Financial Conduct Authority

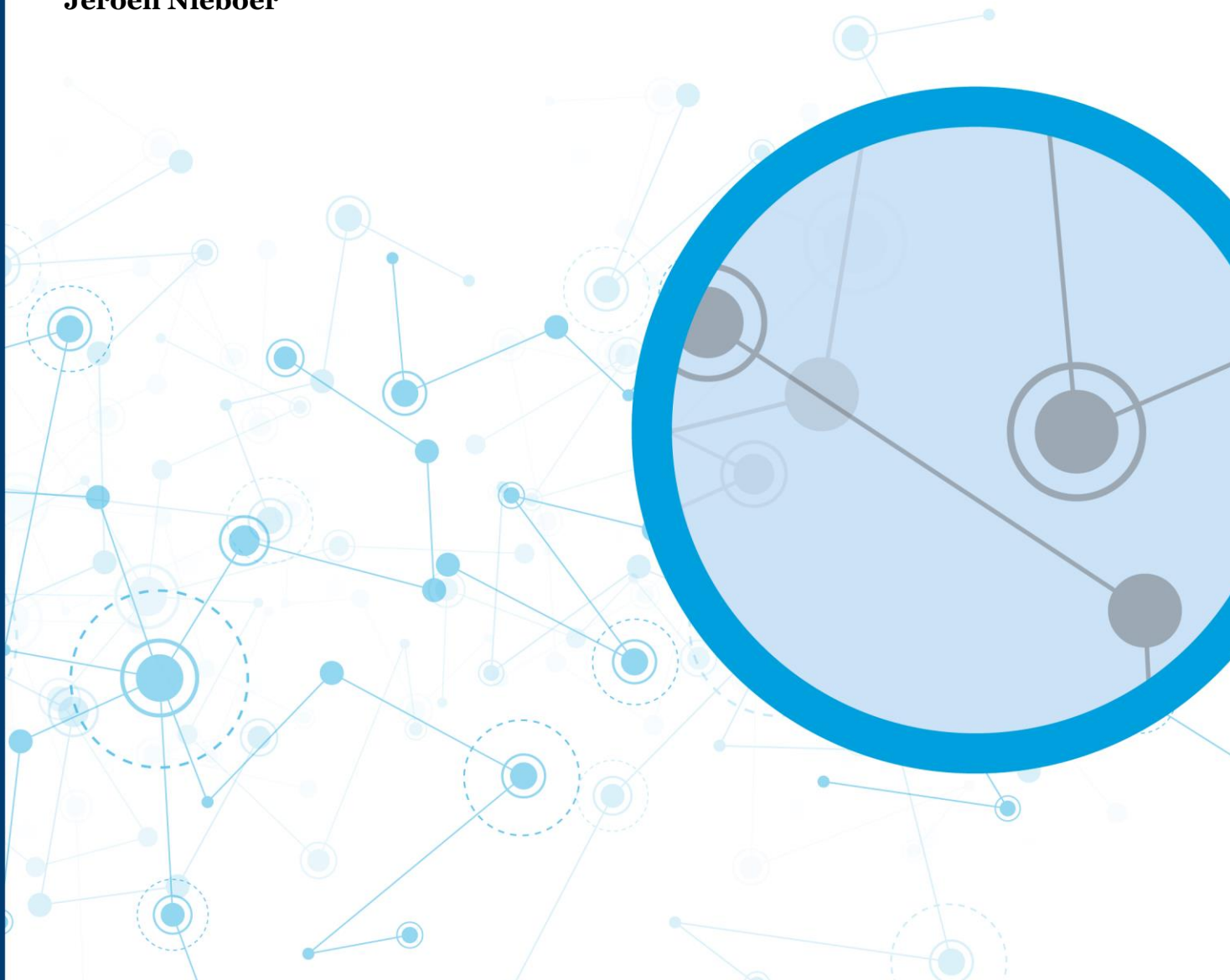
51

Occasional Paper

February 2020

Using online experiments for behaviourally informed consumer policy

Jeroen Nieboer



FCA occasional papers in financial regulation

The FCA occasional papers

The FCA is committed to encouraging debate on all aspects of financial regulation and to creating rigorous evidence to support its decision-making. To facilitate this, we publish a series of Occasional Papers, extending across economics and other disciplines.

The main factor in accepting papers is that they should make substantial contributions to knowledge and understanding of financial regulation. If you want to contribute to this series or comment on these papers, please contact Kevin James (kevin.james@fca.org.uk) or Karen Croxson (karen.croxson@fca.org.uk)

Disclaimer

Occasional Papers contribute to the work of the FCA by providing rigorous research results and stimulating debate. While they may not necessarily represent the position of the FCA, they are one source of evidence that the FCA may use while discharging its functions and to inform its views. The FCA endeavours to ensure that research outputs are correct, through checks including independent referee reports, but the nature of such research and choice of research methods is a matter for the authors using their expert judgement. To the extent that Occasional Papers contain any errors or omissions, they should be attributed to the individual authors, rather than to the FCA.

Authors

Jeroen Nieboer works in the FCA's Economics and Design Unit and is a visiting fellow at the London School of Economics.

Acknowledgements

The author would like to thank Paul Adams, Nafa Ben Haddej, Chris Burke, Alexandra Chesterfield, Karen Croxson, Adam Giles, Benedict Guttman-Kenney, Lucy Hayes, Stefan Hunt, Zanna Iscenko, Jesse Leary, Daniel Mittendorf, Gianandrea Staffiero and other current and former colleagues at the FCA for their insightful comments. This paper has also benefited from inspiring conversations with Colin Camerer, Matteo Galizzi, Anouar El Haji, Andrew Ivchenko, Joana Sousa-Lourenco, Pete Lunn, Eric Johnson, Janna Ter Meer and Shane Timmons. The FCA is grateful to Nick Bardsley for his academic review of the paper.

All our publications are available to download from www.fca.org.uk. If you would like to receive this paper in an alternative format, please call 020 7066 9644 or email publications_graphics @fca.org.uk or write to Editorial and Digital Department, Financial Conduct Authority, 12 Endeavour Square, London E20 1JN.

Executive summary

In this paper, we discuss the merits of online experiments for behaviourally informed policy making. The online experiments we have in mind combine elements of laboratory and natural field experiments (RCTs), with a focus on external validity. These experiments are increasingly used to investigate consumer behaviour in different settings, including the testing of prospective consumer policies.

Most of these experiments have focussed on the demand side of consumer markets, for reasons we discuss in the paper. Although many demand-side policy interventions in recent years have been so-called behavioural interventions or *nudges*, prospective policies tested in an experiment do not have to be. An experiment may test a policy that reduces search cost or provides information, for example. The policy decision will still be behaviourally informed in the sense that the policy was tested with real people in an environment designed to replicate the real world.

Designing an experiment

After establishing that online experiments are well-suited to certain types of consumer policy research, we discuss how to design such experiments. We pay special attention to the key challenges that the policy researcher encounters when designing an experiment: recruiting and rewarding participants, structuring the experimental task and deciding whether to incorporate elements that diverge from the target field context ('abstractions') into the design. Drawing on examples from recent FCA research and the wider literature, we discuss the rationale behind key design choices and explain how these choices affect the interpretation of the experimental findings.

Experimental validity

We also review the academic literature on experimental validity and reflect on its implications for online experiments by policy researchers. We consider internal validity with respect to the relevant policy question as a necessary condition for conducting an experiment, whereas external validity is a harder-to-assess criterion that the policy researcher should nevertheless aim to maximise. We recognise it may not always be possible, or desirable, to fully replicate the field setting. When some aspects of the field cannot be replicated, the policy researcher needs to carefully consider whether departing from reality on these aspects might affect the generality of the experimental findings.

Where next?

We believe that online experiments have good prospects to become an important part of the policy research toolkit and that their use will grow in years to come. We hope this paper serves as a useful guide to the policy researcher looking to design an online experiment. As a quick reference for the researcher, we have summarised the design process of an online experiment in 10 key steps – see **Box 1** on the next page.

Box 1: Online experiments for consumer policy in 10 steps

1. Recruiting participants (p. 14)

Convenience sample or representative? Targeted recruitment of consumers based on market research? Use pre-screening questions?

2. Defining the task (p. 17)

Look at market research and consumer journey. What is the nature of the task? Where does the task take place in the field? Should certain parts of the task be excluded?

3. Deciding what information to provide (p. 17)

Make information available to sufficiently motivated participants. Avoid misinformation. Legal or regulatory disclosure? Industry standards? Can consumers use heuristics?

4. Setting situational parameters (p. 18)

Which situational parameters matter for consumer welfare? Demographics? Any market research on situational parameters? Vignettes, profiles or real parameters?

5. Selecting outcome variables (p. 19)

Which variables are important for consumer welfare? Choice, valuation, consumer understanding? Which choice process and demographic variables?

6. Constructing the choice set (p. 20)

Remove or control certain product attributes? Impose more structure on the choice problem (inferring or imposing preferences, counting mistakes)?

7. Rewarding participants (p. 21)

May be partly determined by recruitment options. Are participants sufficiently motivated by hypothetical rewards? Do task-related rewards narrow the research question?

8. Designing for experimental control (p. 13, p. 23)

Ensure no contamination between treatments. Measure and address attrition. Instruct clearly and gather early data on participant understanding. Is distraction a problem? Do further controls simply reduce noise? Do further controls make the choice environment artifical?

9. Assessing abstractions and experimental validity (p. 15, p. 23)

Ensure internal validity. Gather a large enough sample. Is there any field data to compare against? How might any abstractions affect external validity?

10. Communicating your findings

Describe participant sample and recruitment. Use screenshots to show similarity to field. Explain role of situational parameters. Focus on key outcome variables.

Contents

1 Overview

Purpose

A key focus of applied behavioural economics has been testing in natural environments, leading to policy that is 'behaviourally informed' rather than based on implausible assumptions on human rationality. In this paper, we consider how the goal of behaviourally informed consumer policy can be supported by using online experiments.

The kind of online experiment we have in mind can be seen as an intermediate between a conventional laboratory experiment and a field experiment. It will be designed with a focus on external validity. Since the design of such experiments requires many judgement calls (participant recruitment, payment, task design, etcetera), we have written this paper as a guide. Drawing on a range of examples, we discuss the choices faced by policy researchers at different stages of experimental design. Along the way, we identify some important outstanding questions on the generalisability of consumer behaviour. We also briefly review the literature on experimental validity, within in the context of design choices faced by the policy researcher.

Since the FCA and other policy making bodies regularly make use of online experiments, we hope that the discussion herein sparks further debate among practitioners on best practice and the value of experimental evidence. We believe that online experiments are valuable research tools for policy, provided they are carefully designed to answer the policy question of interest. By having open and honest conversations about the strengths and weaknesses of our research designs, we will be able to design better policy for consumers.

Audience

This paper is intended, first and foremost, for the policy researcher looking to design an online experiment and/or decide whether an online experiment will adequately answer a question of policy interest. It may also be of interest to policymakers deciding whether to commission such work, or academic scientists wishing to investigate aspects of consumer behaviour with potential policy implications.

2 Online experiments for policy

Defining online experiments

In the kind of experiment we are concerned with in this paper, volunteer participants are recruited from, and participate in, an online environment. These experiments are in some sense an evolution of laboratory experiments in social science, in which participants are recruited from a physical environment and come to a physical space for participation (e.g. a room on a university campus). Because the online environment broadens the participant pool, one of the obvious attractions of online experiments is that the researcher can sample from a larger, more representative population.

Table 1 shows an adaptation of the social science experiments taxonomy by Harrison and List (2004), which helps us think more systematically about the contributions of online experiments. Of the four types of experiment in this taxonomy, *framed field experiments* best describe the online experiments commonly used in consumer policy research. These experiments feature participants representative of decision makers in the field (e.g. car buyers) and mimic relevant features of the field context in ways that are recognisable to participants (e.g. showing car mileage statistics and engine specifications).

The language in the taxonomy seems to imply a sharp distinction between laboratory and other types of experiments. For our purposes, however, the significant distinction is between (i) contextualised on-line experiments with volunteer participants and (ii) field experiments with real market participants ("natural field experiments" in Table 1). The key question is: given our online participants are volunteers, are aware they are being observed and do not experience the real-world consequences of their decisions, how closely does their behaviour match what we might observe in a field experiment?

Table 1: Taxonomy of experiments

Type of experiment	Participants	Description
Conventional laboratory experiment	Volunteers (university students)	An abstract framing, an imposed set of rules.
Artefactual field experiment	Volunteers (representative)	An abstract framing, an imposed set of rules.
Framed field experiment	Volunteers (representative)	Field context in either the commodity, task and/or information set that the participants can use.
Natural field experiment (often called Randomised Controlled Trials or RCTs)*	Real market participants	Natural task environment with real consequences, participants often unaware they are in an experiment.

Notes: Adapted from Harrison and List (2004). * = Strictly speaking, every type of experiment is a Randomised Controlled Trial, but scientists and policy makers typically reserve this term for field experiments.

Why the comparison with field experiments? The latter have recently become an integral part of the policy researcher's toolkit and are now somewhat of a reference point for evidence-based policy. The reason is that there is no better way of measuring first-order policy impact than by randomly assigning representative recipients to a policy treatment or control group in the natural task environment. If the aim is to test the immediate impact of a proposed policy on consumers, and it is possible to conduct a field experiment that faithfully replicates the policy, this approach will provide the strongest evidence.

There are, however, situations in which a field experiment is not feasible or desirable. A recent FCA publication outlined some key conditions for a policy organisation deciding whether to conduct field experiments for testing prospective consumer policy (FCA, 2018). We focus here on whether a field experiment is "possible", "appropriate" and "proportionate" (FCA, p. 2). For various reasons, these thresholds may not be met:

- It can be practically impossible to conduct a natural field experiment for regulatory, legal, political, financial, technological or operational reasons. A special case is ethical concerns, for example due to the collection of sensitive data or the potential for harmful outcomes in one of the experimental treatments.
- There are cases where a natural field experiment is possible, but the policy maker fears it may not "sufficiently recreate the policy intervention on the relevant population" (FCA, p. 8). Perhaps macroeconomic conditions are a-typical, the experimental treatment cannot be delivered as the policymaker intended, or the policy intervention is thought to require certain support factors (e.g. sufficient adoption of new digital technology by consumers).[1]
- There may be methodological reasons for not conducting a natural field experiment, for example when a key variable cannot be measured, the same participant could end up in both treatment and control group and/or when there is likely to be interaction between participants in the control and treatment groups.
- Finally, there simply may be too much to test. If there are 15 candidate designs for a new consumer disclosure policy, it will be necessary to narrow the field before proceeding to test the designs in a field experiment.

In these cases, an online experiment may well be the policy researcher's best option for estimating the proposed policy's impact. However, the strength of evidence generated by such experiments depends on the design choices made by the researcher. A first-order concern is to design what Camerer (2015) calls "special features" of the experiment to more closely mimic the field context. In the next section of this paper, we discuss how.

Policy researchers typically do not have the luxury of defining the relevant field context for their experiment – it is defined by the policy question at hand. Paradoxically, this constraint often makes the policy researcher's job easier. When she is asked to evaluate the effect of risk warnings on the sale of structured investment products, the relevant field context is the natural environment in which retail investors typically learn about and purchase these products – say, an investment platform or an advisor-advisee

[1] It does not automatically follow that policy appraisal should now proceed without evidence from field experiments, since other methods may not yield better predictions for policy. One approach we would like to see more of is supplementing the field experiment with methods that plausibly increase the domain of the experiment's findings. A natural field experiment can be modified to generate the variance necessary for estimation of robust structural models, or a field experiment may be combined with an online experiment to cover a wider range of scenarios. See also the discussion in Cartwright and Deaton (2018).

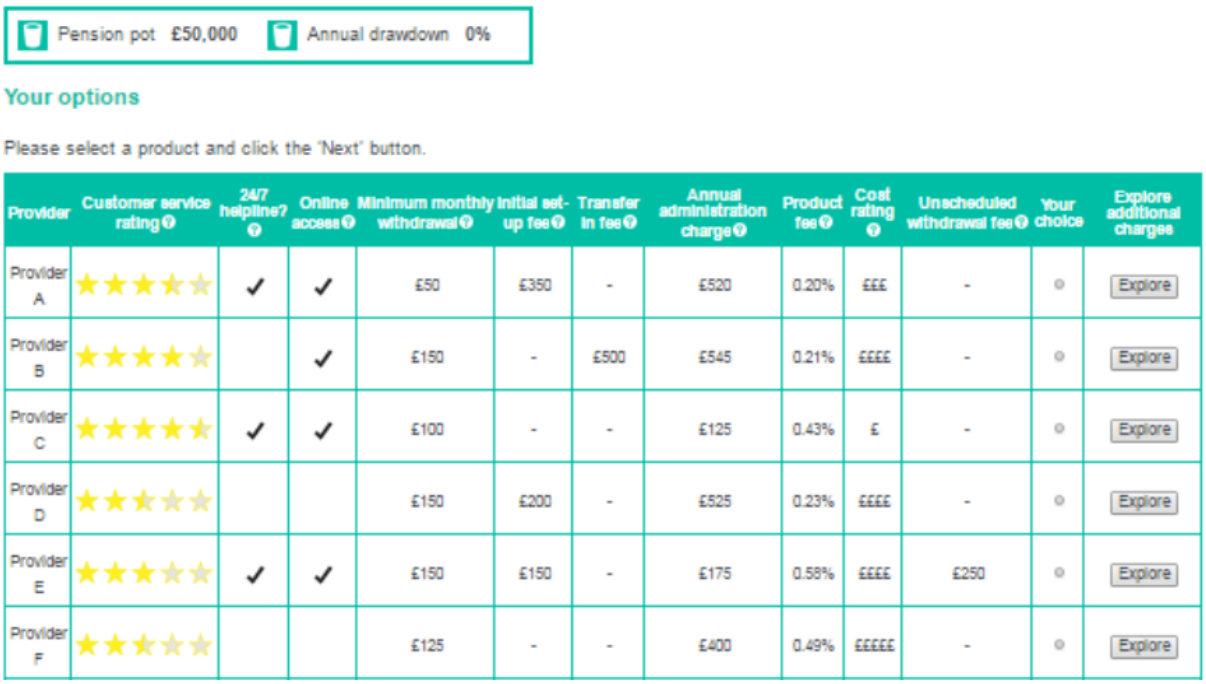
relationship. When she is testing different formats of calorie labels on food packaging, the relevant field context is the outlet where the food is sold, etcetera.

Examples of online experiments for policy

At this point, it might be helpful to give some examples of online experiments used for consumer policy. The FCA has conducted several of these in recent years. We will refer to three of these as case studies throughout this paper:

1. **Add-on insurance:** The first such experiment was Iscenko et al.'s (2014) study on selling insurance as an add-on product. The study measures shopping around and sub-optimal choices by consumers in different scenarios, with a focus on when and how in the purchase process the insurance product was presented to the consumer. Rather than test a regulatory policy, the experiment was intended to diagnose how practices in the market were affecting consumer outcomes.
2. **Income drawdown products:** In 2017, the FCA conducted a policy testing experiment on summary cost metrics for income drawdown products (Oxera, 2017). Income drawdown products are contracts for pension decumulation with complex cost structures, like annuities but with funds remaining invested until they are withdrawn. The FCA therefore conducted an experiment in which participants were asked to select the cheapest product, with different summary cost metrics to help them. The environment was designed to look like a price comparison website (see Figure 1).
3. **Investment funds:** More recently, the FCA ran a policy testing experiment on different risk warnings for investment fund charges (Hayes et al., 2018). The experiment was designed to provide evidence for policies to help draw consumer's attention to fund charges. The experiment tested several forms of disclosure in an environment designed to mimic an online investment platform.

Figure 1: Screenshot of FCA experiment on income drawdown products



Source: Oxera (2017, p. 7).

These are not the only online experiments conducted by the FCA, however. In 2016, the FCA conducted an online experiment on prompting consumers to shop around for annuities (Oxera, 2016). Smart (2016) reports two policy testing experiments that tested designs for payday loan price comparison websites and information prompts to encourage shopping around for annuities. Ter Meer et al. (2018) compare different ways of describing investment pathways for pensions. Mullett et al. (2018) test risk warnings on social media advertisements for financial products. In addition, the FCA conducted discrete choice experiments (also known as conjoint analysis) on consumer preferences for its mortgage and investment platform market studies (FCA, 2018ab).

Looking at other policy makers use' of online experiments for consumer policy, we note that financial decision-making features heavily. An early example is the European Commission's experiments on self-directed and advised investment decisions, in which participants made decisions between pairs of investment products (Chater et al., 2010). For advised investments, Chater et al. also tested several types of disclosure on advisor commissions. Another test of investment disclosure is the German Ministry of Finance's online experiment on unregulated investment risk warnings, in which participants choose a proportion of a hypothetical lump sum to invest (Artinger et al., 2018). More recently, the European Commission conducted experiments that measured consumers comparing and choosing personal current accounts and loans (European Commission, 2019). The research also tested several potential policy remedies, such as shopping around prompts, warnings, information disclosure and cost calculators.

Other domains of consumer decision-making featured in online experiments by policymaking bodies are the energy use of electrical appliances (German Government - Artinger et al., 2017), the impact of carbon emissions on vehicle choice (European Commission - Codagnone et al., 2016), hospital quality (King's Fund - Boyce et al., 2010) and online gambling (European Commission - Codagnone et al., 2014) and consumers' understanding of Terms and Conditions (Behavioural Insights Team, 2019a). Another promising use of online experiments is to select the most promising treatments for inclusion in natural field experiments, for example for utility switching trials (Behavioural Insights Team, 2019b).

Types of consumer policy research

What differentiates consumer policy research from other types of research? The answer is straightforward: the key ingredient is consumer welfare. Either directly or indirectly, consumer policy research is motivated by some concern for consumer welfare (which may also be described as consumer outcomes, wellbeing or harm).

The main concern in the three FCA experiments introduced earlier, for example, is that consumers are paying too much for their insurance, drawdown product or investment funds. In other domains, the policy maker may believe that consumers' welfare evaluations need to be adjusted for some missing costs or benefits. The policymaker may be worried about externalities, as in the case of consumers buying polluting vehicles. Or the policymaker may be worried that consumers are not properly taking significant risks into account, as in the case of online gambling and unregulated investments.

Experiments can play an important role in measuring consumer welfare, through replicating key aspects of the natural choice environment and, crucially, the ability to

compare welfare across different scenarios. By a *scenario*, we effectively mean any combination of factors that may affect a consumer's decision: consumer education, product characteristics, information provided, choice architecture, laws, regulatory constraints, etcetera.

It will be helpful to think about consumer policy research as breaking down into two broad categories: **description and diagnosis** (how do consumer make decisions in markets?) and **policy testing** (how would a mandated change in scenario (a policy) affect consumers' decisions in markets?). We now discuss these in turn.

Description and diagnosis

Descriptive and diagnostic research attempts to capture current consumer behaviour in markets. The most commonly used methods for this type of research are surveys, interviews, focus groups and collecting observational data (such as administrative records). Research can be entirely exploratory, in which case the researcher may not even compute a measure of consumer welfare but simply record what consumers (say they) are doing. Online experiments are rarely used for such exploratory research, although arguably they can be of use when the researcher suspects that survey responses would be affected by limited recall or reporting biases.[2] Unlike surveys, experiments capture participant responses without relying on participants' ability to predict their own behaviour in counterfactual settings – they instead rely on participants to act as if they would if they were placed in the counterfactual setting.

Experiments can be especially useful when the researcher wants to impose some structure on the problem, for example by checking whether observed behaviour matches a theoretical benchmark or consumers' stated preferences, so she can make inferences on consumer welfare. Experiments can also be used to measure responses to exogenous variation in contextual features in ways that may not possible with naturally occurring observational data (quasi-experiments, discontinuities in treatment, etcetera).

Because experiments also allow the researcher to observe behaviour in scenarios that do not exactly match the current market conditions, experiments provide estimates of consumer outcomes in counterfactual scenarios. This can be especially useful when the researcher is trying to disentangle the contribution of multiple factors that differ between two possible scenarios – a diagnostic problem. In the FCA's add-on insurance experiment, for example, the practice of selling insurance as an add-on product is experimentally deconstructed into (i) the presentation of the add-on product with or without easily accessible alternatives and (ii) timing of the add-on presentation (at the point of sale or earlier). Such exercises are especially useful when the researcher is trying to diagnose which institutions and practices in markets are driving lower consumer welfare.

Policy testing

A common challenge is to predict the effect of a new policy intervention. Many of the examples given earlier were tests of prospective consumer policy – typically a prompt or tool to help consumers choose 'better value' products. To conduct a valid policy test, the policy researcher must compare consumer welfare under each proposed policy regime to a plausible alternative scenario (typically the status quo). One way of making such

[2] For example, when it comes to choices made in the 'heat of the moment' such as gambling or drinking. Other alternatives to surveys are unobtrusive observation or ethnography, although obtaining large samples with these methods can be expensive.

comparisons is to survey consumers through individual interviews or focus groups. Such evidence may, however, be biased towards an optimistic assessment of the policy. Survey respondents may feel the need to respond positively to the policy initiative or may overestimate the likelihood they will use the information reported in the disclosure. Experiments can reduce such biases by representing the different scenarios explicitly as treatment and control conditions, within which participants directly engage with the task.

In essence, the policy regime being tested is much like any other scenario change outside the consumer's control. But the fact that the policy maker is contemplating intervention raises the bar for the standard of evidence required. The researcher testing a policy needs to be confident that she can conclusively detect differences in consumer welfare between the status quo and the prospective policy regime - under some reasonable assumptions about consumer preferences. This does not mean that the policy maker requires strictly positive differences in welfare; in some cases, ensuring that there are no negative welfare effects from a new policy is sufficient.[3]

Demand-side focus and 'nudges'

The examples given earlier in this section focus on the role of the demand-side in consumer markets: the decision maker in the experiment is the consumer and there is no strategic interaction with vendors, advisors or corporations. We believe this focus is partly driven by recent enthusiasm for "demand-side remedies" (Fletcher, 2016), itself in no small part due to the success of "nudges" (Thaler and Sunstein, 2008) and the rise of behaviourally informed policy making (Chetty, 2015; Shafir, 2013).

Does this mean that online experiments are only suited to test nudges or behavioural insights? Absolutely not. Although one of the pillars of behaviourally informed policy making is that policies need to be tested by observing real people make real decisions, the way we think about consumers' decisions may not necessarily be 'behavioural'. The consumers in the FCA experiment on add-on insurance may not shop around for alternatives because they have high search costs, for example. In the income drawdown experiment, the summary cost metrics may simply reduce the effort required by the consumer to calculate the cheapest option.

By focussing on the consumer only, the experiments we discuss here effectively take the supply side as given and do not attempt to capture market dynamics. This is a potential limitation that affects online experiments just as much as field experiments, although online experiments have the advantage of being able to present participants with alternative supply-side conditions that may represent the supply side's response to changing consumer behaviour. Could we go one better and have supply-side participants in the online experiment, interacting with consumers? This seems ambitious, given that many suppliers are large organisations with considerable infrastructure (technology, knowledge, governance, networks, etcetera) and incentives, not to mention ample time to formulate a response. It is unclear whether participants in an experiment would act as suppliers would, let alone whether outcomes thus obtained would represent the best possible estimate of a market equilibrium. This may be less of a concern for studying short-term interactions where the supply side's strategy space is arguably constrained (e.g. financial advisors recommending investments to captive clients).

[3] Consider the case where the policy in question is based on a broader public objective (for example, transparency, efficiency, fairness or looking after the most vulnerable in society) but effects on consumer decisions are not understood.

3 Designing an experiment

Imagine we have a consumer policy research question, the answering of which requires evidence on how consumers behave in different scenarios. Perhaps the question is one of measurement, such as finding out if inexperienced consumers have a harder time than experienced ones picking a suitable product. Or perhaps we want to predict the effect of a proposed policy, like a certification mark, a disclosure requirement or a tool to support consumers' decision-making. Now we can start thinking about whether an online experiment is the right tool for the job, and how we would design it.

Setting the scene

In the remainder of this section, we discuss the features of an experiment in more or less chronological order of execution. This does not mean that features should be chosen 'as we go along', however. Quite the opposite: it is important that the entire experiment is designed prior to execution of any of the steps. The relevance and validity of the design can only be judged on the overall package.

Lab or online?

Experiments can be conducted in a field location, a dedicated laboratory space or on-line. Field locations are rarely used for experiments with volunteer participants, although this may be the only way the researcher can find hard-to-recruit participants or catch participants in the right frame of mind.[4]

A more common dilemma is whether to conduct an experiment in a physical laboratory or through an online panel or platform. Are there any reasons to favour one over the other? If recruiting a large, diverse and/or representative consumer sample is important, then online platforms are cheaper and more convenient. In some cases, it may not even be possible to convince a large enough number of representative participants to travel to a physical laboratory. In fact, unless a physical laboratory is readily available and no online infrastructure has been set up, online experiments seem to offer lower costs and more convenience to policy researchers.

One reason the policy researcher may prefer a physical laboratory is because it offers greater control over participants' interactions with the task. The researcher can verify participants' identity, instruct participants directly and enforce rules such as the prohibition of communication between participants during the task (although participants may still communicate outside the laboratory). Another advantage of the physical environment is greater transparency, which can give greater credibility to probabilistic outcomes (for example, a physical device representing the resolution of an insurable event in an insurance choice experiment) and strategic interactions (for example, publicly announcing the set of available products that an advisor can recommend to a consumer).

[4] Examples of the former are the experiments with sports card traders by List (2001, 2003) or financial professionals by Cohn et al. (2014, 2015). An example of the latter, which is effectively the use of targeted recruitment to introduce participants whose information set matches that of the relevant field context, is the field experiment with sugar cane farmers by Mani et al. (2013).

Finally, a laboratory space allows participants to physically engage with people, objects and environments in ways that would not be possible online.

A unique feature of laboratory experiments is the physical presence of an experimenter. The experimenter does not have to be the researcher; in fact, it may aid replicability if these two roles are fulfilled by different people. But since someone will be physically present to facilitate the experiment and participants will see this person as the experimenter, we can ask if this makes a difference. Researchers have long been concerned that participants might alter their behaviour to align with the experimenter's presumed intentions, so-called demand effects (Bardsley, 2008; Zizzo, 2010; De Quidt et al., 2018). An intriguing empirical question is whether the size of this effect is a function of the experimenter's presence. Arguably, the experimenter's role is more conspicuous in a physical laboratory and the effect will therefore be stronger. The question could even be broadened to whether people who attend physical laboratories have different mind-sets than people registered for online consumer panels. Of course, only controlled experimental evidence can address these questions.

Taking back control in online experiments

Relative to the tranquillity of the physical laboratory, conducting experiments online can look like a risky business. Some of the perceived risks, particularly those stemming from less control of participant behaviour, can be mitigated by design.

An oft-heard concern is that bans on communication between participants are impossible to enforce in online studies. To make it less likely that participants communicate, the researcher can use unique participant invitation codes (or hyperlinks) instead of running an open platform. Combined with randomising invitations across time and geography, this drastically reduces the chance that two participants to the study will be able to communicate before or during the study. Especially if the researcher uses a large recruitment pool, the chances of two participants knowing each other is arguably lower than with subjects recruited on a university campus.

A second concern is that participants in on-line experiments are more likely to be distracted, or at least that their level of distraction is unobservable. This is a subtle issue, which may or may not matter to the researcher. First, a heightened probability of distraction may reflect consumers' decisions in the real world well. It may even be true that the level to which people let themselves be distracted (e.g. by not switching off their phone) corresponds to their level of distraction in everyday decision-making environments. Especially for testing policy remedies, it may therefore be desirable to allow for some distraction. Otherwise, how can the researcher know that the policy remedy will work outside the focussed silence of the laboratory? Second, the researcher can increase participants' motivation and focus by using incentives – monetary or otherwise. Third, the researcher can use certain outcome variables to remove observations of distracted participants from the dataset. It is common practice to remove those who complete the experiment too quickly, slowly, or inconsistently.

A related issue is that online experiments often have substantial rates of *attrition* - participants not completing the experimental task. If attrition is non-random then the experimental findings may be affected by attrition bias. The main concern here is correlation with the experimental treatments, but correlation with behaviour or demographics may also matter.

How likely is it that attrition is non-random? Arechar et al. (2018) provide the first evidence that, even though attrition rates differ between lab and online experiments on public goods games, attrition is random. Although encouraging, these findings do not guarantee that other online studies will not suffer from attrition bias. An initial check on the data of any online experiment should therefore start with the following two questions: (i) do attrition rates differ significantly between treatments and (ii) do the kind of participants that do not complete the task differ between treatments in any meaningful way? The answers to these questions should be reported alongside the findings, as they are crucial to the experiment's claim to internal validity. For the latter question, it may be useful to collect demographic or behavioural data at the start or prior to the experiment. As an example: in the income drawdown experiment introduced earlier, it may be that greater ease of use of some summary cost metrics made it more likely that a participant finished the experimental task. How do we account for this?

Should the researcher find evidence of non-random attrition, she may want to adjust her analysis to deal with the potential bias this introduces. Possible adjustments range from formal modelling of the attrition process (Hausman and Wise, 1979) to simpler tools and procedures (Flick, 1988). If there is a credible default outcome, it may be possible to deal with attrition by substituting missing outcome variables and computing an Intent-To-Treat (ITT) estimate. Many of these procedures, such as coding missing data as pre-specified defaults, worst-case outcomes or zeroes, are most convincing if the researcher pre-registers them as part of the design.

Finally, a virtually costless improvement to any online experiment is a free-form text field at the end of the experiment, asking the participants what they thought the experiment was about and if they had any difficulties understanding the task. This provides the researcher with valuable insights on the design, especially when pilot testing, much in the same way as a laboratory experimenter can ask testers for feedback after a laboratory session. Some researchers may want to use this data to exclude participants that did not engage with the task, although this is not recommended unless the exclusion criterion can be demonstrated to have been chosen ex ante.

Whom to invite?

The traditional "convenience sample" in social science experiments is a group of university students – in fact, observations of students underpin most of the established experimental findings in fields such as psychology and economics. Since most university students have at least some experience being a consumer, it seems innocuous to use convenience samples for consumer policy research. Several marketing researchers have, however, documented important differences between student samples and the general adult population (Burnett and Dune, 1986; Peterson, 2001; Peterson and Merunka, 2014). With the advent of online recruitment and experimental platforms, researchers may therefore look beyond convenience samples and instead recruit representative consumer samples.

An oft-heard reason for using a convenience sample is that, by virtue of their age and academic performance, university students can be used to construct a benchmark of optimal choice (an 'upper bound'). If this is true, these participants will be helpful for identifying challenging decisions and limitations of potential policy remedies. After all, if even highly educated young people have problems making decisions in these contexts,

then what hope does the average consumer have?[5] On the other hand, as discussed above, students often lack experience. In settings where experience is likely to play an important role in choosing well, university students seem less suitable as a benchmark. Unfortunately, it is hard to know ex ante what mix of youth, education and experience would be associated with optimal choices. One possibility is to recruit a (convenience) sample of highly educated students alongside a representative sample. An example of this approach is Johnson et al.'s (2013) comparison of representative consumers and MBA students in choosing healthcare insurance. MBA students may represent the sweet spot of sufficient consumer experience, technical ability and motivation to find the optimal answer.

Those who hypothetically consume a good in an online experiment and those actually in the market may still differ substantially. Recruiting participants that are (or recently were) in the market could go some way to close such gaps between hypothetical and real choices. To recruit these participants, consumer panels allow for pre-screening. The FCA has used this approach to recruit consumers who were nearing retirement for its experiments on income drawdown products, those with previous investing experience for its experiment on investment funds and consumers looking for payday loans (reported in Smart, 2016).

It is an open question whether consumers participating in online experiments are more motivated or digitally savvy than the average consumer. If so, this may be an issue if the objective of the experiment is to test new policy remedies. For example, if the intervention is an online calculator that works well in the experimental sample, it does not automatically follow that less digitally savvy consumers can use it. Again, targeted recruitment and participant pre-screening through consumer panels may go some way to alleviate these concerns.

The task

Key to a successful experiment is the design of the task participants are asked to complete. Here, we discuss some principles that will be helpful to the consumer policy researcher picking the features of this task. We start by discussing the issue of abstractions, which the researcher may be faced with several times during design.

Abstractions

Online experiments can never entirely replicate the target field context. Some features of the target field context will not be specified. Other features will be held constant across all treatments. Such deviations from the target field context, so-called *abstractions*, are often necessary to keep the experimental task manageable for participants and reduce the number of factors influencing participants' choices. Furthermore, since many details of the choice environment are unlikely to affect treatment effects, they can be safely omitted or held constant. Participants do not need to be told what day of the week it is before they choose an insurance policy, for example. The reason is that it seems highly unlikely that the treatment effect of, say, a new insurance cost disclosure policy will vary

[5] On the other hand, it will be harder to argue that good decisions by university students (potentially as a result of some policy measure) will be representative of decisions in the wider population.

with the day of the week. The policy researcher will be justified in leaving the day of the week out of the experimental design.

What if we are concerned that abstractions will interact with treatment effects? Is this problematic? Let's consider this with an example: consumers comparing the costs of mortgage products with a calculator tool (treatment) and without (control). The specific abstraction we have in mind is removing mortgage lenders' brand names. We can speculate about the ways in which this abstraction may interact with treatment:

- Removing brand names may **counteract** the calculator's usefulness, for example because it makes the task less realistic to participants who subsequently care less about picking a low-cost product. Assume this is the only way in which the abstraction interacts with treatment. If we now find a positive treatment effect, then the calculator tool is still likely to be of use when the abstraction is removed. Should we find no treatment effect, however, this presents a problem. Is the calculator truly ineffective, or is the lack of a treatment effect caused by the abstraction?
- Removing brand names may **strengthen** the calculator's usefulness, for example because participants focus more on actual costs when not distracted by brand perceptions (e.g. some brands marketing themselves as 'premium'). Assume this is the only way in which the abstraction interacts with treatment. If we now find a treatment effect, its size may be inflated by the abstraction. But if we do find no treatment effect, we can at least conclude that removing the abstraction would not change this fact.

Where does this leave us? Conditional on a positive treatment effect being measured, we may prefer counteracting abstractions – they indicate the treatment effect has survived a tough test. This situation can arise when the researcher has a strong reason to believe the treatment effect will obtain, perhaps based on prior research. Conditional on a null result (no treatment effect), we may prefer strengthening abstractions. But what if we do not have a clear expectation of whether a treatment effect will obtain? It may be problematic to say we prefer strengthening abstractions, as this could create an incentive to test policy interventions in particularly favourable conditions. Counteracting abstractions seem less problematic, so long as the policy researcher respects any null findings – it cannot be argued ex post that the policy treatment would have been effective had it not been for the abstraction.

Of course, the above reasoning only applies where we have good reason to believe the abstractions in the experiment interact with the treatment effect in a particular direction. As the mortgage cost calculator example illustrates, it is difficult to make such predictions, especially ex ante. It is sometimes argued that abstractions represent the 'stripping out' of features resulting in a policy test that isolates only the features that matter, giving the researcher a better chance of detecting a positive treatment effect where one exists. This is obviously true when abstractions simply reduce noise in the data without affecting treatment effects. However, the combined effect of abstractions may remove meaning and key institutional detail from the choice environment in ways that the researcher cannot foresee.[6] We therefore urge caution in introducing abstractions into policy testing experiments.

[6] We concur with Lynch (1982, p. 225) that "the most important determinant of the external validity of an experiment is the researcher's understanding or lack of understanding of the determinants of the behaviour in question". See also Bardsley (2010) and Greenwood (1982) for more detailed arguments on how alteration of the environment may affect external validity.

Defining the task

The experimental task needs to correspond to the target field context: the situation(s) representative of consumers' real-world choice environment. The first aspect of this situation is the kind of task: an information search, a product renewal, a product comparison, etcetera. The second aspect is where the task takes place: a store, a website, a price comparison platform, etcetera. To inform the second aspect, market research data may be needed. It cannot simply be assumed that a certain context will generate decisions that are representative of what consumers do in all contexts. For example, it's unclear whether a risk warning on a physical document has the same effect as the same warning on an online platform.

Visualising the consumer's situation as a consumer 'journey' or 'funnel' can be a helpful way of thinking about the task. This allows the researcher to break down the decision into different stages and decide which of those need to be represented in the experiment. For example, if the experiment tests different food labels in a supermarket, then the decision to go food shopping does not need to be represented in the experiment – participants can simply be told they are shopping for food.[7] Decision making in the real world is complex, of course, but not all this complexity needs to be replicated in the experiment.

Some judgement is involved here. Certain decisions are so straightforward in the real world (for example, buying concert tickets online) that an experimental task can encapsulate the entire decision-making process. For other decisions, like choosing how to decumulate one's pension savings, the decision is so complex that it would be unrealistic to capture everything in a single experiment. Often, it will be helpful to assume that the consumer has already taken one or more steps that precede the task. The FCA's income drawdown experiment assumed that consumers had already decided to pick an income drawdown product (instead of an annuity). Such assumptions can be made more plausible by targeted recruitment of participants, e.g. those who are approaching pensionable age and have pension pots that are easily converted into an income drawdown product.

Providing information

To ensure experimental validity, participants need to be provided with sufficient information to make their decision. What this exactly means differs from experiment to experiment, but a treatment effect driven by a fundamental misunderstanding of the experiment's rules is no success. Imagine a social norm 'nudge' experiment in which participants mistakenly think their job is to guess the majority choice.

Unless the experiment aims to investigate misinformation, lack of information or lack of understanding, a good principle to follow is that the rules of the experiment should be clear to a sufficiently engaged participant. That is, participants should have access to all the information required to figure out how their decisions map onto potential outcomes. This does not mean that all the information should be easier to access or interpret than it is in the field; it simply means that the information is available to sufficiently motivated

[7] A subtler question arises in other settings, where self-selection into the experimental task interacts with the scenarios tested. For example, if the effectiveness of a price-comparison table in a consumer communication is driven by those who would never read the communication in the first place (perhaps because the firm sends out the communication through an on-line platform that these consumers never use), then an experiment will overestimate the table's effectiveness. The answer is not always to include the self-selection stage into the experiment; the researcher can also ask the participants a (screening) question to correct for self-selection bias.

participants in the same way it is accessible in the field. The FCA's experiments on income drawdown products and investment funds, for example, both featured a pop-up button for each available product that displayed full cost schedules when clicked on, in the same way as an online investment platform or comparison website might do.

The policy researcher should look to the target field context to determine which information should be provided and how it should be presented. In consumer markets, there will often be legal or regulatory requirements that prescribe minimal information disclosure. Firms may also have developed certain ways of presenting information to consumers, or even agreed industry standards. Rules and standards do not necessarily lead to greater accessibility – in some markets, the relevant information can only be parsed by highly experienced and knowledgeable consumers.[8] In some cases, the researcher may wish to enhance the information so she can rule out deficient information or understanding as an explanation for her findings. But in most cases, the researcher will want to display the information as consumers encounter it in the real world.

Consumers will often consult various external sources of information before they make a decision. Not all of these are easy to recreate in an experiment – advice from friends and family is one such example. But other sources of information, such as websites and brochures, can be recreated or simply be made available to participants in an experiment. For online experiments, it may even be impractical to insist that participants do not make use of readily available information elsewhere.

Finally, the researcher will want to consider "information that [participants] bring to the task" (Harrison and List, 2004, p. 1022). This effectively means that experience with and attitude towards the decision domain may strongly differ depending on which participants are recruited. If representative participants are recruited, then their behaviour may still differ depending on whether they can make use of familiar heuristics. To allow consumers to use the heuristics they would use for decisions in the field, the information provided in the task should not be unnecessarily cut down, simplified or set to unrealistic values. Broadly speaking, this places some restrictions on ways of presenting information, the size of the choice set and features of options in the choice set. If a consumer wants to use a realistic choice rule such as "compare all TVs within my budget above a certain size; eliminate those with poor screen size/price ratios; pick product with longest warranty" then she should be able to do so in the experiment. Unrealistic or outdated information will make it less likely that such heuristics will be used and may thus affect the external validity of the experimental findings.

Situational parameters: vignettes, profiles or real?

Knowing the situation participants find themselves in is key to interpreting decisions in an experiment. What we might call situational parameters relate to how a consumer intends to (or will) use a product. In the FCA's experiment on income drawdown products, for example, participants' choices could only be evaluated in light of how much money participants would place in the product, their plans for regular withdrawals and their need for ad-hoc withdrawals. For some experiments, situational parameters may include some aspect of the consumer that is thought to be welfare relevant in a wider

[8] Just because information is available does not mean it will be in an understandable or most useful format - a prominent example is use of the 'miles per gallon' metric to compare cars on fuel efficiency (Larrick and Soll, 2008). Misunderstanding of the information provided may be as fundamental as consumers not knowing how to interpret probabilistic information (Gigerenzer, 2013, 2015).

sense (such as their income or socio-economic status). Without knowing consumers' situational parameters, it will be hard to say anything about their welfare.

Although participants may in fact have different degrees of awareness of their situational parameters, the experimenter needs to be clear on what the relevant parameters are. There are different ways of setting these parameters:

- **Vignettes** are descriptions of situations that are presented to participants. An example is the FCA's experiment on add-on insurance, that simply informed participants they were about to buy a main good and would have the option to buy insurance as an add-on product.
- **Profiles** are descriptions of several situations that participants can choose between/are allocated to, based on how closely the profiles correspond to the participant's own situation. An example is the FCA's experiment on income drawdown charges, where participants chose one out of six profiles that best matched their plans for withdrawing their pension savings.
- Participants may also be asked to enter **real parameters**, which will then be used as situational parameters for their decisions. Assuming that participants are reporting their parameters truthfully, their situation in the experiment should therefore closely mirror the real world (although the decision itself may still be unfamiliar). In the FCA's experiment on investment funds, participants were asked to make their decisions based on the amount of money they would typically invest in a fund.

The choice between the above approaches depends on several trade-offs. Real parameters promise the strongest claim to external validity, but they require larger sample sizes and may produce outcomes that cannot be evaluated in terms of decision quality (ie, because the parameters entered do not suggest an unambiguous optimal choice). Profiles and vignettes offer the researcher more control over possible outcomes, but they restrict the richness of the data and may reduce external validity. As a rule of thumb, profiles are a safe bet for policy testing (provided the profiles are based on realistic and relevant parameter sets), whereas descriptive research may benefit from the finer detail that real parameters can offer.

Some researchers may prefer real parameters on the basis that they allow for precise quantification of consumer welfare. Although appealing, such quantification requires assumptions on the external validity of experimental findings that are strong – perhaps too strong. A sufficient condition for calculations of consumer welfare in absolute terms (as done in a Cost Benefit Analysis) is that participants make exactly the same choices in the experimental setting as they would in the real world. Unfortunately, the current evidence on external validity of experiments does not seem to support such a strong assumption (Kessler and Vesterlund, 2015).

What are we measuring?

In most well-conceived experiments, the outcome variable(s) can be derived straight from the research question. For consumer policy experiments, this is typically some metric derived from participants' product choices. Other possible outcomes are valuations (through rankings, ratings, WTA/WTP, etcetera), consumer understanding (of the product chosen or the choice process) and use of relevant information. Sometimes the researcher

will want to know the relative importance of product attributes, which she may ask participants for directly or infer from repeated product choice or product rankings.[9]

It can be useful to measure multiple outcomes. Checking consistency between outcomes can then be informative, for example to verify whether participants did not choose the cheapest product out of the choice set by applying a heuristic that would lead them astray in other situations. Furthermore, policy testing often requires that multiple outcomes are monitored. For example, a new policy may be deemed an improvement on the status quo only if it improves consumer welfare without sacrificing consumer understanding. If all the relevant criteria for judging outcomes can be spelled out clearly before the experiment, this will make the policy researcher's job easier.

Besides outcomes, the researcher will often want to measure other variables. We have already discussed situational parameters, especially those important to welfare judgements. In the FCA experiment on income drawdown products, for example, participants were asked to report the approximate size of their pension pot. Demographic variables are often important, as they reflect a concern for certain sub-groups of the population. There is also a growing trend towards measuring indicators of decision-impairing factors, such as poor numeracy or language proficiency.

The researcher may also want to collect variables than help explain why certain outcomes are observed. Especially useful is data on information accessed, such as clicking on information boxes or opening new screens, which tells the researcher which participants did not make this information available to themselves. This decision can be used as a proxy for search effort or an indicator of whether the information is deemed useful. Other process variables measured during the choice task are response time or time spent on a particular screen, as well as participant characteristics like risk preferences or cognitive styles. Since the latter variable type is often measured after the experimental task, a key assumption is that its value is independent of the experimental treatment.[10] Finally, there are variables that the researcher will want to use as covariates in a regression – either to reduce noise or to correct imbalances between treatments.

Constructing a choice set

Choosing the right products for the experiment's choice set is important. To illustrate, let's say that the researcher aims to model the choice set after products available in the real world. First, she removes some product attributes, specifically those she has good reason to believe do not interact with the treatment effect. She also decides to remove brand names, to reduce noise and thus increase statistical power for a given sample size. Because she suspects that brand names may be correlated with (perceived) product quality, she decides to include a measure of quality as a product attribute. Now, having participants rank or make choices between products can tell her something about how participants trade off quality against other attributes, such as cost.[11]

A crucial assumption for measuring the above trade-off is that third attributes, such as the warranty period, are independent of the quality and cost attributes. If not, the value

[9] Examples of the former are Conjoint Analysis and Discrete Choice Experiments, which may also be used to construct relative product valuations. See Louviere et al. (2010) for more detail on these methods and their underlying assumptions.

[10] If this assumption does not hold, this may lead to erroneous conclusions: does a risk warning make people more likely to factor in their risk preferences when choosing, or does it make people report risk preferences in line with their choice?

[11] Unless the attributes traded off against one another are purely financial, identifying such trade-offs does not help the researcher measure (differences in) consumer welfare. It does provide evidence that consumers are putting more weight on certain attributes than others, which may explain choice patterns observed in the real world.

placed on the correlated third attribute may affect the quality-cost trade-off. To avoid this issue the researcher may omit the third attribute from the product description, hold its value constant or let it vary randomly (as is done in Conjoint Analysis). It seems evident that this assumption can come at the cost of some external validity, as trade-offs in the field may very well be influenced by such correlated third attributes.[12]

A more common variant of the above problem is when the researcher wants to find out how a certain aspect of product choice, such as product cost or energy efficiency, varies with treatment. Again, she may want to omit, keep constant or randomly vary other attributes. She can thus ensure that improvements in product choice (e.g. lower cost or higher energy efficiency) are not off-set by other attributes (e.g. lower quality). That way, the researcher can make statements about consumer welfare between different experimental treatments: for example, consumers spent more money in treatment B than in A but the extra spending did not buy them higher quality items.

It can be difficult to draw strong conclusions on consumer welfare from observed choices, hence the appeal of making some product attributes vary independently of others. Another approach is to impose more structure on the choice problem, either by (i) inferring preferences, (ii) imposing preferences, or (iii) counting mistakes. A particularly refined example of the first technique is found in Harrison and Ng (2016), who measure participants' risk preferences and use these to calculate consumer welfare in a subsequent insurance choice task. Imposing preferences can be done in different ways – the researcher essentially assumes there is some function that translates product attributes into a ranking of products that is independent of treatment. A straightforward example is assuming that participants only care about monetary cost, as reflected by (hypothetical) expenditure on products in the experiment.

Counting mistakes uses the concept of dominance, where one product is at least as good as another product on all attributes and better on some. The choice set can be constructed in such a way that one product dominates all others and is therefore the optimal choice: not choosing it would be a mistake. In the FCA experiment on income drawdown products, one product in the choice set dominated all others and a key metric was the percentage of participants that successfully identified this product.[13] Alternatively, one or more products can be dominated by others and therefore choosing them would be a mistake. This is the approach taken in the FCA's experiment on investment funds, as well as prior work by Choi et al (2009).

Rewarding participants

Volunteer participants in consumer policy experiments need to be rewarded for their participation. In principle, these rewards could be entirely intrinsic – people may be happy to donate their time for the good of society or science – but practical considerations often necessitate material incentives. Participants recruited through online platforms typically receive money or vouchers to compensate for their time (or the chance to win a large prize that more than compensates for their time). The minimum reward required to obtain a large enough sample may be lower if the policy researcher can advertise the study as having a policy objective, although advertising the study as such may attract a participant pool with different characteristics and motivations.

[12] More fundamentally, consumers do not necessarily choose by trading off attributes (Dieckmann et al., 2009),

[13] This intuitive metric does not directly describe welfare, as it does not account for the degree of sub-optimality of different types of mistakes. Further assumptions may therefore be required to calculate welfare differences between treatments.

A straightforward way of rewarding participants is to pay a flat participation fee upon finishing the experiment. Because actions in the experiment do not influence participants' reward in any way, we say that such experiments have hypothetical incentives. In these experiments, the policy researcher assumes that participants will be able to instinctively choose courses of action that would match those in the target field context.[14] If some cognitive effort is required to compare alternatives, the policy researcher assumes that participants are sufficiently motivated (either intrinsically, in reciprocation of the participation fee and/or because they appreciate the results of the research depend on the representativeness of their behaviour) to put in as much effort as they would in the real world. In the FCA experiment on investment funds, for example, participants were asked to imagine choosing funds as if they were investing their own money.

Another option is to have task-related rewards in place of (or in addition to) the participation fee. In some consumer choice experiments, the researcher may reward the participants with the actual product they chose.[15] But in most settings, monetary payment is the norm. How to structure these payments is inextricably linked to the experiment's objectives and the choice set design. If the policy researcher's objective is to measure cost to the consumer, for example, she can provide participants with 'house money' to buy a product and let all financial consequences of the purchase play out in the experiment. This approach was taken in the FCA's experiment on add-on insurance, where participants pay an insurance premium if they buy a product and may suffer a loss (with a given probability) if they choose to remain uninsured and an insurable event happens. Participants thus have a monetary incentive to consider the potential loss up front, as one might presume they would when choosing actual insurance products.

Another type of task-related reward is a performance payment. Participants may be paid for identifying the product with the lowest cost, the highest efficiency, features that best meet stated needs or some other metric of interest to the policy researcher. This approach is often taken when there is a concern that volunteer participants will not be sufficiently sensitive to hypothetical stakes. In the FCA experiment on income drawdown, the overwhelming complexity of comparing products on cost may have made it likely that participants would choose products at random when spending hypothetical money. By paying participants a bonus for identifying the cheapest product, the design provided them with an incentive to take the cost comparison task seriously. This approach may be justified by the experiment's objective (to evaluate different summary cost metrics) and the high stakes involved in real-world income drawdown choices.

Experiments with task-related monetary rewards thus address specific concerns about participants' motivation or lack of appreciation of the real-world stakes. But these concerns may be unwarranted (see Read, 2005) and our solution may in fact answer a rather narrow question – how well participants can maximise the experimenter-designed metric – instead of providing insight into multi-attribute consumer choice. The policy researcher will need to resolve this conundrum in light of her objectives, although no formal framework is available and empirical evidence is scarce. The concepts central to this decision are those relating to experimental validity, the topic to which we now turn.

[14] Observed behaviour with hypothetical rewards may be biased, such as participants being more willing to part with money (Harrison and Rutstrom, 2008; List and Gallet, 2001; Murphy et al., 2005) and less sensitive to risk (Holt and Laura, 2002; 2005).

[15] Alternatives are coupons or contingent reimbursement (for a novel approach using multi-unit auctions, see El Haji et al, 2017).

4 Experimental validity

What does it mean to say that an experiment is valid? Experimenters in the social sciences often use the criteria of internal validity and external validity to evaluate their findings. We now discuss these in turn.

Internal validity

If experimental findings have internal validity, the experiment measures what it claims to measure. In the words of Guala (2003, p. 1198): "the result of an experiment E is internally valid if the experimenter attributes the production of an effect B to a factor (or set of factors) A, and A really is the (or a) cause of B in E". Practically speaking, this means that measured treatment effects passing a threshold of statistical significance are caused by the experimental treatment and not something else. The researcher can ensure these conditions by a controlled (random) assignment to treatments, combined with ex post checks on compliance, attrition and sample balance across treatments.

For the findings of an experiment to be credible (for policy or otherwise), there should be no doubt that they are internally valid with respect to the scenarios tested. There should not be any evidence or plausible suspicion that factors other than the experimenter's chosen scenarios can explain systematic differences between treatments.

The internal validity requirement does not mean that the policy researcher always needs to be able to decompose observed differences in behaviour between experimental treatments into indivisible factors (such as psychological mechanisms or specific features of the choice architecture). The need to decompose treatment effects, or not, depends on the goal of the research. It may be instructive to decompose the 'why' behind sub-optimal choices to identify potential policy remedies, for example in the FCA's experiment on add-on insurance sold at the point of sale. But it is not always necessary to disentangle all the mechanisms behind changes in behaviour. For example, in measuring the effect of calorie labels on food consumption, a researcher does not have to separately identify the effect of (i) having a label on the product and (ii) having a message on the packaging that draws the consumer's attention to calories.

The role of experimental control

Experimental control, which should not be confused with having a control group, is the researcher's ability to prevent aspects of the experiment from varying as they naturally might. Such aspects range from participants' perceptions of the quality of a good in the experiment, to the stated purpose of the task, to participants' ability to use their mobile phone during the experiment.

Researchers often use experimental control to ensure that minimum conditions are met in all treatments. An example is the use of control questions to check that participants have read and understood the experimental instructions. Such control is important for internal validity, as it rules out confusion and misinterpretation as explanations for

observed behaviour. Examples of similar types of experimental control are similarity of design and wording across treatments, similarity of response mechanisms and validation of participants' inputs into experimental software.

In policy research, experimental control often needs to be balanced against the need to represent the target field context in a natural way. Consider communication between a consumer and a salesperson. An 'uncontrolled' implementation would be unstructured communication with some legal and ethical constraints; much like a conversation in the field. But some researchers may be concerned that some salespeople are more persuasive than others and decide to control for this by limiting the communication to a set of pre-selected phrases. This would reduce the noise in decision-making, thereby increasing the statistical power of the experiment for a given sample size. But the extra control may come at a cost. Consumers' behaviour may no longer reflect the field context, as communication has become artificial and constrained. It is not clear whether this implementation of experimental control would be worth the increase in statistical power, which could have also been achieved by gathering a larger sample.

There is a more general point here: it seems unhelpful to characterise the noise-reducing function of experimental control as increasing internal validity, except in the narrow sense of increasing statistical power for a given sample size. A second misconception is that exercising experimental control requires keeping everything constant. Letting some parameters of the experimental design vary randomly is fine in many settings – take, for example, the random draws from a normal distribution that determine prices of insurance products in the FCA's experiment on add-on insurance. As long as the varying parameters are not thought to be correlated with treatment effects in the target field context, the internal validity of the experiment will be unaffected by this 'loss of control'.

External validity

External validity is the extent to which the researcher can claim that the causal forces identified in an experiment will also apply in other contexts. Returning to Guala (2003): "[the result of an experiment] is *externally* valid if A causes B not only in E, but also in a set of other circumstances of interest, F, G, H, etc." (p. 1198, emphasis in original). Compared to internal validity, evidence to support a claim to external validity is harder to come by. The researcher can compare different contexts for similarity, but she cannot be sure that people will behave identically in them. Ideally, she would be able to collect statistical evidence by conducting the same experiment in different contexts, although for policy research this is often not possible.[16] Nevertheless, there are some steps that the researcher can take to support the claim to external validity of her findings.

First, the researcher may be able to gather data that directly indicates whether participants' behaviour in the experiment matches their behaviour in the target field context. For example, a recent FCA experiment with credit card borrowers found that hypothetical credit card repayment decisions by participants closely correspond to repayments in matched individual administrative data from their credit card provider (Guttman-Kenney et al., 2018). Since these participants were assigned to a control condition that represented the market's status quo, the researchers interpreted this as

[16] Comparisons between different contexts, especially laboratory/online and field, are a valuable public good for policy researchers as they provide direct evidence on generalisability. But this public good is underprovided because the individual policy researcher has no incentive to collect data in contexts besides the one promising the highest level of external validity.

evidence that participants' response to varying institutions in the experiment would at least be qualitatively similar in the real world. Although it may not always be possible to match responses to detailed individual data, policy researchers can also survey participants after the experiment or check experimental responses against aggregate statistics.

Furthermore, there is a consensus that external validity of policy testing experiments increases as these experiments more closely resemble the target field context (Camerer, 2012; Levitt and List, 2005; 2007ab; Lunn and Ní Choisdealbha, 2018). In the terms of the Harrison and List taxonomy introduced earlier, this can be done by bringing in a representative participant pool, a commodity, task and/or a participant information set from the field. In practical terms, the policy researcher may take the target field context as a starting point and make the experiment resemble it as much as possible.

Although ultimately an empirical question, the argument that participants are more likely to behave like they do in the field in more natural settings is plausible. This explains why the policy researcher will often encounter the opinion that experimental designs that mirror the level of detail observed in the field are more credible. After all, it is easy for anyone interpreting the results from an experiment with some abstractions to point out a gap in realism vis-à-vis the field. The more gaps, the more room for suspicion that these gaps may affect behaviour in non-trivial ways. And the more room for suspicion, the harder a job the policy researcher will have explaining her experimental findings.

Debating external validity

In a series of papers, Levitt and List (2005, 2007a, 2007b) discuss several obstacles to generalising findings from laboratory experiments (or more generally, experiments not carried out in the field with real market participants). Echoing arguments in Harrison and List, (2004), the obstacles they discuss mostly derive from participants knowing they are being observed or relate to how much field context the experimenter will be able to re-create in a non-field setting. The authors argue that contributions of such experiments may in fact be limited to providing qualitative insights, suggesting mechanisms, proof-of-concept findings (showing "what can happen") and identifying "empirical puzzles".

In reply, Camerer (2015) argues that none of the obstacles identified by Levitt and List are inherent to the type of experiments they critique, except for participants knowing they are in an experiment.[17] He also points out that the potential distortion due to participants being observed may well be a price worth paying for being able to closely observe participants and having more replicable research designs. Although Camerer clearly states that experiments aiming to answer policy questions should "resemble the target external setting", he also provides a long list of experimental findings that appear to generalize well from more abstract settings to the field. Based on these examples, he claims that economists should maintain a default assumption that results in laboratory experiments will generalize to field contexts, unless proven otherwise.

The disagreement on external validity might be resolved easily, suggest Kessler and Vesterlund (2015), if both sides recognise that it is not the quantitative but the

[17] Knowing that you are being observed may affect both selection into and behaviour within the experiment. Note, however, that even in field experiments it may not be feasible or ethical to completely conceal the experimental nature of treatment from participants. At the very least, participants in natural field experiments for policy purposes should not be prevented from knowing they are being experimented upon, be able to opt out of treatment and/or revise the choice they made under treatment. In many cases, this can be achieved by telling all participants that their enrolment into treatment is part of a trial with selected customers. As Loewenstein et al. (2015) show in the context of nudging, such disclosure does not have to undermine treatment effectiveness.

qualitative findings of experiments that should be considered. In other words: it is the direction of an effect, not the size, that matters. This argument is appealing and seems a fair yardstick against which to judge the claims that laboratory researchers typically make about their work. It does introduce some challenges for policy researchers, however.

First, since some policy questions need specific estimates of effect sizes, even if only a lower or upper bound estimate, some quantitative model may still be required to translate experimental findings to the target field setting. Second, Kessler and Vesterlund's analogy of laboratory (or online) experiments as a "wind tunnel" could be problematic for policy testing. If the manipulation of a variable does not affect behaviour in a controlled experiment, they write, "then it is unlikely that the manipulated variable will affect behavior in a more complicated external environment" (p. 7). This argument seems to presuppose that abstractions introduced by the controlled experimental context either strengthen the treatment effect or reduce noise. As we discussed in the previous section, this is not a trivial claim to make. Consider an appeal to good citizenship that fails to reduce tax cheating in an abstract experimental environment, where the tax authority is represented by a computerised institution that may detect cheating and reduce the participants' experimental earnings with some probability. Do we conclusively write off the policy because it did not work in this abstract environment?

Although experimenters may claim to test "general principles of behaviour", this claim is hard to sustain without some supporting empirical evidence. The evidence provided by Camerer (2015) provides some support for the notion that academic scientists have been quite adept at uncovering behaviours that generalise.[18] But this does not mean that the same applies for experiments on consumer behaviour, the topics of which cannot necessarily be hand-picked for their likely generalisability. For want of a body of empirical evidence on the types of consumer behaviour more or less likely to generalise, it seems that external validity will remain an important issue to the policy researcher.

It is worth pointing out that the policy researcher may, generally speaking, be less concerned than the academic scientist about whether her findings will generalise broadly. If she is testing a warning label for risky investment products in her jurisdiction, it does not matter whether the label also works for financial products generally, whether it works in other languages, jurisdictions, etcetera. Her target field context is the policy domain she is responsible for and she will typically want to conduct an experiment that resembles this context as much as possible. She does not ignore generalisability entirely, as she will want the behaviour she observes in her experiment to match that of actual consumers confronted with the risk warning in real life. But this is the primary sense in which she cares about external validity.

Is there an internal-external validity trade-off?

It is often stated that experimental design requires a trade-off between internal and external validity. This puts highly controlled (laboratory or online) experiments at one end of the scale, high on internal validity but low on external validity, and field experiments on the other, high on external validity and low on internal validity. This is a mischaracterisation, particularly so for consumer policy experiments. Although controlled

[18] But see also Galizzi and Navarro-Martínez (2018) and references therein.

experiments offer the experimenter more statistical power for a given sample size, this does not necessarily mean that conducting a similar experiment in the field or in more natural settings implies a substantial sacrifice of internal validity. Internal validity is not a property of a particular experimental paradigm but derives from how well the design can answer the research question.

If the research question is whether consumers heed risk warnings, for example, then it is difficult to precisely state how a controlled, abstract experiment with a convenience sample would provide more internal validity than a well-conducted field experiment with a sufficiently large sample. The thought experiment of a like-for-like comparison of the two designs is of little practical use, for two reasons. First, it is highly unlikely that the distribution of the outcomes of interest in the control group, as well as the distribution of treatment effects, will be comparable in the two experiments. Second, the two designs are also quite different in kind. The field experiment requires a much larger, fixed investment up front but the marginal participant cost is often effectively zero. Given this cost structure, it seems unlikely that an experienced designer of field experiments will skimp on participant numbers.

There is therefore no general rule that means higher (perceived) external validity reduces internal validity in a meaningful sense. Or, to phrase the point in terms of the Harrison and List taxonomy introduced earlier: we do not expect that internal validity will suffer from the introduction of a non-standard participant pool, a commodity, task and/or participant information set from the field context. As long as a sufficiently large participant sample can be recruited, we therefore encourage the policy researcher to stay as close to the field as possible.

The above advice notwithstanding, there is one type of research question that may require the policy researcher to sacrifice some external validity. The researcher sometimes wants to measure whether consumer decisions in some scenario are better than in others, but do so without making strong assumptions on consumer preferences. In the previous section, we discussed how the choice set of products in the experiment can be constructed such that some products strictly dominate others, regardless of the consumers' preferences. Although this construction of the choice set will inevitably involve some departure from the target field setting, it will give the policy researcher an unambiguous measure of consumer welfare.

5 Discussion

It seems that online experiments have good prospects to become an important part of the policy research toolkit. Technology has improved, costs have come down and challenges to experimental control no longer seem insurmountable. Furthermore, the increasing digitalisation in many domains of consumer decision-making means that more and more decisions are made online. As a result, the external validity of online experiments may increase over time.

We believe these are positive developments, but the challenges of online policy experimentation should not be underestimated. Nor should the need for empirical evidence to assess claims to generalisability. Throughout this paper, we have made the point that external validity is of first order importance to consumer policy research. Any element of an experiment's design that diverges from the target field context needs to have been considered carefully and its benefits weighed against the potential loss in external validity. We hope that our discussion of topics such as participant recruitment, abstractions and choice sets will be of use to the policy researcher looking to design such experiments.

In the process of designing an experiment, the researcher will usually develop her own mental model of how the experiment relates to the real world. The model will include her beliefs on the relationship between experimental and field variables, the influence of various experimental treatments, abstractions, selection into the experiment, etcetera. At a minimum, this model should be internally consistent. Ideally, this model will also be informed by empirical evidence. For example, if repeated experiments show that student participants make better choices of broadband providers in an online experiment than in the real world, and we know that students make better choices than the average consumer in the real world, then we may deduce that students' choices in an online experiment are an upper bound on the average consumer's decision in the real world. Spelling out such mental models may be a first step towards more formal definitions of models that translate experimental results to the real world.

Many policy testing experiments imply a claim that a positive treatment effect observed in an online experiment will generalise to the field. Such claims rely on certain assumptions about the interaction between abstractions from the field and the treatment effect. As we have argued in this paper, it is important for the researcher to keep these assumptions in mind as she designs the experiment. First, it may be difficult to know exactly how abstractions introduced in the design interact with treatment effects, as abstractions may alter the meaning of the task and the choice environment to participants. Second, these experiments should be designed as rigorous tests of policy, with researchers resisting the temptation to create experimental conditions that are so favourable to the proposed policy that they are no longer representative of the target field context.

It is probably fair to say that, when it comes to questions of consumer policy, there is still a dearth of evidence on whether experimental findings can be extrapolated to the

field. We have mentioned some of the open empirical questions in this paper, but it is worth restating the main topic areas here as promising avenues for further work.

First, more evidence on differences between participant samples in online and field experiments would be welcome. It is increasingly possible to avoid selection bias by recruiting representative samples for online experiments, such as users of a particular service or prospective buyers for a particular good. Do these participants make choices that are closer to the field than convenience or nationally representative samples? Another aspect is the effect of participants knowing that they are being observed. Is this effect of predictable direction and size for consumer decision-making experiments?

Second, there is little systematic evidence in consumer decision-making settings on the importance of individual features of experimental designs as they more closely mimic the target field context. The list of potential research questions under this header is long indeed. What is the effect of removing brand names? What happens when we simplify, remove or randomly set product characteristics? Does it matter whether participants' information sets are vignettes, profiles, or real? Do we find different optimal policy regimes depending on whether we use hypothetical or task-related rewards? The answers to some of these questions may be specific to a particular type of product, but for others there may be a more general pattern. Research on these topics should give us a better understanding of which consumer behaviours are least sensitive to context. In the long run, we may hope to have some taxonomy of consumer behaviours that tells us which behaviours are fundamental, which are more context-sensitive, and what aspects of context have the strongest effects.

A third and final topic area concerns the limits to online experimentation. Simply put, there may be certain decisions that are not easily replicated in an artificial online environment, despite the experimenter's best efforts. Iscenko et al. (2014) mention the challenges of modelling decisions that play out over longer periods of time, involve learning and that have an emotional component. This list could be extended to include decisions with complex interdependencies, decisions that require unstructured input from several agents, tasks with an important tactile or visceral component and perhaps others. In some of these domains, researchers may have to be content with establishing "proof of concept" findings in experiments, rather than generalisable knowledge. More empirical evidence on the kind of experiments and domains that are less likely to generalise, and why, would greatly help the cause of policy experimentation.

In closing, we repeat our wish that this paper will stimulate further debate and methodological development in consumer policy research. Online experimentation is likely to grow even more in years to come; we are optimistic it will play a big part in improving consumer outcomes through better policy.

References

Afif, Zeina; Islan, William Wade; Calvo-Gonzalez, Oscar; Dalton, Abigail Goodnow. 2019. Behavioral Science Around the World: Profiles of 10 Countries (English). eMBeD brief. Washington, D.C.: World Bank Group. Retrieved from: http://documents.worldbank.org/curated/en/710771543609067500/pdf/132610-REVISED-00-COUNTRY-PROFILES-dig.pdf

Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, *21*(1), 99-131.

Artinger, S., Baltes, S., Jarchow, C. Petersen, M., and Schneider, A.M. (2017). Lifespan label for electrical products. German Federal Ministry for the Environment, Nature and Nuclear Safety. Retrieved from: https://www.bundesregierung.de/breg-en/issues/wirksam-regieren-with-citizens-for-citizens/with-citizens-for-citizens/lifespan-label-for-electrical-products-323362

Artinger, S., Baltes, S., Jarchow, C. Petersen, M., and Schneider, A.M. (2018). Warning note for the protection of small investors. German Federal Ministry of Finance. Retrieved from: https://www.bundesregierung.de/breg-en/issues/wirksam-regieren-with-citizens-for-citizens/warning-to-protect-small-investors-392160

Bardsley, N. (2008). Dictator game giving: altruism or artefact?. *Experimental Economics*, 11(2), 122-133.

Bardsley, N. (2010). Sociality and external validity in experimental economics. *Mind & Society*, *9*(2), 119-138.

Bardsley, N., Cubitt, R., Loomes, G., Moffat, P., Starmer, C., & Sugden, R. (2010). *Experimental economics: Rethinking the rules*. Princeton University Press.

Behavioural Insights Team (2019a). Improving consumer understanding of contractual terms and privacy policies: evidence-based actions for businesses. Best practice guide. Retrieved from: https://www.bi.team/wp-content/uploads/2019/07/Final-TCs-Best-Practice-Guide-July-2019-compressed.pdf

Behavioural Insights Team (2019b). Annual Report 2017-18. Retrieved from: https://www.bi.team/wp-content/uploads/2019/01/Annual-update-report-BIT-2017-2018.pdf

Boyce, T., Dixon, A., Fasolo, B., and Reutskaja, E. (2010). Choosing a high-quality hospital: The role of nudges, scorecard design and information. The King's Fund. Available at: https://www.kingsfund.org.uk/sites/default/files/field/field_publication_file/Choosing-high-quality-hospital-role-report-Tammy-Boyce-Anna-Dixon-November2010.pdf

Burnett, J. J., & Dune, P. M. (1986). An appraisal of the use of student subjects in marketing research. *Journal of Business Research*, *14*(4), 329-343.

Camerer, C. F. (2015). The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Reply to Levitt and List. S. 249-295 in: Guillaume R. Fréchette und Andrew Schotter (Eds.), *Handbook of Experimental Economic Methodology*.

Chater, N., Huck, S., & Inderst, R. (2010). Consumer decision-making in retail investment services: A behavioural economics perspective. Report to the European Commission/DG SANCO.

Chetty, R. (2015). Behavioral economics and public policy: A pragmatic perspective. *American Economic Review*, *105*(5), 1-33.

Choi, J. J., Laibson, D., & Madrian, B. C. (2009). Why does the law of one price fail? An experiment on index mutual funds. *Review of Financial Studies*, 23(4), 1405-1432.

Codagnone, C., Bogliacino, F., Ivchenko, A., Veltri, G., and Gaskell, G. (2014). Study on online gambling and adequate measures for the protection of consumers of gambling services. Report for the European Commission. Retrieved from: http://ec. europa. eu/internal_market/gambling/docs/initiatives/140714-study-on-onlinegambling-final-report_en. pdf

Codagnone, C., Veltri, G. A., Bogliacino, F., Lupiáñez-Villanueva, F., Gaskell, G., Ivchenko, A., and Mureddu, F. (2016). Labels as nudges? An experimental study of car eco-labels. *Economia Politica*, *33*(3), 403-432.

Cohn, A., Fehr, E., & Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature*, 516(7529), 86.

Cohn, A., Engelmann, J., Fehr, E., & Maréchal, M. A. (2015). Evidence for countercyclical risk aversion: An experiment with financial professionals. *American Economic Review*, 105(2), 860-85.

Deaton, A., and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2-21.

De Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and bounding experimenter demand. American Economic Review, 108(11), 3266-3302.

Dieckmann, A., Dippold, K., & Dietrich, H. (2009). Compensatory versus noncompensatory models for predicting consumer preferences.

El Haji, A., Lusk, J., and Onderstal, S. (2017). Online experimental auctions. *Mimeo*.

European Commission (2019). Behavioural study on the digitalisation of the marketing and distance selling of retail financial services. Final report. Retrieved from: https://ec.europa.eu/info/sites/info/files/live_work_travel_in_the_eu/consumers/digitalisation_of_financial_services_-_main_report.pdf

Financial Conduct Authority (2018), When and how we use field trials. Corporate publication. Retrieved from: https://www.fca.org.uk/publication/corporate/how-when-we-use-field-trials.pdf

Fletcher, A. (2016). The role of demand-side remedies in driving effective competition. Report for Which?. Retrieved from: https://www.regulation.org.uk/library/2016-CCP-Demand_Side_Remedies.pdf

Flick, S. N. (1988). Managing attrition in clinical research. *Clinical Psychology Review*, *8*(5), 499-515.

Galizzi, M. M., & Navarro-Martínez, D. (2018). On the external validity of social preference games: a systematic lab-field study. *Management Science*, *65*(3), 976-1002.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of science*, *70*(5), 1195-1205.

Greenwood, J. D. (1982). On the relation between laboratory experiments and social behaviour: Causal explanation and generalization. *Journal for the Theory of Social Behaviour*.

Guttman-Kenney, B., Leary, J., and Stewart, N. (2018). Weighing anchor on credit card debt. *Financial Conduct Authority Occasional Paper Series*, No. 43. London, UK.

Harrison, G. W., & Rutström, E. E. (2008). Experimental evidence on the existence of hypothetical bias in value elicitation methods. *Handbook of experimental economics results*, *1*, 752-767.

Harrison, G. W., and List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*(4), 1009-1055.

Harrison, G. W., & Ng, J. M. (2016). Evaluating the expected welfare gain from insurance. *Journal of Risk and Insurance*, 83(1), 91-120.

Hayes, L., Lee, W., and Thakrar, A. (2018). Now you see it: drawing attention to charges in the asset management industry. *Financial Conduct Authority Occasional Paper Series*, No. 32. London, UK.

Hausman, J. A., & Wise, D. A. (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica*, 455-473.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644-1655.

Holt, C. A., & Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. *American Economic Review*, *95*(3), 902-912.

Iscenko, Z., Duke, C., Huck, S., and Wallace, B. (2014). How does selling insurance as an add-on affect consumer decisions? *Financial Conduct Authority Occasional Paper Series*, No. 3. London, UK.

Johnson, E. J., Hassin, R., Baker, T., Bajger, A. T., & Treuer, G. (2013). Can consumers make affordable care affordable? The value of choice architecture. *PloS one*, *8*(12), e81521.

Kessler, J., & Vesterlund, L. (2015). The external validity of laboratory experiments: The misleading emphasis on quantitative effects. In: *Handbook of Experimental Economic Methodology*, Frechette, G.R., and Schotter, A. (Eds.). Oxford, UK: Oxford University Press.

Levitt, S. D., & List, J. A. (2005). What do laboratory experiments tell us about the real world. In *Journal of Economic Perspectives*.

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

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

List, J. A. (2001). Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards. *American Economic Review*, *91*(5), 1498-1507.

List, J. A. (2003). Does market experience eliminate market anomalies?. *Quarterly Journal of Economics*, *118*(1), 41-71.

List, J. A., & Gallet, C. A. (2001). What experimental protocol influence disparities between actual and hypothetical stated values?. *Environmental and resource economics*, *20*(3), 241-254.

Loewenstein, G., Bryce, C., Hagmann, D., & Rajpal, S. (2015). Warning: You are about to be nudged. *Behavioral Science & Policy*, *1*(1), 35-42.

Louviere, J. J., Flynn, T. N., & Carson, R. T. (2010). Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling*, *3*(3), 57-72.

Lunn, P. D., & Choisdealbha, Á. N. (2018). The case for laboratory experiments in behavioural public policy. *Behavioural Public Policy*, *2*(1), 22-40.

Lynch Jr, J. G. (1982). On the external validity of experiments in consumer research. *Journal of consumer Research*, *9*(3), 225-239.

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *Science*, *341*(6149), 976-980.

Murphy, J. J., Allen, P. G., Stevens, T. H., & Weatherhead, D. (2005). A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics*, *30*(3), 313-325.

OECD (2017). Behavioural insights and public policy: Lessons from around the world, OECD publishing, Paris. Retrieved from: https://read.oecd-ilibrary.org/governance/behavioural-insights-and-public-policy_9789264270480-en#page50

Oxera (2018). Identifying metrics to aid consumer choice in the income drawdown market. *Report prepared for the Financial Conduct Authority's Retirement Outcome Review.* Available at: https://www.fca.org.uk/publication/market-studies/retirement-outcomes-review-interim-report-annex5.pdf

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of consumer research*, *28*(3), 450-461.

Peterson, R. A., & Merunka, D. R. (2014). Convenience samples of college students and research reproducibility. *Journal of Business Research*, 67(5), 1035-1041.

Read, D. (2005). Monetary incentives, what are they good for?. *Journal of Economic Methodology*, *12*(2), 265-276.

Shafir, E. (Ed.). (2013). *The behavioral foundations of public policy*. Princeton (NJ): Princeton University Press.

Smart, L. (2016). Full disclosure: a round-up of FCA experimental research into giving information. *Financial Conduct Authority Occasional Paper Series*, No. 23. London, UK.

Ter Meer, J., Mottershaw, A., Merriam, S., Ní Chonaire, A., and Martin, L. (2018). Improving guidance for retirement planning. Behavioural Insights Team. *Report prepared for the Financial Conduct Authority.* Retrieved from: https://www.fca.org.uk/publication/research/increasing-comprehension-of-investment-pathways-for-retirement.pdf

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1), 75-98.

12 Endeavour Square, London E20 1JN
Telephone: +44 (0)20 7066 1000 Website: www.fca.org.uk