Behavioural Public Policy (2018), 2: 1, 22–40
© *Cambridge University Press* doi:10.1017/bpp.2016.6

The case for laboratory experiments in behavioural public policy

PETER D. LUNN*

Economic and Social Research Institute (ESRI) & Department of Economics, Trinity College Dublin, Ireland

ÁINE NÍ CHOISDEALBHA

Economic and Social Research Institute, Dublin, Ireland

Abstract: Behavioural science is increasingly applied to policy in many countries. While the empirical approach to policy development is welcome, we argue with reference to existing literature that laboratory experiments are presently underused in this domain, relative to field studies. Assumptions that field experiments, including randomised controlled trials, produce more generalisable results than laboratory experiments are often misplaced. This is because the experimental control offered by the laboratory allows underlying psychological mechanisms to be isolated and tested. We use examples from recent research on energy efficiency and financial decision-making to argue that mechanism-focused laboratory research is often not only complementary to field research, but also necessary to interpreting field results, and that such research can have direct policy implications. The issues discussed illustrate that in some policy contexts a well-designed laboratory study can be a good – perhaps the best – way to answer the kinds of research questions that policy-makers ask.

Submitted 30 September 2016; accepted 19 October 2016

Introduction

Just over a decade ago, a group of distinguished behavioural scientists lamented the "painful and frustrating" failure of behavioural science to influence public policy (Amir *et al.*, 2005, p. 444). It is indisputable that times have changed. The issue is no longer whether behavioural science will – or ought to – influence policy. The apparent success and international growth of 'behavioural insights' teams has settled that. The debate is now on how behavioural science can best be exploited by policy-makers.

* Correspondence to: ESRI, Whitaker Square, Sir John Rogerson's Quay, Dublin 2, Ireland. Email: pete.lunn@esri.ie

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

The present paper contributes to this debate by addressing a specific issue, namely the potential for laboratory experiments to make direct and telling contributions to policy development. Underlying the argument is a concern that, in the now international rush to apply behavioural insights to policy, policy-makers and researchers sometimes bestow an unwarranted degree of priority on field experiments and randomised controlled trials (RCTs). Some complex yet important issues surrounding the interpretation and generalisation of results from behavioural studies may not be fully appreciated. The consequence is that some questions that could be best answered by laboratory research are either unaddressed or addressed less effectively by field research. Our argument is not oppositional, but rather promotes consideration of how laboratory and field studies can complement each other. We nevertheless conclude that laboratory experiments ought to play a greater role in evidence-based policy development.

Some initial qualifications are needed. Our research team is funded by Irish government departments and agencies to undertake behavioural research, mostly using laboratory and field experiments. The scope of the article is limited to the comparison between the two, where the key distinction is whether the impact of a manipulation on behaviour is tested in an artificially created environment or in people's everyday environment.[1] Field experiments therefore include experimental trials of policies conducted in the field, including RCTs. This narrow scope is not intended to imply superiority of these over other research methods, but reflects limitations of space and allows us to draw on our experience in conducting experiments for policy to illustrate the relative advantages of laboratory work. This article is not a comprehensive review; examples of research are selected to illustrate arguments. One of us recently undertook an international review of how behavioural economics is being applied to policy (Lunn, 2014), but the pace of change is rapid, and much relevant activity, especially that of public officials, goes unrecorded. Thus, while we have striven to make the argument dispassionate and rigorous, some of the motivation behind it and the arguments we put forward are born of our own experiences of interfacing with policy-makers in Ireland, the UK and elsewhere.

1 Harrison and List's (2004, pp. 1011–1114) taxonomy of field experiments in economics distinguishes them from laboratory experiments along six criteria: subject pool; nature of the information provision; commodity traded; trading rules; size of stakes; and naturalistic environment. Space does not permit a full discussion of this, but we adhere to the simpler view that the artificial versus natural environment constitutes the distinction proper. While the other criteria listed match common characteristics of designs in experimental economics, they are not inherent aspects of laboratory experiments; undertaking the experiment in an artificial environment is inherent.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

A common refrain is that field trials of policy interventions, especially RCTs, represent a superior form of direct evidence for policy development, because they show whether a proposed intervention actually works in the real world. The UK's Behavioural Insights Team (2012a) simply states that RCTs "are the best way of determining whether a policy is working" (p. 4). Similarly, at the interface between policy-makers and behavioural researchers, it is frequently stated that while laboratory experiments are useful for scientists to test new ideas and to develop theories of behaviour, the small samples and unrealistic laboratory environment mean that one cannot generalise results to the real world. Combining these views, van Bavel *et al.* (2013) argue that RCTs are "the purest and most accurate observation of behaviour, unlike experiments which take place in a laboratory" (p. 14). The Behavioural Insights Team concentrates overwhelmingly on field trials of behaviourally informed policy interventions, with 17 of its 20 academic publications describing RCTs or field trials.[2] The distribution of recent behavioural research for policy is also weighted towards field experiments, especially RCTs. Work listed on the Scopus database of peer-reviewed literature from 2010 to the present, in journals with 'policy' in the title, using the search term 'behaviour', returns just 24 results for additional search terms 'lab experiment' and 'laboratory experiment'. In contrast, 97 results are returned for 'field experiment' and 'randomised controlled trial' (41 and 56, respectively). Broadening the search to all relevant abstracts whose titles include 'experiment' or 'trial' raises these numbers to 30 and 107.[3]

According to the arguments we present, there exist policy contexts in which the familiar assertions of the superiority of field trials are essentially valid, and other contexts in which they are highly contestable and, potentially, seriously awry. If it is accepted that different methods are better in different contexts, then generating evidence for policy should be about selecting the right tool for the job at hand. Our aim is to assist in this by provoking both researchers and policy-makers to consider when a laboratory experiment is the most appropriate way to generate evidence for policy. Although focused on similar themes, the nature of the research questions asked by scientists and policy-makers differs. We argue that laboratory studies with rigorous experimental control will often be more useful than field studies for both types of research question. This is primarily because laboratory experiments have the

2 Accessed at www.behaviouralinsights.co.uk/academic-publications/, 26 September 2016.

3 The categorisation was based on the judgement of the authors, applying the definition of laboratory versus field experiment or RCT. Abstracts referencing non-human behaviour only are excluded. American variants of spelling were included.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

capacity to isolate and test the properties of psychological and behavioural mechanisms that are likely to operate across multiple contexts.

We begin by briefly assessing the pros and cons of laboratory and field experiments, using both highly familiar and, perhaps, some less familiar arguments. This forms the backdrop for the main argument, which is that there are contexts in which laboratory experiments are more likely to produce results with direct policy application. Specifically, we argue that the superior experimental control offered by the laboratory allows psychological mechanisms to be better isolated and tested, with potentially strong policy implications. To illustrate, we consider two specific policy contexts. In the first of these – the energy efficiency gap – we argue that the interpretability of field results suffers due to insufficient, rigorously controlled laboratory research. In the second – financial decision-making – we illustrate how laboratory studies can provide direct evidence for policy. Having presented and illustrated our core argument about generalising from experiments, we consider the types of research questions policy-makers face and offer some guidance on when laboratory and field research are more or less likely to deliver answers for policy development.

Generalisation and causal mechanisms

In revisiting the contrast between field and laboratory experiments, rehearsal of familiar pros and cons can inhibit us from noting more subtle elements. Although there are many issues regarding ethics and experimenter effects, the core argument can be summarised as follows. The power of field experiments arises from access to naturalistic behaviour; environmental validity is ensured because the experiment takes place in the environment itself. When assignment to treatment and control groups is entirely random, as in an RCT, individual differences are unlikely to explain any differences in results. Conversely, the power of laboratory experiments is that behaviour can be measured under multiple conditions in which the environment is under complete experimental control, allowing the effects of precise experimental manipulations to be inferred. Randomised allocation to groups is straightforward. The same participant's behaviour can be measured repeatedly under multiple conditions, increasing statistical power. However, differences between the controlled laboratory environment and people's everyday environment render environmental validity questionable. This implies that laboratory experiments may be good for testing scientific hypotheses, but that field experiments, including RCTs, have greater environmental validity and are therefore more informative for policy purposes.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

Generalisability of field experiments

Unless strongly qualified, this familiar analysis attributes a misplaced generalisability to field experiments, especially RCTs. RCTs are narrow in scope by definition (Cartwright, 2007). They exist to evaluate interventions, and a positive effect in a well-designed RCT tells us only that the exact intervention worked in the context in which it was undertaken, including the specific population and the agent who tested it. This means that an RCT is not generalisable unless we make additional assumptions about the causal mechanism underpinning the result. Whether the same intervention will work when carried out by someone else, subsequently, on different people or elsewhere depends on that mechanism being unaffected by these differences. Cartwright and Hardie (2012) discuss the example of a Tennessee study in the 1980s that found improved student performance following reductions in class size. Implemented in California, the same intervention did not work. This was likely due to a lack of space and qualified teachers, indicating that the benefits of classroom size may interact with classroom quality. One implication is that had the class size trial been conducted first in California, the null result may have deterred Tennessee from reducing class sizes. With multiple mechanisms at play, an RCT can miss or underestimate a potentially important effect on which a good policy could be built.

RCTs in public policy are fundamentally different from their equivalent in medicine, despite the oft-repeated argument that the success of the technique in medicine implies that they should also be used for public policy (e.g. Duflo & Kremer, 2005; Haynes *et al.*, 2012; van Bavel *et al.*, 2013). The locus of the mechanism in a drug trial is the human body; the locus in a policy trial is a subset of human society. If causal mechanisms that produce societal outcomes differ across subsets of society more than causal mechanisms that produce medical outcomes differ across human bodies, the analogy is flawed. Additionally, a policy trial is often carried out by a public organisation or unit that has volunteered to conduct it. If such willingness is correlated with how well different units can implement new policies, national roll-out following a positive RCT result may prove disappointing.

These arguments notwithstanding, field trials are useful and often the best way to gather evidence for a new policy, or to evaluate an existing one. It is also necessary to note that many RCTs and field experiments are explicitly based on theory (e.g. Boudet *et al.*, 2016) or explicitly test mechanisms (List, 2004; Lynham *et al.*, 2016). The generalisability issue, therefore, is not strictly one of field versus laboratory, but of study design. The complex nature of human behaviour and the innumerable factors influencing it require inferences to be drawn about effect sizes and how much weight to give evidence for or

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

against specific mechanisms and interactions. The relative benefit of RCTs is that they tell us whether a policy works in context; however, if we wish to generalise research beyond its immediate context, then we need to test mechanisms. Some field experiments test mechanisms, but laboratory studies exist explicitly for this purpose.

Mechanisms in the laboratory

Prior to discussing mechanisms, it is important to recognise a long-running debate on the external validity of laboratory and field studies in economics. Levitt and List (2007a, 2007b) critique laboratory studies for not generalising to real markets in the way that field studies conducted in actual markets (e.g. sports card markets) do. In counter-argument, Camerer (2011) and Kessler and Vesterlund (2014) argue that laboratory experiments reveal the general way in which behaviour is influenced by experimental factors. Kessler and Vesterlund (2014) refer to this as qualitative external validity, meaning that the relationship between factors is valid, although its precise manifestation may differ in the field due to additional factors. However, this debate centres on how to test economic theory, not how to inform policy. The criticisms of laboratory studies primarily question whether preferences and game-theoretic strategies elicited in the laboratory will be reflected in market behaviour, but laboratory experiments do much more than elicit preferences and strategies. They can also be used to determine, amongst other things, the scope, capability, capacity, consistency, speed and accuracy of whatever psychological mechanisms drive decision-making and behaviour, as well as how these properties vary according to situational variables. Human beings can alter preferences or strategies in different environments at will; altering abilities is alas a different matter. Yet what humans are and are not capable of is often crucial for policy.

Consequently, one means of understanding why a policy works (or does not) is to isolate the psychological mechanism(s) it operates on. Well-designed experiments do this, identifying the mechanisms driving the behaviour of interest and illuminating how they operate. Where prior evidence and inference suggest candidate causal mechanisms upon which a policy could be based, policy-makers may gain better evidence from funding experiments designed to isolate and assess precisely defined mechanisms than from running RCTs.

Similar points have been made previously (Deaton, 2010; Ludwig *et al.*, 2011). The latter argue that a mechanism experiment can sometimes be more informative for policy than an evaluation experiment, because it gives more generalisable insight into the operation of a specific candidate mechanism. The mechanism experiments discussed by Ludwig *et al.* (2011) are field experiments, yet the same logic applies to laboratory experiments. Once it is

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

accepted that not only what works, but how it works, matters then the superiority of RCTs over other field experiments, or field experiments over laboratory experiments, cannot be straightforwardly asserted.

When it comes to convincingly and repeatedly demonstrating a behavioural effect, the shortcoming of questionable environmental validity must be weighed up against some distinct advantages that the laboratory environment conveys. Taking multiple measurements from the same participant increases statistical power, permits multiple conditions to be tested efficiently and allows inference that the mechanism-targeting manipulation drives the hypothesised differences found, even among a small sample (provided that the experiment is well designed, with factors other than the manipulation held constant and order effects controlled for). It is straightforward to affirm effects via replication or to test additional manipulations in order to further investigate effect sizes, unhypothesised effects or alternative explanations.

These advantages are what we mean by isolating the mechanism. They can rarely be matched by field experiments, which predominantly rely on between-subjects measurements and frequently require the cooperation of a private company, public body or voluntary organisation. The result is that sampling is always an issue[4]; statistical tests are generally less powerful, while precise replication and further manipulation take a long time and may be impossible if a collaborating entity is unwilling to repeat the exercise.

Laboratory studies also have potential for spill-over effects. Some cognitive and perceptual mechanisms are essentially universal, modulated by context but not context specific. Isolating these mechanisms and measuring their properties implies a greater likelihood that the research conducted will be of benefit not only to the policy-maker directly engaged with the research, but also to other policy-making domains. This is in contrast to the narrower scope of field experiments and RCTs. Instances in which mechanism-focused laboratory experiments are appropriate are discussed next.

Some examples

The energy efficiency gap

A gap between actual and optimal energy consumption arises because economic agents do not make sufficient investments in energy-efficient technologies (Jaffe & Stavins, 1994; Allcott & Greenstone, 2012). This energy efficiency gap is of considerable interest to policy-makers in multiple countries,

4 Sampling can be an issue for laboratory experiments, too, especially where student samples are used to save costs (see Henrich *et al.*, 2010, for an in-depth discussion).

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

many of whom are struggling with challenging climate change targets. The problem, if resolved, could result in a reduction in energy use that benefits both consumers' wallets and the global climate, yet successful ways to encourage advantageous behaviour remain elusive. In some cases, interventions targeted at altering frequent energy use behaviours (Boudet *et al.*, 2016) or promoting home energy use monitoring (Allcott & Rogers, 2014; Lynham *et al.*, 2016) have been developed. Here, we focus on a single behaviour that has repeated environmental impacts following an initial choice – the purchase of energy-using appliances. Behavioural research in this field illustrates how challenging it is to develop effective behavioural policy without an understanding of the causal mechanisms involved.

Studies generally test whether preferences for or purchases of energy-using appliances are altered by provision of information about annual, multi-annual or lifetime operating costs. Underlying these designs are assumptions about the causal mechanism(s) driving behaviour, namely that consumers are unaware of or inattentive to operating costs at the time of purchase. Recent research indicates that these mechanisms are only partial explanations (Allcott & Taubinsky, 2015). The growing body of field and survey experiments presents inconsistent results, creating difficulties for policy-makers interested in evidence-based policy. For example, in an experiment conducted with a German online appliance retailer, Deutsch (2010) found that disclosure of lifetime operating cost information decreased the mean energy use of chosen washing machines. In a study conducted in Ireland, however, the addition of five-year energy cost labels for tumble dryers did not significantly reduce the average energy consumption of machines purchased in retail outlets (Carroll *et al.*, 2014). More energy-efficient dryers were bought in a field experiment conducted in Finnish retail outlets when a lifetime energy cost label was combined with staff training, but no effect was found for the same intervention with fridge–freezers, nor was an effect found for either the label or training intervention alone (Kallbekken *et al.*, 2013).

Stated-preference studies present similarly contradictory evidence. In one study, provision of lifetime operating costs was successful in pushing survey respondents to prefer more expensive but more efficient televisions, whereas provision of annual operating costs had the opposite effect (Heinzle, 2012). In a similar study, labels showing annual operating costs were found to have no effects on preferences for boilers. However, when individuals' time preferences were taken into account, such labels did prompt respondents to give greater weight to future operating costs (Newell & Siikamäki, 2014), suggesting that although current labelling standards result in an energy efficiency gap, labels are nonetheless effective. However, the extant problem is that the gap exists at all.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

Our intention here is not to criticise individual studies, which are often thorough and well designed, but to highlight how little we can infer from the available results. The studies all provide consumers with information on likely future costs incurred by the purchase, but the provision of financial information is not universally effective. The narrowness of scope of field trials is evident. We cannot know whether differences in results between studies arise from differences in consumers, countries, retailers, staff training or labels. Most studies assess different appliances, creating further difficulties in comparing results and inferring mechanisms. Furthermore, we cannot replicate or manipulate these studies to hold these factors constant while another is varied, because the field environment concedes experimental control. The essential problem for policy is that the underlying causal mechanisms are insufficiently understood, the methods do not uniquely identify candidate mechanisms and, consequently, results cannot be generalised. It is difficult to justify the preponderance of policy-focused field experiments and RCTs relative to controlled laboratory research in the ongoing absence of strong evidence about causal mechanisms.

Nonetheless, the studies gesture towards multiple testable mechanisms. One such mechanism may be the manner in which people process numbers. The fact that results differ by appliance suggests that the range of upfront and operating costs, and the proportions of these to one another, could be driving mechanisms. Extensive laboratory research in experimental psychology indicates that when humans reason about numbers, the noise surrounding estimations is proportionate to the number (Whalen *et al.*, 1999; Feigenson *et al.*, 2004). Manipulating these factors is difficult in a field trial, as retailers cannot alter the prices of different appliances for different consumers, nor hold the attributes of an appliance constant while upfront and operating costs are varied. A laboratory study allows these factors to be manipulated. Multiple trials and within-subject comparison means that preferences can be tracked as the manner in which operating costs are presented – and their proportionality to upfront costs – are manipulated.

Even within the narrow domain of appliance purchases, the energy efficiency gap is likely to be affected by multiple factors, and the translation of laboratory results to an intervention needs to account for these. For example, the success of the information intervention only with staff training (Kallbekken *et al.*, 2013) indicates that further social or attentional mechanisms contribute to the behaviour. Nevertheless, we argue that isolating and measuring generalisable mechanisms in a laboratory study is likely to make a direct and probably stronger contribution to developing effective policy by determining how universal psychological mechanisms affect behaviour.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

Choice of loans

This approach, in which hypothesised psychological mechanisms are investigated via repeated behavioural measures obtained across multiple conditions in the laboratory, was adopted in a recent study in our laboratory, which investigated how consumers choose personal loans (Lunn *et al.*, 2016). The experiments were funded by four economic regulators in Ireland[5] and designed to inform regulatory policy. The study was conducted in close cooperation with the Consumer Protection Directorate of the Central Bank of Ireland, which was concerned about consumer decision-making in the market for personal loans. We present a subset of results here to illustrate some of our core arguments about using laboratory experiments for policy.

Part of the study involved a choice experiment. A sample of consumers made multiple choices between pairs of loans, which varied in term and interest rate [annual percentage rate (APR)]. They were instructed to choose which of the two loans they preferred, assuming that they had to make monthly repayments out of regular income. The experiment was incentivised such that participants were rewarded for consistent choices. The usual questions of external validity apply, but the laboratory setting permitted exhaustive manipulation of the interest rates, terms of the loans and any additional information presented. Of particular interest was whether and how decisions are affected by the presence or otherwise of explicit information on the financial cost of the loans (the cost of credit). European regulations require providers to specify a representative example that shows how the key variables of a credit product are related, but there is no obligation to display the financial cost of each offering. In practice, providers vary with respect to whether, with what prominence and at which decision points they provide financial cost information. The research question of interest to policy was whether this variation in explicitness of information affects consumers' choices.

Figure 1 displays a subset of the results for two conditions when the pair of loans differed in term by one year. The vertical axis shows the probability of opting for the longer of the two loans, all else equal. The solid line relates to a 'monthly repayment' condition in which both offerings contained three pieces of information: term, APR and monthly repayment. The dashed line relates to a 'full information' condition in which financial cost information was provided explicitly. Although the product offerings were the same, whether the financial cost was implicit or explicit had a strong impact on

5 The Central Bank of Ireland, the Competition and Consumer Protection Commission, the Commission for Energy Regulation and the Commission for Communications Regulation.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

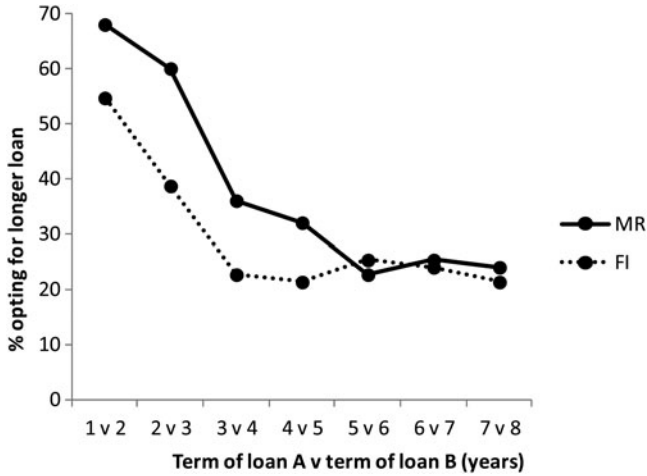

Figure 1. Results of a choice experiment on loans. Provision of explicit financial cost information substantially reduces the probability of opting for the longer of two loans, especially for loan terms of five years or less. FI = 'full information' condition; MR = 'monthly repayment' condition.
Source: Lunn *et al.* (2016).

choices. Explicit financial cost information pushed consumers towards shorter loans.

The data suggest likely properties of the psychological mechanisms underpinning the result. Choosing a length of loan requires consumers to trade off burdensome monthly repayments against financial cost. The former decrease with longer terms, while the latter increases. The relationship is strongly non-linear in the range where we recorded the greatest effect on choices. Opting for a shorter loan can reduce financial cost substantially with a relatively modest effect on monthly repayments. The data imply that, unless the trade-off is made explicit, consumers fail to appreciate this non-linearity. Based on this insight into the mechanism, a second experiment designed and tested a potential nudge that consisted of an example table designed to make consumers notice the non-linear nature of the relationships. Exposure to the table successfully reduced the discrepancy between choices in different conditions.

Consider these findings in the context of the familiar argument about generalising from laboratory experiments. In the loans study, participants voluntarily signed up, were incentivised to respond consistently, were able to pay full attention and saw key information only, free from extraneous marketing material and other persuasive influences. Despite these advantageous

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

circumstances, choices were altered by the primary experimental manipulation, with a large effect size that was systematically related to other experimental variables. The implication, in our view, is that the experiment isolated a psychological mechanism that is likely to operate outside the laboratory. This inference regarding generalisation of the experimental results may be contrasted with the inference one might make regarding the nudge. The example table was successful in the laboratory, further suggesting that difficulty appreciating non-linear relationships affects choices of personal loan, but might or might not be a useful policy intervention. In this case, the familiar refrain about generalising from laboratory experiments applies, since mandating such a table would have little effect if it were insufficiently prominent, failed to attract attention, became swamped by other information and so on.

This study is not an isolated example. A similar logic underpins work on price frames conducted by the UK Office of Fair Trading (Huck & Wallace, 2010). Relative differences in consumer decisions across multiple price frames revealed which frames generated the most disadvantageous purchases. Drip pricing (starting with a base price and adding additional charges at later points) emerged as a price frame of particular concern. This finding led to a change in UK regulations on surcharges. Like the loans study, the research was conducted directly for policy-makers and illustrates circumstances in which laboratory research is likely to generalise, because if consumers make errors or alter decisions on the basis of how information is presented in an idealised environment in which they are paying attention, they are unlikely to do better in the real world. Such studies tap into the limits of cognitive capability.

Researchers and policy-makers who consider studies like these may take different positions on the robustness of the design, the effectiveness of the incentive and the strength of the evidence for the policy. For present purposes, the studies illustrate our main arguments. It would probably be impossible to isolate the same psychological mechanisms through field experiments. In the case of loans, even if a loan provider were willing to cooperate, the control over information provision and variation in offerings and choices required would be impossible to implement with sufficient statistical power. Ethical approval would be hard to come by, since real choices involving financial risk could be influenced in potentially disadvantageous directions. Suppose these barriers were overcome and, following a successful field experiment, other researchers formed an alternative hypothesis about the mechanisms involved. Precise replication or manipulation of the original experimental design would probably be impossible, confounding the test. In contrast, the experimental control and statistical power offered by the laboratory allows potential mechanisms to be isolated and tested, repeatedly if necessary.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

The loans study also shows how laboratory studies can be generalisable beyond the specific policy context. Consumers remained inclined to give greater weight to explicit (compared to implicit) information even once they had substantial experience of the relationships between the key variables of the loans. This echoes other laboratory research on judgement that shows how cue fluency, defined as ease of processing, increases cue weights in multiple cue judgements (Shah & Oppenheimer, 2007). This mechanism, which would again be difficult to isolate outside the laboratory, may be of interest to multiple consumer policy areas.

Choosing the right tool

In deciding which method to use, the starting point is the empirical research question. The questions policy-makers ask are different from those scientists ask. For scientists, the research question is, generally, whether a hypothesis holds and, consequently, in which direction the weight of evidence for a theory is tipped. A well-designed experiment identifies and tests the hypothesised effect, distinguishing between alternative theories. For policy-makers, the research question is how to increase desirable and decrease undesirable outcomes, and perhaps how best to identify and measure these outcomes. Input from behavioural science might consist of theories offering candidate answers to the research question, in addition to existing policies and practices in other jurisdictions or domains that are built explicitly or implicitly on behavioural theories. In many cases, existing empirical evidence may be limited.

This difference between research questions is crucial. Suppose a policy-maker has a defined behavioural outcome, such as reducing teenage smoking or encouraging investment in home insulation. One option is to exploit the existing body of findings in behavioural science and, perhaps in collaboration with researchers offering 'behavioural insights', to select some candidate behavioural interventions based on a list of potentially relevant phenomena (e.g. Dolan *et al.*, 2010). The most promising of these off-the-shelf interventions can be tested in field trials, then adopted, adapted or discarded according to results. This is how the UK Behavioural Insights Team and its emulators generally operate (Behavioural Insights Team, 2012b, p. 2). Success depends on the strength and generalisability of the mechanisms behind candidate interventions. Much depends, therefore, on the initial selection of ideas. If mechanisms are insufficiently generalisable and powerful, resources and time can be spent producing null or contradictory results and chasing small effect sizes. As described in the 'Some examples' section, this characterises field trials of interventions intended to narrow the energy efficiency gap. Where mechanisms are not well understood, field trials can be wasteful shots in the dark.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

To avoid this, policy-makers need to consider the research question carefully. Are they in a position to ask whether a candidate policy will produce a desired behaviour, rather than needing to ask why an observed pattern of behaviour arises? In short, is the right research question 'whether' or 'why'? If the policy problem is clearly diagnosed and the causal mechanisms underlying behaviour are sufficiently understood, the question may be of the 'whether' type, and field trials (including RCTs) of candidate interventions might be appropriate. If the mechanisms are not well understood, the most informative question to ask might be of the 'why' type, in which case laboratory research may be most appropriate.

There are many policy areas where the right research question for policy-makers to ask is not whether but why. In these cases, behaviourally informed policy drawing on multiple relevant theories can be developed via advanced knowledge of the behavioural literature and then trialled. Yet without the generation and testing of specific hypotheses, this approach is unlikely to answer the question of why the relevant behaviour occurs. As described in detail in the 'Generalisability of field experiments' section, if the answer to whether the policy works is positive, the relative contribution of each theory or established mechanism will not be clear and generalising from the result will be problematic; if the answer is negative, specific theories and mechanisms will not be ruled out.

The superior experimental control offered by the laboratory makes it a powerful method for addressing 'why' questions. Consider concerns about households taking on high-interest credit products. We can investigate potential mechanisms with simple laboratory experiments, say by testing whether a sample of householders who take on high-interest credit products understand the compounding of interest payments and the fee structure of credit products. This is important not only for considering possible regulatory policy responses, but also for understanding whether choices observed in the market are truly disadvantageous (i.e. negatively influenced by modes of information provision), rather than being the result of genuine preferences. The research in the 'Choice of loans' section succeeds in demonstrating this via a within-subject design that is not possible in field research. To base good policy on the premise that decisions are disadvantageous, the psychological mechanism driving those decisions needs to be understood. Why the behaviour occurs matters.

As another example, Crosetto *et al.* (2016) asked why many consumers fail to make healthy food choices despite the provision of nutritional information. Participants in their laboratory study were incentivised to purchase a daily menu that met predetermined nutritional goals in an online shopping environment. Laboratory control allowed for systematic variation in the nutritional

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

label and the time allowed for the task, so that information-processing mechanisms could be tested. When participants had limited time to shop, they found it easier to satisfy the nutritional goals if labels were of the simple traffic light type than if labels displayed more complete information on guideline daily amounts. This is an important finding for policy, because it implies that the integration of information from guideline daily amount labels is too complex for consumers to purchase food for a balanced diet, even when focused on doing so.

Although we argue that the laboratory environment is helpful for addressing 'why' questions, there are contexts in which 'whether' questions should be asked in the laboratory and 'why' questions in the field. In some cases, the mechanisms being targeted are necessarily environmental, and a field test might be necessary to rule out a particular mechanism. For example, one might attempt to answer why poor diets are common in low-income areas by implementing an extreme form of grocery store subsidies for fruit and vegetable purchases by providing free fruit and vegetables to a sample of families (Ludwig *et al.*, 2011). If this extreme test produces null results, it suggests that availability is not the answer to the 'why' question. Furthermore, to streamline an energy-saving intervention, one might ask why smart meters work – is it learning about energy use or the saliency of real-time feedback (Lynham *et al.*, 2016)? To answer this question, medium- or long-term energy-saving behaviour in response to smart meters needs to be measured, and therefore a field study or RCT is necessary.

Laboratory tests of information-based consumer interventions (e.g. mandated disclosures, warnings and advice pages) can answer 'whether' questions. If the intervention does not alter within-subject choices in the desired direction for an incentivised sample of consumers who are able to pay full attention to their decision, then it is unlikely that the intervention will be effective outside the laboratory. However, these laboratory-'whether' and field-'why' questions are special cases that arise for particular kinds of mechanism and policy.

Ideally, laboratory and field studies should complement one another. Studies of discrimination in hiring are a case in point. Field experiments that use correspondence tests, in which group membership is signalled by the name on otherwise identical job applications, have revealed the extent of discrimination (Riach & Rich, 2002). However, they provide little insight into why discrimination remains so strong. Rooth (2010) conducted a correspondence test and enlisted the same recruiters to undertake a subsequent laboratory test of implicit stereotyping. The results showed that the probability of hiring ethnic minority candidates was strongly related to implicit stereotypes, but not explicit attitudes. This insight about the psychological mechanism is important for policy, as interventions to measure and combat implicit stereotypes differ from other anti-discrimination policies (Hardin & Banaji, 2013).

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

Although laboratory studies permit the greatest experimental control, field experiments and other methods can be used to test mechanisms, and we do not wish to imply otherwise.[6] Nevertheless, we contend that the nature of the key research question for policy, in particular how often it is a 'why' rather than a 'whether' question, ought to imply greater use of the laboratory for generating evidence for policy.

Conclusions

This article began by acknowledging the great strides that have been made in deploying behavioural science for policy use in recent years. This is positive – it implies greater use of empirical evidence in the design, development and evaluation of policies, as well as a move towards more realistic models of behaviour and away from highly generalised and often inaccurate microeconomic models. However, despite the successful application of behavioural insights to an increasing range of policy problems, the argument presented here is, in part, a critique. Our concern is that one approach, in which off-the-shelf behaviourally informed interventions are selected for field trials, may be too dominant in the use of behavioural science for policy. This approach is often supported by the claim that field experiments, especially RCTs and similar field trials, produce evidence that is more applicable to the real world than laboratory studies.

Our aim has been to offer a focused overview of how scientific results generalise to policy contexts. With reference to existing arguments and to recent examples from the literature, we have tried to show how and why laboratory experiments can be good – and often the best – ways to provide evidence for policy. The narrowness of scope of field trials means that results can be specific to the context of the experiment and difficult to generalise to other contexts and policy areas. In contrast, while accepting that environmental validity is always an issue, the experimental control offered by laboratory experiments allows researchers to isolate and test psychological mechanisms that are likely to generalise to multiple contexts. Candidate mechanisms can also be assessed in beneficial circumstances in the laboratory, such that failure to produce a successful outcome renders it highly unlikely that the mechanism can form the basis for a successful policy.

It is uncontroversial to conclude that research methods should complement each other, with the methods chosen in different cases being appropriate to the

6 Good examples in the domain of discrimination are List (2004) and Gneezy *et al.* (2012), who use a combination of field and laboratory experiments to tease apart statistical and taste-based discrimination in the marketplace.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

nature of the research question. The argument we have provided here was designed to provoke greater thought regarding how to put these conclusions into practice. Based on this argument, we contend that increased use of laboratory studies to investigate the behavioural mechanisms underlying issues of direct interest would be of benefit to policy-makers. The appetite for applying behavioural science to policy that has emerged since the lament of Amir *et al.* (2005) offers hope that further benefits can be reaped from well-designed laboratory research for policy.

Acknowledgements

The authors would like to thank Paolo Crosetto, Stefan Hunt, Caitríona Logue, Féidhlim McGowan and Jason Somerville for helpful comments, feedback and discussion, as well as a seminar audience at the Irish Behavioural Science and Policy Network.

References

Allcott, H., and M. Greenstone (2012), 'Is there an energy efficiency gap?', *The Journal of Economic Perspectives*, **26**(1): 3–28.

Allcott, H., and T. Rogers (2014), 'The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation', *The American Economic Review*, **104** (10): 3003–3037.

Allcott, H., and D. Taubinsky (2015), 'Evaluating behaviorally motivated policy: experimental evidence from the lightbulb market', *The American Economic Review*, **105**(8): 2501–2538.

Amir, O., D. Ariely, A. Cooke, D. Dunning, N. Epley, U. Gneezy, B. Köszegi, D. Lichtenstein, N. Mazar, S. Mullainathan, D. Prelec, E. Shafir and J. Silva (2005), 'Psychology, behavioural economics, and public policy', *Marketing Letters*, **16**: 443–454.

Behavioural Insights Team (2012a), *Test, Learn, Adapt*, London: Cabinet Office.

Behavioural Insights Team (2012b), *Annual Update 2011–2012*, London: Cabinet Office.

Boudet, H., N. M. Ardoin, J. Flora, K. C. Armel, M. Desai and T. N. Robinson (2016), 'Effects of a behaviour change intervention for Girl Scouts on child and parent energy-saving behaviours', *Nature Energy*, **1**: 16091.

Camerer, C. (2011), *The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List* (SSRN 1977749).

Carroll, J., S. Lyons and E. Denny (2014), 'Reducing household electricity demand through smart metering: The role of improved information about energy saving', *Energy Economics*, **45**: 234–43.

Cartwright, N. (2007), 'Are RCTs the gold standard?', *BioSocieties*, **2**(1): 11–20.

Cartwright, N. and J. Hardie (2012), *Evidence Based Policy: A Practical Guide to Doing it Better*, Oxford: Oxford University Press.

Crosetto, P., L. Muller and B. Ruffieux (2016), 'Helping consumers with a front-of-pack label: Numbers or colors? Experimental comparison between Guideline Daily Amount and Traffic Light in a diet-building exercise', *Journal of Economic Psychology*, **55**: 30–50.

Deaton, A. (2010), 'Instruments, Randomization, and Learning about Development', *Journal of Economic Literature*, **48**: 424–455.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

Deutsch, M. (2010), 'Life cycle cost disclosure, consumer behavior, and business implications', *Journal of Industrial Ecology*, **14**: (1), 103–120.

Dolan, P., M. Hallsworth, D. Halpern, D. King and I. Vlaev (2010), *MINDSPACE: Influencing behaviour through public policy*, London: The Cabinet Office/Institute for Government.

Duflo, E. and M. Kremer (2005), 'Use of Randomization in the Evaluation of Development Effectiveness', in O. Feinstein, G. K. Ingram and G. K. Pitman (eds), *Evaluating Development Effectiveness*, New Brunswick, New Jersey and London: Transaction Publishers.

Feigenson, L., S. Dehaene and E. Spelke (2004), 'Core systems of number', *Trends in Cognitive Sciences*, **8**(7): 307–314.

Gneezy, U., J. List and M. K. Price (2012), *Toward an Understanding of why People Discriminate: Evidence from a series of Natural Field Experiments*. NBER Working Paper 17855.

Hardin, C. D. and M. R. Banaji (2013), 'The Nature of Implicit Prejudice', in E. Shafir (ed.), *The Behavioral Foundations of Public Policy*, Princeton NJ: Princeton University Press, 13–31.

Harrison, G. W. and J. L. List (2004), 'Field Experiments', *Journal of Economic Literature*, **42**(4): 1009–1055.

Haynes, L., B. Goldacre and D. Torgerson (2012), Test, learn, adapt: developing public policy with randomised controlled trials. London: Cabinet Office.

Heinzle, S. L. (2012), 'Disclosure of energy operating cost information: A silver bullet for overcoming the energy-efficiency gap?', *Journal of Consumer Policy*, **35**(1): 43–64.

Henrich, J., S. J. Heine and A. Norenzayan (2010), 'Beyond WEIRD: Towards a broad-based behavioral science', *Behavioral and Brain Sciences*, **33**(2–3): 111–135.

Huck, S., and B. Wallace (2010), *The impact of price frames on consumer decision making* (Office of Fair Trading 1226).

Jaffe, A. B., & R. N. Stavins (1994), 'The energy-efficiency gap What does it mean?', *Energy Policy*, **22**(10): 804–810.

Kallbekken, S., Sælen, H., & Hermansen, E. A. (2013), 'Bridging the energy efficiency gap: A field experiment on lifetime energy costs and household appliances', *Journal of Consumer Policy*, **36**(1): 1–16.

Kessler, J., and L. Vesterlund (2014), 'External Validity of Laboratory Experiments: The Misleading Emphasis on Quantitative Effects'. in G. Frechette and A. Schotter (eds), *The Methods of Modern Experimental Economics*, Oxford: Oxford University Press.

Levitt, S. D., and J. A. List (2007a), 'Viewpoint: On the generalizability of lab behaviour to the field', *Canadian Journal of Economics/Revue canadienne d'économique*, **40**(2): 347–370.

Levitt, S. D., and J. A. List (2007b), 'What do laboratory experiments measuring social preferences reveal about the real world?', *The Journal of Economic Perspectives*, **21**(2): 153–174.

List, J. (2004), 'The nature and extent of discrimination in the marketplace: evidence from the field', *Quarterly Journal of Economics*, **119**: 49–89.

Ludwig, J., J. R. Kling and S. Mullainathan (2011), 'Mechanism experiments and policy evaluations', *The Journal of Economic Perspectives*, **25**(3): 17–38.

Lunn, P. D. (2014), *Regulatory Policy and Behavioural Economics*, OECD Publishing.

Lunn, P. D., M. Bohaçek and A. Rybicki (2016), *An experimental investigation of personal loan choices*, Dublin: ESRI.

Lynham, J., K. Nitta, T. Saijo and N. Tarui (2016), 'Why does real-time information reduce energy consumption?', *Energy Economics*, **54**: 173–181.

Newell, R. G., and J. V. Siikamäki (2014), Nudging energy efficiency behavior: The role of information labels, *Journal of the Association of Environmental and Resource Economists*, **1**(4): 555–598.

Riach, P. and J. Rich (2002), 'Field experiments of discrimination in the market place', *The Economic Journal*, **112**: F480–F518.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press

Rooth, D-O. (2010), 'Automatic associations and discrimination in hiring: real world evidence', *Labour Economics*, **17**(3): 523–534.

Shah, A. K. and D. M. Oppenheimer (2007), 'Easy does it: The role of fluency in cue weighting', *Judgment and Decision Making*, **2**(6): 371–379.

van Bavel, R., B. Hermann, G. Esposito and A. Proestakis (2013), *Applying Behavioural Sciences to EU Policy-making. JRC Scientific and Policy Report*, Brussels: European Commission.

Whalen, J., C. R. Gallistel, and R. Gelman (1999), 'Nonverbal counting in humans: The psychophysics of number representation', *Psychological Science*, **10**(2): 130–137.

https://doi.org/10.1017/bpp.2016.6 Published online by Cambridge University Press