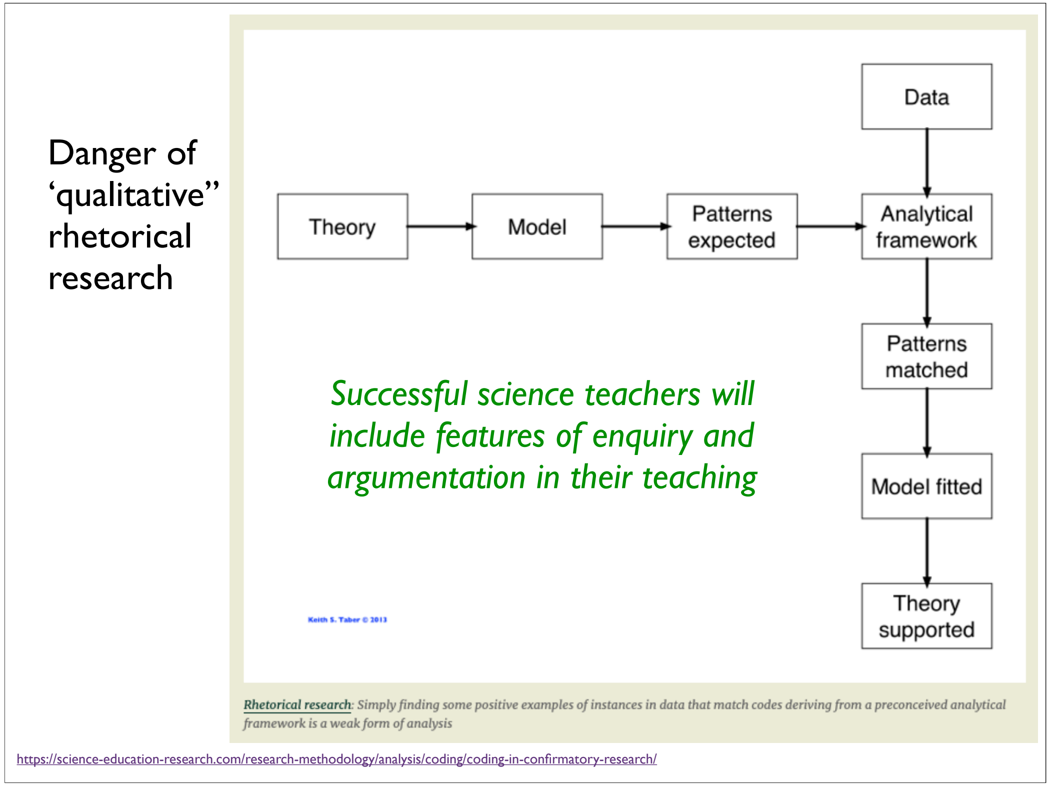
Rhetorical research: Simply finding some positive examples of instances in data that match codes deriving from a preconceived analytical framework is a weak form of analysis

https://science-education-research.com/research-methodology/analysis/coding/coding-in-confirmatory-research/

It is very easy for qualitative studies to become rhetorical.

We look for something. We find evidence.

https://science-education-research.com/research-methodology/analysis/coding/coding-in-confirmatory-research/



Danger of ‘qualitative” rhetorical research

Theory → Model → Patterns expected → Analytical framework

Data → Analytical framework → Patterns matched → Model fitted → Theory supported

Successful science teachers will include features of enquiry and argumentation in their teaching

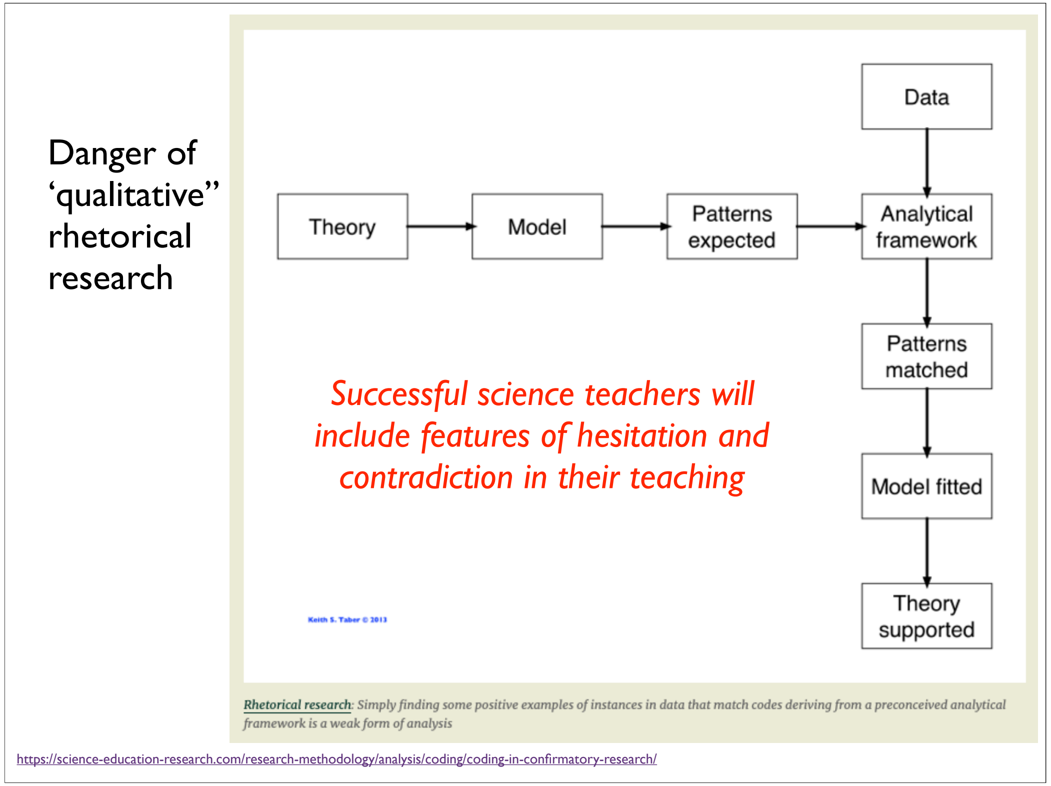
Keith S. Taber © 2013

Rhetorical research: Simply finding some positive examples of instances in data that match codes deriving from a preconceived analytical framework is a weak form of analysis

https://science-education-research.com/research-methodology/analysis/coding/coding-in-confirmatory-research/

If we believe that enquiry and argumentation are essential to good science teaching; and devise an observation schedule based on indicators of enquiry and argumentation; and then observe the lessons of good science teachers, we are likely to find indicators of enquiry and argumentation.

The question is, so what?



Danger of ‘qualitative” rhetorical research

Theory → Model → Patterns expected → Analytical framework

Data → Analytical framework → Patterns matched → Model fitted → Theory supported

Successful science teachers will include features of hesitation and contradiction in their teaching

Keith S. Taber © 2013

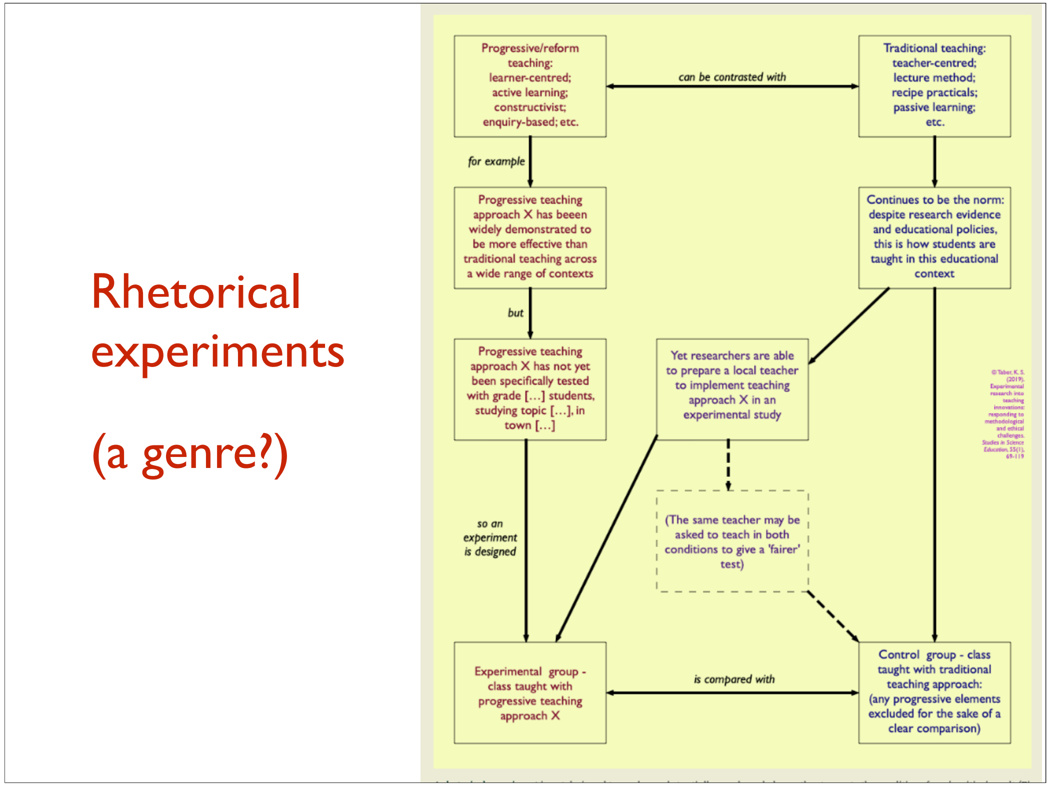
Rhetorical research: Simply finding some positive examples of instances in data that match codes deriving from a preconceived analytical framework is a weak form of analysis

https://science-education-research.com/research-methodology/analysis/coding/coding-in-confirmatory-research/

If instead, we believed that hesitation and contradiction were essential to good science teaching; and devised an observation schedule to spot incidents of hesitation and contradiction; and then observed the lessons of good science teachers, I suspect we could also find evidence of some level of hesitation and contradiction. So, what?

That’s one reason why quantitative, experimental studies have more kudos.

Simple forms of qualitative studies are good at identifying potentially interesting points, but not suitable for testing substantive hypotheses.



Rhetorical experiments

(a genre?)

Progressive/reform teaching: learner-centred; active learning; constructivist; enquiry-based; etc.

can be contrasted with

Traditional teaching: teacher-centred; lecture method; recipe practicals; passive learning; etc.

for example

Progressive teaching approach X has beeen widely demonstrated to be more effective than traditional teaching across a wide range of contexts

Continues to be the norm: despite research evidence and educational policies, this is how students are taught in this educational context

but

Progressive teaching approach X has not yet been specifically tested with grade […] students, studying topic […], in town […]

Yet researchers are able to prepare a local teacher to implement teaching approach X in an experimental study

so an experiment is designed

(The same teacher may be asked to teach in both conditions to give a 'fairer' test)

Experimental group - class taught with progressive teaching approach X

is compared with

Control group - class taught with traditional teaching approach: (any progressive elements excluded for the sake of a clear comparison)

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

But there is a genre of rhetorical experiments as well.

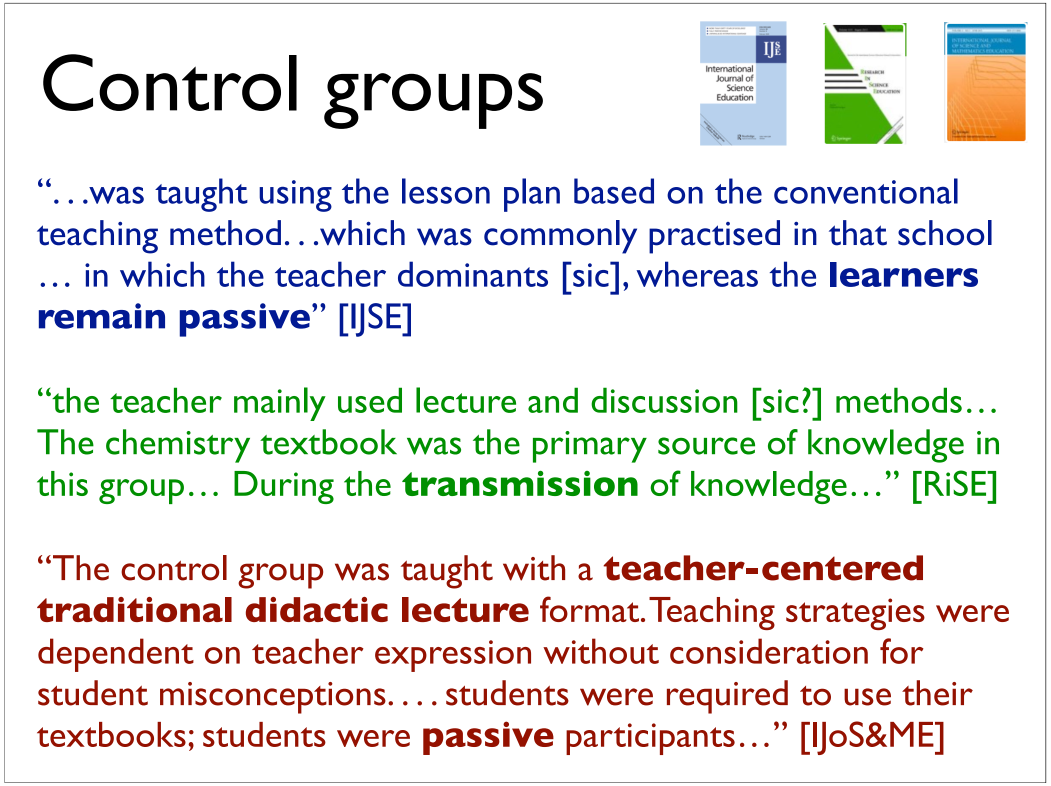
These papers usually start by making very strong claims about the merits of constructivism and some particular pedagogy associated with it. These papers will offer strong, convincing, theoretical arguments why such teaching is so much more effective than what may be labelled 'traditional' teaching (though by now I would have hoped constructivist-minded teaching would be traditional, but anyway.)

They then report a whole raft of prior studies which have shown the superiority of the focal pedagogy in a wide range of contexts - across geographical locations, age ranges, topics, languages of instruction, school cultures, etc.

But, they tell us, no one has yet tested the effectiveness of this pedagogy with, say, 13-14 year-olds studying the specific topic of the periodic table, in rural schools in South Cambridgeshire.

Of course, from everything that has been reported, there is every reason to think this pedagogy will be effective in such a context. No sense of Karl Popper's notion of scientists testing bold conjectures, here: we only test dead-cert hypotheses.

But to make absolutely sure that the experiment works, we compare our preferred pedagogy with a teacher lecturing a class. There may be some 'discussion' in the sense of the teacher answering questions, but we do not allow group-work or any real dialogue, practical work, or access to digital technologies; and we restrict written work to exercises at the lower end of Bloom's taxonomy: recall, comprehension, and some application. The teacher teaches as if any educational thinking post about 1870 never happened. That is the comparison condition for our theoretically-sound, widely-proven, pedagogy.



Control groups

"…was taught using the lesson plan based on the conventional teaching method…which was commonly practised in that school … in which the teacher dominants [sic], whereas the **learners remain passive**" [IJSE]

"the teacher mainly used lecture and discussion [sic?] methods… The chemistry textbook was the primary source of knowledge in this group… During the **transmission** of knowledge…" [RiSE]

"The control group was taught with a **teacher-centered traditional didactic lecture** format. Teaching strategies were dependent on teacher expression without consideration for student misconceptions.…students were required to use their textbooks; students were **passive** participants…" [IJoS&ME]

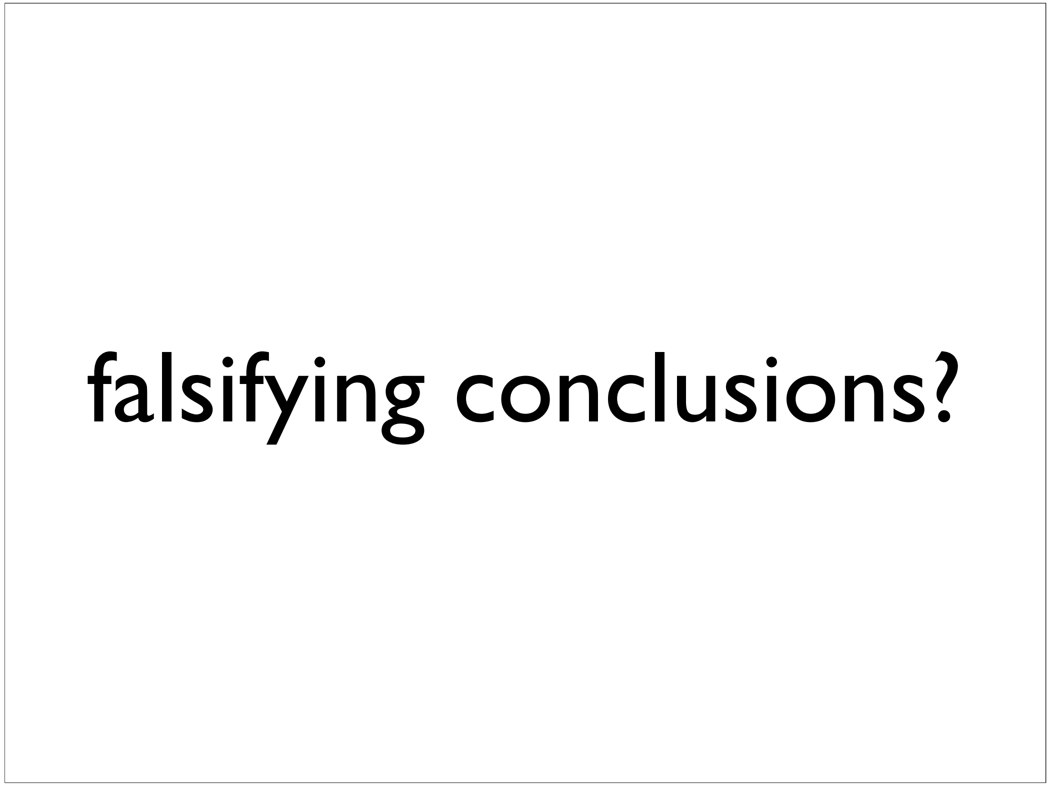
I find this utterly unethical.

It is unethical in at least two senses.

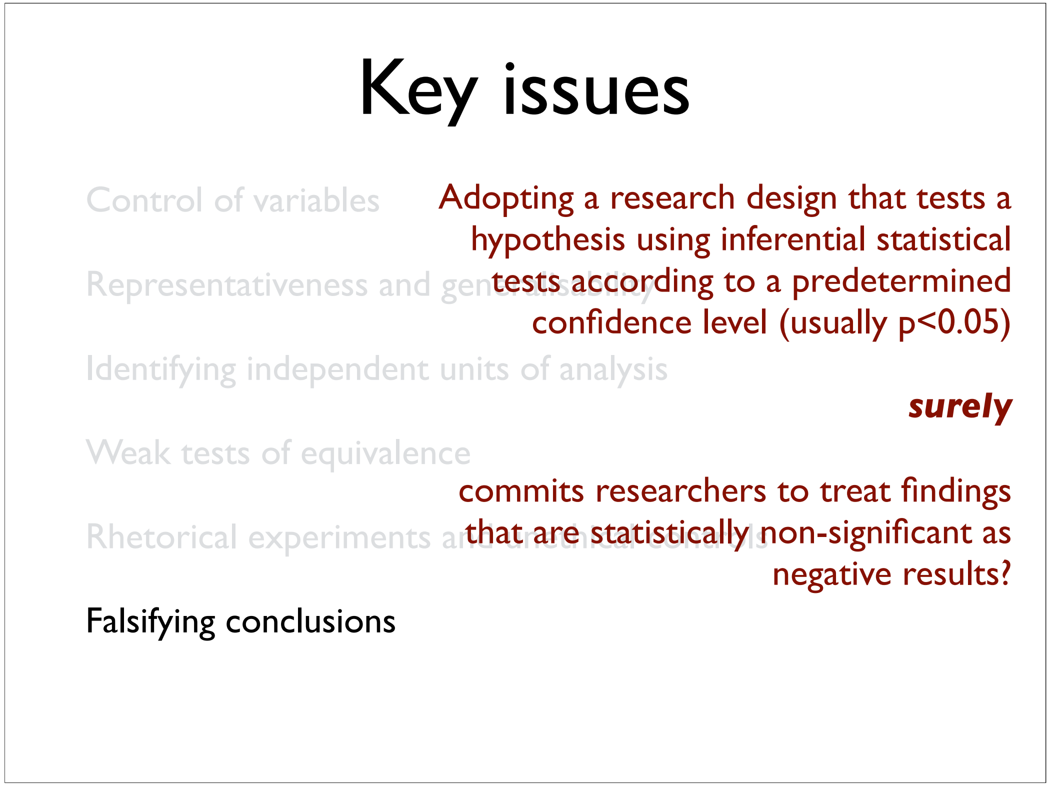
- Any research which inconveniences people for no good reason, is pointless and so a waste of resource. A demonstration posing as an experiment falls into this category.
- Asking control teachers to deliberately avoid good practice in their classroom, so to ensure a positive study outcome, is also clearly wrong.

And here I have seen quite a few of these studies, including some published in good journals.

How come reviewers do not spot that this is not genuine enquiry, and it is not an acceptable way to treat study participants? I can only assume we are all blinded by the assumed superiority of the experiment in science.



falsifying conclusions?



Key issues

Control of variables

Representativeness and generalisability

Identifying independent units of analysis

Weak tests of equivalence

Rhetorical experiments and unethical controls

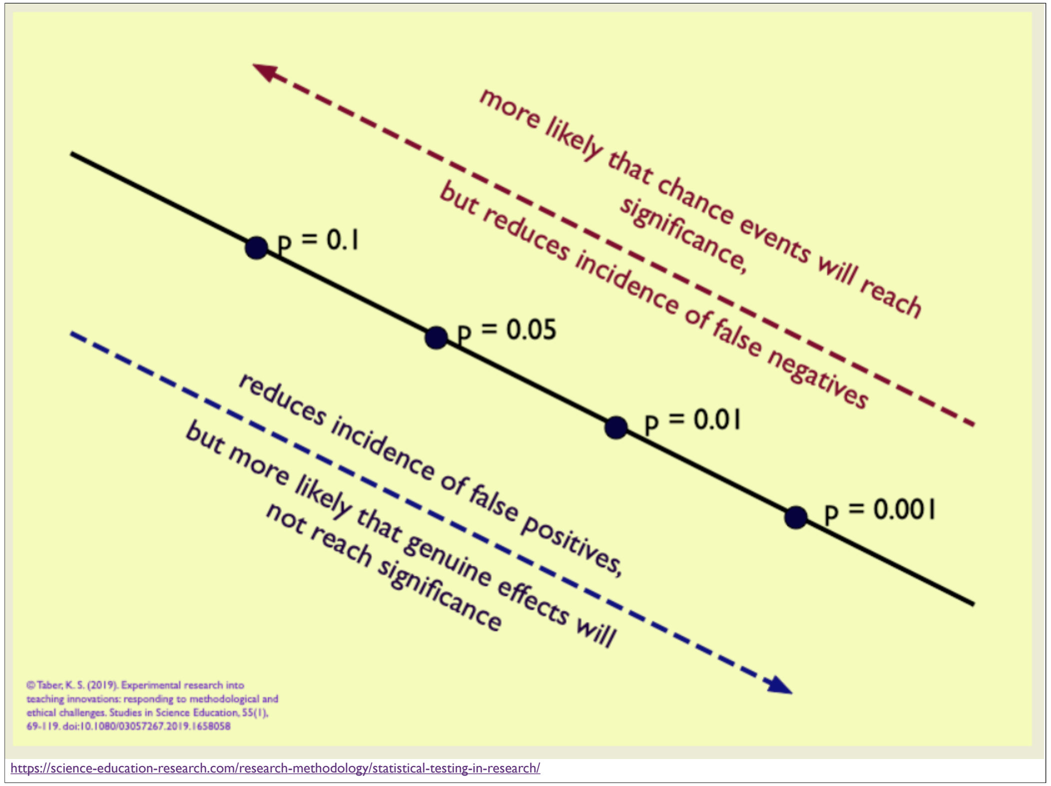
Falsifying conclusions

Adopting a research design that tests a hypothesis using inferential statistical tests according to a predetermined confidence level (usually $p<0.05$)

surely

commits researchers to treat findings that are statistically non-significant as negative results?

My final issue also reflects this idea, in that sometimes authors are so convinced about their expectations that they fail to observe the logic of their experiment.



more likely that chance events will reach significance,
but reduces incidence of false negatives

reduces incidence of false positives,
but more likely that genuine effects will not reach significance

©Taber, K. S. (2019). 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

https://science-education-research.com/research-methodology/statistical-testing-in-research/

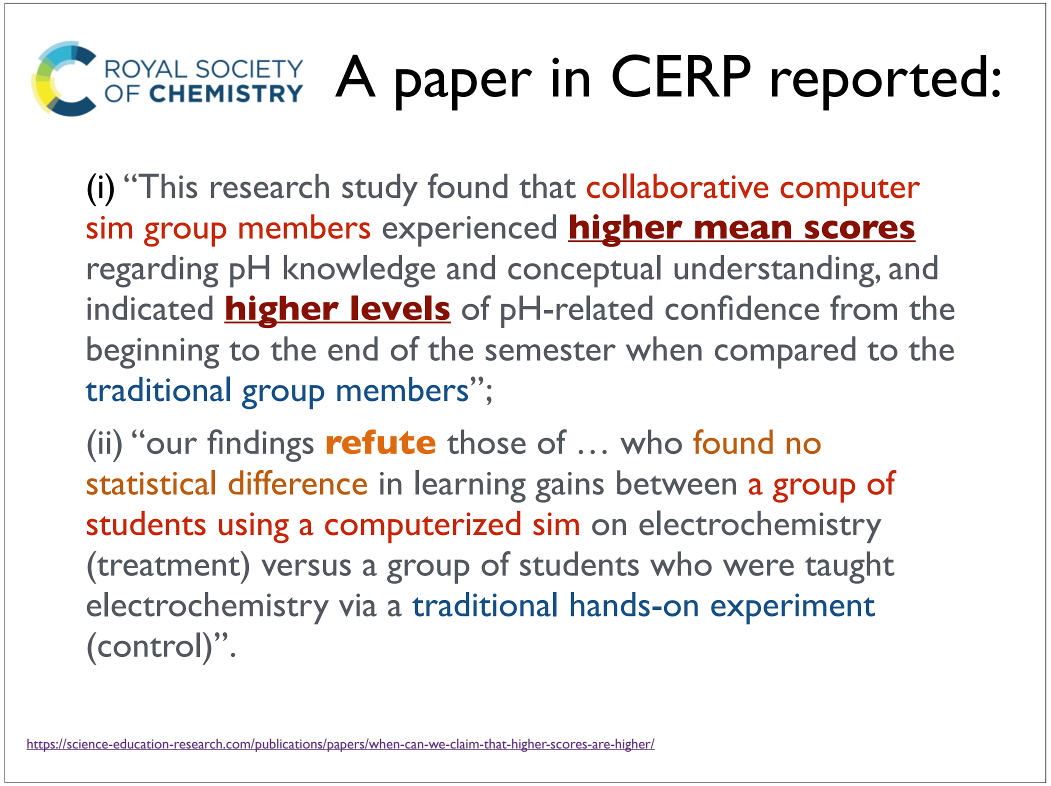
The confidence level chosen to determine whether differences between outcomes are statistically significant is arbitrary. Nearly always, we use a cut-off of 0.05: but there is nothing magical about that value. I suspect aliens with different numbers of fingers to us may have chosen a different critical value.

Whatever critical value we choose, the outcome is only an indicator. We are always subject to false positives where results are found significant, but unbeknown to us are actually due to chance effects; and false negatives, where results are found to be non-significant despite there being some kind of causal effect too small to reach significance in small samples.

But, if we are going to use inferential stats, then the conclusions of our experiments, certainly subject to appropriate caveats, should be determined by the outcomes of the analysis undertaken.

I am going to briefly refer to two papers, one published in each of the two most important chemistry education journals.

https://science-education-research.com/research-methodology/statistical-testing-in-research/



ROYAL SOCIETY OF CHEMISTRY

A paper in CERP reported:

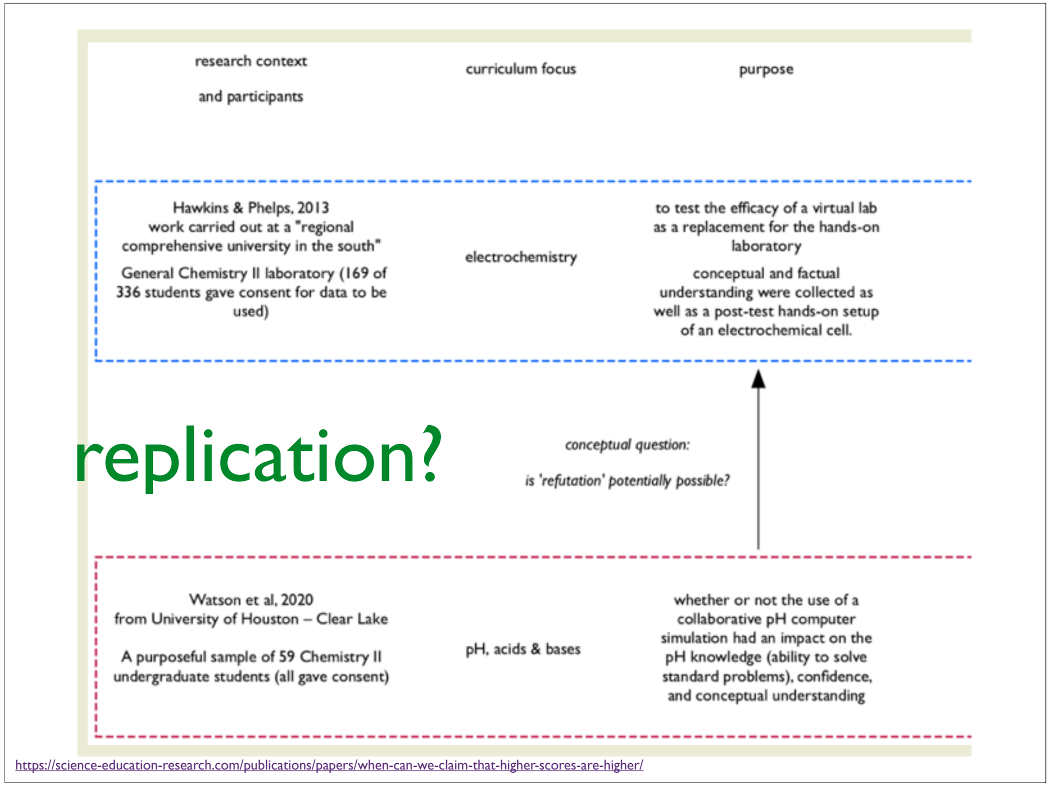
(i) “This research study found that collaborative computer sim group members experienced **higher mean scores** regarding pH knowledge and conceptual understanding, and indicated **higher levels** of pH-related confidence from the beginning to the end of the semester when compared to the traditional group members”;

(ii) “our findings **refute** those of … who found no statistical difference in learning gains between a group of students using a computerized sim on electrochemistry (treatment) versus a group of students who were taught electrochemistry via a traditional hands-on experiment (control)”.

https://science-education-research.com/publications/papers/when-can-we-claim-that-higher-scores-are-higher/

A paper in *Chemistry Education Research and Practice*, reported positive outcomes from a study into a learning approach, and made a point of suggesting this refuted a previous study which had not found a significant effect.

https://science-education-research.com/publications/papers/when-can-we-claim-that-higher-scores-are-higher



research context and participants	curriculum focus	purpose
Hawkins & Phelps, 2013 work carried out at a "regional comprehensive university in the south" General Chemistry II laboratory (169 of 336 students gave consent for data to be used)	electrochemistry	to test the efficacy of a virtual lab as a replacement for the hands-on laboratory conceptual and factual understanding were collected as well as a post-test hands-on setup of an electrochemical cell.

replication?

conceptual question:

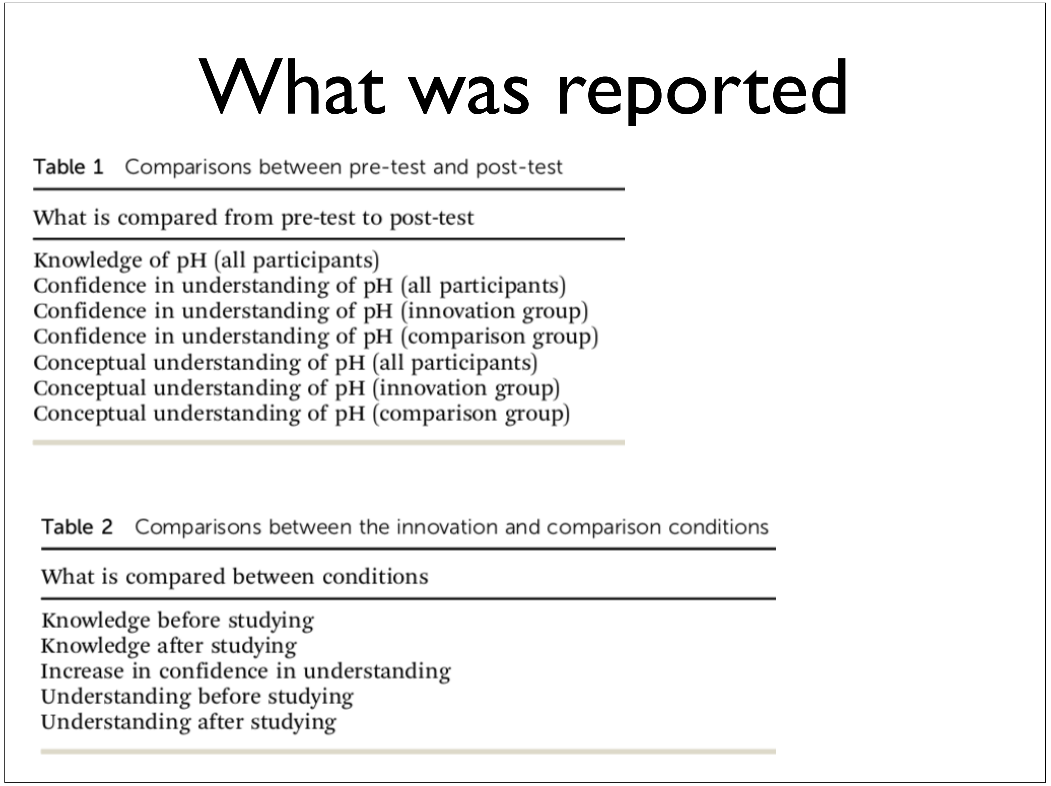
is 'refutation' potentially possible?

Watson et al, 2020 from University of Houston – Clear Lake A purposeful sample of 59 Chemistry II undergraduate students (all gave consent)	pH, acids & bases	whether or not the use of a collaborative pH computer simulation had an impact on the pH knowledge (ability to solve standard problems), confidence, and conceptual understanding

https://science-education-research.com/publications/papers/when-can-we-claim-that-higher-scores-are-higher/

The whole question of replication is worthy of a talk of its own, but I am merely going to comment here that it was questionable whether the two studies were similar enough for one to be able to refute the other.

https://science-education-research.com/publications/papers/when-can-we-claim-that-higher-scores-are-higher/



What was reported

Table 1 Comparisons between pre-test and post-test

What is compared from pre-test to post-test
Knowledge of pH (all participants)
Confidence in understanding of pH (all participants)
Confidence in understanding of pH (innovation group)
Confidence in understanding of pH (comparison group)
Conceptual understanding of pH (all participants)
Conceptual understanding of pH (innovation group)
Conceptual understanding of pH (comparison group)

Table 2 Comparisons between the innovation and comparison conditions

What is compared between conditions
Knowledge before studying
Knowledge after studying
Increase in confidence in understanding
Understanding before studying
Understanding after studying

The study presented a range of results, comparing pre-test to post-test, and between the experimental and comparison groups.

Results

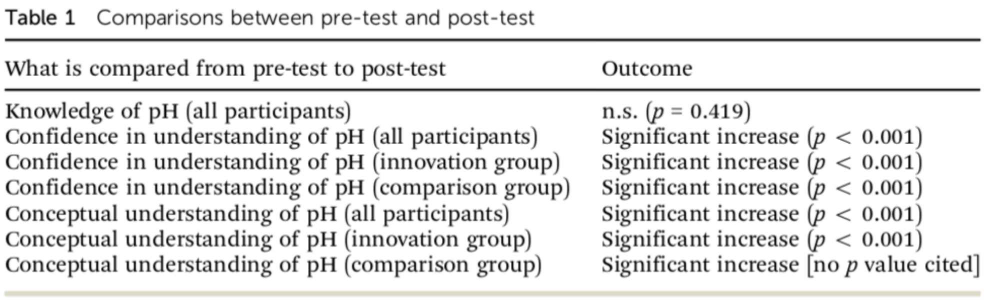


Table 1 Comparisons between pre-test and post-test

What is compared from pre-test to post-test	Outcome
Knowledge of pH (all participants)	n.s. ($p = 0.419$)
Confidence in understanding of pH (all participants)	Significant increase ($p < 0.001$)
Confidence in understanding of pH (innovation group)	Significant increase ($p < 0.001$)
Confidence in understanding of pH (comparison group)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (all participants)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (innovation group)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (comparison group)	Significant increase [no *p* value cited]

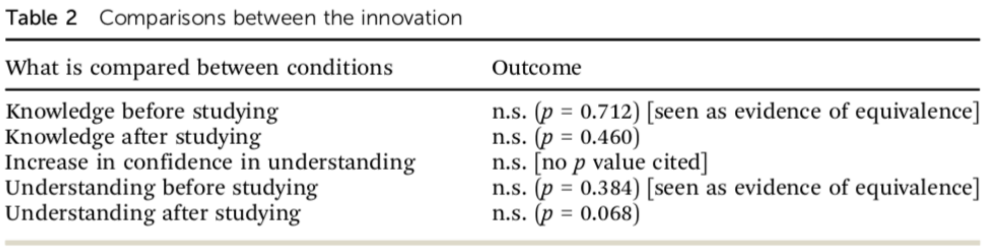


Table 2 Comparisons between the innovation

What is compared between conditions	Outcome
Knowledge before studying	n.s. ($p = 0.712$) [seen as evidence of equivalence]
Knowledge after studying	n.s. ($p = 0.460$)
Increase in confidence in understanding	n.s. [no *p* value cited]
Understanding before studying	n.s. ($p = 0.384$) [seen as evidence of equivalence]
Understanding after studying	n.s. ($p = 0.068$)

Results

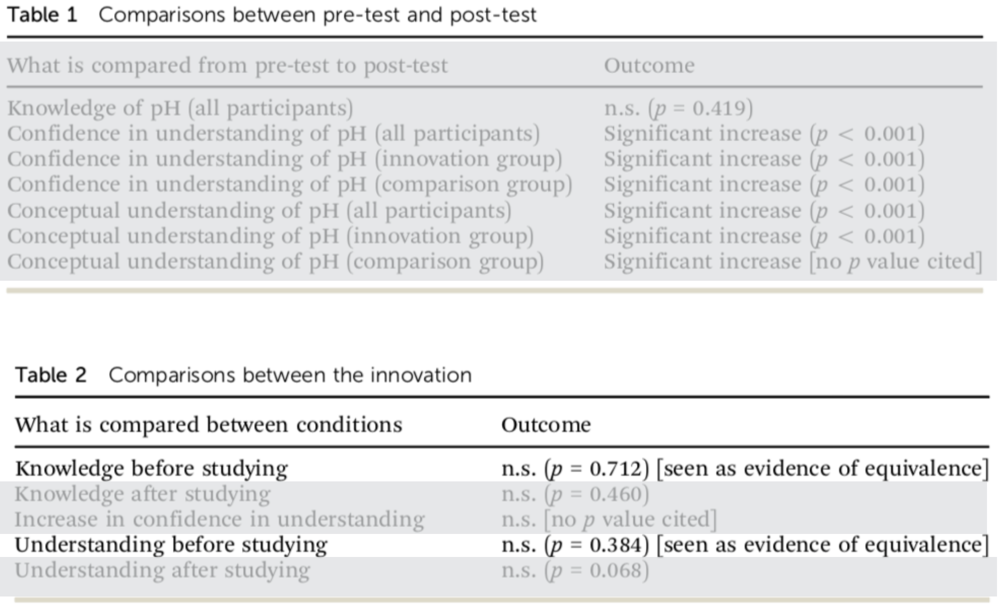


Table 1 Comparisons between pre-test and post-test

What is compared from pre-test to post-test	Outcome
Knowledge of pH (all participants)	n.s. (p = 0.419)
Confidence in understanding of pH (all participants)	Significant increase ($p < 0.001$)
Confidence in understanding of pH (innovation group)	Significant increase ($p < 0.001$)
Confidence in understanding of pH (comparison group)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (all participants)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (innovation group)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (comparison group)	Significant increase [no *p* value cited]

Table 2 Comparisons between the innovation

What is compared between conditions	Outcome
Knowledge before studying	n.s. (p = 0.712) [seen as evidence of equivalence]
Knowledge after studying	n.s. (p = 0.460)
Increase in confidence in understanding	n.s. [no *p* value cited]
Understanding before studying	n.s. (p = 0.384) [seen as evidence of equivalence]
Understanding after studying	n.s. (p = 0.068)

The two groups were considered equivalent at pretest, for the familiar reason that the differences were not so different as to reach statistical significance. My interpretation of a p value of 0.384 is that the differences between the two groups probably were NOT due to chance effects, but I think I have already said enough about that.

Results

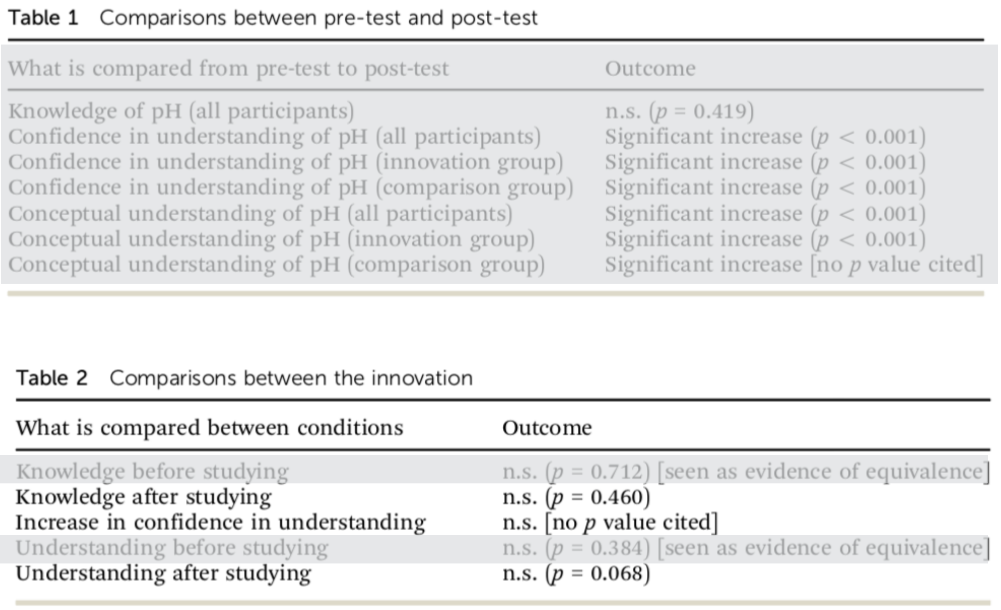


Table 1 Comparisons between pre-test and post-test

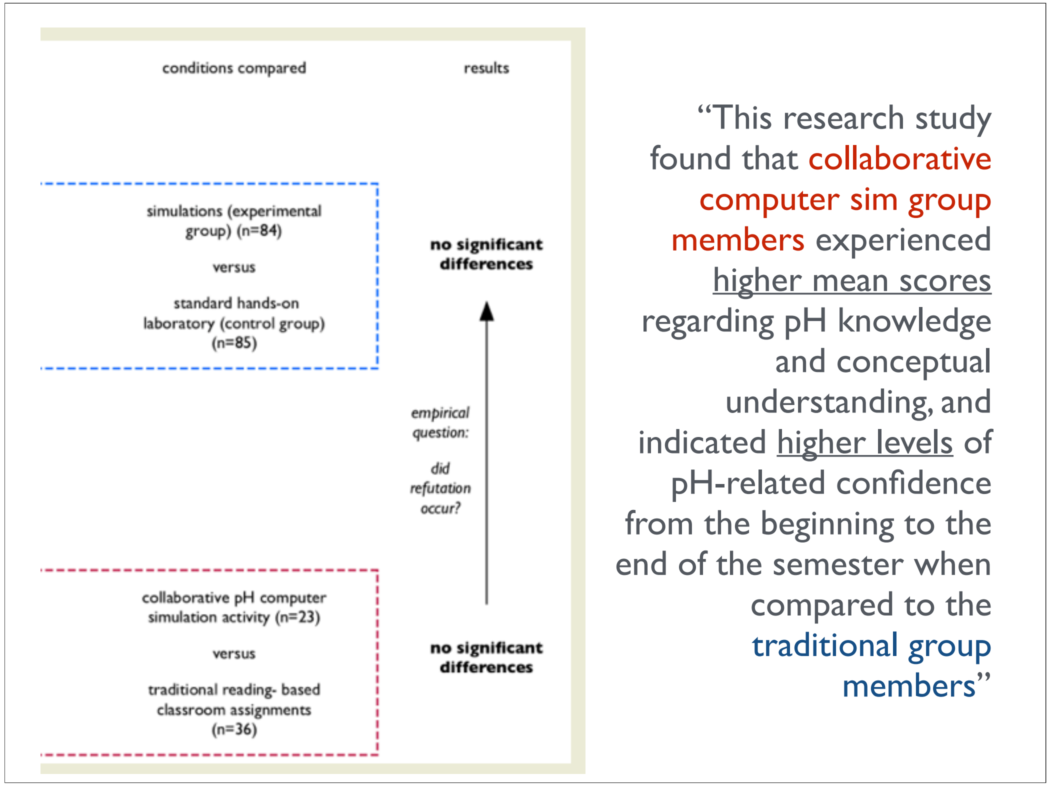
What is compared from pre-test to post-test	Outcome
Knowledge of pH (all participants)	n.s. ($p = 0.419$)
Confidence in understanding of pH (all participants)	Significant increase ($p < 0.001$)
Confidence in understanding of pH (innovation group)	Significant increase ($p < 0.001$)
Confidence in understanding of pH (comparison group)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (all participants)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (innovation group)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (comparison group)	Significant increase [no p value cited]

Table 2 Comparisons between the innovation

What is compared between conditions	Outcome
Knowledge before studying	n.s. ($p = 0.712$) [seen as evidence of equivalence]
Knowledge after studying	n.s. ($p = 0.460$)
Increase in confidence in understanding	n.s. [no p value cited]
Understanding before studying	n.s. ($p = 0.384$) [seen as evidence of equivalence]
Understanding after studying	n.s. ($p = 0.068$)

But surely that is not really important in this study, as there was not a significant difference between the two groups after the intervention. Any differences in scores between the two groups were not sufficiently unlikely so as be considered statistically significant.

So, in terms of the experimental design set out before collecting data, this is a negative result.



conditions compared

results

simulations (experimental group) (n=84)

versus

standard hands-on laboratory (control group) (n=85)

no significant differences

empirical question:

did refutation occur?

collaborative pH computer simulation activity (n=23)

versus

traditional reading- based classroom assignments (n=36)

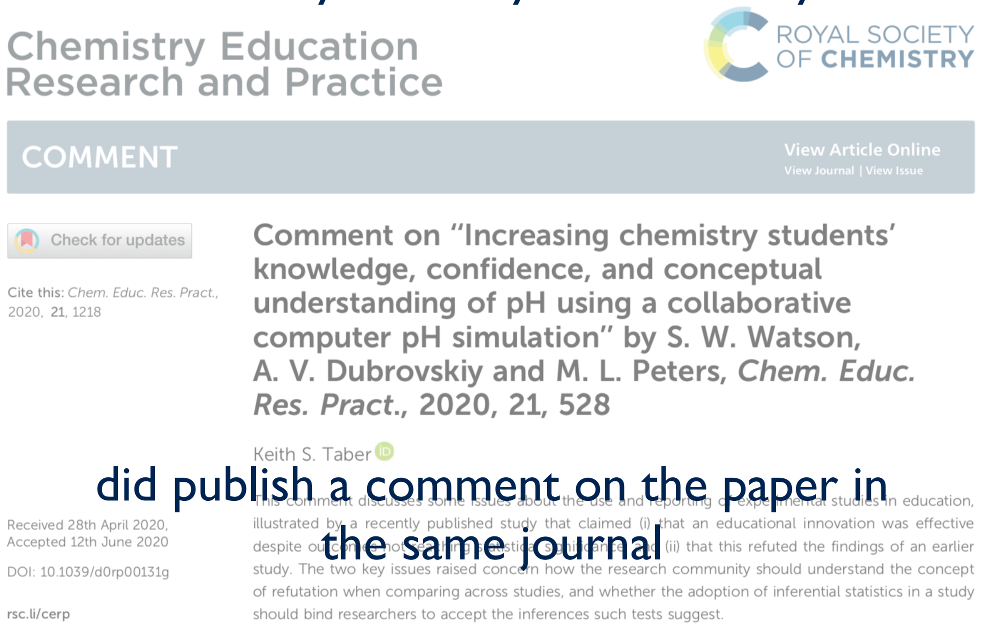
no significant differences

“This research study found that collaborative computer sim group members experienced higher mean scores regarding pH knowledge and conceptual understanding, and indicated higher levels of pH-related confidence from the beginning to the end of the semester when compared to the traditional group members”

There is certainly nothing wrong with reporting a negative result, and indeed there is strong belief that research literature is distorted by a bias towards authors submitting, and journals preferentially publishing, positive outcomes.

But, these authors are claiming to refute a study that did not find significant differences by having carried out another study that ALSO did not find significant differences. Here the negative result is imply ignored when presenting the study conclusions as being positive outcomes.

The Royal Society of Chemistry



Chemistry Education Research and Practice

ROYAL SOCIETY OF CHEMISTRY

COMMENT

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Educ. Res. Pract.*, 2020, **21**, 1218

Comment on "Increasing chemistry students' knowledge, confidence, and conceptual understanding of pH using a collaborative computer pH simulation" by S. W. Watson, A. V. Dubrovskiy and M. L. Peters, *Chem. Educ. Res. Pract.*, 2020, 21, 528

Keith S. Taber

Received 28th April 2020,
Accepted 12th June 2020

DOI: 10.1039/d0rp00131g

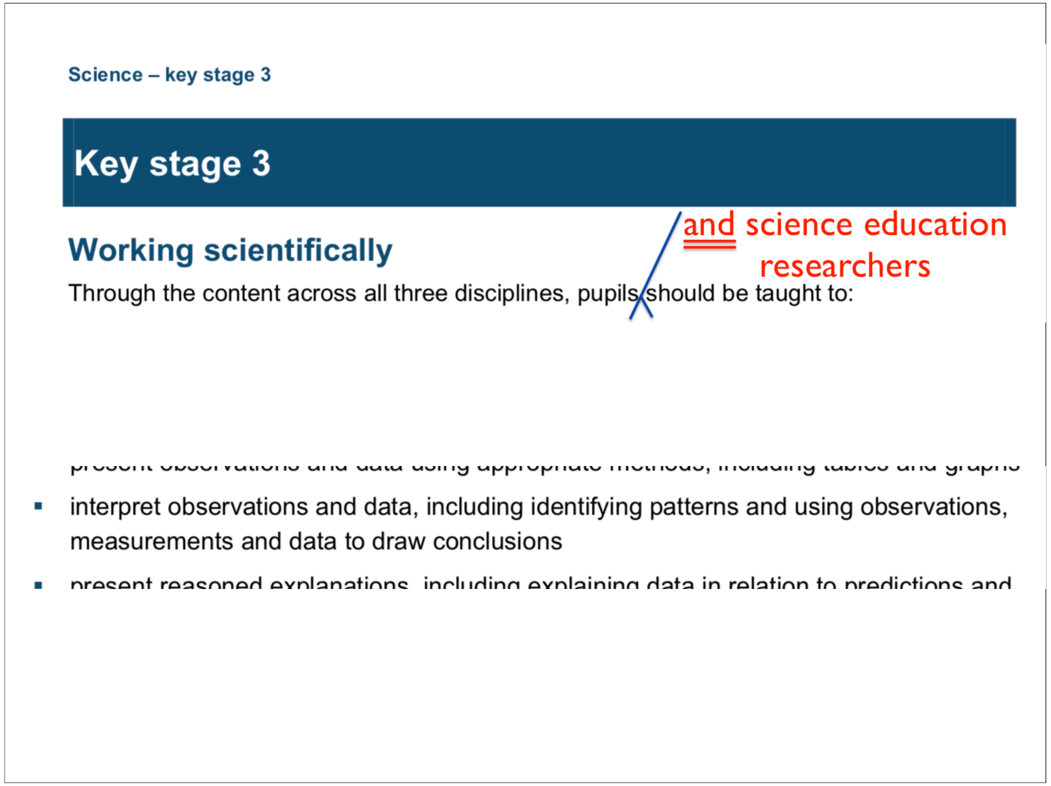
rsc.li/cerp

This comment discusses some issues about the use and reporting of experimental studies in education, illustrated by a recently published study that claimed (i) that an educational innovation was effective despite outcomes not reaching statistical significance and (ii) that this refuted the findings of an earlier study. The two key issues raised concern how the research community should understand the concept of refutation when comparing across studies, and whether the adoption of inferential statistics in a study should bind researchers to accept the inferences such tests suggest.

did publish a comment on the paper in the same journal

Sadly, having noticed this, I felt it necessary to write up a comment - which to be fair to the journal, was published.

[Taber, K. S. (2020). Comment on "Increasing chemistry students' knowledge, confidence, and conceptual understanding of pH using a collaborative computer pH simulation" by S.W. Watson, A.V. Dubrovskiy and M. L. Peters, Chem. Educ. Res. Pract., 2020, 21, 528. Chemistry Education Research and Practice. doi:10.1039/D0RP00131G]



Science – key stage 3

Key stage 3

Working scientifically

and science education
researchers

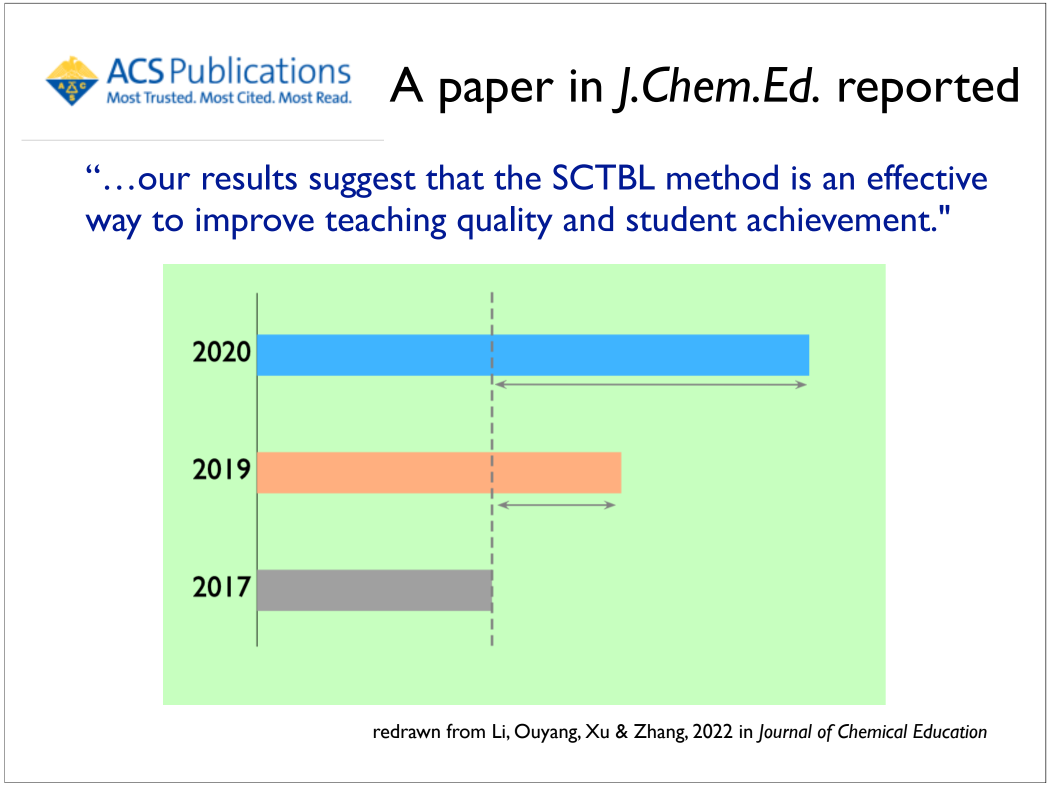
Through the content across all three disciplines, pupils should be taught to:

present observations and data using appropriate methods, including tables and graphs

- interpret observations and data, including identifying patterns and using observations, measurements and data to draw conclusions
- present reasoned explanations, including explaining data in relation to predictions and

Again, we would not accept this kind of sloppy work from school children.

For my final example I turn to the other top chemistry education specific journal, the Journal of Chemical Education published by the American Chemical Society, whose journals are, according to the American Chemical Society itself, at least, 'most trusted'.



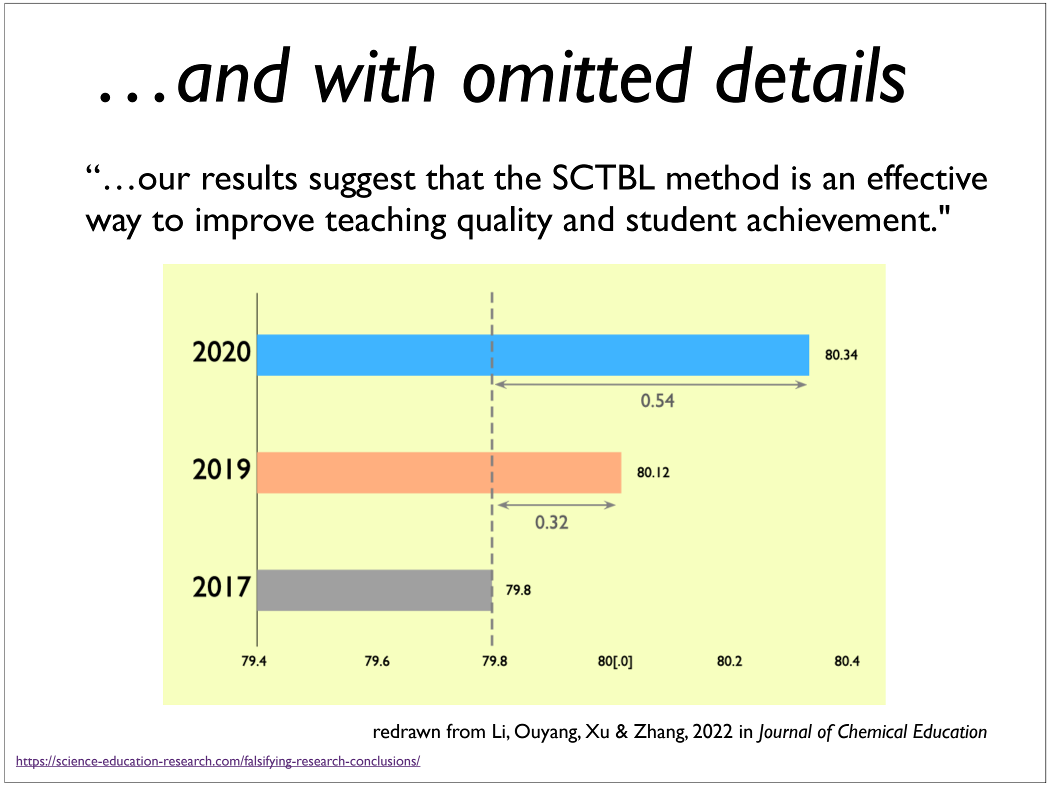
A paper in *J.Chem.Ed.* reported

"…our results suggest that the SCTBL method is an effective way to improve teaching quality and student achievement."

redrawn from Li, Ouyang, Xu & Zhang, 2022 in *Journal of Chemical Education*

This is based on a figure showing results reported in a study in the *Journal of Chemical Education*. After 2017, the researchers modified their medicinal chemistry course, by implementing what they called 'Student-Centred Team-Based Learning Teaching Method', and as you can see student results improved thereafter.

You will notice I have failed to include any numbers on this figure, which is disingenuous of me, because the original did include the numbers.



…and with omitted details

"…our results suggest that the SCTBL method is an effective way to improve teaching quality and student achievement."

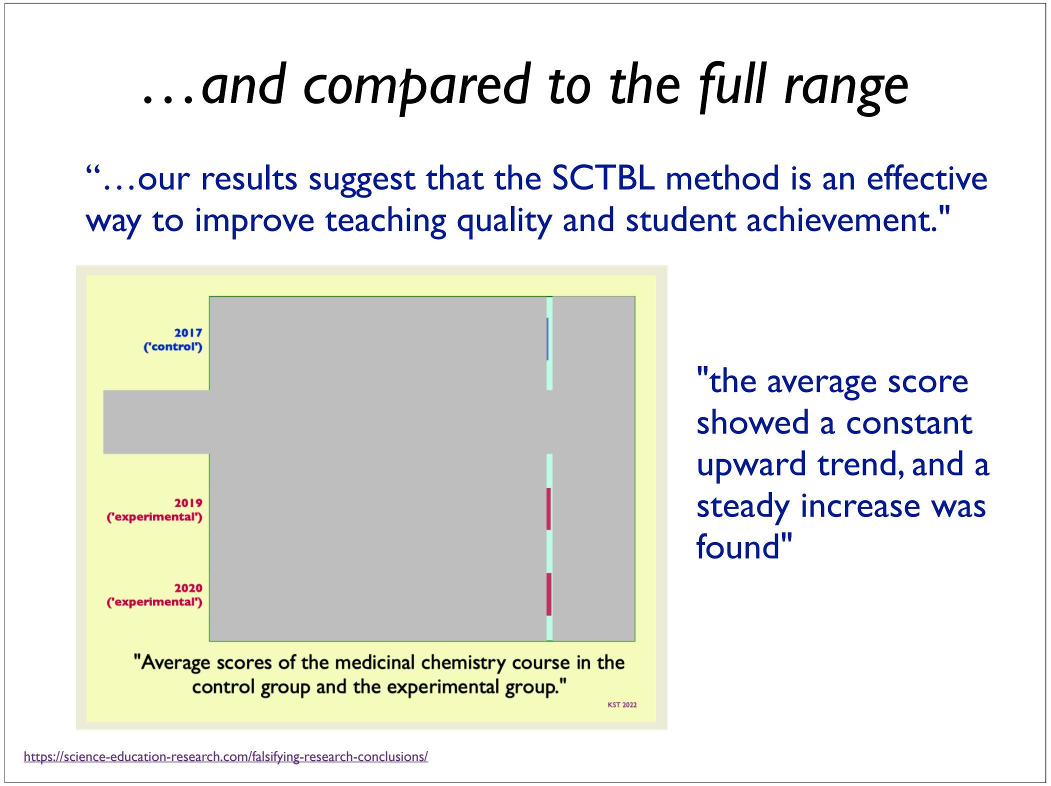
redrawn from Li, Ouyang, Xu & Zhang, 2022 in *Journal of Chemical Education*

https://science-education-research.com/falsifying-research-conclusions/

And with the numbers we see that we are viewing a graph with a truncated axis.

That is a perfectly valid technique used to emphasise differences, but I wondered if here they might have over-emphasised the differences?

https://science-education-research.com/falsifying-research-conclusions/



…and compared to the full range

"…our results suggest that the SCTBL method is an effective way to improve teaching quality and student achievement."

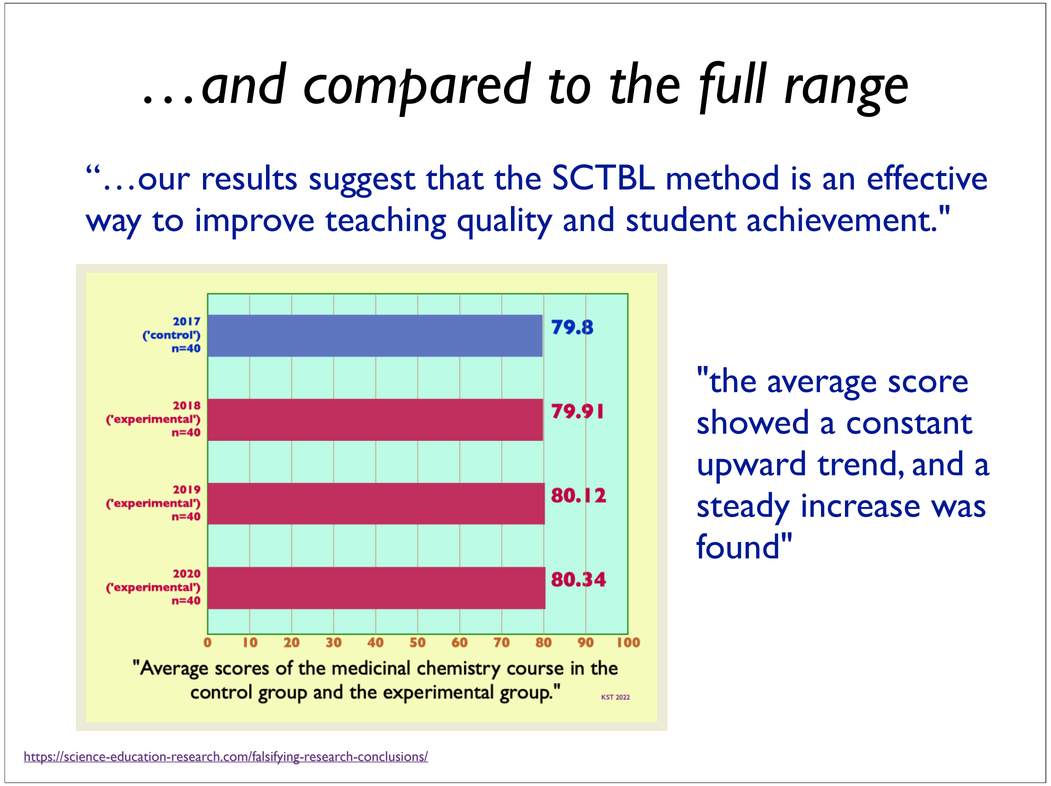
"Average scores of the medicinal chemistry course in the control group and the experimental group."

"the average score showed a constant upward trend, and a steady increase was found"

https://science-education-research.com/falsifying-research-conclusions/

They are just focusing on a small range of values.

https://science-education-research.com/falsifying-research-conclusions/



…and compared to the full range

“…our results suggest that the SCTBL method is an effective way to improve teaching quality and student achievement."

"Average scores of the medicinal chemistry course in the control group and the experimental group."

"the average score showed a constant upward trend, and a steady increase was found"

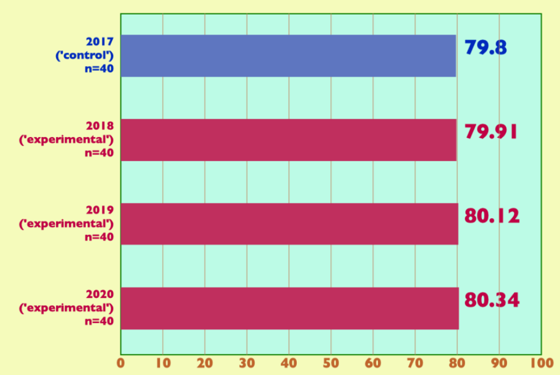
https://science-education-research.com/falsifying-research-conclusions/

And if I present the whole graph as it might have been drawn, the change seems less impressive.

https://science-education-research.com/falsifying-research-conclusions/

…and compared to the full range

"the average score showed a constant upward trend, and a steady increase was found"



"Average scores of the medicinal chemistry course in the control group and the experimental group."

KST 2022

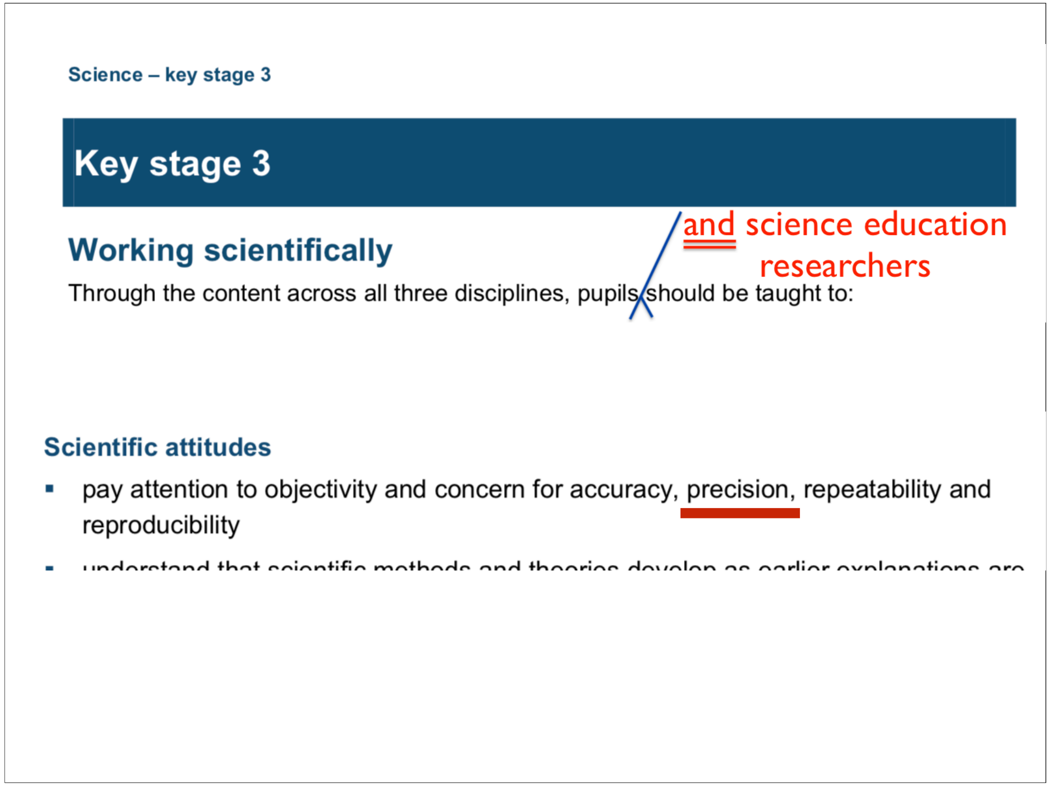
What might we suggest to our school age students about precisions of measurement?

https://science-education-research.com/falsifying-research-conclusions/

Perhaps some of you are used to marking student work, and perhaps you feel your marking is entirely objective and very precise.

But, given how cohorts shift from year to year, I really do not think average course scores to two places of decimals are meaningful.

https://science-education-research.com/falsifying-research-conclusions/



Science – key stage 3

Key stage 3

Working scientifically

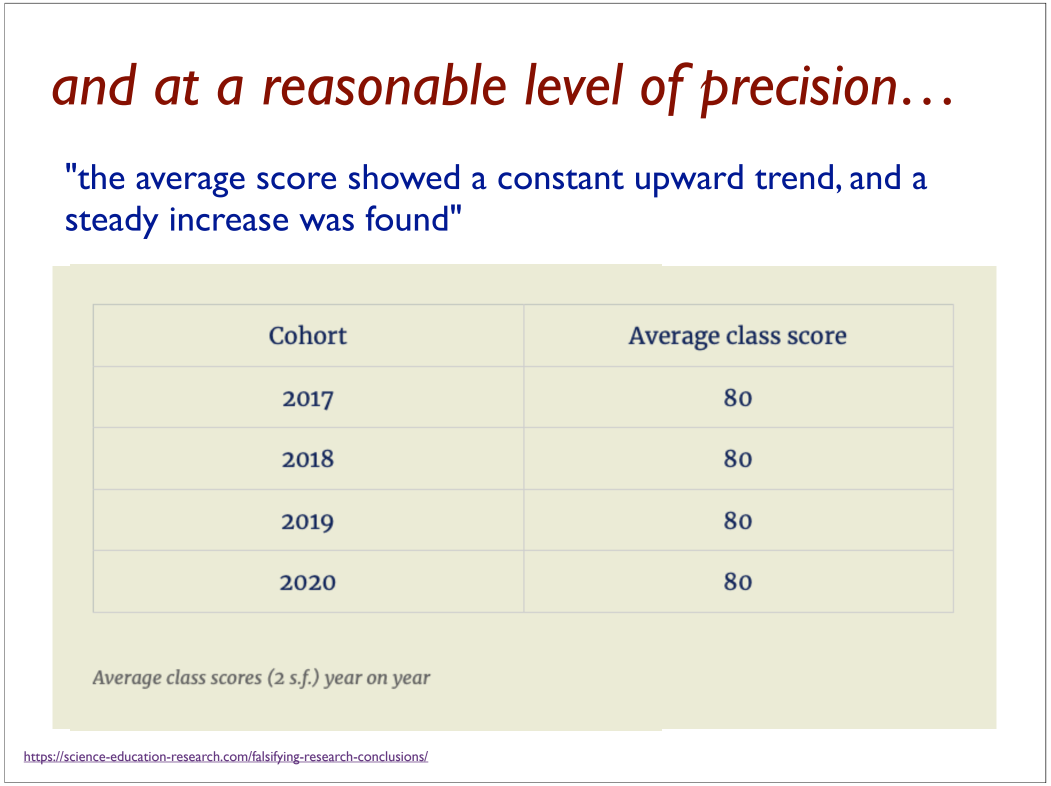
and science education researchers

Through the content across all three disciplines, pupils should be taught to:

Scientific attitudes

- pay attention to objectivity and concern for accuracy, precision, repeatability and reproducibility
- understand that scientific methods and theories develop as earlier explanations are

We would not accept this from school children.



and at a reasonable level of precision…

"the average score showed a constant upward trend, and a steady increase was found"

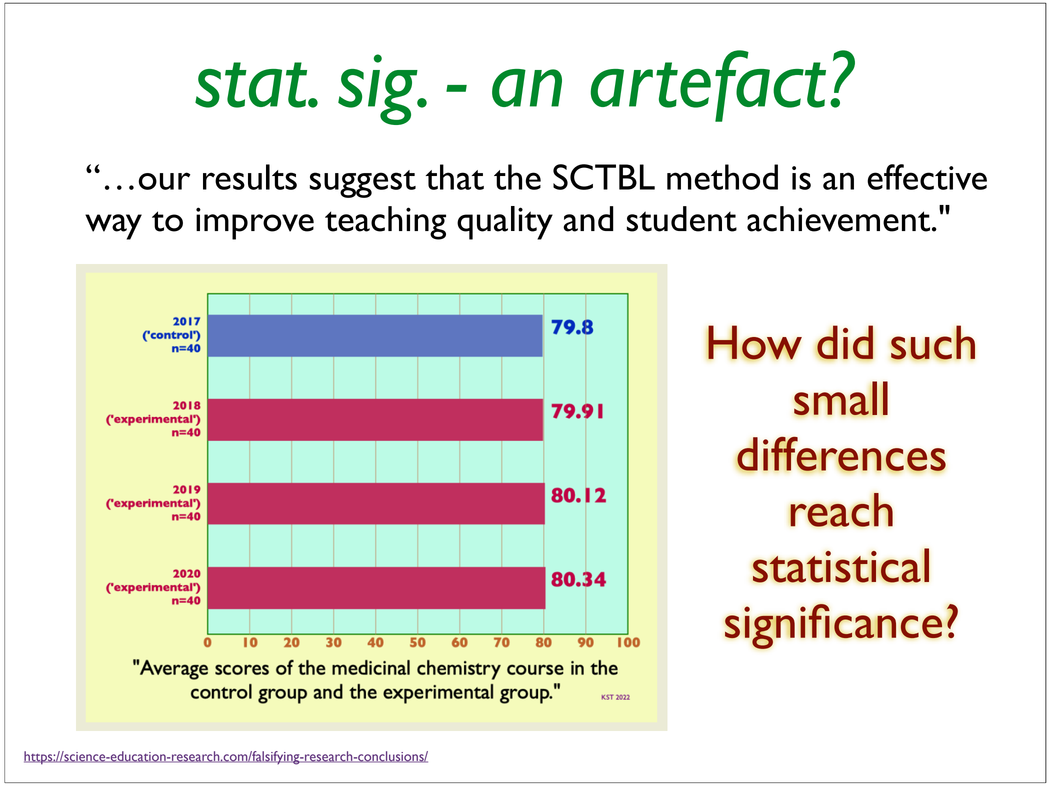
Cohort	Average class score
2017	80
2018	80
2019	80
2020	80

Average class scores (2 s.f.) year on year

https://science-education-research.com/falsifying-research-conclusions/

So, I have reworked their results.

https://science-education-research.com/falsifying-research-conclusions/



stat. sig. - an artefact?

"…our results suggest that the SCTBL method is an effective way to improve teaching quality and student achievement."

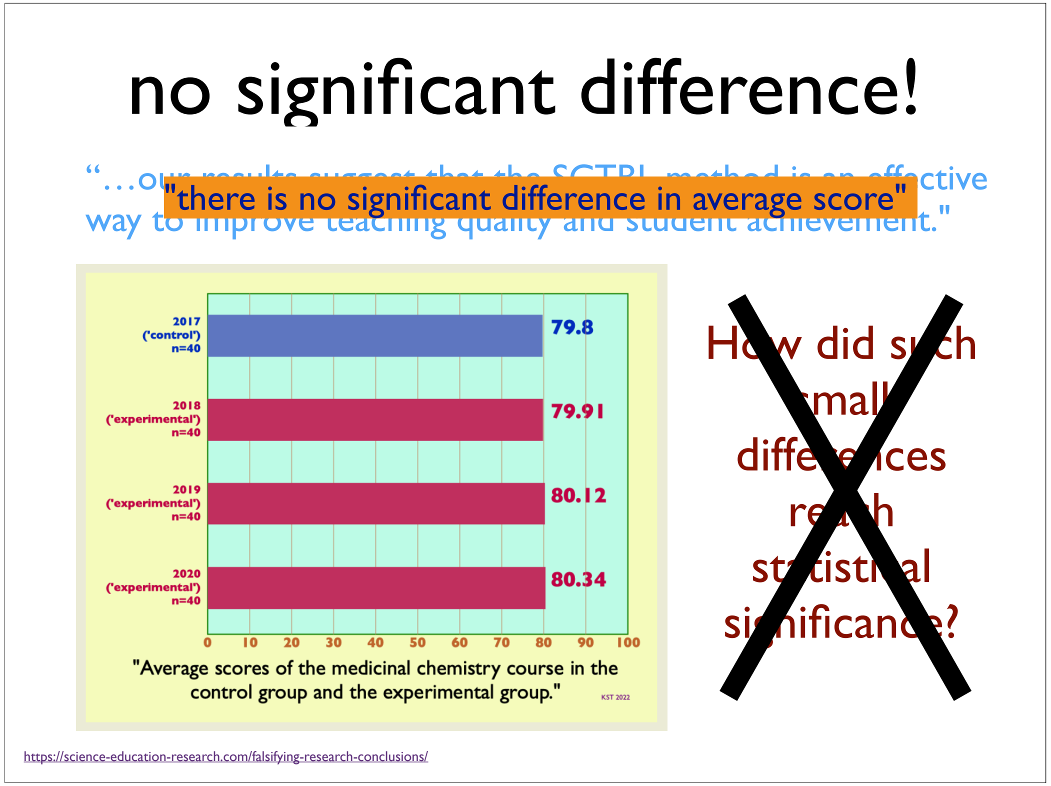
"Average scores of the medicinal chemistry course in the control group and the experimental group."

How did such small differences reach statistical significance?

https://science-education-research.com/falsifying-research-conclusions/

But even if I am being too fussy, and you think this level of precision can be justified, we might wonder how such small differences led to a statistically significant outcome.

https://science-education-research.com/falsifying-research-conclusions/



no significant difference!

“…our results suggest that the SCTBL method is an effective way to improve teaching quality and student achievement."

"there is no significant difference in average score"

"Average scores of the medicinal chemistry course in the control group and the experimental group."

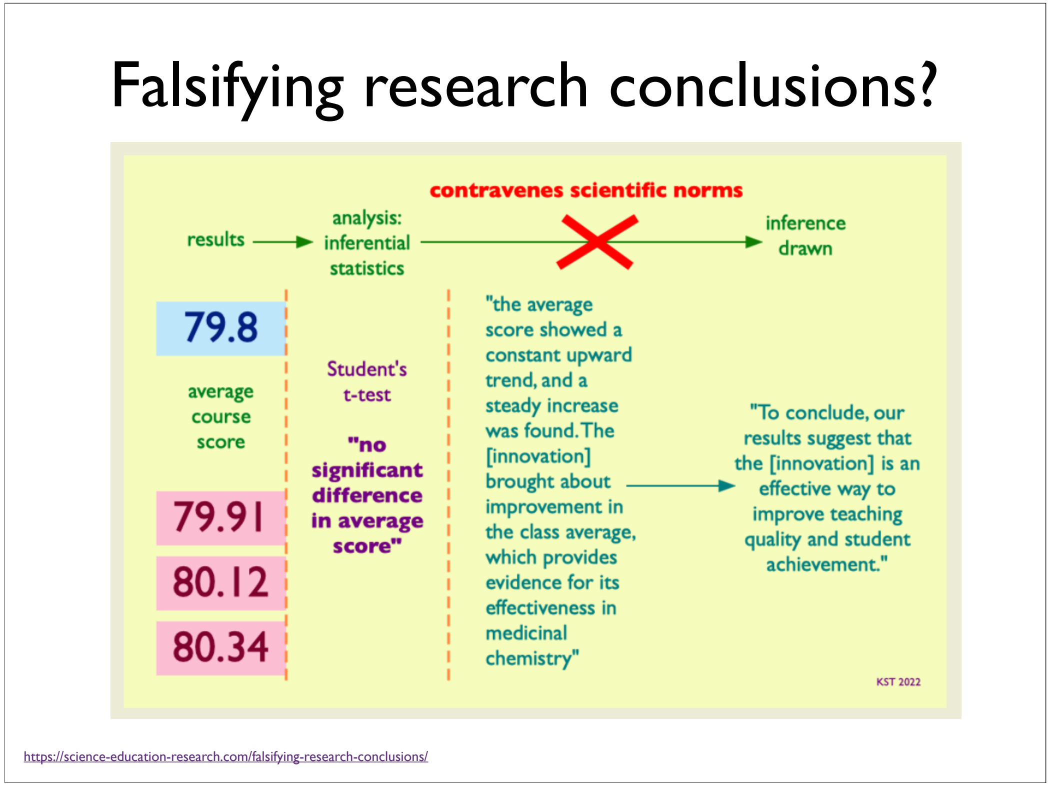
~~How did such small differences reach statistical significance?~~

https://science-education-research.com/falsifying-research-conclusions/

Of course, they did not.

The authors did the analysis, and reported there was no significant difference between the scores before and after the new approach was implemented.

https://science-education-research.com/falsifying-research-conclusions/



Falsifying research conclusions?

contravenes scientific norms

results → analysis: inferential statistics → inference drawn

79.8

average course score

79.91

80.12

80.34

Student's t-test

"no significant difference in average score"

"the average score showed a constant upward trend, and a steady increase was found. The [innovation] brought about improvement in the class average, which provides evidence for its effectiveness in medicinal chemistry" → "To conclude, our results suggest that the [innovation] is an effective way to improve teaching quality and student achievement."

KST 2022

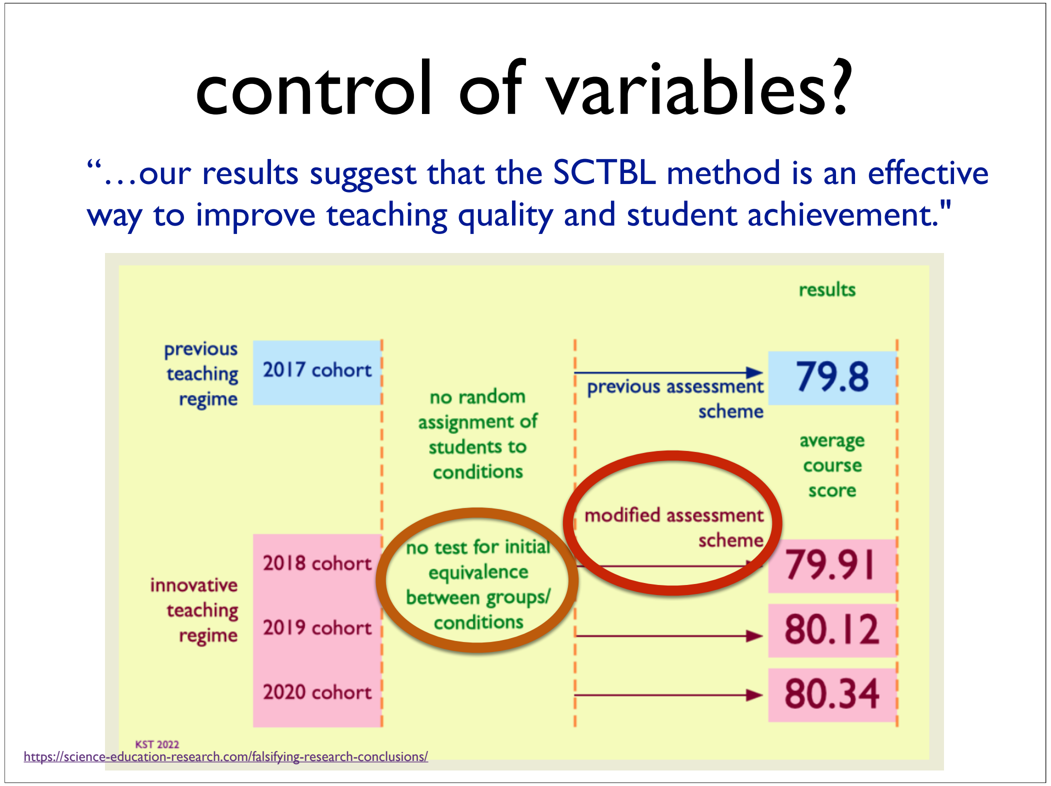
https://science-education-research.com/falsifying-research-conclusions/

Yet, these authors felt they could ignore this analysis and reach a positive conclusion.

Presumably the peer reviewers, and the editor, thought this was fine.

I think this goes completely against proper scientific practice.

https://science-education-research.com/falsifying-research-conclusions/



control of variables?

“…our results suggest that the SCTBL method is an effective way to improve teaching quality and student achievement."

results

previous teaching regime

2017 cohort

no random assignment of students to conditions

previous assessment scheme

79.8

average course score

modified assessment scheme

2018 cohort

no test for initial equivalence between groups/ conditions

79.91

innovative teaching regime

2019 cohort

80.12

2020 cohort

80.34

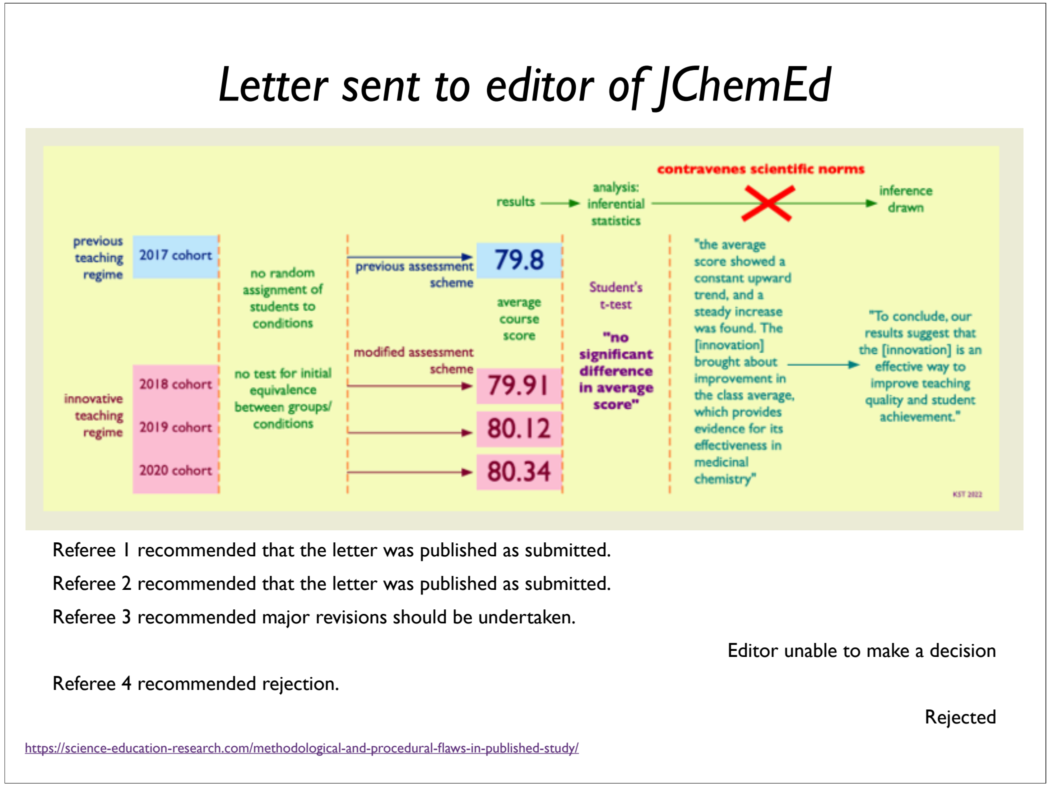
KST 2022

https://science-education-research.com/falsifying-research-conclusions/

This is leaving aside such a comparison only makes sense if we think subsequent cohorts can be considered equivalent. The authors did not even use a weak test of equivalence of the kind I have discussed earlier, because they did not use any test of equivalence. That just assumed that the 40 students admitted to the course each year could be treated as equivalent.

It was also clear from details in the paper that there had been some necessary modification in the assessment process in moving to the new teaching approach. This may have only effected a minor component of the final score, but that seems relevant when they think they are measuring to a hundredth of a percentage point.

I just cannot see how peer reviewers thought this was okay.



Letter sent to editor of JChemEd

contravenes scientific norms
results
analysis: inferential statistics
inference drawn
previous teaching regime
2017 cohort
no random assignment of students to conditions
previous assessment scheme
79.8
average course score
Student's t-test
"no significant difference in average score"
"the average score showed a constant upward trend, and a steady increase was found. The [innovation] brought about improvement in the class average, which provides evidence for its effectiveness in medicinal chemistry"
"To conclude, our results suggest that the [innovation] is an effective way to improve teaching quality and student achievement."
modified assessment scheme
innovative teaching regime
2018 cohort
2019 cohort
2020 cohort
no test for initial equivalence between groups/ conditions
79.91
80.12
80.34
KST 2022

Referee 1 recommended that the letter was published as submitted.

Referee 2 recommended that the letter was published as submitted.

Referee 3 recommended major revisions should be undertaken.

Editor unable to make a decision

Referee 4 recommended rejection.

Rejected

https://science-education-research.com/methodological-and-procedural-flaws-in-published-study/

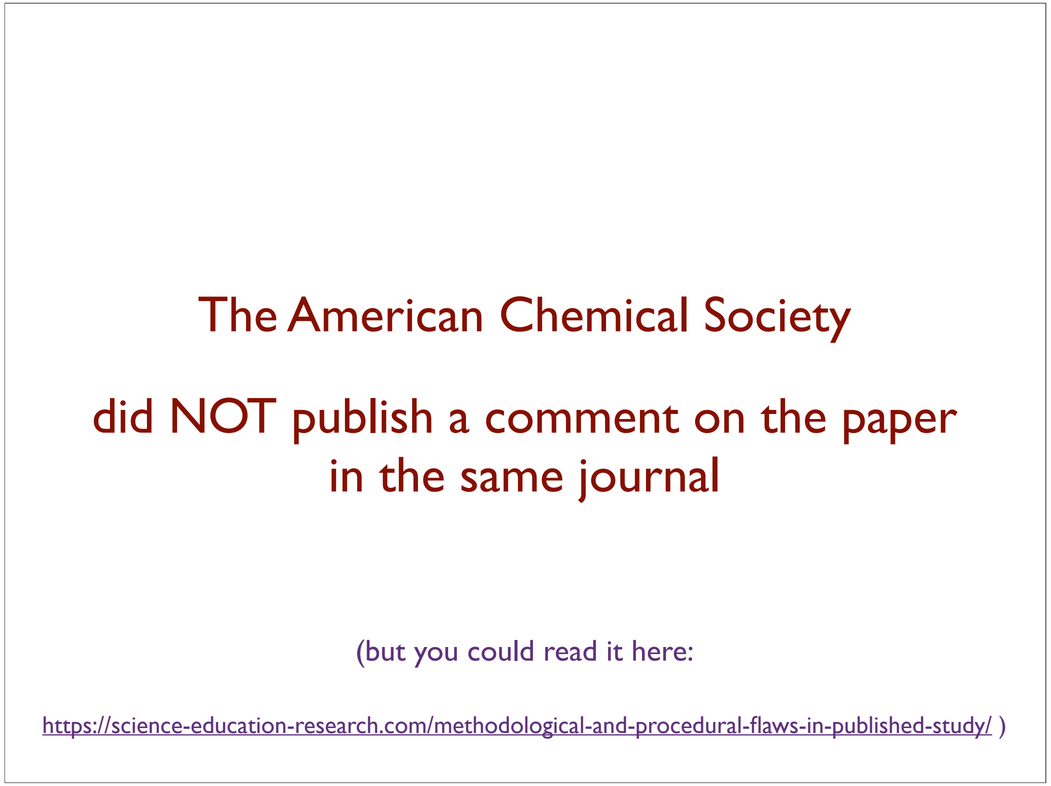
So, I wrote to the editor about it. I was told my comments could be considered for publication if I submitted them as a formal submission. So, I did.

Apparently, the editor initially asked three reviewers to read my submission. Two thought it should be published. One thought some changes were needed before it could be published. Now, I have been a journal editor, so I am pretty clear how I would respond to that review profile as an editor. But this editor was unsure.

The editor then asked a fourth person, who thought the comment should be rejected. And so it was.

I was told I could prepare a resubmission, but that if I did so it would be better to focus on general issues and not the specific paper I wanted to critique. That, of course, would have been a completely different article. I must admit to having been shocked by this outcome.

https://science-education-research.com/falsifying-research-conclusions/



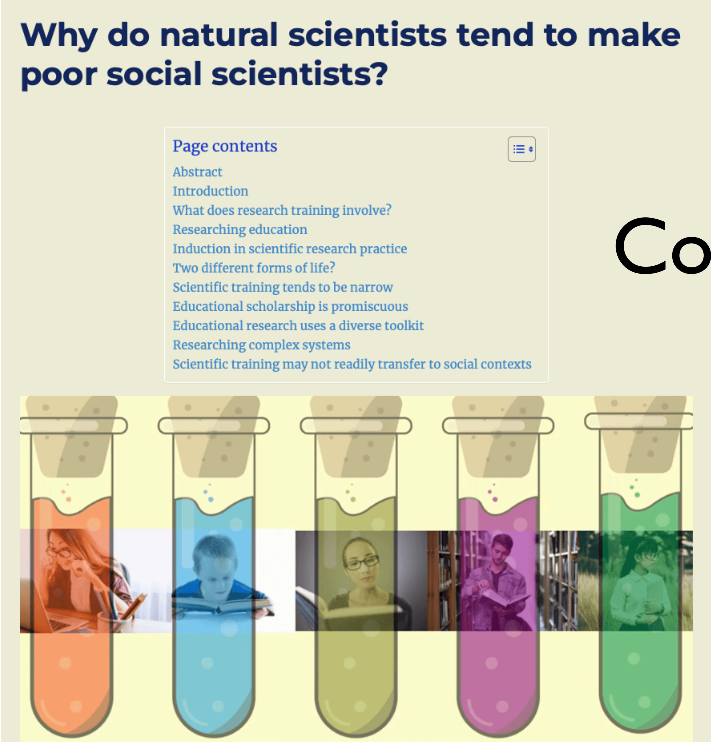
The American Chemical Society

did NOT publish a comment on the paper in the same journal

(but you could read it here:

https://science-education-research.com/methodological-and-procedural-flaws-in-published-study/)

https://science-education-research.com/methodological-and-procedural-flaws-in-published-study



Why do natural scientists tend to make poor social scientists?

Page contents

- Abstract
- Introduction
- What does research training involve?
- Researching education
- Induction in scientific research practice
- Two different forms of life?
- Scientific training tends to be narrow
- Educational scholarship is promiscuous
- Educational research uses a diverse toolkit
- Researching complex systems
- Scientific training may not readily transfer to social contexts

Conclusion

https://science-education-research.com/publications/miscellaneous/why-do-natural-scientists-tend-to-make-poor-social-scientists/

To conclude. We are scientists, or we like to think we are, and scientists do experiments. I know from my time working in science teacher preparation that generally science graduates tend to think experiment is the method of choice when we ask them to undertake small-scale enquiry into their teaching.

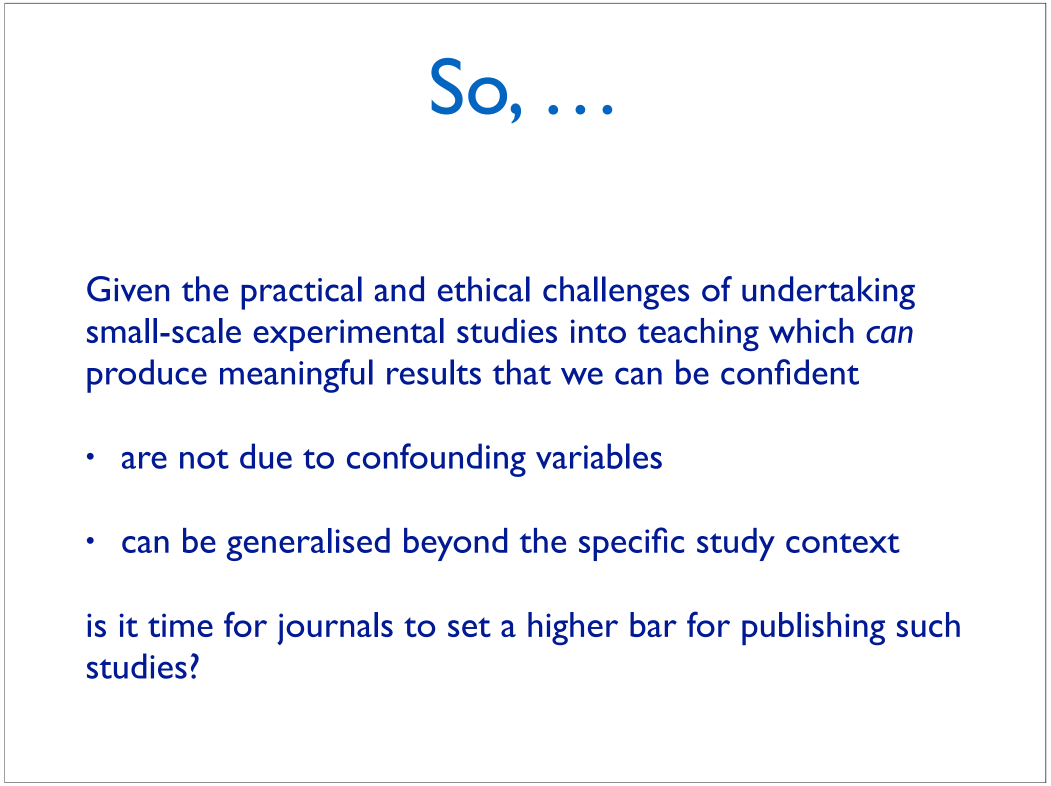
Perhaps our scientific training so promotes the merits of experimental procedures that science educators have an implicit bias that is strong enough to overcome any concerns about features that invalidate so many educational experiments of this kind.

Perhaps our scientific education leads us to think that the world can be organised into natural kinds such that one copper wire of a certain gauge is assumed to be able to stand for any other copper wire of the same dimensions; and so one teacher, or one class of fifteen year-olds, can stand for any other?

Perhaps the scientific mindset that objectifies the natural world is so strong that we see people as experimental subjects that respond to our treatments without regard to the inter-subjective nature of our interactions with them and what they might think of us and our experiments?

Perhaps this is also why we tend to see classes as arrays of individuals and forget that people in groups interact and influence each other.

https://science-education-research.com/publications/miscellaneous/why-do-natural-scientists-tend-to-make-poor-social-scientists/



So, …

Given the practical and ethical challenges of undertaking small-scale experimental studies into teaching which *can* produce meaningful results that we can be confident

- are not due to confounding variables
- can be generalised beyond the specific study context

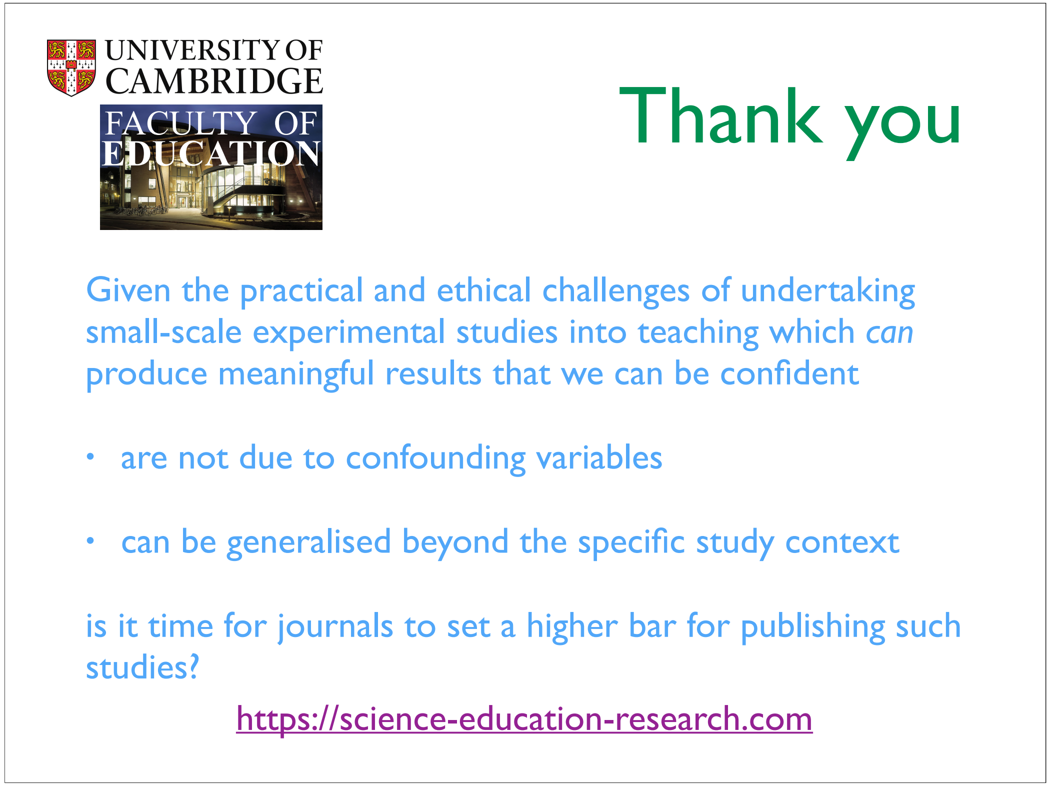
is it time for journals to set a higher bar for publishing such studies?

I would suggest that when a valid experiment is possible, it is usually to be preferred.
But if we cannot do a valid experiment, we should not do an experiment at all.

An invalid experiment is not scientific, and is a waste of valuable resource - including researcher time and participant goodwill.

An invalid experiment carried out by a researcher undermines their claim to be a competent scientist.

So, *what does the science education community's propensity for publishing invalid experiments in its journals say about us collectively?*



UNIVERSITY OF CAMBRIDGE
FACULTY OF EDUCATION

Thank you

Given the practical and ethical challenges of undertaking small-scale experimental studies into teaching which *can* produce meaningful results that we can be confident

- are not due to confounding variables
- can be generalised beyond the specific study context

is it time for journals to set a higher bar for publishing such studies?

https://science-education-research.com

Thank you.

Keith S. Taber
Emeritus Professor of Science Education
University of Cambridge

https://science-education-research.com