https://science-education-research.com

This is the author's manuscript version.

The version of record is:

Taber, K. S. (2013). Non-random thoughts about research. *Chemistry Education Research and Practice*, 14(4), 359-362. https://doi.org/10.1039/c3rp90009f . (This can be downloaded freely from the journal website at https://pubs.rsc.org/en/content/articlelanding/2013/rp/c3rp90009f.)

Non-random thoughts about research

Keith S Taber

There has been a lot of talk recently in the educational research community about randomised trials in education - their relative merits and limitations, and the difficulties of organising them in educational settings. Indeed the Department for Education in the context where I work (England) brought out a paper on the topic a few months ago suggesting teaching could learn from medicine in its regular use of randomised trials (Goldacre, 2013). Medicine, it seems, employs research-based practice, and someone at the Department of Education apparently felt that those of us working in education might like to consider taking a lead from medicine in moving teaching towards becoming a research-based profession.

Interestingly they did not ask a professor of education to make this case (presumably anyone who knows about teaching or educational research would find it difficult to present this as anything like a novel idea), but a respected medical doctor-cum-journalist, thus explaining the well-intended, but ultimately rather condescending, message that education needs to learn about research from another academic area. I expect that actually there is a good deal we could learn about some forms of research from medicine (and many other fields) - but I will not be looking to that source for suggestions about tactful approaches to cross-disciplinary fertilisation.

One of the central features of random trials is of course *randomisation*. This is a critical process in some forms of research. Many published studies reports findings from research undertaken in

specific classrooms, and/or with particular teachers and/or learners, yet also claim to have implications well beyond those few contexts or people. Research submitted for publication needs to show some kind of representativeness, generalisability or wider relevance to be considered as having potential to inform the work of other educators.

A case study, for example, is research in the idiographic tradition (Windelbrand, 1894/1980), that values learning about the individual case due to its own inherent interest. Cases are one instance among many, and cases may be selected for study either because they are considered typical in some way of many other instances, *or* because they are seen as in some way different and special (Taber, 2013). So although in an *instrumental* case study, the case is chosen because it is considered to be representative, in some sense, of a wider population of cases; in an *intrinsic* case study it may be the perceived peculiarities of the case which lead to its choice for study. We may seek to hear about the school science experiences of the Nobel laureate or observe the classroom of the teacher who achieves national recognition for excellent practice because they are special cases. Yet even in this situation - the deliberate choice of a single example considered atypical in some way - there is an implicit assumption that this instance has enough in common with other instances such that what is learned from this study can help us better understand other cases. If we know what encouraged the Nobel laureate into science; if we know how the expert teacher goes about her work; then perhaps we can change aspects of practice to benefit others.

In case study work we would not select the case at random as the logic of this type of research requires us to chose our cases purposefully on a principled basis. However, in some other types of research randomisation is essential for a rigorous study. This is the case in surveys and experimental work.

Surveying a population

In a survey we are usually sampling from a population. In some situations we can almost reach an ideal situation where the sample approximates the population. If we are interested in what the undergraduate students in one particular university department think about some topic - perhaps their ratings of the usefulness of an array of on-line learning resources - it may be quite feasible to ask all of them. A few may be ill during the survey period or may not complete the task for some other reason - but if we had a 95% return giving some clear patterns of results then there would be little gained by statistical tests to estimate 'errors' in the measurements due to sampling.

In this kind of enquiry we can not generalise beyond the single department where the work was undertaken to suggest that the same result would apply elsewhere (such as another university department with a different admissions profile, course structure and set of teaching norms) - but that is less important in *context-directed research* where the purpose is to inform practice in the particular context of that institution, rather than *theory-directed research* where we are looking to develop public knowledge that can inform the field more widely (Taber, 2013). This does not necessarily mean a report of the research will be of no interest to others: but sufficient contextual information needs to be offered to allow readers to judge if the context of the research is sufficiently similar to their own teaching context to consider the study relevant to their own work (Kvale, 1996).

More commonly, we sample a small proportion of the population. Perhaps we are interested in knowing which practicals are demonstrated by the high school chemistry teachers in a country. There may clearly be too many teachers to ask them all, so we may look to survey, say, 5% of the population. If we collect data from 5% of the teachers chosen at random then statistics can give us estimates of sampling error - the extent to which the frequencies of responses in different categories from the sample are likely to misrepresent the results we would have got if we had data from every high school chemistry teacher in the country. Perhaps we will find that 85% of our sample report demonstrating the reactions between lithium, sodium and potassium with water: and the statistics tell us it is likely the figure we would have got if we had asked all eligible teachers would have been 85%±4% - somewhere in the range 81-89%.

However this is based on the assumption that our sample *is* representative of the full population in the sense that each member of the population had the same chance of being included in the sample. We hope the sample is representative in terms of whatever factors may influence answers to the particular survey questions - years of teaching experience may perhaps be relevant; hair colour is less likely to be. Often we can only decide what might be relevant by introducing theoretical considerations - but a truly random selection will avoid having to decide which factors need to be considered. So key questions become:

- Are we able to identify all the eligible teachers and access them?
- Will all the teachers who are asked to complete the questionnaire do so?
- Can we make a random selection of the population for our sample?

Perhaps our data base is incomplete and we only have contact details for a proportion of the high school chemistry teachers. That would not matter if the teachers missing from the records are

typical of the wider population. However if there are systematic differences between the teachers where records are available and those where they are not, then this undermines the ability of the sample to fully reflect the population. Perhaps teachers in the database tend to be those who are well established in their teaching appointment and/or are members of a teaching association. Those that are not on the record may move school regularly, or even drift in and out of teaching, and not have membership of professional organisations. We might suspect that this group could have teaching behaviours that are not well aligned with those who are on our list. (As an example, we *could* have theoretical reasons to suspect teachers missing from the record do less demonstrations and instead set more textbook based work.)

Even if we do have a full list of the members of the population to be sampled, we may have to accept that not all those we ask to complete our questionnaire will do so. Many surveys have quite low response rates - people are often busy and may not be motivated to assist in our research. So if we select 5% of the teachers at random, we may only get returns from 40% of our sample (and that would be very impressive compared with many surveys), so only 2% of the population. We could survey a higher proportion (at greater cost) to get a higher response with the same response rate. Perhaps if we sent questionnaires to 12% of the teachers we might get responses from our target of 5% of the population. Again however we can only trust inferential statistics if the 5% from whom we have data are no different (in ways relevant to the research) from the 7% who failed to return the questionnaire. If we suspect there is a systematic bias - for example, teachers who value demonstrations highly and use them a lot are more likely to complete and return a questionnaire on demonstrations - then a basic assumption of the statistical tests is undermined.

If we do have a complete list of eligible teachers, and we are able to be confident that respondents are not systematically different from non-respondents, it is still important that the sample has been randomly selected from the population. This would seem to be the least problematic aspect of the process - but I wonder how many readers know how they would *actually* go about randomly selecting, say, 500 teachers from a list of 10 000?

Testing a hypothesis

The other major type of research strategy where randomisation is important is in experimental design. Here we use statistical tests to tell us whether the differences found in outcomes in two conditions are likely to be down to the difference in the manipulated variable, rather than other incidental or accidental factors - by finding out how likely our result would be if it was just due to

chance. In this type of research we use tests that are only valid when there has been a randomisation process, and they are completely undermined if we introduce some systematic bias into the comparison by the way we, for example, assign classes, teachers, institutions, or students.

Randomised trials involve a large number of institutions or classes or teachers that are likely to vary in a wide range of ways. The investigators may have to accept they cannot control for class size, socioeconomic factors, teaching style, length of lessons, setting strategies (i.e., the extent to which students are 'mixed-ability' or grouped according to attainment or perceived ability) and so on. For example, the investigators may be asking teachers in the intervention condition to take on some novel approach in their teaching - and different teachers will vary in their flexibility; how quickly they adapt to changing their teaching; how their teaching style and philosophy fits with the intervention; and how motivated they are to cooperate with the requests of researchers. We do not eliminate such teacher differences by randomly assigning teachers to the intervention ('experimental') or comparison (cf. control) conditions, but we at least ensure there are no *systematic* differences between the two groups of teachers. Of course, randomisation does nothing to help with other threats to validity, such as poorer performance from teachers working out of their comfort zone in seeking to implement unfamiliar ideas or teaching materials, or the potential for students to be engaged more than usual (or feel threatened) by something a little different from standard classroom fare. (That is, in education, randomised trials may introduce unintended systematic differences that relate to the adoption of something that is different to the norm, as well as the specific intended nature of that novelty. Double blind designs and placebo treatments do not readily transfer from the medical model).

Given the nature of random events, such trials are only likely to be informative when they involve enough teachers (or institutions or classes or whatever the unit of analysis may be) such that randomisation is likely to lead to two heterogenous groups that overall contain a similar mixture of teachers (or classes or institutions) likely to collectively be representative of teachers (or classes or institutions) more widely. With smaller samples, real differences in outcomes may not show as statistically significant: that is, we will often get a 'type 2 error' or 'false negative'.

Small scale studies

In practice, large scale trials are difficult to organise and much of the research undertaken in chemistry education is on a smaller scale. Many studies submitted to *Chemistry Educational Research and Practice* are carried out within an individual institution, and indeed often by researchers

enquiring into their own home university, college or school. Such studies may lack any sense of a statistically representative sampling of a wider population - and when reported well there is no attempt to frame them in such a way. Clearly such work does not share all the affordances of randomised trials encompassing diverse teaching contexts: but just as clearly it offers some advantages in allowing detailed investigation of a particular context, and reporting that provides 'thick description' (Geertz, 1973) of that context that may offer readers insights into the 'hows' and 'whys' as well as the 'what' of the research outcomes. The field progresses by drawing on the complementary strengths of different kinds of studies (Taber, 2009).

Sometimes when research students are investigating classroom teaching and learning they need to sample a small number of the learners (or lessons) to be interviewed or observed in detail. Some research students in this situation will assume that by making random choices they make their research more rigorous: forgetting that a random choice can actually lead to an unrepresentative sample - perhaps not observing any of the lessons with class practical work; or not interviewing any girls in the class, or any of the least able students, simply because they were not picked at random (Taber, 2013). Just as randomisation can be an essential part of some research designs, making *principled selections* on grounds of representativeness rather than random sampling can be important to the integrity of other studies.

A relatively weak form of research is the use of quasi-experimental designs that compare only two teachers or classes, one of which is undertaking some intervention. Here, two teachers working with 'parallel' classes, say, are compared, with one teacher adopting some kind of novel educational treatment and the other continuing with the familiar way of doing things. This situation usually involves existing differences between the classes that have to be accepted by the investigator: perhaps the timetable means the classes have their lessons at different times of day or week or perhaps they are taught in different rooms that contribute differently to classroom atmospheres. There may be attempts to ensure that the teachers involved are of similar levels of qualification and teaching experience (but we all know that teachers, just like classes, are all 'one-offs' - each somewhat different from any other teacher).

Occasionally the researchers may be able to assign students randomly to the two different conditions, but it is much more common to work with classes that already exist. These may or may not have originally been based on random assignment of students - but have often since developed their own identity and history - and particular character. Often the best the researcher can do in these circumstances is to randomly assign the teachers/classes to the two conditions. This is the

minimum requirement to be able to use inferential statistics, but is clearly a very weak approach as the randomisation of two teachers/classes to two conditions does nothing to address inherent differences: but rather just makes it equally likely that each of the unique teacher/class combinations will become the intervention group.

However studies are sometimes seen which do not even use randomisation at this minimal level - but rather perhaps assign the teacher who was most willing to take on the intervention to that condition. In these circumstances there is no possibility of using inferential statistics to tell us how likely any measured difference between the outcomes in the two conditions is to be a genuine effect of the intervention - all the research can validly tell us is the nature of the difference actually measured between the two particular classes and their teachers. Even if statistics tell us the differences in measured outcomes are very unlikely to happen by chance, we already know that the assignment to conditions was itself not chance - so that offers a potential explanation for differences found as an alternative to the factor being manipulated in the two conditions. It becomes almost meaningless to cite statistical significance in such circumstances (and indeed misleading unless the statistics are offered with the appropriate provisos).

Coherence in research design

Sometimes randomisation is fundamental to a research design - sometimes it is the last thing we would want. Ultimately the decision to use randomisation or not is linked to the nature of the research design. In 'Quantitative' work where inferential statistics are to be employed, randomisation may be crucial for valid application of statistical tests. In some other types of study, randomisation does not only fail to increase rigour, but can actually undermine the type of principled sampling necessary for many interpretative studies. In grounded theory informed approaches (Glaser & Strauss, 1967; Taber, 2000), for example, a key feature is 'theoretical sampling' where sampling decisions are informed by the developing analysis, to deliberately test out emerging hypotheses (for example by seeking informants with particular characteristics).

Journal reviewers and editors therefore need to be alert to where randomisation is, and is not, claimed in a paper submitted for consideration for publication. We should also be aware that there are common misconceptions about the nature of randomness and what random distributions might be like (Batanero & Serrano, 1999; Garvin-Doxas & Klymkowsky, 2008). We might expect researchers submitting studies for publication would understand randomness, and in particular how to make random selections, but occasional examples are met at the level of 'I randomly

selected the first teacher I came across', or 'I randomly selected every third child from the class register'!

This makes one wonder whether articles in the educational research literature that make an unelaborated reference to random sampling or randomisation in quasi-experimental studies always do actually refer to a valid randomisation process rather than some form of convenience sampling or tacit application of an implicit sampling frame that the investigators considered random, or 'as good as' random, at the time. Such doubt can be avoided if when we write-up our research we always think to follow any claim about randomisation by a sentence of two explaining the technique we used to ensure that sampling or assignment to conditions was indeed random. Then readers of our work can know that 'random' means random, and not just that we asked the teacher who happened to be in the staffroom when we were getting a coffee. Randomisation may only be relevant in some types of study, but, when it is required, readers of research reports need to know that analyses are based on a genuinely random sample, or a valid randomisation to the conditions being compared. Authors who simply refer to selection at random without further elaboration may find that when their submissions are returned for revision they are asked to add a brief explanation of *how* they ensured 'random' meant random in their particular study.

References

Batanero, C., & Serrano, L. (1999). The Meaning of Randomness for Secondary School Students. *Journal for Research in Mathematics Education, 30*(5), 558-567. doi: 10.2307/749774

Garvin-Doxas, K., & Klymkowsky, M. W. (2008). Understanding Randomness and its Impact on Student Learning: Lessons Learned from Building the Biology Concept Inventory (BCI). *CBE-Life Sciences Education, 7*(2), 227-233. doi: 10.1187/cbe.07-08-0063

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

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: strategies for qualitative research*. New York: Aldine de Gruyter.

Goldacre, B. (2013). Building Evidence into Education. London: Department for Education.

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

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

Taber, K. S. (2009). The positive heuristic: directions for progressing the field *Progressing Science Education: Constructing the scientific research programme into the contingent nature of learning science* (pp. 325-356). Dordrecht: Springer.

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

Windelbrand, W. (1894/1980). History and Natural Science: Rectorial address, Strassbourg, 1894. *History and Theory, 19*(2), 169-185.

Further publications can be accessed at:

https://science-education-research.com/publications/