Martyn Hammersley

AGAINST 'GOLD STANDARDS' IN RESEARCH: ON THE PROBLEM OF ASSESSMENT CRITERIA[1]

The problem of assessment criteria is an issue about which there has been a great deal of debate. Much of it has been concerned with the question of whether the same criteria are appropriate for quantitative and qualitative research, and indeed whether there can be any criteria for qualitative inquiry.[2] For the most part these discussions have been concerned with how *researchers* should assess methods and findings. But there is an equally important question about how lay audiences – notably funders, sponsors, gatekeepers, and stakeholders – do, and should, assess research. And this is, of course, likely to be particularly important in the context of evaluation research.[3] It is important to remember that these audiences are likely to approach the task of assessing research findings rather differently from researchers.

The notion of a 'gold standard' is sometimes seen as solving, or at least as greatly easing, both these aspects of the problem of assessment; and perhaps especially engagement with lay audiences. What this involves, of course, is treating one method as superior: either as the standard by which all others should be judged, or as the only effective method. In recent times, in influential quarters, randomised controlled trials (RCTs) have frequently been given this status. This largely derives from the influence of the evidence-based medicine movement, but the idea has now, of course, been extended to social science research, especially where this is concerned with directly informing policymaking and other forms of practice.

It is not difficult to find declarations that RCTs are the gold standard. Here are a couple of examples:

> 'The "gold-standard" research method for addressing the question "what works?" in evidence-informed policy-making and practice is

[1] Paper given at 'Was heißt hier eigentlich "Evidenz"?', Frühjahrstagung 2015 des AK Methoden in der Evaluation Gesellschaft für Evaluation (DeGEval), Fakultät für Sozialwissenschaften, Hochschule für Technik und Wirtschaft des Saarlandes, Saarbrücken, May 2015.

[2] See, for discussion and references, Hammersley 1992:ch4, 2008 and 2009.

[3] I am assuming for the purposes of argument that evaluation is a form of research. I am aware that there is a controversial issue here. For what it's worth, my view is that a relatively sharp distinction needs to be drawn between research and other activities, including the evaluation of whether policies are good or bad, beneficial or detrimental, and the provision of practical or policy advice. My position is broadly a Weberian one, see Hammersley 2014:ch3.

> the randomised controlled trial.' (Torgerson & Torgerson, 2008, p.1)
>
> 'The gold standard methodology in the social sciences is called "random assignment."' (http://www.edchoice.org/Research/Gold-Standard-Studies)

It is worth remembering that gold standard-type arguments are not restricted to advocacy of RCTs, though it represents one of the more extreme cases. Other examples include advocacy of:

- Participant Observation (Becker and Geer 1957)
- Audio or video-recordings of 'naturally occurring' social interaction (Potter and Hepburn 2005).
- In-depth interviews as the method for Interpretative Phenomenological Analysis (Smith et al 2009)
- Mixed methods (Creswell and Plano Clark 2007:5; but see Symonds and Gorard 2010).

My argument will be against all gold standard arguments, in principle, but I will focus particularly on the case of RCTs.

Before explaining why I don't think that RCTs can serve as a gold standard, I perhaps ought to sketch in a little more of the background to recent attempts to give them this status. The evidence-based medicine movement began in the 1980s, championed particularly by clinical epidemiologists (Pope 2003). In its most radical and newsworthy form it treated RCTs, or the synthesis of their findings in *systematic reviews,* as the only effective way of determining 'what works'. Clinicians were required to access research evidence of this kind, and to use only those treatments that had been scientifically validated.[4]

The idea of evidence-based practice came to be supported by health service managers and by government policymakers; and was extended to new areas, including social fields. One reason for this was that it fitted with the 'new public management' that became influential in the 1990s, and that continues to shape government today, with its concern to make public sector professionals more 'transparently' accountable (Pollitt 1990; Lynn 2006; Head 2008).

[4] This radical or classical version of evidence-based practice has often been qualified or liberalized in the face of criticism (this sometimes being signaled by a shift to 'evidence-informed practice'), as regards both what counts as evidence and what role research should play in relation to practice. However, the more this is done, the less distinctive the position becomes.

The privileging of evidence from RCTs has been particularly strong in the US. Under George W. Bush, in some fields (notably education), the use of RCTs became virtually a requirement for any research receiving federal funding. More recently, the Obama administration announced its commitment to evidence-based social programs, which the Brookings Institution, a major promoter of RCTs, has dubbed “the most important initiative in the history of federal attempts to use evidence to improve social programs.”

In the UK, while there were signs of a waning of the evidence-based practice movement towards the end of the first decade of the twenty-first century, there has been a recent resurgence. The UK Government’s ‘behavioural insights team’ (Haynes et al 2012) [popularly known as the ‘nudge unit’, because it draws heavily on the ideas of Thaler and Sunstein 2008] promotes RCTs in Government policy, and increased numbers are being funded in education, social work, and criminology. Torgerson (2014) claims ‘a growing realisation by policy-makers that other research designs cannot effectively answer the “What works?” question’.

So that’s the background. What I want to suggest is that the idea of RCTs as a gold standard is a case of overselling. But first I must emphasise that I am not denying that RCTs can be a very useful method. I’m not dismissing them as of no value, simply questioning the claim that they are of *superior* value overall. Their distinctive value lies in the fact that random allocation to groups receiving different ‘treatments’ greatly reduces the danger of selection bias: the danger that the distribution of background differences among cases will generate misleading effects or will obscure treatment effects. A key point is that it does this for variables whose significance is not known as well as those that researchers suspect might act as confounding factors; and it does not require measurement of any of these factors. In this specific respect, RCTs are superior to non-experimental quantitative research and to qualitative work. My point, though, is that this neither makes them superior in all respects to other methods, nor guarantees that their findings will be valid (because there are other potential sources of error).

So, while RCTs are a very useful method:

1. They are not *essential* in reaching sound conclusions.
2. Nor do they eliminate all threats to validity, even in theory, and certainly not in practice.

3. Furthermore, there are often severe problems, and limits, in applying them in the social field.[5]

Problems with RCTs

Internal versus external validity

It is widely recognised that, while RCTs can maximise likely internal validity, this does not guarantee external validity, because 'in the wild' the treatment will be operating amidst many of the confounding factors whose effects the trial minimised (see Cartwright 2007). There is also a tension in running trials between what is required to achieve internal and external validity. Thus, in medicine it is common for additional forms of control, besides randomisation, to be exercised over potentially confounding factors: for example, patients with other conditions besides the one to which the trial relates, or those taking other medicines, may be ruled out, so as to avoid these affecting the trial results. This stems from recognition that randomization cannot deal with all background factors. However, when the treatment comes to be applied 'in the wild', these other factors will operate, and may significantly affect the effectiveness of the treatment.[6]

The context in which RCTs are particularly valuable is in testing the effectiveness, and side effects, of drugs. Here there may be a good case for arguing that it is the gold standard. However, as we move away from that context, even in the field of medicine, the power of RCTs weakens. Before explaining why, it is worth remembering that in the drug field, RCTs are used as a complement to laboratory work, which will have produced a considerable body of knowledge about the drug. By contrast, in social fields RCTs are usually expected to provide the whole scientific knowledge base for the 'treatment' (other social science work is rejected on the grounds that it doesn't meet the 'gold standard'). What this laboratory work does is to provide knowledge about the mechanisms that underpin the effectiveness of the drug. Yet while RCTs may be able to demonstrate an empirical pattern, they cannot tell us what are the causal mechanisms involved, and the conditions under which these operate. I suggest that this becomes a serious problem in the social field, where we have much less reliable knowledge of causal mechanisms.

[5] You will find more detailed presentations of these arguments in Hammersley 2002 and 2013.

[6] There are also issues about the effectiveness of randomization and the need for it, see Worrall 2007.

The difference between theory and practice

While, in principle, RCTs provide considerable control over variables, in the social field especially there can be major practical problems in establishing this (see the discussion in Gueron 2002). There may be resistance to random allocation, from those who feel that they are being disadvantaged, and on broader ethical grounds.[7] Even aside from this, there are also often problems in ensuring compliance on the part of the large number of practitioners 'delivering' the treatment. Also, whereas in many trials in medicine there is double blinding – neither doctors nor patients know who is receiving which treatment – with many social interventions this is impossible: the differences in treatment may be obvious to all participants. This introduces a potentially serious threat to validity.

Measurement problems

In the social field the problems involved in measuring the outcomes in which there is interest are usually even more severe than they are in the field of medicine. Indeed, almost all social outcomes of policy relevance are extremely difficult to measure accurately. This is not a problem that is unique to RCTs, of course, but it does damage their claim to provide convincing evidence about 'what works' as regards social policies and practices. While I am not going to discuss this issue in detail here, it is a major problem.[8]

The problem of standardising 'treatments'

In the context of drug trials considerable control can be exercised over the treatment that is administered to each patient included in the trial, as regards both its constituents and dosage. Once we move away from this context, standardisation is much harder to achieve, and this introduces potential bias into the findings, which could be random and/or systematic. This is a particular problem in the social field because the 'treatments' involved take the form of processes of social interaction, in which those 'delivering' the treatment will necessarily adapt and adjust to the particular people they are interacting with: a counsellor will vary her or

[7] There are certainly ethical issues to be addressed as regards RCTs, as with other methods. I have not addressed these here because RCTs are not usually promoted as a gold standard in ethical terms.

[8] For a more detailed discussion of my views on measurement, see Hammersley 2010.

his behaviour according to relevant characteristics of the client, a teacher to those of a class, and so on. Moreover, the more successful those administering an RCT are in standardising the treatment the greater the likelihood that the findings will lack external validity, because 'in the wild' those supposedly applying the treatment will not act in standard ways. Furthermore, the RCT is unlikely to be able to indicate all significant variations in 'treatment'.

A matrix not a hierarchy

As I indicated, my argument here is not that RCTs are of no value, simply that they have weaknesses as well as strengths. In this general sense they are the same as all other methods. And the conclusion I draw from this is that rather than thinking of methods in terms of a hierarchy it is more appropriate to think in terms of a matrix. This would indicate that each method has significantly different strengths and weaknesses, arising in large part from the threats to validity to which they are subject, and the likely strength of these. In selecting methods we must make judgments about what seems likely to be the best method, or combination of methods, *for a particular project*, in light of the research questions being addressed and the context in which it is being carried out.

In other words, my argument is that methods vary independently in their susceptibility to each of the various threats to validity that plague social research. For this reason, and for more practical reasons, all methods have distinctive strengths and weaknesses. As a result, an overall ranking is not meaningful. So, in selecting a method we are forced to trade-off some advantages against others. This is true even when we *combine* methods, since combining methods does not automatically cancel out validity threats.

Not only must we deal with multiple threats to the validity of our conclusions, that no single method or combination of methods can entirely eradicate, but also there are some fundamental disagreements among social scientists about the nature and relative seriousness of these validity threats. I think it is a mistake to ignore these, as is largely done in advocacy of RCTs as a gold standard. One thing that seems to be ignored here is the past history of evaluation research. If we go back to the 1960s, in the US and elsewhere, we find demands that new policies and programs be subjected to large-scale quantitative evaluation, sometimes involving RCTs. The experience of carrying these out led to recognition of the problems and weaknesses involved in such evaluations. One outcome was the emergence of various forms of qualitative evaluation. It would not be unreasonable to argue that in some respects this amounted to an overreaction to the problems, and that there was a failure to recognise that the new approaches also had serious weaknesses. But it is important to try to learn from the experience of the past, and to engage with the arguments involved, for example about the distinctive features of carrying out research on human social behaviour, as against studying the chemical effects of drugs on human bodies.

In summary, then, in the first part of this paper I have argued that the idea of a gold standard sitting atop a hierarchy of methods ranked in terms of their scientific validity is a fallacy. While any assessment of the likely validity of research findings must take account of the research methods used, what will be required is careful judgment about the likely seriousness of particular validity threats, and their significance in the context of the research concerned. One feature of the argument that RCTs represent a gold standard, as with claims for other quantitative techniques, has been the idea that they eliminate 'subjective judgment', since they involve the application of 'transparent procedures'. As a result, so it is claimed, the validity of findings is open to assessment by lay people (Oakley 2000). However, while the nature or degree of judgment involved can vary in research, it can never be eradicated (Gorard 2006). Furthermore, judgment is not necessarily bad: only bad judgment is bad![9]

Pragmatic or rhetorical issues

Up to now I have been solely concerned with the question of whether there can be any justification for the notion of a gold standard in research methodology, and I have concluded that there is no justification. However, there is a more pragmatic aspect to this question that, I suspect, is especially important in the context of evaluation research.

In dealing with lay audiences, there may well be rhetorical advantages associated with the gold standard argument about RCTs. This is because it implies that:

- Research can produce demonstrably sound conclusions.
- These are authoritative because they derive from the technical expertise of researchers.

[9] I have argued this in relation to measurement issues, see Hammersley 2010.

- This expertise is ‘transparent’ because it involves the use of explicit procedures and thereby allows lay assessment, since likely validity of findings can be determined by asking whether or not an RCT was employed.

These rhetorical advantages of RCTs may well seem appealing given the particular challenges that evaluation researchers face. Even more than other social scientists, they are caught in a serious dilemma: they must satisfy two audiences with somewhat different requirements – those of stakeholders and those of fellow researchers. Moreover, satisfying stakeholders must be done in a context where not only is it assumed that research should be able to supply what is wanted, when sometimes it cannot, but also there is competition from unscrupulous traders in ‘evidence’, such as 'think tanks' and pressure groups.

This context seems to me to generate potential dilemmas: tensions between what is judged to be necessary to produce sound findings, from a research point of view, on the one hand, and what stakeholders will find acceptable, plus their usually very tight time-scales, on the other. There may also be tensions between both of these and what the researcher believes will serve the common good, or be politically desirable in some other sense.

In this context, there may well be demands from funders or stakeholders for RCTs, or it may be that the gold standard rationale can help evaluation researchers in dealing with these audiences. But for the reasons I have explained, I do not believe that this is a sound strategy.

It would be helpful if I could recommend some alternative, more effective general way of dealing with these problems, but I’m sorry to say that I can’t. Indeed, I don’t believe that there can be any single solution, or perhaps any ‘solution’ at all. Unfortunately, these are problems of a kind that is often referred to as ‘wicked’ (Churchman 1967). Rather than their being open to resolution via appeal to a gold standard, or any other purportedly ‘transparent’ criterion, they must be dealt with on a case-by-case basis, in a way that meets the requirements of particular projects: *there is no general solution*. Pragmatic judgment is unavoidable (but need not be simply ‘arbitrary’ or ‘subjective’).[10]

[10] We need to subvert the spurious contrast between procedure and judgment, objectivity and subjectivity (Hammersley 2011).

Conclusion: Getting the balance right

Unfortunately, we must face the fact that research cannot always provide answers to questions that are seen as pressing by policymakers and practitioners. *Nor can we guarantee the validity of our findings*. And we need to make this clear to lay audiences. So, there is a danger of over-promising or over-claiming. At the same time we must not under-sell the contribution of research. This is undoubtedly a difficult balancing act to sustain. However, the idea that there is a methodological gold standard does not help with this. It is a delusion and a deceit.

References

Becker, H. S. and Geer, B. (1957) 'Participant observation and interviewing: a comparison', *Human Organization*, 16, 3, pp28-32.

Cartwright, N. (2007) 'Are RCTs the gold standard?', *Biosocieties*, 2, 1, pp11-20. (Available at: http://www.lse.ac.uk/CPNSS/research/concludedResearchProjects/ContingencyDissentInScience/DP/Cartwright.pdf.)

Cartwright, N. and Hardie, J. (2012) *Evidence-Based Policy*, Oxford, Oxford University Press.

Churchman, C.W. (1967) 'Wicked problems', *Management Science*, 14, 4, pp141-2.

Creswell, J. W. and V. L. Plano Clark (2007). *Designing and conducting mixed methods research*. Thousand Oaks, California, London, Sage.

Gorard, S. (2006) 'Towards a judgement-based statistical analysis', *British Journal of Sociology of Education*, 27, 1, pp67-80.

Gorard, S. and Symonds, J. (2010) 'Death of mixed methods? Or the rebirth of research as a craft', *Evaluation and Research in Education*, 236, 2, pp121-36.

Gueron, J. (2002) 'The Politics of Random Assignment', in Mosteller, F. and Boruch, R. F. (eds.) *Evidence Matters: Randomized trials in education research*, Washington D.C., Brookings Institution Press.

Hammersley, M. (1992) *What's Wrong With Ethnography?* London, Routledge.

Hammersley, M. (2002) *Educational Research, Policymaking and Practice*, London, Paul Chapman/Sage.

Hammersley, M. (2005) ‘Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers’ case for research-based policymaking and practice’, *Evidence and Policy*, vol. 1, no. 1, pp1-16.

Hammersley, M. (2008) ‘Assessing validity in social research’, in Alasuutari, P., Bickman, L. and Brannen, J. (eds.) *Handbook of Social Research Methods*, London, Sage.

Hammersley, M. (2009) ‘Closing down the conversation? A reply to Smith and Hodkinson’, *Qualitative Inquiry*, 15, 1, pp40-8.

Hammersley, M. (2010) ‘Is social measurement possible, or desirable?’, in Walford, G. Tucker, E., and Viswanathan, M. (eds.) *The Sage Handbook of Measurement: How Social Scientists Generate, Modify, and Validate Indicators and Scales, London,* Sage.

Hammersley, M. (2011) ‘Objectivity: a reconceptualisation’, in Williams, M. and Vogt, P. (eds.) *The Sage Handbook of Methodological Innovation*, London, Sage.

Hammersley, M. (2013) *The Myth of Research-Based Policy and Practice*, London, Sage.

Hammersley, M. (2014) *The Limits of Social Science*, London, Sage.

Haynes, L., Service, O., Goldacre, B., and Torgerson, D. (2012) *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials*, London, Behavioural Insights Team, Cabinet Office, UK Government.

Head, B. (2008) ‘Three lenses of evidence-based policy’, *Australian Journal of Public Administration*, 67, 1, pp1–11.

Lynn, L. (2006) *Public Management: Old and New*, London, Routledge.
Oakley, A. (2000) *Experiments in Knowing*, Bristol, Polity.

Pollitt, C. (1990) *Managerialism and the Public Services*, Oxford, Blackwell.

Pope, C. (2003) ‘Resisting evidence: evidence-based medicine as a contemporary social movement’, *Health: An Interdisciplinary Journal*, 7, 3, pp267–282.

Potter, J. and Hepburn, A. (2005) ‘Interviews in qualitative research: problems and possibilities’, *Qualitative Research in Psychology*, 2, 281-307.

Smith, J., Flowers, P., and Larkin, M. (2009) *Interpretative Phenomenological Analysis*, London, Sage.

Thaler, R. and Sunstein, C. (2008) *Nudge: Improving decisions about health, wealth, and happiness*, New Haven, CT, Yale University Press.

Torgerson, C. (2014) ‘What works…and who listens? Encouraging the experimental evidence base in education and the social sciences’. Inaugural lecture, Durham University.

Torgerson, D. and Torgerson, C. (2008) *Designing Randomised Trials in Health, Education, and the Social Sciences*, Basingstoke, Palgrave Macmillan.

Worrall, J. (2002) ‘What Evidence in Evidence-Based Medicine?’, *Philosophy of Science*, 69, pp. S316-330.

Worrall, J. (2007) ‘Why there’s no cause to randomize’, *British Journal of Philosophy of Science*, 58, 3, pp451-88. (Earlier version available at: http://www.lse.ac.uk/CPNSS/pdf/DP_withCover_Causality/CTR%2024-04-C.pdf)