Do Randomised Controlled trials Offer a Solution to 'Low Quality' Transport Research?

Dr Steve Melia

Senior Lecturer, Centre for Transport & Society

University of the West of England

Abstract

This study examines the claims made by several articles in recent years that much transport research is of poor quality and that a hierarchy of methods favouring experimental methods should be applied to meta-studies and the evaluation of transport interventions. These arguments are most frequently applied to the evaluation of voluntary travel behaviour change (VTBC, or 'smarter choices' in the UK) interventions or programmes.

The article begins by reviewing the studies which have applied experimental methods to the evaluation of VTBC and those which have proposed evidence hierarchies as the solution to the perceived shortcomings of other studies. It discusses the potential for, and reviews the evidence of, various forms of bias in the 'before and after' studies typically used to evaluate VTBC programmes. A new method for assessing the robustness of 'before and after' evaluations of area-based VTBC programmes is proposed as one possible solution in some circumstances.

It analyses arguments for and against evidence hierarchies and proposes five criteria under which experimental methods can be demonstrated to produce more robust findings than other method. It concludes that experimental methods have only limited applicability to transport research and that evidence hierarchies favouring experimental methods risk misleading policymakers.

1. Introduction

The selection of research methodologies to inform policy is a contested area, amongst researchers policy advisers and policy makers. The debate revolves around two broad questions:

1. What are the most appropriate methods to address different research challenges?
2. How should the choice of methods (as well as their application) influence the assessment of research findings in meta-studies and literature views intended to inform policy?

The first of these questions primarily concerns researchers but the interventions of policy makers/advisors around the second question have implications for debate around the first one. The spectrum of perspectives on both questions can be characterised as a choice between two approaches. The first approach, favoured by most transport researchers and some policy makers/advisors (e.g. Tavistock Institute and AECOM, 2010) seeks to select the most appropriate method based on the nature of the research questions with no generalised preference for one method over others. The second approach favours a methodological hierarchy, usually with Randomised Controlled Trials (RCTs) at the top. In response to the first question, this approach would recommend a method from the highest possible level in the hierarchy to address each research question. In response to question 2 it would attach greater weight to findings generated by methods at the higher levels of the hierarchy.

Methodological hierarchies (or 'evidence hierarchies' when applied to meta-studies) have become more influential in transport policy and research in recent years as academics and professionals from a health background have engaged more with transport issues (e.g. Graham-Rowe *et al.,* 2011, Rowland *et al.,* 2003). They have been advocated by policy advisors to some governments in contexts which encompass transport, such as social policy

in Australia (Leigh, 2009) and promoting physical activity amongst children in the UK (NICE, 2007). They have also influenced national policy in the UK on Voluntary Travel Behaviour Change programmes (VTBC – or 'smarter choices' in British terminology), as discussed below.

A substantial literature exists on the issues of research design and the choice of methodology for answering different research questions in different contexts: it is not the intention to address the breadth of that issue here. This article will begin with a brief discussion of the spectrum of research designs used to inform evidence-based policy making. It will then briefly review the debate around one area where methodological hierarchies have been proposed as a solution to perceived shortcomings: the effectiveness of VTBC programmes.

The rest of the article will focus on a principal area of contention in this debate: the circumstances under which experimental methods can and should be used in transport research. Section 5 will propose 5 criteria (Table 4) for the application of experimental methods in circumstances where interventions must change human behaviour to be effective, and where those interventions must be generalisable to wider populations in order to inform policy. It will argue that the advantages of experimental methods, and the breadth of their applicability to transport research are not as great as advocates of methodological hierarchies claim. It will then consider the implications of this for evidence-based transport policy.

2. Evidence Based Policy and the Spectrum of Research Designs

The principles of evidence-based policy have been central to transport studies since it emerged as distinct discipline, but the focus of evidence-gathering was, until fairly recently, rather narrow. Whilst government-commissioned studies examine issues such as traffic flow and road safety, as recently as 2000 Terry wrote, in a UK context:

> In the case of research supported through universities and research councils, little of such work is, nor is intended to be relevant to policy.
>
> (Terry, 2000 p. 188)

Several factors have changed that situation since then. In academia, there has been a greater emphasis on achieving external impact, whilst governments have turned to academic and specialist research in pursuit of environmental and (particularly since the recession) macro-economic objectives. The different forms of cost benefit analysis used in many developed countries (Mackie and Worsley, 2013) depend upon inputs from empirical studies. At the macro level contested evidence of 'what works' is central to debates around transport and the built environment, the economic impact of transport infrastructure and the potential for, and best means of achieving, modal shift. In this context, the methodologies used to generate research evidence have been increasingly scrutinised in recent years.

The terminology in the literature is not always consistent but six methodological approaches to transport evaluations can be outlined as follows. The term '**experimental** study' or 'true experiment' usually refers to studies which are both randomised and 'controlled' i.e. subjects are randomly assigned to either an experimental or a control group. This narrow definition is synonymous with RCTs: there are a number of variations around the basic RCT design but they all share those two characteristics . **Quasi-experimental** approaches use a comparator group, as similar as possible to the intervention group instead of random allocation, which may not be possible in many circumstances. **Theory-based** approaches are non-experimental methods which test hypotheses about the causes as well as the impacts of interventions in real world situations. **Outcome or observation studies** observe or measure the impacts of interventions without seeking to test causal hypotheses. These usually contain elements of quantification and may also overlap with the fifth category, **judgemental approaches**, where experts or even programme administrators assess the impacts of interventions or policies. This category would usually be considered a last resort, where no other method could be applied (Rossi *et al.,* 1999), although some element of 'expert judgement' may be unavoidable when considering the broadest research questions

such as: what combinations of policy measures are effective in achieving long-term modal shift? As a sixth category, studies may **combine** more than one of the other five.

Tavistock Institute and AECOM (2010) was commissioned by the Social Research and Evaluation unit of the UK Department for Transport to give guidance to public bodies evaluating transport interventions. It discusses the advantages and limitations of each approach but does not suggest a hierarchy. It includes several flow charts suggesting the most appropriate approach depending on: the focus of the evaluation, nature of the intervention and feasibility of the approach. It provides a useful starting point for researchers and practitioners, but has some limitations. The study behind the report was written by consultants with input from academics in the field. The guidance draws on a mixture of theory and practical experience: it is not always clear where a recommendation is being made for theoretical, practical or cost reasons. Some of the criteria under the 'nature of intervention' heading are debatable, (Tavistock Institute and AECOM, 2010 Figure 6). The requirement for causal pathways to be short and straightforward, for example, cannot be demonstrated from first principles. The recommendation that experimental methods should only be used where "expected outcomes are small or medium sized" appears to be related to cost: where impacts are large, a "simple before and after study is likely to suffice" (Tavistock Institute and AECOM 2010, p. 41). This recommendation avoids the key question: would experimental methods generate more or less reliable evidence than other alternatives in those contexts? (The word 'reliable' is used in this article to mean 'evidence which can be relied upon for policy' rather than its more specific definition in research terminology).

As discussed in Section 6, many of the bigger questions in transport research could clearly not be addressed by experimental methods. These include the impacts of major infrastructure projects, longer-term changes to the built environment and the impacts of interventions or programmes on national economies. As a result, the debate around methodological hierarchies in transport research has tended to focus around one area: VTBC, where there has been a vigorous debate about the reliability of evaluation studies.

3. Voluntary Travel Behaviour Change and Problems of Research Bias

The terms VTBC and smarter choices have been used in different ways, but generally refer to interventions which seek to change travel behaviour by 'management and marketing rather than operations and investment' (Sloman *et al.,* 2010), though many VTBC programmes include or accompany elements of infrastructural change (e.g. new cycle routes). A principal aim of these measures is usually to reduce single occupancy vehicle driving. There have been many published evaluations of VTBC programmes, usually based on 'before and after' self-reported travel surveys, which are susceptible to several forms of bias, tending to overstate the effectiveness of interventions. These include: social approval bias (Bonsall, 2009), expectation bias or the 'good subject effect' (Morton and Mees, 2010 following Orne, 1970), non-response bias, where those with a positive story to tell are more likely to complete the 'after' surveys (Chatterjee, 2009). Some of the studies have been conducted by organisations or individuals with an interest in promoting VTBC, creating the risk of reporting or retrieval bias, where greater prominence is given to positive results (Möser and Bamberg, 2008).

Many of these effects have been recognised in other fields for some time and whilst their influence on VTBC programmes is easy to identify in theory, there is little specific evidence of their importance in practice. Morton and Mees (2010) cite the evaluation of the Travelsmart programme in Alamein, Melbourne, as one example where these effects allegedly distorted the results. Ker (2011) defends the original evaluation and rejects Morton and Mees' criticisms. From the evidence presented in both papers (citing local factors, variations in weather etc.) it is not possible to determine whether the reported modal shift was overstated but there are reasons for believing that it *might* have been. The response rate of the 'after' survey was considerably lower than the 'before' survey and the self-reported travel in the after survey may have been susceptible to the good subject effect.

Although the reasons have not been studied, there are some examples which suggest that, depending on survey design, self-reported travel surveys may substantially distort findings in

this area. Following a strategy to increase cycling, using both infrastructure improvements and marketing measures in York, UK, Harrison (2001) reported that cycling as the usual mode of travel to work had risen from 15% of working people in the 1991 Census to 18.6% in a self-completed household survey conducted in 2000. The 2001 Census, published the following year, showed that the share of cycling had, in fact, fallen to 12% (ONS, 2009: Table CS121). In an evaluation of the UK's Cycling City and Towns programme, a face-to-face survey in 2009 appeared to show a substantial *increase* in physical *in*activity, from 37% to 26% of respondents, compared to a baseline telephone survey, which asked the same question (Chatterjee and Hardin, 2011). The researchers ascribe this difference to respondents' greater honesty in face-to-face interviews than on the telephone. This explanation suggests that social approval bias influenced the baseline telephone survey; it also suggests that social disapproval of dishonesty outweighed any influence of the good subject effect in the final face-to-face survey.

There is an emerging consensus on some of the methodological measures which could help to address – though not entirely eradicate – these concerns. These have focussed particularly on the use of more objective data, such as traffic or pedestrian counts or GPS-based tracking (Bonsall, 2009, Chatterjee, 2009), and clearer separation of evaluation studies from the advocates and implementers of VTBC schemes (Morton and Mees, 2010). Passive observation measures such as traffic counts should solve problems of social approval bias, non-response bias and the good subject effect, but they introduce other problems, such as how to ensure that the counting points are representative of the wider area. GPS-based measures avoid reporting biases associated with reporting, but are not immune from the good subject effect, if participants' awareness of the survey purpose exerts an unconscious effect on actual behaviour.

Some studies have sought to 'triangulate' self-reported travel data with objective measures of traffic volumes, for example (e.g. Sloman *et al.,* 2010). These studies inevitably rely on interpretive judgements because general traffic levels are influenced by many external unmeasurable factors. Melia (2013) suggested a new method, which could help to provide a more objective basis for such judgements. This would involve trip counters on self-contained networks of residential streets, where all movements in and out can be measured and compared to self-reported data (including visitors and deliveries). The residents of these areas may not be typical of the whole study area, so other methods such as travel diaries may still be needed across the wider area. By measuring traffic volumes in circumstances where they can be precisely compared to self-reported data, this method would enable the researchers to quantify the effect of any self-reporting biases.

Three studies have recommended RCTs as the specific solution to these problems, in evaluating VTBC programmes in general (Graham-Rowe *et al.,* 2011, Möser and Bamberg, 2008), and evaluating school travel plans in particular (Rowland *et al.,* 2003). Möser and Bamberg (2008) use a methodological hierarchy in a meta-study assessing the effectiveness of VTBC programmes. They conclude that the mainly "weak quasi-experimental" evaluation studies they reviewed "may underestimate but more probably overestimate the true causal car reduction effect" of VTBC measures.

Graham-Rowe *et al.* (2011) reviewed 77 evaluations of transport interventions designed to reduce car use, most of which could be described as VTBC interventions. They classify the evaluations into five levels of research quality, which follow a methodological hierarchy rather than any assessment of how effectively the methods were applied. Those evaluations classified as high quality used 'rigorous experimental designs' (only five were RCTs). Those classified as 'low quality' used "weak designs without control groups, from which we cannot draw methodologically valid inferences." The authors recognise that "rigorous experimental designs are challenging in field studies". They suggest that 'weaker' research methods have tended to exaggerate the effectiveness of certain interventions but that valid evidence does support the effectiveness of some interventions. They make a plea for "more robust evaluation methods…and in particular that RCTs are adopted wherever possible"

Rowland et al. (2003) conducted a RCT of school travel plans, which found no significant modal shift in travel to schools in London. They go further than others in arguing that positive evidence from RCTs should be a condition for continued public funding of school travel plans.

This debate has directly influenced national policy on VTBC in the UK. The UK Department for Transport (DfT, 2012) cites Möser and Bamberg (2008) in national guidance which had the effect of ascribing only limited potential benefits to VTBC programmes when appraising the cost-benefit ratios of transport projects seeking public funding.

4. Causal Relationships and Sources of Bias

The justifications for methodological hierarchies are not always clear in the meta-studies and policy advice papers which use or advocate them, although they can sometimes be inferred. The UK National Institute for Clinical Excellence (2007) states, that without RCTs "it is impossible to demonstrate causality" – thus by implication they believe that RCTs can demonstrate causality. NICE also suggests that other methods are more susceptible to bias. The use of monetised benefits in cost benefit analysis depends upon the quantification of impacts. Meta-studies such as Graham-Rowe *et al.* (2011) and Möser and Bamberg (2008) imply that RCTs can more reliably quantify the impact of VTBC programmes. It is not always clear where these authors are comparing the ideal forms of each research approach, or the effectiveness of their application in practice. As the following analysis will demonstrate, *ideal* RCTs can demonstrate causality and quantify impacts more reliably than other methods, but the inference that this also applies to *real* RCTs depends on several assumptions, which may or may not hold in the real world.

The conditions for establishing causality are complex and contested. For clarity of explanation, table 1 uses one of the simpler versions:

- cause and effect are statistically associated (association)
- cause precedes effect (time order)
- no third factor creates an accidental or spurious relationship between the variables (non-spuriousness) and:
- the mechanism by which the cause influences the effect is known (causal mechanism)

Table 1: conditions for establishing causality in a policy context (Singleton and Straits 1999, cited in: Handy, Cao and Mokhtarian, 2005)

By design (and practical constraints in most cases) many quantitative studies in transport and related fields establish only association. In the interpretation of the findings, the other three criteria are sometimes overlooked.

By introducing an intervention and observing subsequent changes, RCTs would satisfy the time order criterion. This can also be satisfied by other methods – statistical and qualitative – providing 'before and after' data is available.

The main advantages of RCTs relate to the non-spuriousness condition. In most research situations, the outcome under investigation (e.g. modal share of driving) is influenced by some known and many unknown factors. The random allocation of subjects to an experimental or control group in a RCT reduces the risk of an imbalance of unknown factors influencing the outcome – providing the randomisation is effectively performed and sample sizes are large enough.

This condition can also be satisfied by non-experimental methods through statistical analysis of factors hypothesised to influence the dependent variable. The net influence of unknown or unobserved factors can be assessed through measures of the goodness of fit of a model. The key difference is that the known factors – the independent variables (e.g. household car ownership or neighbourhood population density) may be acting partly as proxies for other unknown or unobserved factors (e.g. personal preferences).

As the purpose of a RCT is simply to test the response to a defined intervention, the internal validity of the findings (i.e. 'did that intervention cause that effect in the experiment?') is not affected by unobserved factors. This advantage of RCTs depends on the ability of the

researchers to ensure that the intervention is the only relevant change which affects the experimental group differently from the control group. In a laboratory situation, or where the issue under study is physiological, properly conducted randomisation will satisfy that requirement. Where the issue under study concerns human behaviour and where the subjects are interacting with a wider society this condition may be violated by a range of circumstances encountered in the real world.

The caveat 'in a policy context' under **Table 1** is necessary to justify the fourth condition. The first three conditions are sufficient to demonstrate causality in a purely experimental situation (i.e. they can demonstrate internal validity). It may be possible, in other words, to discover that something works without understanding why. But to generalise from experimental findings to a wider population (i.e. to demonstrate external validity), we would need to assume (following: Cartwright, 2010) that the policy change would:

- change the independent variables in the real world in the same way as the intervention did in the experiment, and
- leave the causal relationships in the experiment unchanged

In a medical context, where the causal mechanisms affect the human body, it may be reasonable to assume that these two conditions would hold. Where the intervention concerns human behaviour in a social context this cannot be assumed. If the causal mechanisms are unknown, there is no way of assessing whether the second of those conditions is likely to hold or not: they may apply serendipitously, but relying upon such a possibility could generate misleading advice. Examples of the challenges posed by these conditions in a transport context will be discussed in Section 5.

Various forms of bias may affect either or both of the association and the non-spuriousness criteria. Jadad and Enkin (2007) list over twenty forms of bias which can and sometimes do affect RCTs at different stages. Most of these could affect any other research method in a similar way but some are likely to affect RCTs differently from non-experimental alternatives.

In a non-experimental study, selection bias refers to the process of selecting a sample from a wider population. Self-selection or non-response bias is a sub-set of this problem, where people who respond to a survey (for example) are atypical of the population under study. Jadad and Enkin draw a distinction between *population selection bias,* which refers to the selection *of* the study population, and *selection bias* referring to the treatment of subjects *within* the study population. Randomisation, where properly administered, solves the second problem but not the first. If the study population is itself a biased subset of a wider population, and if the study is used to inform policy relating to that wider population, the findings will be misleading. The search for willing participants in a RCT may, in some circumstances, increase the risk of this form of bias, compared to the alternative method of studying similar behaviour amongst a larger population in a 'real life' situation. An example of this will be discussed below.

5. Criteria for the Use of Experimental Methods to Inform Transport Policy

Whatever their advantages and disadvantages, it seems the contribution of RCTs to knowledge in transport studies has been fairly limited, so far. There are two possible explanations for this: that transport researchers have been neglecting a method which could improve the quality of their work, as implied by Graham-Rowe *et al.*, (2011) and Möser and Bamberg (2008) or that RCTs are of limited use in answering 'real world' transport questions. To assess those two possibilities, this section will consider the conditions under which RCTs can be used, with comparisons to other methods.

Starting from the principles discussed so far, **Table 2**, lists five criteria for the application of experimental methods to assess interventions which seek to alter human behaviour, and where the intention is to apply 'successful' interventions more widely.

1. The main focus of the research is to test (but not explain) a hypothesised cause-

effect relationship.

2. A representative study population of a sufficient size can be obtained from the target population to whom the intervention would be applied.
3. The intervention can be applied selectively to an experimental group within the study population.
4. No other factors with a significant influence on the outcome would impact the experimental and control groups differently during the experiment.
5. Wider application of the intervention would replicate the causal relationships which applied during the experiment.

Table 2: Criteria for the use of experimental methods to inform policy (all must be satisfied)

Criterion 1 implies that existing knowledge must be sufficient to construct a hypothesis where the intervention is believed to affect a limited number of known and measurable outcomes. Another criterion from Tavistock Institute and AECOM (2010) suggests that experimental methods should only be used to measure interventions with "a single outcome goal" but there is no reason in principle why a RCT cannot measure more than one outcome. Similar considerations would apply to quantitative analysis of 'real world' data. Where the intention is to test and explain, a combined method, using both quantitative and qualitative methods is likely to be most appropriate.

Two elements of Criterion 2 could affect the generalisability of findings: the representativeness of the sample and the sample size. Similar challenges may affect other survey-based evaluations, but it may be more difficult to obtain volunteers for an experiment than respondents to a survey about something which was already happening. Where researchers or commissioners have a strong preference for RCTs this may lead to RCTs where this criterion is not satisfied. Rowland *et al.* (2003) provides one example. From a starting population of 42 schools in one London Borough, half agreed to participate in a RCT measuring the effect of school travel plans. 10 were allocated to the control group and 11 to the intervention group, of whom 2 withdrew after randomisation: the requirements for the control group were more onerous than those of the intervention group. This produced a study with a relatively small sample and also introduced the risk of both self-selection bias (from the two who dropped out) and population selection bias from both the decision to focus on one borough, and the 50% of schools who declined to participate.

An alternative approach might have involved a wider study of schools where travel plans were introduced over a longer period of time, using before and after counts, but without the experimental element. Which would generate more reliable findings to inform policy would be impossible to determine from looking at research design alone.

Criterion 3 rules out using a RCT to assess something which has already been applied across the country. It also rules it out for assessing the impact of context-specific infrastructure such as new roads or railway lines. It effectively rules out measures which are introduced over specific geographic areas, unless the areas themselves become the unit of randomisation in a cluster RCT, in which case the second and fourth criteria would become much more onerous. Cluster trials applied to larger geographical areas, even where practically possible, could violate Criterion 4.

An analogous issue was considered by Sloman *et al.* (2010) in evaluating the Sustainable Travel Demonstration Towns, a three year VTBC programme in England. Although a RCT would not be possible (the towns were selected based on bids to government) the authors considered and rejected a quasi-experimental approach because the municipal leaders of comparator towns were likely to respond to the programme by making changes of their own. The existence of an experiment might also change the nature of interactions between external bodies, such as the national Department for Transport, and the two groups of towns: experimental and comparator.

Whether criterion No. 4 is satisfied or not would be difficult to prove in most situations: it would require a judgement on each occasion. The criteria proposed by Tavistock Institute

and AECOM (2010) such as 'short timescales' and 'political stability during the trial' may be considered judgements about the likely implications of such factors on this criterion.

As Goodwin (2011) has argued, the relationship between transport interventions and outcomes is generally characterised by synergies between measures, delayed and imperfectly reversible effects and feedback loops, both positive and negative. All of these factors affect Criterion number 5.

The combination of Criteria 2 and 5 would be very difficult to *fully* satisfy in a transport context. If, for example, the purpose of the experiment was to test an intervention, which might become a national policy, to satisfy Criterion No. 2 the study population would be randomly drawn from the national population. But if social norms and the behaviour of external bodies had any influence on the target behaviour, introducing the intervention as a national policy would violate Criterion 5, because a small group of randomly distributed individuals would have no impact on those factors, whereas a national policy probably would.

Similar challenges may also affect interventions aimed at more specific groups. For example, Fuji and Kitamura (2003) distributed a free one month bus pass to a small random sample of students and found only a limited effect on travel behaviour. If such an intervention was then introduced as a policy, applying to an entire cohort on one campus, Criterion 5 would be violated by the effect on social norms and possibly on the bus operator, who might respond to an increase in patronage by changing the frequency of services, for example. Several studies have shown that perceived social norms do indeed influence modal choice in situations where VTBC might apply (e.g. Bamberg, 2003, Melia, In Press).

Cluster trials might address this problem in some circumstances, but not all. To take the issue tested by Rowland *et al.* (2003) introducing travel plans in a handful of randomly distributed schools is unlikely to have the same effect on social norms around 'the school run' and active travel by children as a city-wide or national programme of school travel plans.

Criterion 5 will be more onerous where the experiment is used to estimate the *quantitative* impacts of an intervention. If an experiment demonstrates some change in behaviour, then it may be reasonable to draw conclusions about the likely direction of change from introducing similar measures as a wider policy: it would be less reasonable to estimate the *magnitude* of any change based on such findings. A failure to reject a nul hypothesis based on a small-scale experiment conducted over a short time period could also be misleading for policy purposes, where synergies and lagged effects are significant.

Several studies have found that car ownership acts as a mediating variable between a range of independent variables and car use (e.g. Van Acker and Witlox, 2010). The effects of independent variables such as income are lagged and asymmetrical (Dargay, 2001), all of which suggests that policies introduced over a longer time period are likely to have a significantly different effect from experiments conducted over a shorter period of time. It is also one of several reasons why synergies between policies are likely to be particularly important in a transport context.

One issue which illustrates these effects can be found in the literature on cycling infrastructure. A number of studies of localised improvements in cycling infrastructure have found a limited, or no significant effect on overall cycling numbers, leading some to reject "the hypothesis that cycle use is suppressed by the absence of routes and networks" (Harland, 1993: a quasi-experimental before and after study, see also Brand *et al.,* 2014). A meta-study based on 139 studies of infrastructure and policies designed to increase cycling found that cities which experienced large increases in cycling had all made substantial investments in cycling infrastructure, as well as a range of other policies (Pucher *et al.,* 2010). The study concluded that infrastructure investment is an essential part of a "comprehensive approach", enhanced by synergies. One of these synergies relates to the growth over time of 'cycling cultures'. The relationship between pro-cycling municipal policies, infrastructure improvements and the cycling culture of a city would be possible to investigate but difficult, if not impossible, to accurately quantify. If the synergies are as important as Pucher *et al.* suggest, then RCTs which attempted to quantify individual effects would yield misleading results. Following the criteria in table 4, they would also be unsuitable for testing the strength of such synergies.

The same argument applies to overall transport policy at national or municipal levels. Several studies – mainly descriptive – have suggested synergies between different policies (e.g. on road network design, parking policy, public transport and cycling) have contributed to the success of cities such as Freiburg in reducing the modal share of driving at a time when it was rising elsewhere (e.g. Melia, 2006). For investigating these types of multi-faceted policy issues, observation studies are likely to be the most appropriate method (discussed further in Tavistock Institute and AECOM 2010).

In their list of biases potentially affecting RCTs, Jadad and Murray (2007) include: 'choice-of-question bias'. Decisions over research objectives and questions may be influenced by vested interests, the personal agendas of researchers and/or constraints related to cost and convenience. This form of bias can affect any research project, but although RCTs are no more or less likely to be affected, a methodological hierarchy designed to inform policy, particularly where it is linked to public funding is likely to exacerbate the problem. It would incentivise researchers to focus on narrow questions such as 'how does the construction of a few cycle paths in area x affect rates of cycling?' Evidence from such studies would then be given greater weight than the broader longer-term observation studies which suggest that networks of dedicated infrastructure can indeed increase rates of cycling.

6. Conclusions

Table 4 sets out the criteria for the generation of reliable findings from a RCT. Where a RCT entirely satisfies all five criteria it can be demonstrated from first principles that it will generate more reliable findings than other research methods. The discussion and examples suggest that such circumstances are likely to be rare in transport research. Whether RCTs which *partially* satisfy the conditions in Table 4 produce more reliable results than other methods is an empirical question: it cannot be demonstrated from first principles nor deduced by comparing research designs, as required by methodological hierarchies. To test which method more accurately quantified the impact of an intervention would require comparisons between results generated by RCTs and by other methods, in a context where the 'right answer' was known. As it is difficult to conceive of such a circumstance applying to a practical transport question, researchers choosing an appropriate methodology, and policymakers/advisors interpreting research evidence, must fall back on what Flyvbjerg (2001) calls *phronesis*, or 'practical wisdom'. Although it is not possible to demonstrate from first principles which method would generate more reliable findings, the above discussion suggests that timescales of the cause-effect relationships, the importance of social and cultural influences on the target behaviour and similarities or differences between the experimental and policy target populations, are all relevant when assessing the relative advantages of experimental methods and non-experimental alternatives.

RCTs are likely to be more reliable in testing whether a cause-effect relationship exists. The conditions required for generalising findings about the *magnitude* of such effects are more onerous. They are unlikely to apply in contexts where social and institutional influences on travel behaviour are significant, as in most VTBC programmes. Although the reasons for its recommendations are not always clear, Tavistock and AECOM (2010) provides a useful starting-point for researchers conducting transport impact assessments. They suggest that experimental methods are best suited to evaluating relatively straightforward small-scale interventions with a short time frame. This cannot be demonstrated from first principles but it is a reasonable response to the uncertainties discussed in this article.

What may seem a rather technical issue for researchers does have potentially important implications for transport policy. The UK national guidance on VTBC programmes in transport appraisals (DfT 2012) chose to follow a methodological hierarchy, giving greater weight to experimental findings. This effectively reduced public support for VTBC in the UK. As VTBC programmes are generally introduced in geographical areas or nationwide, the implication that experimental methods can more accurately quantify the impacts of VTBC programmes is, following the analysis in this article, misplaced.

This article has challenged the arguments in favour of evidence hierarchies and argued that the application of such hierarchies increases the risk of 'choice of question bias'. Several of the articles reviewed here assume but do not demonstrate that the RCTs they or others have conducted satisfy the criteria for generalisation. Some reasonable concerns have been

expressed about the reliability of VTBC evaluations. Section 4 discussed some of the ways in which such concerns might be overcome, but ultimately all decisions about methodology and the implications of research findings rely on judgements. Evidence hierarchies create an incentive for researchers to make unwarranted assumptions about the generalisability of findings, creating an illusion of quantitative precision, misleading for policymakers.

References:

Bamberg, S. (2003) How does environmental concern influence specific environmentally related behaviors? A new answer to an old question. *Journal of Environmental Psychology*. 23 (1), pp. 21-32.

Bonsall, P.W. (2009) Do we know whether personal travel planning really works? *Transport Policy*. 16 (6), pp. 306-314.

Brand, C., Goodman, A. and Ogilvie, D. (2014) Evaluating the impacts of new walking and cycling infrastructure on carbon dioxide emissions from motorized travel: A controlled longitudinal study. *Applied Energy*. 128 (0), pp. 284-295.

Cartwright, N. (2010) What are randomised controlled trials good for? *Philosophical Studies*. 147 (1), pp. 59-70.

Chatterjee, K. (2009) A comparative evaluation of large-scale personal travel planning projects in England. *Transport Policy*. 16 (6), pp. 293-305.

Chatterjee, K. and Hardin, J. (2011) The contribution of walking and cycling to achieving recommended levels of physical activity. In: Anon. (2011) *43rd Conference of the Universities Transport Study Group* [online]. January.

Dargay, J. (2001) The effect of income on car ownership: evidence of asymmetry. *Transportation Research Part A: Policy and Practice*. 35 (9), pp. 807-821.

DfT, (2012) *TAG UNIT 3.10.6 Modelling Smarter Choices*. www.dft.gov.uk/webtag: Department of Transport.

Flyvbjerg, B. (2001) *Making Social Science Matter : Why Social Inquiry Fails and how it can Succeed again*. Cambridge: Cambridge University Press.

Fujii, S. and Kitamura, R. (2003) What does a one-month free bus ticket do to habitual drivers? An experimental analysis of habit and attitude change. *Transportation*. 30 (1), pp. 81-95.

Goodwin, P. (2011) Can Travel's Random Elements be Controlled? *Local Transport Today*. 564 .

Graham-Rowe, E., Skippon, S., Gardner, B. and Abraham, C. (2011) Can we reduce car use and, if so, how? A review of available evidence. *Transportation Research Part A: Policy and Practice*. 45 (5), pp. 401-418.

Harland, G. (1993) *Cycle Routes*. Crowthorne: Transport Research Laboratory, Safety Resource Centre.

Harrison, J. (2001)
Planning for more cycling: the York experience bucks the trend. *World Transport Policy & Practice*. 7 (3), pp. 21-27.

Jadad, A.R. and Enkin, M. (2007) *Randomized Controlled Trials : Questions, Answers, and Musings* [online]. 2nd ed. Oxford: Blackwell.

Ker, I. (2011) Too good to be true? An assessment of the Melbourne travel behaviour modification pilot. *World Transport Policy & Practice*. 17 (1), pp. 14-26.

Leigh, A. (2009) What evidence should social policymakers use? *Australian Treasury Economic Roundup*. 1 pp. 27-43.

Mackie, P. and Worsley, T., (2013) *International Comparisons of Transport Appraisal Practice*. Report number: PPRO 04/03/31.Leeds: Institute for Transport Studies.

Melia, S. (In Press) Alternatives to Private Car Use by Mobile NHS Professionals. *Journal of Sustainable Transportation*.

Melia, S. (2013) No sign of Smart Travel Towns in Census. *Local Transport Today*. April 19th (620), .

Melia, S., (2006) *On the Road to Sustainability - Transport and Carfree Living in Frieburg*. [Available online:] www.stevemelia.co.uk/vauban.htm: University of the West of England W.H.O. Healthy Cities Collaborating Centre.

Morton, A. and Mees, P. (2010) Too good to be true? An assessment of the Melbourne travel behaviour modification pilot. *World Transport Policy & Practice*. 16 (2), .

Möser, G. and Bamberg, S. (2008) The effectiveness of soft transport policy measures: A critical assessment and meta-analysis of empirical evidence. *Journal of Environmental Psychology*. 28 (1), pp. 10-26.

NICE, (2007) *Promoting Physical Activity for Children: Active Travel Interventions* [online]. Report number: 5.National Institute for Clinical Excellence. [Accessed February 12th 2014].

ONS (2009) *2001 Census.* Available from: www.nomisweb.gov.uk .

Pucher, J., Dill, J. and Handy, S. (2010) Infrastructure, programs, and policies to increase bicycling: An international review. *Preventive Medicine*. 50, Supplement (0), pp. S106-S125.

Rossi, P.H., Freeman, H.E. and Lipsey, M.W. (1999) *Evaluation: A Systematic Approach*. 6th ed. London: Sage.

Rowland, D., DiGuiseppi, C., Gross, M., Afolabi, E. and Roberts, I. (2003) Randomised controlled trial of site specific advice on school travel patterns. *Archives of Disease in Childhood*. 88 (1), pp. 8-11.

Sloman, L., Cairns, S., Newson, C., Anable, J., Pridmore, A. and Goodwin, P. (2010) *The Effects of Smarter Choice Programmes in the Sustainable Travel Towns* [online]. London: Dept. for Transport.

Tavistock Institute and AECOM (2010) *Guidance for Transport Impact Evaluations : Choosing an Evaluation Approach to Achieve Better Attribution* [online]. London: Department for Transport.

Terry, F. (2000) Transport: Beyond Predict and Provide. In: Davies, H.T.O., Nutley, S. and Smith, P., eds. (2000) *What Works? : Evidence-Based Policy and Practice in Public Services* [online]. Bristol: The Policy Press, pp. 187-206.

Van Acker, V. and Witlox, F. (2010) Car ownership as a mediating variable in car travel behaviour research using a structural equation modelling approach to identify its dual relationship. *Journal of Transport Geography*. 18 (1), pp. 65-74.