A proposal for judging the trustworthiness of research findings

Stephen Gorard

Abstract

This paper offers a procedure for, and a description of the elements involved in judging how trustworthy a research finding is. The idea is of value to the users of research evidence and to researchers themselves when creating a synthesis of existing evidence (i.e. in a literature review. The focus here is on active designs to address causal research questions, but the ideas can easily be extended to other types of research. Other than design, the elements suggested are sample size and quality, data quality, fidelity of intervention, and threats to validity. These are combined in a kind of 'sieve' to produce a judgement-based star-rating for the believability of a piece of research. Trustworthiness of research findings is currently an area with too little focus for the development of new researchers.

Introduction

If social science research is to have useful and warranted impact, for example in public policy, the prospective research user really needs to know three things. These are:

- the 'effect' size or strength of any pattern or finding that is being reported,
- the costs, benefits and possible dangers of using that finding in practice,
- and how trustworthy the findings are.

This assumes, of course, that the findings are truly reported and that the research has integrity. If the research is fabricated, distorted or exaggerated, for example, then none of the other three things matter. The focus of ethics committees ought to be much more on keeping

research honest than it currently is. The effect size or summary of the finding is a relatively straightforward technical issue dealt with well in many methods resources. Cost:benefit analyses are generally well-covered by economists. However, the trustworthiness of findings is less well-developed as a theme for researchers.

The same issue of trustworthiness arises when researchers are considering each other's work. A synthesis of existing evidence must take more than the effect size into account, otherwise weak evidence will be bundled along with strong evidence, leading to invalid and possibly dangerously misleading conclusions. 'Strong' and 'weak' here refer not to size of the difference, pattern or trend uncovered but to how convincing the evidence for it is. Meta-analyses, systematic and narrative reviews of existing evidence will be invalid if each study is merely given equal weight. But they will also be invalid if only some studies are included while others are rejected as not reaching some threshold of trustworthiness. 'Trustworthiness' here is something like how convincing the finding is, or even how much one would be prepared to bet on it being true or replicable. How can researchers portray the trustworthiness of their results, or judge those of others? This is the question addressed by this paper.

Both reasons for needing an assessment of research trustworthiness currently face similar problems. Policy-makers, practitioners, advisers, think tanks and other research users generally do not understand enough about the reported research base in their area to make the kinds of judgement necessary. And the main reason for this is that researchers do not present their findings and the evidence for them in a form that others can readily understand. Researchers do the same to each other, presenting results with undigested output from statistical software, long paragraphs and sentences, and needless verbiage and neologisms (Gorard, 2013). Partly because of this some researchers think that they are providing a comprehensible estimate of trustworthiness when they are not.

A clear example is the reporting of 'confidence' intervals. Confidence intervals around a research finding are widely misunderstood and misinterpreted by the researchers themselves, meaning that there is little chance that research users will understand them but a good chance they will be misled (Gorard 2014). Confidence intervals are therefore dangerous (Matthews 1998). They are not an estimate of how much 'confidence' to have in the result; nor do they offer a likelihood that the 'true' result will lie within that interval. Their true definition is an ideal, and it is recursive (involving itself) and reversed in logic

(modus tollens). Confidence intervals are frequently misapplied to situations not involving true and complete random samples. They take no account of design bias, missing data, measurement error or any of the myriad things that really matter when judging the trustworthiness of a research finding. But they are erroneously presented by many writers as if they could do these magical things. What is offered in this paper is not to be used 'instead' of misguided approaches like confidence intervals because confidence intervals do not address the same issues at all. Confidence intervals should not be used to judge the trustworthiness of findings.

The elements of trustworthiness

It must be assumed for the purposes of this paper that any study being considered in terms of trustworthiness has been fully and clearly explained. If the elements discussed below, such as sample size or respondent dropout, are not clear then the reader is justified in assuming the worst. Put simply, a poorly described study cannot be trusted. Once the design and methods used in a study are clear, then it is possible to begin judging how believable the findings ought to be. The proposal in this paper is that a number of related issues need to be considered, stemming from the design of the study and its subsequent conduct. The design needs to fit the research question(s) being addressed (White 2009). The discussion that follows is based around causal questions, since the designs for these are the most complex, and therefore the hardest for other researchers to assess.

Design

Although a valid causal model may start with an association, and may include an explanatory model and a proposed sequence of events, it is fundamentally based on a comparison between two or more groups of cases (Gorard et al., 2011). One (or more groups) is exposed to one level of the purported cause, and another is exposed to a clearly different level of the cause. All other things being equal, a difference in the effect between the two groups can be interpreted as evidence of cause: effect. One key element of such a model is that the comparison between groups is a fair one. This fairness can be achieved in a number of ways. One is to randomly allocate all participating cases to the initial groups, in a randomised controlled trial (RCT). Another is to allocate cases to groups in terms of a threshold or cut-off point, in a regression discontinuity design (RDD). The technique of matching cases between the two groups, in terms of their known characteristics, can achieve superficial balance but is intrinsically more likely to lead to bias or imbalance than either RCT or RDD. Not matching cases and

simply having two (or more) naturally occurring groups is clearly even weaker. And weaker still is to have only one group and to compare before and after data. There is a hierarchy of designs for causal questions. No researcher can be blamed for not using a stronger design if it is genuinely not possible, and research does not become useless simply because it uses an inferior design. But an inferior design must then limit the kinds of claims made by the researcher, and the trustworthiness with which the findings are viewed by others.

A real-life example might be a synthesis in which an intervention was tested in one piece of research using an RCT with 100 randomly-allocated participants receiving the intervention and a further 100 randomly-allocated participants not receiving it. In another piece of research the same intervention was tested with 200 cases. All were asked to volunteer and 100 did so. The results of the 100 volunteers were then compared with those of the other 100. It would be quite wrong to treat the evidence of the second study as anything like as important as that of the first (Gorard and See 2013).

Scale

A second consideration of trustworthiness would be scale. In general and all other things being equal, a larger study is more impressive than a smaller one. A study wishing to make a causal claim, but comparing two groups of 10 cases each, for example, would be trivial. It could also be seen as unethical – wasting the time of all concerned. In making a causal claim there are other factors to consider, as described below, but the comparison at the heart of the claim would usually need several hundred cases. A case is the unit that would be randomised or otherwise allocated to the two (or more) groups but not necessarily the unit from which data is collected. For example, if 60 hospitals were randomly allocated to two groups, and then data was collected from all patients, the cases would be the hospitals not the patients. 'N' would be 60.

A claim is commonly made that in some ways research in the social sciences is harder than in natural science because the cases are more variable and less inherently predictable (Nash 2004, Gorard 2004). This may be so, but it is seldom pursued to its logical conclusion. In order to make believable claims, social science research would therefore need a larger number of cases than used in other areas of investigation.

Attrition

As important as scale is the completeness of the cases involved in the study. If the design has worked thus far, the research has two sizeable and very similar groups of cases ready for a comparison to take place after only one group has received the 'treatment' level of the purported cause. Any dropout from the study is serious after the cases have been allocated to comparison groups, because there is no reason to believe that the dropout will be either random or balanced (Hansen and Hurwitz 1946, Sheikh and Mattingly 1981). In social science, it can often be the knowledge of which group a case has been allocated to that creates the imbalance. For example, if the cases are people, knowing that being in the treatment group involves some effort might make cases in that group more likely to drop out. Or, if being in the treatment group is seen as exciting or beneficial there may be more attrition from the other group. Either way, those dropping out may be busier, more mobile, more likely to be homeless, less motivated, less literate, less concerned and so on. Despite lax WWC guidance to the contrary (http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_draft_standards_handbook.pdf), bias can easily arise even if the amount of dropout is equal between the groups. Imagine a study looking at a remedial reading intervention for primary school children. The higher attaining children allocated to the treatment group might find the intervention programme patronising and unnecessary, making them more likely to drop out. The lower attaining children in the other group may be demoralised at not getting the extra help given to their peers in the treatment group. They may be less likely to continue to co-operate with the study. Even if the numbers dropping out are small and identical between the groups, the fact that the kinds of children dropping out in each group tend to differ will then create considerable bias for the results.

For example, an RCT with 100 randomly-allocated participants receiving the intervention and a further 100 randomly-allocated participants not receiving it, and 100% completion rate has N=200. An RCT with 200 randomly-allocated participants receiving the intervention and a further 200 randomly-allocated participants not receiving it, but only a 75% completion rate has N=300. In any synthesis of evidence, the first study must be treated as far more trustworthy than the second despite the reported N being smaller. The missing 100 cases in the second study could completely transform the findings if their data was available. In fact this is a sensible approach to considering the possible impact of attrition – how different would

the data from the missing cases have to be to negate the apparent finding from the cases that are available?

Quality of data

Whatever the outcome of the intervention (or similar) is intended to be, the evidence presented for it needs to be of high quality. There are many factors to consider here. Where cases remain involved in the study but have missing data of any kind, this should be treated as part of attrition (above). A research report needs to clarify the N for each and every analysis, since N is rarely a constant in practice.

The data on outcomes needs to be reliable in its true senses of being repeatable/replicable and of being judged to be the same by different observers (rather than merely internally consistent). In this respect, real-life measures such as length tend to be better than counts such as how many cases had a certain clear characteristic (Gorard 2010). Worse in turn are standardised tests of attainment which tend to be somewhat better than questionnaires used to estimate latent concepts such as motivation. Weakest of all will be impression data (although as with all forms of data collection this will have other advantages, but just not for trustworthiness – the subject of this paper).

It is important that the outcome(s) of interest is specified and made clear before the study is conducted if at all possible. This is to prevent researchers or users subsequently dredging the results for success or failure. The outcome also needs to be independent of the intervention itself. A key threat, especially when the outcome measure is tied in any way to the intervention, is that the treatment group might practice the post-test or a close proxy for it.

Further threats to safe findings

There are a large number of further issues that could enhance or reduce the trustworthiness of research results. They are bundled together here because in practice (see below) the overall level of research quality is already set by this point. This is because the various elements of quality are related in practice to some extent. For example, it is unlikely that a design based on a very weak comparison group would bother with whether the participants were 'blind' as to which group they were in when the outcome data was collected. Similarly it would be rare for a large RCT not to have pre-specified outcomes.

Perhaps the greatest amorphous threat to any study is a conflict of interest for anyone involved. Traditionally this has been interpreted as

concern where stakeholders stand to gain financially from the results of the study. However, COIs are wider than this. Researchers can have prestige or prior claims wrapped up in a study intended to test their own, perhaps well-known, theory. This might make them reluctant to face a robust and independent test of their claims. Practitioners can become unreasonably enthusiastic about an intervention even though ostensibly they have nothing to gain from an untrustworthy finding. The solution is, of course, that evaluators must be unconcerned about the nature of the results other than their quality, and that all interested parties should be 'blinded' as far as is feasible.

Other threats to validity include having so many possible outcomes that some must be positive, the unintentional experimenter effect, accidental diffusion of treatment between allocated groups, post-allocation demoralisation, and regression to the mean (Shaddish et al. 2002).

An aid to judging trustworthiness

All of these ideas are summarised in Table 1, and associated with a simple star rating system. This can be used to help assess the trustworthiness of any relevant study. The ratings are from 4★, the best kind of evidence that could be expected from a single large study, to 0, a study that adds little or nothing to the evidence base. A suggested procedure would be to start with the first column, reading down the design descriptions (for addressing a causal question) until the study is at least as good as the descriptor in that row. An RCT or RDD, for example, might lead to row 1. A propensity score matched design might lead to row 2. If the design is not reported or there is no comparator this would lead immediately to row 5. Staying in the row achieved for the design, move to the next column and read down the scale descriptions until the study is at least as good as the descriptor in that row. An RCT with only 12 cases in each group would end up in row 5 at this stage. Then repeat this process for each column, moving down (never up) the rows until the study is at least as good as the descriptor in that row. The final column in the table gives the estimated star rating for that study.

Table 1: A 'sieve' to assist in the estimation of trustworthiness

Design	Scale	Dropout	Outcomes	Fidelity	Validity	Rating
Fair design for comparison	Large number of cases per comparison group	Minimal attrition, no evidence of impact on findings	Standardised pre-specified independent outcome	Clear intervention, uniform delivery	No evidence of diffusion or other threat	4★
Balanced comparison	Medium number of cases per comparison group	Some initial imbalance or attrition	Pre-specified outcome, not standardised or not independent	Clear intervention, unintended variation in delivery	Little evidence of diffusion or other threat	3★
Matched comparison	Small number of cases per comparison group	Initial imbalance or moderate attrition	Not pre-specified but valid outcome	Unclear intervention, with variation in delivery	Evidence of experimente r effect, diffusion or other threat	2★
Comparison with poor or no equivalence	Very small number of cases per comparison group	Substantial imbalance and/or high attrition	Outcome with issues of validity or appropriateness	Poorly specified intervention	Strong indication of experimente r effect, diffusion or other threat	1★
No report of comparator	A trivial scale of study, or N unclear	Attrition not reported or too high for any comparison	Too many outcomes, weak measures, or poor reliability	No clearly defined intervention	No consideratio n of threats to validity	0

This means that an evaluation will be judged to be as good as the lowest classification it has achieved for each of the six categories. For any column, if it is not possible to discern the quality of the study from the available report(s) then the rating must be placed in the lowest (0) category. In using this aid, the emphasis throughout is intended to be on judgement. The ratings represent how much one might be prepared to stake on an intervention working or not, based on a single evaluation, in the same context or setting again.

The ratings should take no account of whether the intervention itself was deemed successful. That is part of the impact assessment. Nor should they take into account the practicalities, or otherwise, of the intervention. That is part of the cost : benefit analysis. A low rating should not be interpreted as necessarily the 'fault' of the researcher –

who will often be faced with practical, resource and ethical constraints. The researcher can however be deemed at fault in three common ways:

- if the rating is low because the reporting of research is poor or incomplete, or cannot be understood by the audience for which it is intended

- if there is clearly a better, simpler or more powerful approach the researcher could have used with the same resources

- or if the researcher tries to make claims or draw conclusions unwarranted by their study.

A few examples can help illustrate how the approach is used, although they are presented only in summary here. Two recent evaluations of literacy catch-up programmes for 11-year-olds are of contrasting quality.

An evaluation of Switch-on Reading was based on 314 individually randomised Year 7 pupils who were struggling with literacy (http://educationendowmentfoundation.org.uk/projects/switch-on-reading/). The pupils were taught individually in their own schools meaning that diffusion was just about impossible. They were pre-tested before randomisation to one of two groups of 157 (i.e. blind), and re-tested on-line. The test was standard, independent of the intervention and agreed as a fair test by the programme developers. Six pupils (under 2%) dropped out for a number of valid reasons including leaving the country between pre and post-test. None had extreme scores. After the evaluation the control pupils also received the intervention. The design of the study was a fair one, the randomisation produced two reasonable sized and balanced groups, and the attrition was too low to have affected the substantive result. Using the sieve, this study would be 4★ (or perhaps 3★ if 150+ is not considered a 'large' number of cases in each group).

An evaluation of Response to Intervention, used for the same reasons and with the same kind of pupils as Switch-on, involved 61 schools randomised to the intervention or not (http://educationendowmentfoundation.org.uk/projects/response-to-intervention/). This is immediately a weaker study, despite also being an RCT, because however many pupils took part in each school the number of cases is really only 30 (schools) per comparison group. In practice, 12 schools dropped out or provided no post-test scores after

being randomised to groups. In total, attrition was near 25%, mostly from the control group. The study is therefore 1★. It starts as a fair design, loses a step in Table 1 because of scale, and then loses further steps because of attrition. By that stage the other columns cease to matter very much. In fact, the quality of the rest of the study including the conduct of the intervention itself was considerably better than 1★. Nevertheless, in any synthesis of evidence the evaluation of RTI must be weighted much lower than that for Switch-on irrespective of their reported results and effect sizes, and the fact that both were RCTs led by the same research team with roughly equivalent resources.

A third recent evaluation was of an attempt by a partnership of schools to enhance the use of feedback by teaching staff (http://educationendowmentfoundation.org.uk/projects/anglican-schools-partnership/). All nine primary schools in the partnership formed the intervention group. Comparisons were provided by a set of five neighbouring schools not in the partnership (compared in terms of data collected specifically for the evaluation), and all other state-funded schools in the same local authority (compared in terms of official Key Stage 2 results, and of their published value-added scores). The study was conducted like this because it was a large pilot (around 3,000 pupils), but it is, at best, a 'matched comparison' and so the study drops through the sieve to 2★ in the first column. It then stays in that row, because in all other respects the study was at least as good as the descriptions of 2★. For example, attrition was actually quite low given the scale.

These three examples are simply illustrations. But they do show that the same kind of design can lead to more or less trustworthy results depending upon what happens in practice. They show that the trustworthiness of a study is not a function of its researchers, funders, institutions, or the outlet in which it is published. They also show that despite the so-called 'gold standard' of RCTs, other designs can turn out to be more trustworthy in practice.

Conclusion

The procedure here is intended to be as inclusive as possible. It is deliberately non-specific about the kind of data involved in any study since the latter is independent of issues like design, scale and data quality. An RCT can have any kind of outcome from standardised test to differences in impression. A trustworthy number of cases for any claim would be the same whether the data was collected face-to-face, via Skype, email, or survey for example. The procedure can be

extended to descriptive research, rather than the kinds of causal investigations discussed here, using mostly the same factors such as scale, attrition, or data quality. The same basic idea of the methodological warrant would apply to descriptive studies as much as causal ones (i.e. if the conclusion to be drawn were not in fact true how could we explain the apparent evidence for it?). But that will be the subject of a further paper.

There is no technical or push-button solution that will decide whether to include a study in a synthesis of evidence or not, or how much weight to give it if it is included. Either decision is necessarily a judgement (Gorard 2006). This judgement should be justified and made clear to others so that they can decide if they agree, or where exactly it is that they disagree. It is perhaps here that the 'sieve' might be most useful. Unfortunately there is also a danger that it becomes a complex or extensive technical check-list (http://educationendowmentfoundation.org.uk/evaluation/). This is not the intention and not how it will be best used.

The cell descriptions are also deliberately non-specific. This is not lack of care, but passing of control to the user. For example, the phrase 'a large number of cases' might be interpreted rather differently, depending upon the precise context, question or pay-off. There is also an interaction between the simple number of cases, their completeness, representativeness of a wider set of cases, and the integrity of the way they have been allocated to groups. 'A large number of cases' would certainly be in the hundreds, but there is no precise figure such as 400 that can be set, other than as a rough guide. An excellent study might have one case below whatever threshold is suggested (399) and a weak one might have one more (401). Similarly, a true RCT might be considered a 'fair design for comparison' but there will be other designs of equal ability to discriminate between effect and noise. Some may not even have been thought of yet. There is no limit to the ingenuity of research design. An attrition rate of 2% might be crucial if the missing cases all had extreme scores in the same direction, whereas 10% might still yield reasonably secure results if there was an obvious reason for the dropout that was unbiased across groups and types of cases. As with N, there is no clear threshold between 'minimal' attrition and worse that can be defended. It is like the hair: beard argument. There is a clear difference between trivial attrition and non-trivial attrition. But where precisely that difference lies is a matter of judgement, based on what is known about the precise context and the nature of the missing cases and where they appeared in the research process.

Part of what is achieved by the procedure described here is a way of retaining more evidence in an evidence synthesis while being careful and scrupulous about the quality of evidence. Ten studies of reasonable quality each involving 10 cases must be at least as important as one reasonable quality study of 100 cases. Yet many traditional approaches to evidence synthesis would reject each of the ten smaller studies on account of scale (examples from https://eppi.ioe.ac.uk/cms/). Similar rejections often occur because of deficiencies of design or even because of the nature of data collection or analysis. These same syntheses may, rather strangely perhaps, also make great play about including all studies even those that are not published because of the bias caused by the file-drawer problem (Torgerson 2003). All readably reported studies should be considered in a synthesis, published or unpublished and despite deficiencies of scale and quality. All studies can help the aggregation towards the best possible bet on whether a finding is true or whether an approach works or not. At the same time, however, a synthesis cannot be a simple vote-count. The quality of each study needs to be taken into account explicitly, as well as it 'effect' size and costs. It is such judgements of quality that the sieve presented in this paper is intended to assist with.

Acknowledgements

The ESRC and Nuffield Foundation funded work that led, in part, to this procedure. Beng Huat See, Steve Higgins and Camilla Neville all contributed valuable ideas and comments for the star ratings.

References

Gorard, S. (2004) Three abuses of 'theory': an engagement with Nash, *Journal of Educational Enquiry*, 5, 2, 19-29

Gorard, S. (2010) Measuring is more than assigning numbers, pp.389-408 in Walford, G., Tucker, E. and Viswanathan, M. (Eds.) *Sage Handbook of Measurement*, Los Angeles: Sage

Gorard, S. (2013) *Research Design: Robust approaches for the social sciences*, London: SAGE

Gorard, S. (2014) The widespread abuse of statistics by researchers: what is the problem and what is the ethical way forward?, *Psychology of Education Review*, 38, 1, 3-10

Gorard, S. and See BH (2013) *Do parental involvement interventions increase attainment? A review of the evidence*, London, The

Nuffield Foundation, http://www.nuffieldfoundation.org/sites/default/files/files/Do_parental_involvement_interventions_increase_attainment1.pdf

Gorard, S., See, BH and Davies, P. (2011) *Do attitudes and aspirations matter in education?: A review of the research evidence*, Saarbrucken: Lambert Academic Publishing

Hansen, M. and Hurwitz, W. (1946) The problem of non-response in sample surveys, *Journal of the American Statistical Association, 41*, 517–529

Matthews, R. (1998) *Bayesian Critique of Statistics in Health: The great health hoax*, http://www2.isye.gatech.edu/~brani/isyebayes/bank/pvalue.pdf

Nash, R. (2004) Science as a theoretical practice: a response to Gorard from a sceptical cleric, *Journal of Educational Enquiry*, 5, 2, 1-18

Shadish, W., Cook, T. and Campbell, D. (2002) *Experimental and quasi-experimental designs for generalized causal inference*, Belmont: Wadsworth

Sheikh, K. and Mattingly, S. (1981) Investigating nonresponse bias in mail surveys, *Journal of Epidemiology and Community Health, 35*, 293–296

Torgerson, C. (2003) *Systematic reviews*, London: Continuum

White, P. (2009) *Developing research questions: a guide for social scientists*. London: Palgrave

Stephen Gorard, School of Education, Durham University
s.a.c.gorard@durham.ac.uk