Using experiments in consumer research

Research Document

Publication date: 1st March 2010

Foreword

1.1 Why did we decide to commission a report on experiments?

Ofcom has recently been doing some work on behavioural economics and its implications for regulation. Behavioural economics explores insights from psychology on the behaviour of individuals, and uses these to develop a more realistic understanding of how individuals make decisions. Traditional economics assumes that individuals are good at making decisions: behavioural economics relaxes this assumption and highlights that individuals are subject to cognitive limitations, impulses and emotions, which can lead to apparent "errors" or "biases" in decision-making.

A key insight from the work we have been doing on behavioural economics is that evidence on actual consumer behaviour can be helpful in understanding how consumers react to different market features, and what type of remedies are likely to be successful. Behavioural economics has recently garnered a lot of interest in policy-making, and some organisations have started to use experiments to understand how behavioural biases can influence outcomes in particular markets. Experimental-style techniques are well-established in areas such as consumer product research, but have only recently been used more often in regulation and policy-making contexts.

For example, the Federal Trade Commission in the US commissioned an experiment to understand how consumers would react if they were given information on the commission received by mortgage brokers. The experiment yielded the interesting result that with additional information, consumers were less likely to choose the lowest cost product. Additional information in this case appeared to confuse consumers rather than improve consumers' decisions[1]. This is one example of the potential of experiments to uncover effects that traditional theory may not predict, and provide more evidence on the effectiveness of remedies before they are implemented.

Other organisations which have commissioned experiments in the UK include the FSA[2] and the OFT[3]. The OFT and Competition Commission (CC) have recently published a report on *Road-testing of consumer remedies* which describes how testing remedies can help to identify the most effective remedy and to fine-tune its design. The OFT and CC report also discusses different ways of testing remedies, including qualitative methods, simulation, quantitative surveys and economic experiments. These different methods all have strengths and weaknesses, and in some cases, some types will be suitable than others. For example, the report highlights that experiments provide data on actual consumer behaviour and allow the performance of different remedies to be compared. However, the report also notes that care needs to be taken in extrapolating the results of an experiment to the real world, and drawing inferences about how much real world consumer behaviour may actually change. A

[1] *The Effect of Mortgage Broker Compensation Disclosures on Consumers and Competition: A Controlled Experiment,* (Lacko, 2004, for the Federal Trade Commission)

[2] For example, the FSA commissioned an experiment on the influence of insurance sellers on buyers, and the impact of information disclosure. *Information versus Persuasion: Experimental Evidence on Salesmanship, Mandatory Disclosure and the Purchase of Income and Loan Payment Protection Insurance* (De Meza, D., Irlenbusch, B. and Reyniers, D.,2007)

[3] For example, an experiment was undertaken as part of a report commissioned by the OFT on scams, *The psychology of scams: Provoking and committing errors of judgement*, (prepared by the University of Exeter School of Psychology.)

general conclusion of the report is that "which road testing method is employed, how it is employed and whether more than one method is used depends on the particular policy question and what type of information is required to answer that question."

We consider that experiments may have a role in developing policy for markets that Ofcom regulates. We therefore commissioned a report from London Economics and University College London (UCL) to investigate the potential uses of experiments in understanding consumer behaviour, and to develop a better understanding of the potential benefits of experiments if used in Ofcom's work.

1.2 The potential uses of experiments

Experiments test the actual behaviour of individuals under different conditions. In an experiment, individuals may be faced with choices to make under different circumstances, and the experimenter is able to observe how consumers react. The main advantages include:

- Experiments allow you to observe what consumers actually do, not what they say they will do.
- They allow testing that may be impossible or very expensive to carry out in the actual market.
- It may be easier to establish causality as the experimenter can vary one feature at a time.

However, set against this:

- Experiments may not be able to provide information on the likely magnitude of effects in the real world.
- Where information on consumers' beliefs and perceptions are required, other types of testing are likely to be more useful.

Where the effectiveness of a remedy depends on how consumers react, testing (potentially using methods such as experiments) may play a useful role. However, as highlighted in the OFT and CC report on *Road-testing of consumer remedies*, it will not be useful or appropriate to use experiments across all areas of our work. In some cases, other types of research will be more informative. We already undertake substantial research on consumer behaviour, and conduct a number of different types of research. We therefore believe that the use of experiments is an additional tool that complements the work that we already do.

To assess the potential benefits and limitations of experiments in Ofcom's work, we asked London Economics and UCL to carry out an experimental study in the specific context of devising effective ways of providing information on the price of telephone calls.

1.3 Experiment on providing price information

This specific experiment was designed to consider a number of options for improving the understanding of consumers when choosing services funded through call charges.

Consumers are able to find out the cost of a call in many ways. All phone companies are bound, to a greater or lesser extent depending broadly on the services that they offer and to whom, by rules – called General Conditions – which require them to be transparent about

the cost of calls. Where applicable, these General Conditions require them to publish details of call prices and ensure that their customer service staff point consumers to these charges.

There may also be other sources of information where consumers can find out about call tariffs. For example, for certain types of calls such as Premium Rate Services, some information about the cost of calls is published by the provider of that service.

However, consumers in general do not always know the exact cost of a call before they make a call, and we have identified that this uncertainty can lead to consumer detriment.

Ofcom has in the past looked at developing policies to address this potential uncertainty. For example, the National Telephone Numbering Plan (the Numbering Plan) sets out the allocation of telephone numbers to communications providers and the permitted use of those ranges (for example, restrictions on the type of service or BT's retail price).

However, many of the price-related requirements n the Numbering Plan only apply to BT, and other communications providers (CPs) are able to charge their own tariffs for calls to individual number ranges. Consumers therefore often do not know the exact tariff of a call before making that call, unless they have actively looked up their CP's prices.

We are currently engaged in considering options to address identified consumer concerns around price transparency in this area. This review is one of Ofcom's proposed projects in its draft Annual Plan for 2010/11. It will be undertaking an examination of the impact on consumers of relevant aspects of the current regime and will analyse the full costs and benefits of different remedies to address any identified consumer problems.

In the past, Ofcom has considered a number of informational remedies to tackle these problems, such as pre-call announcements at the point-of-call and improved provision of price lists. There are clear costs to any informational remedy. For example, we have found that the cost of pre-call announcements in particular is likely to be very high because of the high degree of variation in charges within a number range. There are also other relevant considerations that need to be made to ensure such interventions are viable, for example, we need to consider the risk of pre-call announcements disrupting certain machine-to-machine calls, including personal, fire and burglar alarms, leading to potential life-threatening consequences[4].

Nonetheless, in assessing the benefits to consumers, we have previously found it very difficult to predict how effective these measures would be if introduced. For example, stakeholders have previously suggested that consumers may find pre-call announcements annoying and, if offered the choice, would switch off the feature shortly after its introduction, thereby limiting the effectiveness of such a facility.

Our experiment therefore sought to understand how different interventions may influence callers to make better decisions when making a call. In this specific experiment, participants performed best when provided with pre-call announcements that stated the exact call price, however, consumers appeared to benefit to some degree from all types of interventions. The experiment also provided some interesting results which might not immediately be predicted by traditional theory. These included:

[4] Such risks were key to Ofcom's previous decisions to withdraw the requirement for pre-call announcements for 070 numbers and proposals for such announcements on 087 numbers: http://www.ofcom.org.uk/consult/condocs/numbering03/070precall/ and http://www.ofcom.org.uk/consult/condocs/0870calls/0870condoc.pdf

- the ability of participants to learn to make better decisions over time, regardless of the intervention; and

- total bill costs can impact future behaviour. However, the impact on future behaviour is driven by the size of the *total* bill, rather than the information provided on the bill on the prices of individual calls and the associated per minute charges.

The experiment has provided some helpful indications of the type of interventions that are most likely to be effective in this situation, although it did not test the full range of possible interventions. The outcome does need to be reviewed with some caution as it only provides us with evidence on one aspect of the issue. As noted above, experiments can be very helpful in indicating the relative effectiveness of different interventions, but care needs to be taken in extrapolating the results to the real world.

For example, the report notes that interventions that do not work well in the laboratory experiment are unlikely to work in the real world. However, interventions that appear to work well in the laboratory experiment may not perform as well in the real world. It is also difficult to predict and quantify the size of real-world welfare gain that different interventions may deliver. We would therefore need to look to other forms of analysis to assess the likely size of real-world welfare gain, and weigh any potential benefits against the full set of costs, which as described earlier, may be substantial.

We will further consider these findings when we review our approach to regulating non geographic call services to ensure that the regulatory framework delivers optimal consumer outcomes in terms of range of services and clarity and appropriateness in costs and charges.

1.4 Conclusions on the use of experiments

Our work on behavioural economics has further highlighted the usefulness of consumer research in developing policy. We consider that experimental-style techniques are likely to act as a useful complement to the substantial consumer research that we already do. However, as with other research methodologies, there are advantages and disadvantages of experiments, and their appropriateness will vary depending on the context. It is unlikely that carrying out experiments will be helpful in all cases.

Therefore, nothing in the report means we will necessarily use experiments in all cases and any action we take, in any particular case but, for example, in enforcing general consumer protection law, is subject to the position under existing law, including rules about evidence, which Ofcom applies.

EXPERIMENTAL ECONOMIC RESEARCH

Final Report

Ofcom

Prepared by

Charlotte Duke[♦]

Steffen Huck[♣]

and

Brian Wallace[♠]

October 2009

[♦] Charlotte Duke is a Senior Economic Consultant at London Economics.

[♣] Steffen Huck is Professor of Economics and Head of Department at University College London.

[♠] Brian Wallace is a Senior Research Fellow with the Economic and Social Research Council Centre for Economic Learning and Social Evolution at University College London

Experimental economic research

Ofcom

Prepared by

Charlotte Duke

Steffen Huck

And

Brian Wallace

Contents

Contents

Tables & Figures

Page

Executive Summary

This report presents a new method for policy development: Controlled experiments in economics.

The purpose of the study is to assist Ofcom to understand the role experiments in economics can have in understanding consumer behaviour, and to provide an example experiment conducted for Ofcom. The experiment investigates consumer behaviour under different methods of providing call price information.

Experiments in economics

Controlled experiments in economics are like those used in the physical sciences in that they have the features of *control*, *treatment* and *replication*.

Control allows the policy-maker to isolate the features of the market or system that are driving human behaviour (consumer and firm behaviour), and to robustly test cause and effect between the individual feature and the outcome. This is different to field data, because in field data it is impossible to remove all confounding impacts on behaviour in order to isolate causality.

Treatment allows the policy-maker to change one feature or a set of features in a systematic way, and therefore to measure the relative change in behaviour due to changes in the system or market. Again this is very difficult to do with field data because, as mentioned above, there are confounding influences on behaviour, and because it would be necessary to implement the intervention in two very similar field settings which can be costly (or impossible if similar settings cannot be found).

Replication, as in the physical sciences, allows the observations from one experiment to be replicated across different parameter sets (environments), different researchers, sample groups, sample sizes, and locations to check that the results are robust and not just some peculiar feature of the specific experiment.

The method uses humans in the experiment setting, and these participants make choices and complete tasks within the experiment. The experiment setting is stylised, and captures those features of the real world field that are of importance to the policy question at hand. They do not capture all the complicated features of the real world.

The degree of abstraction from the real world field is one of judgement. The important features that drive behaviour need to be included, but the confounding and uncontrollable factors should be minimised. Therefore each experiment is different, and it depends on why the experiment is being used.

Different types of experiments for policy development

There are four broad categories of experiments each with different levels of abstraction and control. These categories are the following:

- Conventional laboratory experiments: *Experiments that employ a standard subject pool of students, an abstract framing, and an imposed set of rules within which decisions are made and tasks completed.*[1]

  Conventional laboratory experiments are invariably the cheapest and quickest to implement. Conventional experiments are often the most abstract, in that they hone in on the drivers of behaviour and eliminate all other confounding factors. Conventional laboratory experiments therefore have the highest degree of control and explanatory power (in terms of cause and effect). These experiments are very good at isolating the drivers of behaviour, and the relative importance of these drivers. Conventional laboratory experiments are not as good at determining how behaviour may vary across different types of participants, and cannot easily extrapolate to absolute magnitude of impacts in the field.

- Artefactual field experiments: *Experiments which employ non-standard subject pools.*

  Artefactual experiments have a high degree of control because they are conducted in the controlled laboratory and focus on the main drivers of behaviour eliminating confounding influences. They often use participants drawn from the real world field, as opposed to university students. As such, these experiments can assist policy-makers to assess how behaviour may differ across different types of participants. If the participant sample is of sufficient size, is drawn from the population of interest, and the experiment uses parameters that represent those in the field, then inferences about the absolute size of impacts in the field can be drawn. However, these experiments are still stylised; time is compressed (i.e. the participants undertake tasks and make decisions in a short time frame with no other competing factors on their time or resources), further we cannot observe participants' true valuations for different goods, and as such limitations will still exist when attempting to extrapolate to aggregate field impacts.

- Framed field experiments: *Experiments with field context in either the commodity, task, or information set that the subjects can use.*

[1] These definitions are taken from Carpenter et. al. 2005.

Framed field experiments are (often) conducted outside the traditional laboratory, for example via the internet with subjects at their home or office.[2] Framed field experiments use context in that subjects know the nature of the goods they are trading, or the tasks they are attempting to complete. Some control can be lost in framed field experiments because field participants bring unobservable experiences, objectives and beliefs to the experiments. Further, in some experiments, such as the pollution experiments conducted by Duke (2008), the use of the terms "water", "pollution", "tax" and "subsidy" when testing a proposed environmental policy may have created some bias.[3] This was particularly the case when farmers were used in the experiment, because the farmers may have wanted to signal to government that they did not want a pollution tax introduced in the region.[4] Therefore, the bias from framing may be particularly important when the issue to be tested is politically sensitive or has received a lot of media attention.

As with artefactual field experiments, if the sample size is of sufficient size, participants are drawn from the population of interest, and the parameters used in the experiment represent the true field parameters, then inferences about the magnitude of impact in the field can be drawn. However, the same caveats will apply.

- Natural field experiments: *Experiments that take place in the field environment in which subjects naturally undertake the tasks and where the subjects do not know that they are in an experiment.*

  Natural experiments are similar to field pilots. The researcher observes and collects information on behaviour, and can introduce treatments by selecting two different groups and introducing different incentives into these two (or more) groups. Natural field experiments have the lowest level of control, and replication can be difficult because exactly the same environment/setting may be difficult to find. However, natural field experiments are the best for extrapolating to the aggregate magnitudes in the field.

[2] We use the term "often" conducted outside the laboratory because the definition proposed by Carpenter et al., 2005, does not require that the experiments are held outside the lab, only that they have field context. However, as a useful discrimination between artefactual and framed field experiments we use the taxonomy that framed field experiments are conducted outside the lab.

[3] See section 3.3 for a discussion of this experiment.

[4] There had been a lot of media coverage about the government's intent to increase the costs of production to landholders in order to make landholders incorporate the external costs of their production on the environment.

These four categories of experiments are not mutually exclusive, and features of each can be combined. For example, the experiment implemented for Ofcom in this study is a controlled laboratory experiment with features of a framed (field) experiment.

The use of experiments for policy development

Experiments in economics can be used for a number of purposes. Traditionally laboratory experiments have been used to *test the underlying principles of behaviour* (economic theory), and learnings from these experiments have both confirmed and disputed traditional theory. In some instances experiments have lead to development of theory. One well known area of this is behavioural economics. Behavioural economics is a branch of economics that incorporates learnings from psychology into economic models of behaviour. The area builds upon and complements traditional economic theory, and identifies where human behaviour may not always conform to the fully rationally economic agent. For example, humans can have cognitive limitations and therefore more information does not always help consumers to make better choices.

Artefactual field experiments can be used to test if the underlying behaviour holds across different types of economic agents. For example, across different sexes, education levels, employment and geography.

Similarly, we may want to forecast how behaviour may change if new or different incentives are introduced. Economic experiments allow policy-makers to observe how actual behaviour changes when interventions are introduced.

Further, we may want to compare the outcomes across different types of interventions to assist the policy choice of which incentive should be used, and how the incentive should be designed. Economic experiments allow policy-makers to compare incentive performance against pre-determined policy objectives.

Representativeness of the experiment

The experiments take place in a stylised setting, and abstraction from the "real world" field is necessary. Abstraction is necessary so that control and treatment can be used, and causality (cause and effect) between the incentives (present in the underlying behavioural framework or in the interventions) and behaviour can be investigated. If the experimental setting is too complicated then the explanatory power of the experiment may be reduced.

Different types of experiments have different degrees of abstraction. Laboratory experiments generally have the highest degree of abstraction and invoke the highest level of control. Natural field experiments have the lowest

level of control; participants are simply undertaking their everyday behaviour.

In between there is a spectrum of abstraction levels, and the level of abstraction depends on the motivation for undertaking the experiment and the questions the experiment needs to answer.

The type of participants used in the experiments can also be important. Traditionally university students have been used as experimental subjects. University students are good candidates as they bring to the experiment very few external influences.[5] University students respond to the incentives created in the experiment, and as such they do not muddy results with unobservable and uncontrollable motivations and beliefs.

University students, however, may, on average, have better computational skills than the general population, and as such they may be better at undertaking the tasks posed in the experiments. Researchers have found that the use of university students is very good at identifying what will not work well in the field. For example, if university students fall prey to behavioural biases then it is very likely that the general population will also. However, the relationship is not symmetric, if university students do not fall prey to behavioural biases then this does not necessarily mean that non-university will also not fall prey.

Economists have now opened the laboratory to non-student subject pools, using samples of consumers from target populations, managers from firms and specialists from particular professions.

Overall, if the fundamental principles that underlie the behaviour are robust, then the outcomes across different sample groups are the same. In other words, if the behavioural theory is robust then outcomes will be the same across different types of participants. What is observed, however, is that the magnitude of the effects may differ across different groups. For example, students may be quicker at learning new tasks because they are more adept at studying than those that are not students.

University students are in general the cheapest and easiest sample group to use. However, if there are a priori reasons to expect that the group of interest is fundamentally different from university students then a representative sample should be used. Of course one can conduct the experiment with both a student sample and consumer (or firm) sample to ensure that any differences in behaviour are picked-up.

[5] University students bring few pre-conceptions of how they should behave to the experiment.

The example experiment for Ofcom

The Ofcom experiment concerns the provision to consumers of telephone number price information. The experiment tests consumer behaviour in five different settings or treatments, representing different policy interventions aimed at improving price transparency. These are the following:

- A baseline treatment that studies consumer behaviour in an environment where there is little price information about call prices and price search is not easy.
- A pre-call announcement treatment in which participants hear an announcement at the beginning of the call which informs them of the exact price of the call. Participants can choose to turn off the announcement and turn it back on as they wish.
- Pre-call announcements in which the consumer is informed of the maximum price the call may cost (again with the option to turn the announcement off and on).
- Price information on the monthly bill in the form of a list detailing all call charges for the telephone numbers such that consumers have access to all relevant price information at the end of a month (or cycle in the experiment).
- Telephone short-codes which are short numbers that can be added to the beginning of a telephone number that allow consumers to learn the costs of different calls more easily and at reduced costs.

The baseline and the interventions are tested in two environments, a landline and a mobile phone environment.

The type of experiment used for Ofcom is a controlled laboratory experiment with framing. It is a controlled experiment because incentives are tightly specified (i.e. behaviour is driven by the incentives created in the experiment and by no other unobservable factors), and the underlying behaviour (tested in the baseline treatment) is compared to the four different ways of providing price information in exactly the same environments. Therefore the only thing that changes is the way price information is provided.

This high level of control allows Ofcom to compare consumer welfare between the baseline and the four alternative methods of providing pricing information, and to compare outcomes between the four alternatives themselves. Causality is established because, as previously stated, only the way price information is provided changes in the environment, and therefore changes in behaviour are caused by the information set-ups only.

The experiment is framed, because participants in the experiment know they are making phone calls, and the objective is to find the "best" phone number to use in order to solve their tasks. The tasks have a value to the participant just as making a call and receiving a service or information has value to consumers in the field.

The strength of this framed controlled experiment is in the comparisons of performance between the interventions when compared to one another and when compared to the baseline. As we have used a controlled experiment with a student participant pool of 211 people, it is more difficult to assess how large the real-world gains may be under the different interventions. In order to extrapolate here it would be necessary to increase the number of participants, to use different types of consumers, and parameters that represent the actual costs of making phone calls. These extensions, will allow inference about the aggregate impacts. However, a natural field experiment would invariably be superior in this regard.

The experiment for Ofcom illustrates that interventions that provide price information in a precise form and at the time of making of the call helps participants to make better choices about the best phone number to use to solve the task they have at hand. Pre-call announcements with the actual price incurred per minute perform unambiguously better than the other methods tested; pre-call announcements with maximum call charges, call prices on the monthly bill and short-codes. The observations are consistent in both the landline and mobile phone environment, and across the IQ distribution of participants in the experiment.

Conclusions

Experiments present the policy-maker with a new method that allows the observation of actual human behaviour in a controlled setting such that cause and effect can be isolated, and relative impacts observed. It allows policy-makers to test the underlying behavioural model to see if in fact consumers and firms behave as the framework predicts. The method allows rapid, and relatively cheap, comparisons of interventions such that unexpected outcomes can be identified early on in the process and undesirable outcomes mitigated. Experiments open the box on the economic agent, and test if the complicated human does operate as economics predicts.

1 Introduction

In March 2009, Ofcom commissioned Charlotte Duke of London Economics, and Steffen Huck and Brian Wallace of University College London, to undertake an experimental economics investigation into consumer behaviour. Specifically the objectives of this research are:

- Assist Ofcom to understand the role experiments in economics can have in helping Ofcom understand consumer behaviour.
- Provide illustration of how different types of experiments have been used to assist policy-making across different policy fields.
- Highlight the strengths and weaknesses of experiments for policy-design and testing.
- Implement an economic experiment that studies consumer behaviour in making telephone calls in a baseline environment where there is little information about call prices and price search not easy.
- Analyse how consumer behaviour may change if alternative methods of providing pricing information are used, and if the alternatives improve consumer welfare as compared to the baseline and compared to each other.

This report therefore presents the following:

- What experiments in economics are and their main features that differentiate them from other research methods.
- The different types of economic experiments that can be used for policy-making and examples of how the different types have been used.
- The experiment implemented for Ofcom, which tests consumer behaviour when faced with different information set-ups for call costs.
- Criticisms of experiments in economics.
- Issues of design when using experiments for policy, and when experiments may not be the best testing method to use.

1.1 Experiments in Economics

Economic experiments are a quantitative method that can be used in the following ways to assist policy-making:

- Test the framework which underlies the understanding of consumer and firm behaviours. For example, in competition policy the underlying framework allows competition authorities to model expected price and quantity setting behaviour of firms in a market. Likewise, consumer utility theory, and the behavioural economics framework, allows policy-makers to predict consumer decision-making in the market. Experimental economics allows policy-makers to test and observe if firm and consumer behaviour conform to the framework predictions, and to identify under

what conditions and situations expected behaviour may differ from that predicted by the frameworks.

- Observe how firm and consumer behaviour may change when interventions are introduced into the market. For example, in the study for Ofcom presented in this paper, interventions that aim to improve price transparency of phone calls are tested and whether these interventions improve outcomes for consumers is observed.

- Compare the performance of different interventions. For example, this Ofcom study introduces different ways of providing price information and compares the change in consumer outcomes across the different interventions.

Examples of these different *uses* for policy-making are presented in section 3.

The hallmarks of experiments in economics are *control*, *treatment* and *replication*.

- *Control* means individual decisions made in the experiment are induced by the incentives created in the experiment and by no other factors. [6]

- *Treatment* is the ability to change specific incentives or features of a policy and to identify how individual decisions change as a result, thus, establishing true causality, i.e. why behaviour is changing.

- *Replication* is the ability for the experiment to be conducted multiple times by the same researcher, by different researchers and across different populations, in order to verify the results.

There are four main categories of experiments.[7] These are the following:

- Conventional laboratory experiments: *Experiments that employ a standard subject pool of students, an abstract framing, and an imposed set of rules within which decisions are made and tasks completed.*[8]

  Conventional laboratory experiments are invariably the cheapest and quickest to implement. Conventional experiments are often the most abstract, in that they hone in on the drivers of behaviour and eliminate all other confounding factors. Conventional laboratory experiments

[6] Control is a feature which can be increased or decreased depending on the type of experiment used.

[7] This taxonomy is proposed by Carpenter et al. 2005.

[8] These definitions are taken from Carpenter et. al. 2005.

therefore have the highest degree of control and explanatory power (in terms of cause and effect), and as such the causality between incentives and behaviours can be isolated with high precision. Therefore, these experiments are very good at isolating the drivers of behaviour, and the relative importance of these drivers. Conventional laboratory experiments are not as good at determining how behaviour may vary across different types of participants (because they use student subjects), and cannot easily extrapolate to absolute magnitude of impacts in the field (because the sample size is generally small).

- Artefactual field experiments: *Experiments which employ non-standard subject pools.*

  Artefactual experiments have a high degree of control because they are conducted in the controlled laboratory and focus on the main drivers of behaviour eliminating confounding influences. They often use participants drawn from the real world field, as opposed to university students. As such, these experiments can assist policy-makers to assess how behaviour may differ across different types of participants.[9] If the participant sample is of sufficient size, taken from a representative distribution of the population of interest, and the parameters used in the experiment reflect the true field parameters, then inferences about the absolute size of impacts in the field can be drawn. However, these experiments are still stylised; time is compressed (i.e. the participants undertake tasks and make decisions in a short time frame with no other competing factors on their time or resources), further we cannot observe participants' true valuations for different goods, and as such limitations will still exist when attempting to extrapolate to aggregate field impacts.

- Framed field experiments: *Experiments with field context in either the commodity, task, or information set that the subjects can use.*

  Framed field experiments are (often) conducted outside the traditional laboratory, for example via the internet with subjects at their home or office.[10] Framed field experiments use context in that subjects know the nature of the goods they are trading, or the tasks they are attempting to complete. For example, consumers of credit (store cards,

[9] However, one should be mindful, that some control and therefore explanatory power over causality can be lost due to unobservable and uncontrolled private features of the specific individuals participating in the experiment.

[10] We use the term "often" conducted outside the laboratory because the definition proposed by Carpenter et. al., 2005, does not require that the experiments are held outside the lab, only that they have field context. However, as a useful discrimination between artefactual and framed field experiments we use the taxonomy that framed field experiments are conducted outside the lab.

credit cards) may participate in an experiment in which they buy different types of credit in the experiment but do not actually buy the credit in real life. Some control can be lost in framed field experiments because field participants bring unobservable experiences, objectives and beliefs to the experiments. However, framed field experiments are useful if type of consumer (or firm) is important. Further, because they are framed (i.e. use the name of the actual good as opposed to a fictitious good), they are useful for illustration.[11] Again, if a sufficient sample size is used, participants are drawn from a representative population, and the parameters reflect the field parameters, then inferences about the absolute size of impacts in the field can be drawn. However, the same caveats will apply as for artefactual field experiments.

- Natural field experiments: *Experiments that take place in the field environment in which subjects naturally undertake the tasks and where the subjects do not know that they are in an experiment.*

  Natural experiments are similar to field pilots. The researcher observes and collects information on behaviour, and can introduce treatments by selecting two different groups and introducing different incentives into these two (or more) groups. Natural field experiments have the lowest level of control, and replication can be difficult because exactly the same environment/setting may be difficult to find. Further, field experiments are the most expensive, and take the longest period of time to complete because the intervention or policy is actually implemented in the field (a pilot). However, natural field experiments are the best for extrapolating to the aggregate magnitudes in the field.

These four categories of experiments are not mutually exclusive, and features of each can be combined. For example, the experiment implemented for Ofcom in this study is a controlled laboratory experiment with features of a framed (field) experiment.

Examples of these different categories (*types*) are presented in section 3.

1.2 The Ofcom experiment

The experiment conducted for Ofcom concerns the provision of telephone number price information to consumers. It is a controlled laboratory experiment which is framed; participants in the experiment are university students, and they interact in a controlled environment such that, the only

[11] Note. It is also possible to use framing in controlled laboratory experiments, as is done in this experiment for Ofcom.

features that change within the experiment are the different ways call cost information is revealed to participants across two environments – a mobile phone environment and a landline environment. Participants know they are making phone calls, but they do not actually use a handset, instead they interact with computer screens electing to make calls or searching for price information given the different information set-ups.

Ofcom requested that four different methods of providing price information to consumers were tested in the experiment. The four price revelation methods are:

- Pre-call announcements with the exact price of the call announced to the caller.
- Pre-call announcements with the maximum price the call may cost announced.
- Price information on the monthly bill in form of a list detailing all call charges for all telephone numbers such that, in principle, consumers have easy access to all relevant price information.
- Telephone short-codes such that consumers can learn the costs of different calls easily at greatly reduced costs.

The performances of these interventions in terms of consumer welfare are compared using two different environments, a landline and mobile phone environment. This is because when using a mobile phone it is more difficult for consumers to search and find price information and price dispersion is much bigger creating greater risks.

In addition to the interventions, Ofcom also wanted to learn about consumer behaviour given the current price revelation systems. Specifically, the questions are:

- Do consumers actively seek out price information?
- Do consumers make different decisions depending on whether they know the cost of a call?
- Do consumers learn from bill shock, i.e., from seeing high-cost telephone calls on their telephone bill?

1.3 Broad observations from the Ofcom experiments

The experiments implemented for Ofcom, provide some clear observations on the relative performance of the different interventions. These observations are the following:

- As compared to the baseline in which price information is not readily available, all interventions tested in the experiment increased consumer welfare significantly.
- Comparing across the different interventions, pre-call announcements with the exact price of the call and the possibility to opt out (turn off the announcement) is the best performing intervention in both the landline and mobile phone environments. Consumers learn prices effectively and then typically switch announcements off.
- The second best performing intervention is short-codes.
- Both price lists on the monthly bill and pre-call announcements where the maximum possible call cost is announced perform less well.
- Pre-call announcements with maximum price information is the worst performing intervention in terms of consumer welfare measures – a result according with expectations.

If we consider consumer behaviour in the baseline where price search is expensive but some price information can be garnered from the bill we can address a number of interesting issues:

- Participants in the experiment do actively seek out price information.
- Further, participants do use the price information they find to make better choices when making calls, namely if they search more they call cheaper numbers.
- The positive effects of search activity do however diminish, and therefore there are costs of excessive search.
- Participants do learn over the course of the experiment and 'bill shock' (unexpectedly large bills) plays an important role. Learning from bill shock is not simply driven by the price of a call as measured by the charge per minute, but is a function of the total costs. If the total costs are large, this attracts participants' attention

and leads to a change in behaviour. In other words, bill shock matters and is driven by the total size of the bill.

When we control for differences in intelligence, using a standard IQ questionnaire which was completed at the end of the experiment, we observe that irrespective of IQ the interventions help consumers to make better choices about the calls they make.

1.4 What does it imply for field policy

The question is one of extrapolation, and how much the observations from a controlled experiment can inform policy-makers about the outcomes in the real world field.

Section 5 of this report discusses extrapolation of experiments. In regard to the experiment implemented for Ofcom a number of important issues can be raised.

1. The interventions have been tested in two environments, a mobile and landline environment. The mobile environment has different parameters to the landline environment. The performance of the interventions is the same in the two environments. Therefore the laboratory observations are not sensitive to precise parameter values.

2. The ranking in terms of the intervention performance is stable across the different environments, again reinforcing that the observed behaviours are not subject to the specific parameters used.

3. The performance of the different interventions is consistent with what we would expect from theory. For example, in the pre-call announcement with maximum price information, much less information is revealed to consumers in this case, and therefore we would expect its performance to be poorer as compared to the other interventions.

4. The experiments have been conducted using university students, and as such the observations from the experiment may overestimate the quality of the participants' choices as compared to the general population. In simple terms we have a smart participant group. This means that the controlled experiment is very good at identifying what interventions work poorly – because if the 'smart' group perform poorly than the average consumer is also likely to perform poorly. However, it is more difficult to predict whether interventions that work well in the controlled lab experiment will also work as well in the field.

We are however confident that the two interventions that perform best in the experiment will also work well in the field. This is because all participants benefited from the interventions including those at the lower end of the IQ distribution for our sample of participants. Further, the environment within which we compare the performance of the interventions is simple, and as such, the controlled experiment probably underestimates the effects of the two interventions that make precise price information readily available when it is needed. In the more complex real world field, the benefit of having the precise price information at the right time is probably larger as compared to having imprecise price information or having to wait until the monthly bill arrives.

The experiment is therefore robust in its external validity for the comparative statics: The relative performance of the interventions when compared to one another, and when compared to the baseline. In terms of how large the real-world welfare gains may be under the different interventions, this is more difficult, and given the sample size used in the Ofcom experiment this can not be done with rigour. In order to extrapolate here, it would be necessary to increase the number of subjects used in the experiment, and to use subjects drawn from a representative distribution of the population of interest (those that make the phone calls) such that statistical inferences about the magnitude of the effects across the general consumer population can be made.

Therefore, the strength of the experiment is in its comparative statics of intervention performance.

2 A brief history of experimentation in economics

The earliest experimental paper published in economics is Chamberlin's (1948) study of bilateral trading in a homogenous market with many buyers and many sellers. Chamberlin induced valuations for the fictitious traded good simply by endowing his students with such values (written on a piece of paper). Chamberlin showed how bilateral trading led to systematic deviations from the expected equilibrium prediction. Namely, demand and supply did not equate at the predicted market-clearing price. Rather Chamberlin reported i) substantial price dispersion, ii) systematically more trade than predicted, and iii) systematically lower prices than predicted.

Chamberlin's early experiments illustrated a situation in which firm and consumer behaviour do not conform to the framework predictions. Specifically, Chamberlin observed that when the institution for negotiation between buyers and sellers is weak in terms of information revelation (i.e. a one-to-one negotiation process), then markets may not move to the efficient equilibrium where the combined benefits to consumers and producers are maximised.

While the first ever recorded experiment in economics demonstrates how useful experiments can be in detecting phenomena that existing theory would not predict, experimental economics in the United States developed in a rather different manner. Actually, it was a student of Chamberlin (and a participant in his experiments) who set out a dozen years later to prove that one can make theory work by choosing the right experimental setup. Vernon Smith published his first paper to this effect in 1962 – it became the cornerstone for the Nobel Prize he received forty years later and defined much of the early phase of experimental economics in the US.

Vernon Smith showed that the predictions of the competitive market framework do hold when the institution in which buyers and sellers interact has a high degree of information revelation on both sides of the market (buyers post their bid to buy and sellers post their offer to sell in a public arena).

This early work was used to test the framework which underlies the understanding of consumer and firm behaviour (section 1), and identify under which situations (the rules of trade – the institution) behaviour differs and conforms to that predicted by the theory framework.

There are, however, two further and quite different roots of modern experimental economics. On the one hand, there was a small group of German economists around Heinz Sauermann who, completely

independently (in fact, rather earlier) and unnoticed by the American mainstream developed their own tradition of conducting experiments. One of Sauermann's early assistants with whom he did several oligopoly experiments was Reinhard Selten who received a Nobel Prize in 1994 – together with John Nash and John Harsanyi – for his work in game theory.[12] Selten was always interested in finding out how people really behaved rather than showing that one could make them behave as assumed in theoretical models. Consequently, the German school of experimental economics was always committed to notions of bounded rationality, learning, and what is now called behavioural economics. In 1982 perhaps the most important paper emerged from the German school – a paper that in many ways has been crucial for the rise of behavioural economics into the mainstream. Guth, Schmittberger, and Schwarze (1982) introduced to the world the "ultimatum game", now the most famous game in all of experimental economics.

The ultimatum game is a simple yet powerful construct. Two players negotiate over a fixed amount of money (or pie). The first player states the share they demand. The second player sees this demand and either accepts or rejects it. If player 2 accepts the proposed split then the respective shares are allocated. If, however, player 2 rejects the proposal then both players receive zero. Orthodox economic theory predicts that player 1 should get virtually 100% of the share. This is because player 2 is better off with any positive amount as compared to zero and so should accept any proposed division. Player 1 should anticipate this and only offer player 2 a nominal positive share.

However, as Guth et al., found this result does not hold in the laboratory. Instead they observed notions of fairness where player 2 rejects the offer if the offer is perceived to be not enough (or fair), and strategic anticipation in which player 1 anticipates that player 2 will reject an unfair share (if the offer is perceived to be too aggressive).[13] The ultimatum game literature firmly established that the orthodox view of "homo economicus" as a perfectly rational and completely selfish individual was flawed. This was the beginning of behavioural economics and the recognition that behavioural biases, from the perfectly rational outcome, can arise.

[12] Game theory is branch of economics that takes account of strategic interactions between market players. Namely, an individual's pay-off (surplus) in the market depends on the choices made by other players, such that they engage in a strategic game to maximise the best outcome for themselves given other players' strategies.

[13] Hoffman, McCabe, Shachat, and Smith (1994), attempted to construct an experiment that showed that the rational expectations outcome does in fact exist. However, they were unsuccessful, and the principles of "other regarding behaviour", that individuals do not always care (only) about monetary pay-offs, but also care about other factors such as fairness, has become an important element in understanding human economic behaviour.

The American school was, however, much closer to the mainstream and for a long while determined how mainstream economics perceived the role of experiments. This drastically changed in the late 90s when the evidence for bounded rationality and behavioural models became overwhelming.

Finally, the third root of modern experimental economics is not be found in economics but in psychology, where the work of Amos Tversky and Daniel Kahneman (who shared the 2002 Nobel Prize with Vernon Smith) established many notions such as loss aversion which now form the core of mainstream behavioural economics. Until very recently, the mainstream would typically distinguish between "behaviouralists" and "experimentalists" where the former were proceeding from Tversky and Kahneman and the latter either from Smith (and the other US-pioneer Charlie Plott) or from Sauermann and Selten.

Today these three branches have sufficiently intertwined such that a determination of the ancestry of particular researchers or papers can be much more difficult.

The main period of growth in experimental economics occurred between the mid nineties and the early 2000s – where practically all mainstream economics journals started to publish experimental work on a more or less regular basis. Much of the early growth was driven by theorists who turned to experiments – often simply to see whether their theoretical predictions would work or not and to get inspiration for new alternative models. Insofar, it is still fair to say that the bulk of the experimental economics literature is game-oriented – with clearly structured interaction and game theoretic benchmark solutions. Prominent examples of this kind of research include next to bargaining games, prisoners dilemma and public good games, games of coordination and conflict, oligopoly games, and trust games.

However, the more experiments become part of the mainstream they establish themselves simply as just another tool at the disposal of the economist who wants to tackle whatever question she is interested in – a tool that has strengths and weaknesses as will be discussed in this report in much detail (section 5).

Crucially, experiments have over the last decade helped to build alternative theoretical models that are increasingly often used to understand many important real-world phenomena. The most prominent examples of this kind are perhaps the models of social preferences (essentially created to capture the findings from the vast ultimatum game literature). For example, Fehr and Schmidt (1999), show that in competitive situations a single purely selfish player can encourage a large number of extremely inequity averse (or fair) players to behave in a completely selfish manner, too. Likewise, under certain conditions for the provision of a public good, a single selfish player is capable of inducing all other players to contribute nothing to the public good although the others may care a lot about equity. Fehr and Schmidt also show,

however, that there are circumstances in which the existence of a few inequity averse players creates incentives for a majority of purely selfish types to contribute to the public good. Moreover, the existence of inequity averse types may also induce selfish types to pay wages above the competitive level.

The findings of Fehr and Schmidt have implications in the real world field such as why people vote when it is optional, pay taxes honestly even if the chance of detection for failing to pay is low, participate in unions and protest movements or work hard in teams even if the rewards may be larger if they instead act individually or selfishly.

The latest trend in experimental economics is to open up the laboratory – experiments are conducted with representative household samples or in the field. However, it is easy to predict that the lab experiment with its supreme levels of control, the precise measurements and excellent cost-effectiveness will remain at the core of the experimental agenda.

3 Using experiments in economics for policy

Experiments can be used in three main ways for policy development. As stated in section 1, these are the following:

- To test the underlying framework that predicts the behaviour of consumer and firms, and to identify when and how behaviour may either conform or differ from the framework predictions.
- To observe how behaviours may change when different incentives or interventions are introduced in the markets.
- To compare the performance of different interventions (or designs of an individual intervention) against the pre-established policy objectives. In other words, does the intervention achieve the policy goals, and how do different intervention designs compare?

Experimentation in economics is akin to experimentation in the physical sciences, in that it uses *control* and *treatment* to isolate why and how individual (or group) behaviour may change, and *replication* to verify the results are robust.

The experiments take place under controlled laboratory conditions, which has benefits in terms of being able to vary some conditions while holding others constant, in order to isolate the specific influence of particular factors. The laboratory set-up also allows tests for robustness of results by repeating experiments and checking replicability. This is of particular importance because traditionally, it has not been possible to conduct tests of outcomes ahead of time. And, it is often infeasible, given the costs, to undertake policy actions across the economy to better understand the outcomes.

Experiments provide empirical data on current behaviour and they help to forecast how behaviour may change in the future, in order to assist policy decision-making.

3.1 The main features of experiments in economics and how they differ from other testing methods

The main features of experiments in economics, as stated in section 1, are *control*, *treatment* and *replication*.

- *Control* means individual decisions in the experiment are induced by the motivations and incentives created in the experiment and by no other factors.[14]

- *Treatment* is the ability to change specific incentives or features of a policy and to identify how and why individual decision-making changes as a result, thus establishing *causality* – the isolation of why and how behaviour changes.

- *Replication* is the ability for the experiment to be conducted multiple times by the same researcher, by different researchers and across different populations, in order to verify the results.

Experiments in economics use real people to make economic decisions in controlled environments. Experiments are therefore different to *simulation* methods because experiments make no up-front assumptions about how consumers or firms will behave. In simulation models it is necessary to construct the individual behavioural model (often called the utility or production function) before running the simulation.[15]

Experiments induce behaviour through the use of real monetary incentives and disincentives just as decisions made in the field have real positive and negative outcomes. This feature differentiates experiments from some of the quantitative valuation methods employed by decision-makers. This includes the methods, *revealed* and *stated preference* which are special types of quantitative methods that set up different future scenarios and ask people to reveal their valuations of different outcomes. These methods can suffer from hypothetical biases as the respondent (the person revealing the valuation) does not actually receive or incur the benefits or costs of the outcome, and as such the revealed or stated preference can be poorly conceived.

The use of replication, and the generation of empirical data for analysis, differentiates experiments from qualitative methods such as focus groups, in-depth interviews or case studies. While such methods allow in-depth exploration of issues, they do not easily provide data suitable for econometric analysis and the transfer of observations to other situations and populations can be difficult. This is because the sample sizes are often small and the opinions and beliefs expressed can be impacted by the group setting.

[14] Control is a feature of experiments that can be increased or decreased depending on the type of experiment used. Different types of experiments are presented later in this chapter.

[15] Namely, in simulation, the behavioural algorithms (equations) must be constructed and these equations include assumptions about how the simulated consumer or firm, for example, will optimise their behaviour when faced with different incentives. Of course these assumptions are based on theory, and often calibrated with observational data form the field.

As is well known to policy-makers, no one testing method is perfect and all have strengths and weaknesses. A recent publication by the Office of Fair Trading and the Competition Commission in the UK called Road Testing of Consumer Remedies[16], provides a typography of how different testing methods – qualitative, quantitative surveys, numerical modelling and simulation, and economic experiments – can be used to pre-test interventions in the economy. Particular focus of the OFT publication is the pre-testing of interventions in the economy to improve outcomes for consumers.

The strengths and weaknesses of economic experiments for policy-making are discussed later in this report.

3.2 Different types of experiments for policy-making

As outlined in section 1, there are many different types of economic experiments that can be used to inform policy design. The types of experiments can be broadly categorised in the following way:[17]

- Conventional laboratory experiments;
- Artefactual field experiments;
- Framed field experiments; and
- Natural field experiments

In policy development, all four types of economic experiment are used. Below, these four types are briefly discussed.

Conventional laboratory experiments have the highest level of control and, because of this feature, they are often used to test specific design features of a policy. The experiments are used as a proof of principle that the fundamental incentives upon which an intervention is built hold in practice in a simple and highly controlled setting. Subjects or participants in the experiment are often university students, and the type of goods and services they exchange in the experiment are often not revealed but simply labelled *good 1,2,…..,n* or *goods 'red', 'white'* and *'blue'*. Conventional lab experiments are the quickest and easiest to implement. However, they are sometimes subject to criticism that the observations from conventional experiments cannot be transferred to the real world because lab experiments lack external validity. Levitt and List, 2005, emphasise that conventional laboratory experiments use non-representative and inexperienced subjects (i.e. students), subjects earn/lose

[16] Road Testing of Consumer Remedies, 2009, http://www.competition-commission.org.uk/our_role/analysis/road_testing_report.pdf . A study completed by London Economics (with specialist input by Steffen Huck), for the OFT and CC.

[17] This taxonomy is proposed by Carpenter et. al. 2005.

small stakes as compared to those in the real world, and the settings are artificial and do not readily extrapolate to the real world. On the other hand, there are several studies that show how the findings from laboratory experiments do carry over into real-world settings, see, for example, Abeler and Marklein (2008) in a recent study on biases in consumer behaviour that occur in both an abstract laboratory and a natural field experiment.

While external validity is debated amongst experimental economists, the true strength of experiments is internal validity. By varying only one aspect of the environment at the time experiments can help to establish true causality. If two experiments are identical with the exception of one aspect and different behaviour results, one can be certain that the observed difference in behaviour was a consequence of the variation.

While many experiments in economics focus on identifying the effects of incentive mechanisms several others analyse the role of context and social norms[18] both can be useful for informing policy. Experiments are also very useful for testing small design features as compared to testing the full operating policy.

Artefactual experiments are conducted in the laboratory and use subjects that may, for example, be experienced in undertaking the experimental tasks in a real world setting (for example, bond traders in a financial market experiment). Alternatively, the experiments may use different types of subjects (i.e. men and woman, undergraduate and graduate students, or young and older individuals) to test if the fundamental incentives hold across different groups in our economy. And, if the participant sample is of a sufficient size, inferences with precise statistical properties about the target population can be drawn. Artefactual experiments are useful if the type of individual is considered important or past knowledge and experience is important.

Framed field experiments are conducted outside the traditional laboratory, but not necessarily in the 'real-world' field. Framed field experiments use subjects that are familiar with the setting in which the intervention may be implemented (as can also be the case in artefactual experiments). Further, subjects in the experiments will often know the nature of the good or service that they are exchanging (for example, they know they are buying mortgages or they know they are buying pollution permits). Some control can be lost in framed field experiments as subjects may bring behavioural biases learnt in the 'real-world' to the experimental setting and then make their decisions

[18] For example, the water market experiments conducted by Duke 2006, used conventional laboratory experiments before moving to framed field experiments. This was because water is a highly political topic, and using framed field experiments ran the danger of muddying the results because subjects may have behaved how they thought the experimenter or society believed they should behave, as opposed to responding to the private incentives being tested in the experiment.

according to their real world experiences as opposed to responding to the incentives in the experiment. Framed field experiments are useful for illustrating to stakeholders how new policy designs and incentives operate, and can be used to ease the introduction of new policy by being used as a training tool.

The fundamental feature of **natural field experiments** is that the subjects do not know they are participating in an experiment and the subjects naturally undertake the tasks which the experiment is attempting to observe. A recent example of a natural field experiment on consumer behaviour can be found in Huck and Rasul (2007) who compare different fundraising schemes in an experiment carried out in conjunction with the Bavarian State Opera in Munich. For the participants the experiment just looks like any other fundraising call and they would normally be not aware of the fact that other recipients would receive slightly different letters (that, for example, mention the presence of a lead donor). Natural field experiments have the lowest level of control and treatment.

3.2.1 The Ofcom experiment

The experiment conducted for Ofcom in this study is a controlled laboratory experiment with framing because of the following features:

- The highly controlled design allows direct comparison between different information set-ups, because the set-ups are implemented in exactly the same environments (mobile and landline). In other words only how the price information is revealed to subjects changes nothing else changes.[19]

- The environment is stylised, subjects do not actually use a phone to make a phone call, but instead interact with a computer screen electing to make calls by clicking on a call button, or searching for price information by clicking on search. Likewise they do not actually hear pre-call announcements on price, but instead click on a button and are shown the pre-call announcement while a real time clock simulates the actual time incurred by the consumer if they were hearing the announcement.

- The environment is framed as participants know they are making phone calls and that the service received by making a call has value to them.

[19] The costs associated with different numbers does change in order to control for learning in the experiment. See the experiment design section.

- The experiment uses university students, and not subjects drawn from the wider population, and therefore the subjects are reasonably proficient at undertaking tasks, although they had not participated in similar experiment previously.

The Ofcom experiment could be expanded such that more and more realism is built in. For example, actual phone handsets could be used to simulate the real world feeling of making a phone call. Subjects could be drawn from outside a university student population, to see if the behaviour holds across different types of people, for example different education levels, age or income brackets.

It is unlikely, however, that a natural field experiment could be used, because it would be necessary to introduce different information revelation set-ups in the actual field and for some consumers to receive one type of set-up and for others to receive another. This may not be possible for legal reasons and, if it were, it would be more expensive as would be a framed field experiment. On the other hand, a laboratory experiment allows a simple and robust test of the interventions Ofcom is interested in.

More real world complexity does not always improve the information value of experiments, and the increase in costs associated with increased real world realism is often not justified. The issue of external validity and economic experiments – the ability to extrapolate beyond the laboratory to the field are discussed later - but in the case of Ofcom, and the experiment undertaken in this study, two potential expansions could be considered. The use of different subjects to see if the observed behaviour holds more widely and, the use of a larger subject pool such that greater extrapolation of the aggregate impact on consumer welfare can be undertaken.

3.3 Examples of experiments for public policy design and testing

Experiments in economics and their application to policy design is an expanding area of applied policy development internationally. In this sub-section examples of the different *uses* and different *types* of experiments for policy development are presented.

In summary the different uses are:

- Test the underlying behavioural framework;
- Observe how behaviour changes when different incentives are introduced; and,
- Compare performance of alternative incentive designs.

The different types of experiments can be classified as:

- Conventional laboratory experiment;
- Artefactual field experiment;
- Framed field experiment; and,
- Natural field experiment.

Table 1 presents these examples.

Table 1: Experiments for policy development

Experiment type	Use of experiment	Description
Conventional laboratory experiment to test the impact of information feedback on sellers' trustworthiness, and consumers' ability to select trustworthy sellers e.g. e-bay style information feedback *Learning Trust,* (Bohnet et al., 2005) Student subjects drawn from a university campus.	**Test the underlying behavioural framework:** When a consumer can only learn the quality of the good they have bought after they have bought it, then the buyer (and seller) may suffer the cost of *moral hazard*. **Observe how behaviour changes** (both firm and consumer) if information feedback on seller quality is provided. **Compare performance** of different information feedback in mitigating moral hazard.	When consumers purchase goods over the internet, or purchase hotel rooms or holiday tours they can only observe the quality of the good or service after they have bought it. These goods are often called *experience goods*. Namely, the seller may advertise the quality of their product as 'high quality'. The consumer then chooses whether to believe the seller's high quality claim. If the consumer does believe the seller's claim then they buy the good. Only at point of receipt can the consumer verify the quality of the good. Moral hazard arises because some sellers lie. That is, sellers claim they have high quality but, in fact, they supply low quality because it is cheaper for them to do so. Typically, lying firms will be driven out of business by truthful firms, but this can take some time and consumers can suffer as a consequence. A conventional experiment by Bohnet, Harmgart, Huck and Tyran, 2005, tested the market outcomes under four different information set-ups; (1) no feedback information on sellers' histories in regard to quality or truthfulness of each seller, (2) all buyers received information on sellers' histories, (3) all sellers received information on sellers' history, (4) all buyers and sellers received information on the sellers' histories. Economic theory predicts that providing consumers with information feedback on sellers' trustworthiness does help consumers to select high quality and trustworthy sellers. However, theory does not predict that, when sellers receive information about their own quality and truthfulness, and the quality and truthfulness of their competitors, then all sellers become more trustworthy. In this case, theory is silent; it should not matter if sellers receive feedback. The reason for this effect, which theory does not predict, is that not all sellers realise the positive impact of reputation, and believe that they can continue to operate in an untruthful way. When sellers' receive feedback on their competitors, the untruthful sellers realise their truthful competitors are doing better and thereby change their own behaviour to be more truthful. When both buyers and sellers receive information, buyers are able to trust sellers to a larger extent, and sellers' trustworthiness increases. This is borne out in the emergence of consumer websites such as "tripadvisor.com", and the use e-bay style seller ratings.

Table 1: Experiments for policy development		
Experiment type	**Use of experiment**	**Description**
Conventional laboratory experiment to test the impact of price floors in duopoly markets (two firms in the market). *Price floors and competition* (Dufwenberg et., al. 2007) University students take the role of firms and consumers are simulated through demand functions. This is an example of combining economic experiments with simulation methods.	**Test the underlying behavioural framework:** When at least two firms compete on price for identical goods (Bertrand competition), competition between firms will drive price below the equilibrium outcome for the market. This can lead to a decrease in the number of firms thereby ruining competition in the market and potentially resulting in a situation where only one firm is left in the market. Price floors are sometimes used by regulators to mitigate competition.	Economic theory frameworks predict that the use of price floors set above the market equilibrium (otherwise they would be of no effect), will increase market price from the equilibrium to the higher price floor. The economic experiment implemented by Dufwenberg et. al., finds, however, that the outcome is completely the opposite to that predicted by orthodox theory. In duopoly markets (two firms) price floors actually reduce (not increase) market price even in one-shot interaction (where the firms interact only once) where repeated-game arguments (in which firms interact multiple times), which say collusion is easier to maintain with harsher punishments, do not apply. The key reason for this surprising result is that firms do not reach the competitive equilibrium in the absence of regulation. Rather, many firms manage to collude and sustain prices above marginal costs. However, once the floor is introduced it generates a new equilibrium and with prices above marginal cost collusion breaks down and prices drop from the high collusive level to the lower price floor. This phenomenon does not occur with four-firm oligopolies, however, because the competitive outcome is reached in the oligopolisitic market (firms are unable to collude), and therefore the introduction of a price floor increases market prices. This is an example of how experiments can be used to identify under what conditions behaviour conforms to the predictions of the behavioural framework (theory) and when it differs. This exercise would have been much more difficult (if impossible) in the 'real-world' field because it would have been necessary to find two examples of Bertrand competition, one with two firms and one with four firms and then to introduce price floors into these two markets and observe the behaviour of firms and consumers.

Table 1: Experiments for policy development		
Experiment type	**Use of experiment**	**Description**
Conventional laboratory experiment that investigates the use of price regulation in markets for experience goods. *Pricing and trust* Huck et. al. 2008. University students take the role of firms and consumers.	**Test the underlying behavioural framework; and,** **Observe how behaviour changes when regulation is introduced**. In situations where there is asymmetric information (firms have information about their goods which consumers do not have), and in which consumers cannot determine the quality of a good until after they have bought it (experience goods), then moral hazard may arise and high quality firms are driven out of the market	In situations where the consumer cannot directly observe the quality of a good prior to purchase and consumption, regulators sometimes introduce minimum standards or verification schemes to mitigate this information problem. Such schemes can be costly to introduce and manage, and therefore regulators may want to investigate alternative regulations which achieve similar outcomes (ensure quality is maintained for consumers). Huck, Lunser and Tyran (2008) investigate the use of regulated prices in four-firm oligopolistic markets and in monopolistic markets. In both market set-ups consumers know the history of the sellers (see the first example in this table for a discussion of the effect of this), such that trust can be established. The experiment then tests the impact of introducing regulated prices as compared to allowing firms to set their own prices. A surprising observation is that the quality of goods traded, and the volume of trade, increases in both the monopolistic and oligopolistic settings. Orthodox theory would predict that if price is set above the competitive equilibrium in the oligopolistic market, then (given demand functions are downward sloping), quantity traded will decrease and quality will increase. The reason for this unexpected effect is subtle. Without regulation consumers tend to buy from the cheapest firm as opposed to the firm with the best quality record. This drives price down to marginal cost which has two effects. For firms the provision of high quality goods essentially stops being profitable and for consumers the price is so low that receiving a low quality good is not so bad anymore. In other words, low prices that emerge reduce the incentives for firms to provide high quality and reduce incentives for consumers to be extra careful about quality. On the other hand, with a regulated price firms only compete on their reputations gained from delivering high quality in the past (because they can't compete on price) and consumers can pay full attention to these reputations. Both effects together increase total welfare in these markets.

Table 1: Experiments for policy development		
Experiment type	**Use of experiment**	**Description**
Artefactual field experiment to investigate **a) gender based behavioural differences** **b) employment based behavioural differences** **c) location based behavioural differences** **in regard to the propensity to cheat in transactions** *Gender and Corruption: Insights from an experimental analysis (Alatas et al 2006)*	**Testing the underlying behavioural framework:** Orthodox economic theory is generally silent on the differences in behaviour between different types of people, and people in different locations. However, regulators are often very interested in differences across different types of consumers. Artefactual field experiments use different types of consumers to assess if behaviour is different across types.	Sometimes behavioural differences across different genders, employment types or geographic locations may important. Such behavioural differences could drive the propensity for different people to cheat or behave truthfully. This for example could be important in self reporting regulation schemes such as taxation. Atlas et al 2006, undertook such an experiment to investigate the impact of gender, employment and geography on corruption. Specifically they sought to test: Do men and woman have the same propensity to cheat in a transaction? Is the propensity to cheat the same across different geographic regions? Does employment status have an impact? The authors used male and female university students in three different geographical locations (Australia, India and Indonesia), and public servants and students in the same geographic location (Indonesia). Understanding behavioural differences across gender, employment and geography can help policy-makers to target interventions more effectively. The artefactual field experiment illustrates how economic experiments can be used to evaluate differences between different types of individuals where standard economic theory would not predict such differences. The experiment observations observed that in Australia, women are more likely to punish cheating compared to men. However, in Indonesia and India, there were no observed differences between genders. When observing behavioural differences between public servants and students, public servants were less likely to undertake cheating activity as compared to students, and this is because the public servants were concerned about the overall welfare impact of cheating as compared to students who sought short-term gains more often.

Table 1: Experiments for policy development		
Experiment type	**Use of experiment**	**Description**
Artefactual field experiment to test the performance of pollution taxes University students and water managers take the role of firms emitting pollution into waterways due their production. *Using prices to manage environmental externalities: evidence from a field experiment* Duke 2008.	**Test the underlying behavioural framework** and compare if behaviour differs when different types of participants are used in the experiments.	Market performance across four treatments is first investigated. Treatment 1 is a no-regulation baseline treatment in which subjects trade water contracts with no salinity (pollution) policy operating. Treatment 2 is a salinity tax which is incurred by buyers of water if the water contract is traded into a higher externality zone. Treatment 3 is a salinity subsidy treatment in which a subsidy is paid to sellers of water who sell to buyers located in a lower salinity externality zone, and treatment 4 is a tax plus subsidy treatment. In this combined tax-subsidy treatment buyers of water must pay a tax to the regulator on trades that increase salinity and sellers receive a subsidy from the regulator on trades that reduce salinity. Comparing outcomes across the (four) treatments allows a check that the markets are working as expected; market outcomes are moving towards the predicted equilibrium for each treatment. Once the performance of the system has been examined, the impact of subject pool within a treatment is then investigated. The results suggest that subject pool choice may impact upon the magnitude (but not the direction) of market outcomes in some policy environments. Market knowledge subjects may be hardened against the tax, and are accepting of the subsidy. Landholders must pay many taxes in their business, and a tax on water is not palatable for these economic agents. Student subjects on the other hand, do not run agricultural businesses and do not have experience in the policy environment. Student subjects use less private unobservable information in their decisions as compared to market knowledge players (at least) in the environment and policies tested here. Two context relevant issues may be influencing market knowledge subjects' behaviour. First, loss aversion, a well known concept in both psychology and economics, may be driving market knowledge subjects to value losses of potential surplus from the salinity tax more than potential gains from the salinity subsidy. While student subjects may also exhibit this feature, market knowledge subjects who are familiar with the costs and benefits of agricultural production exhibit a stronger effect. Secondly, the context and use of the words salinity and water probably created a perception in market knowledge subjects that the experiments could be used as an opportunity to inform government about their own (private) opinions surrounding salinity in the Murray. Market knowledge subjects may be signalling to regulators that taxes will not work in the field. This is an example of where student subjects may behave closer to what the underlying theory predicts (because all incentives can be controlled for), while the magnitude of the market players' behaviour is slightly different because they are bringing unobservable private information to the experiment. See the discussion of weaknesses of experiments for a discussion of differences across subjects in experiments.

Table 1: Experiments for policy development		
Experiment type	**Use of experiment**	**Description**
Framed field experiment to test the impact of mandatory information disclosure *The Effect of Mortgage Broker Compensation Disclosures on Consumers and Competition: A Controlled Experiment,* (Lacko, 2004, for the Federal Trade Commission)	**Test the underlying behavioural framework:** Inadequate information supplied to consumers (imperfect information) **Observe the impact of an intervention** to improve information revelation.	How information is provided to consumers is important. Lacko, 2004, for the Federal Trade Commission, used a framed field experiment to test the impact of mandatory information disclosure on broker compensations for consumer mortgages. The experiment tested the regulatory requirement that mortgage brokers would need to prominently disclose to the consumer the commission that they receive on loans made (called a YSP) as compared to revealing only the total cost. This was driven by a concern by policy-makers that brokers were placing borrowers in above market rate interest rate loans without the borrower's knowledge. The imperfect information was feared to lead to a welfare reducing outcome because borrowers could not select the least cost loan. Lacko implemented framed field experiments in a shopping centre setting. Participants knew they were involved in an experiment, and had experience in searching for and purchasing mortgages. In order to test the hypothesis, subjects were presented with two different information disclosure sheets, one that prominently disclosed the YSP, and one that reported a single total cost. Traditional economic theory predicts that in aggregate economic agents rationally use all available information to make optimal decisions. However, the experiments found that additional information disclosure added confusion to borrowers' choice of product and resulted in mistaken choices: A lower proportion of subjects chose the lower cost loan when they were shown the additional information; 60-70% with additional disclosure as compared to 85-94% without additional disclosure.

Table 1: Experiments for policy development		
Experiment type	**Use of experiment**	**Description**
Natural field experiment to test how the presentation of different choices, as opposed to the inherent value of different choices, impacts upon consumer behaviour. *What's Psychology Worth? A Field Experiment in the Consumer Credit Market,* (Bertrand et al. 2006) Participants did not know they were participating in an experiment. They simply received a letter from their lender offering a new loan.	**Test the underlying behavioural framework:** Do consumers make choices rationally by weighing up the costs and benefits of each choice, or do consumers make choices using other psychological information? **Observe if behaviour changes:** Does different presentation, or framing, of loan information change consumer behaviour?	**Treatment 1**. How the loan offer was described: Some borrowers received a letter with a single example of repayment for a given maturity. Other borrowers were given multiple examples of loans for different amounts and maturity. In all cases, borrowers were told that different loan sizes were available, but only in one treatment were different examples provided in the letter. Economic theory predicts that simple presentation of multiple examples should not affect take-up, and may in fact increase it because transaction cost associated with calculating different repayment rates is minimised. Behavioural research suggests that more choice may induce decisional conflict and reduce take-up. In the experiment it was observed that a single example increased take-up as compared to multiple examples, and this effect was large for both lower and higher cost loans. Therefore, providing consumers with additional examples and information does not necessarily help consumers **Treatment 2a**: Provision of comparison rates: Some borrowers received information on other lenders' interest rates, and some did not. **Treatment 2b**. How the comparison rate was framed e.g. "save if you borrow from us" or "lose if you borrow elsewhere". Economic theory predicts that comparisons and framing should have little effect because borrowers should be informed about market prices. Behavioural research, however, suggests that framing manipulations can affect choice. The experiments observed that the provision of comparison interest rates had no statistical impact on loan take-up. However, the framing "you will lose if you borrow elsewhere" increased take-up of the loans for both higher and lower cost loans. **Treatment 3**: Inclusion of a photo with a person smiling (same/different race as the borrower, male/female). Economic theory would predict that the photo would have no impact at all on loan take-up. Behavioural research suggests that source attractiveness and source-recipient similarity are attributed more favourable traits. The experiments observed that including a photo of the same or different race, as compared to the borrower, did not have an impact on take-up. Including a picture of woman increased take-up significantly by male borrowers, However, there was no impact on female borrowers of including either a male or female photo.

3.4 Previous experimental findings on price transparency

Somewhat surprisingly, there has not been much experimental research on the effects of price transparency for consumer behaviour and market outcomes. There are however a few notable exceptions, most of which focus on the strategic behaviour of firms and the resulting impact on consumers as opposed to specifically testing consumer behaviour. Davis and Holt (1996) test the Diamond paradox that predicts that for homogeneous goods the smallest consumer search costs can raise equilibrium prices from the competitive to monopoly levels. This extreme prediction is not borne out in the laboratory data although market prices are increasing in search costs. More support for the Diamond prediction is reported in an experiment by Cason and Friedman (2003). For policy applications this implies that relatively small search costs are less of a worry than theory predicts. Rather, the amount of attention the regulator might want to pay to search costs should be increasing in these search costs. Lynch and Ariely (2000) study the role of price transparency in a field experiment with online wine retailers. They show how lower costs for price searches increases price sensitivity of consumers. However, they also show that this does not necessarily induce ruinous price competition, where firms prefer to exit the market, if it is offset by a simultaneous increase in quality transparency. Therefore, in a policy context, if sellers can differentiate themselves based on quality features/differences, then even if the cost of searching for price information is very low, competitive pressures exerted by consumers do not lead to excessive competition where quality cannot be profitably maintained.

In another field experiment Brown, Hossain and Morgan (mimeo 2007) show that shrouding shipping costs increases revenue for online retailers – to the detriment of consumers. As such, policy-makers may want to require online retailers to clearly publish any shipping costs alongside the costs of the product itself.

Baye et al. (2006) study the introduction of the Euro as a natural field experiment. In online markets the common currency makes price comparisons easier and, thus, increases price transparency. The authors show that, quite counter-intuitively, this can increase prices due to strategic responses by retailers. Essentially, the intuition for this follows three steps. First, more transparency implies more competitive pressure for consumers who shop around by comparing prices. Second, more competitive pressure in this market means reduced profits for firms. Third, reduced profits for targeting consumers who shop around increases the incentives to simply rely on "loyal" consumers and charge higher prices to them.

Remarkably, this is borne out in the data (taken from Kelkoo on gaming consoles, computer games, music CDs, PDAs, printers and scanners). After the introduction of the Euro, the authors document a price increase of 3% that can be attributed to the increased transparency (that is, after controlling for cost, demand and market structure effects). None of the published literature directly addresses the policy questions tested in this study.

4 The Ofcom experiment

In this section we present the controlled laboratory experiment conducted for Ofcom.

4.1 Experimental design

The experiment is designed to capture the essence of what it means to make phone calls in an environment where prices are not readily available and lengths of calls are stochastic in nature. This reflects the situation faced by consumers when making non-geographic phone calls in the field.

There is a landline environment and a mobile phone environment that mainly differ in how costly it is to search for price information and the risk involved due to different degrees of price dispersion. It is three times as costly to search in the mobile phone environment as it is in the landline environment, capturing the real world features of "being on the move" with a mobile phone and therefore facing more difficulty (more costs) to undertake a price search using the mobile as compared to, say, the internet at home.

In both environments, subjects have to complete different "tasks" by making telephone calls. While these tasks are framed in an abstract manner the idea is that they resemble real-life tasks like finding out something from your bank or ordering a pizza from a delivery service. Altogether there are 9 different tasks in the experiment and, for each task, there is a premium that is paid to the subject on completion. This resembles the reward of receiving the desired information from the bank or eating the pizza that was ordered. All 9 tasks together form one task cycle and each of the two parts or phases of the experiment consists of 14 such cycles.

The first part of the experiment, i.e. the first 14 cycles, represents the baseline – a situation where price information about telephone numbers is not readily available. The second part, comprising another 14 cycles, resembles one of four different interventions that provide more price information. These interventions are described in detail in the next subsection.

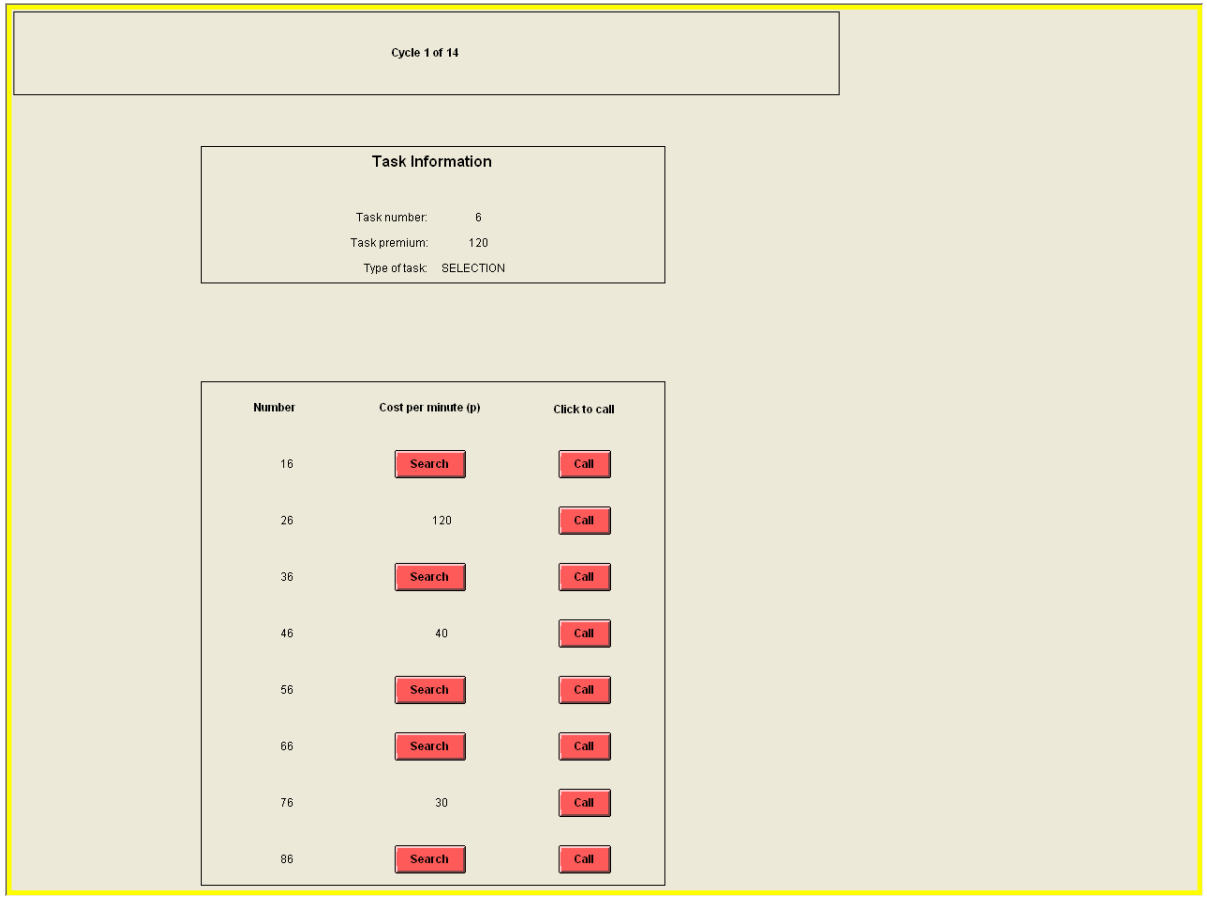
There are two types of task that must be completed by subjects in both parts of the experiment (the baseline and the interventions). The first type of task can be solved by calling one of several different numbers. These are called "selection tasks", and represent a situation where the consumer has the choice of different numbers to call in order to get the same information, such as different airline telephone booking services. The second type of task is where there is no choice of number but where subjects can decide not to call at all. These are called "binary tasks" and mimic a situation such as calling the water supply company or instead using the internet to collect the

information/make an appointment. The problem in selection tasks is to find the cheapest number that solves the task. The problem in binary tasks is to find out whether or not it pays to complete the task at all.

Initially, subjects have very little information about calling costs. They are not told the call charge per minute nor do they know the length of calls, which replicates the situation of real consumers before they begin to search for any information on call costs. While charges are fixed, call durations are stochastic.

In the baseline, subjects can actively search out price information. In the experiment they do this by simply clicking on a search button. On the screen subjects can see all the numbers that can solve the current task. Next to each number are two buttons, a search button and a call button. If subjects click on the search button they incur the search costs and the call charges per minute appear next to the number. If subjects click on the call button, the call to this number is initiated. See Figure 1 that shows a screen shot for one of the selection tasks.

Figure 1: Choosing which number to call



In order to resemble that such search is anything but trivial in real life, monetary search costs are rather high. This mimics, for example, the cost of going to the internet sites of different phone carriers to search and find the information on the different call costs, or seeking out advertising material that has the costs published.

Subjects can perform as many searches as they like. Once they are satisfied with what they know they can pick a number to call in the selection tasks or decide whether or not to call in the binary tasks. They do this by pressing the call button. Once they press this button, the passing of time is simulated by a counter. Time ticks away in 2-second intervals each of which represents 30 seconds that will be charged according to the price of the chosen number. Calls can be terminated at any point in time but if they are, no premium is paid. We opted to include such elements of real time as we considered them an important ingredient of making phone calls in real-life.[20]

At the end of a telephone call subjects know how long it took and, if they knew the call charge per minute, they can, in principle, compute the total costs of the call but they are not shown this information. Only at the end of a completed task cycle will they receive this information on the screen – in the form of their phone bill for that cycle. Bills are itemized and list, for every call made, the duration and total costs. The charge per minute is not shown but can obviously be derived by a simple division – although subjects' ability to perform such arithmetic may vary – just as this ability may vary in real life.

Subjects' earnings have four components: the total premia earned for task completions minus the call charges incurred minus the search costs plus a time premium that falls in the actual time spent on the completion of all task cycles.

The last component enhances the role of real time in the experiment. The passing of time and, in particular, the actual delays that occur when calls take a long time are designed in order to mirror the psychological costs of waiting on the phone in sometimes long queues when nothing really happens. Essentially, one might view this as a design trick to induce a certain impatience into the experiment that we believe plays a role when making phone calls in real life.

As mentioned earlier, there are two types of task, selection and binary tasks. The tasks differ further in call charges, call durations and premia. For example, there is one selection task where all numbers are relatively cheap; another where all numbers are relatively expensive; and yet another where there are large differences between call charges.

[20] Time is used in experiments where it is considered an important feature of the environment to be tested.

The mobile and landline environments are structurally identical but differ in their parameters. Mobile call charges and mobile call premia are twice that of landline charges and premia. Mobile search costs are three times higher, which, as previously mentioned, reflects the idea that while on the move it is extremely difficult to find out call charges, especially when the phone does not provide internet access.

4.2 Treatments

The experiment implements a 2x4 design. On the one hand, there is the landline versus mobile split, and on the other, there are four interventions that increase price transparency. These interventions are

(a) Precise pre-call announcements where subjects are informed about the call charges after clicking on the dial button. Subjects have, however, the possibility to opt-out of this scheme at any time in which case the announcements cease;

(b) Imprecise pre-call announcement (also with opt-outs) where subjects are informed about the maximum call charges after clicking on the dial button;

(c) Price information on telephone bills where subjects can see all the relevant call charges in an annex to the phone bill at the end of each cycle;

(d) Telephone short-codes that can be used to learn about prices that in the experiment simply reduce the search costs dramatically (by 94% and 98% for landlines and mobiles respectively).

In all treatments, subjects first complete a full baseline cycle. Then the intervention takes place. In order to avoid confounds between previous learning and the intervention, there is a new set of telephone numbers in the second phase of the experiment. However, structurally, the intervention phase is identical to the baseline phase and for each number that previously had some particular call charges there is now a new number with exactly the same charges. It is this constancy of the environment that allows us to measure the impact of the interventions precisely. Specifically, we can, for each subject, compare their baseline performance with their performance under the intervention they have been assigned to and then conduct statistical tests on these differences which helps us to filter out subject-specific idiosyncrasies.

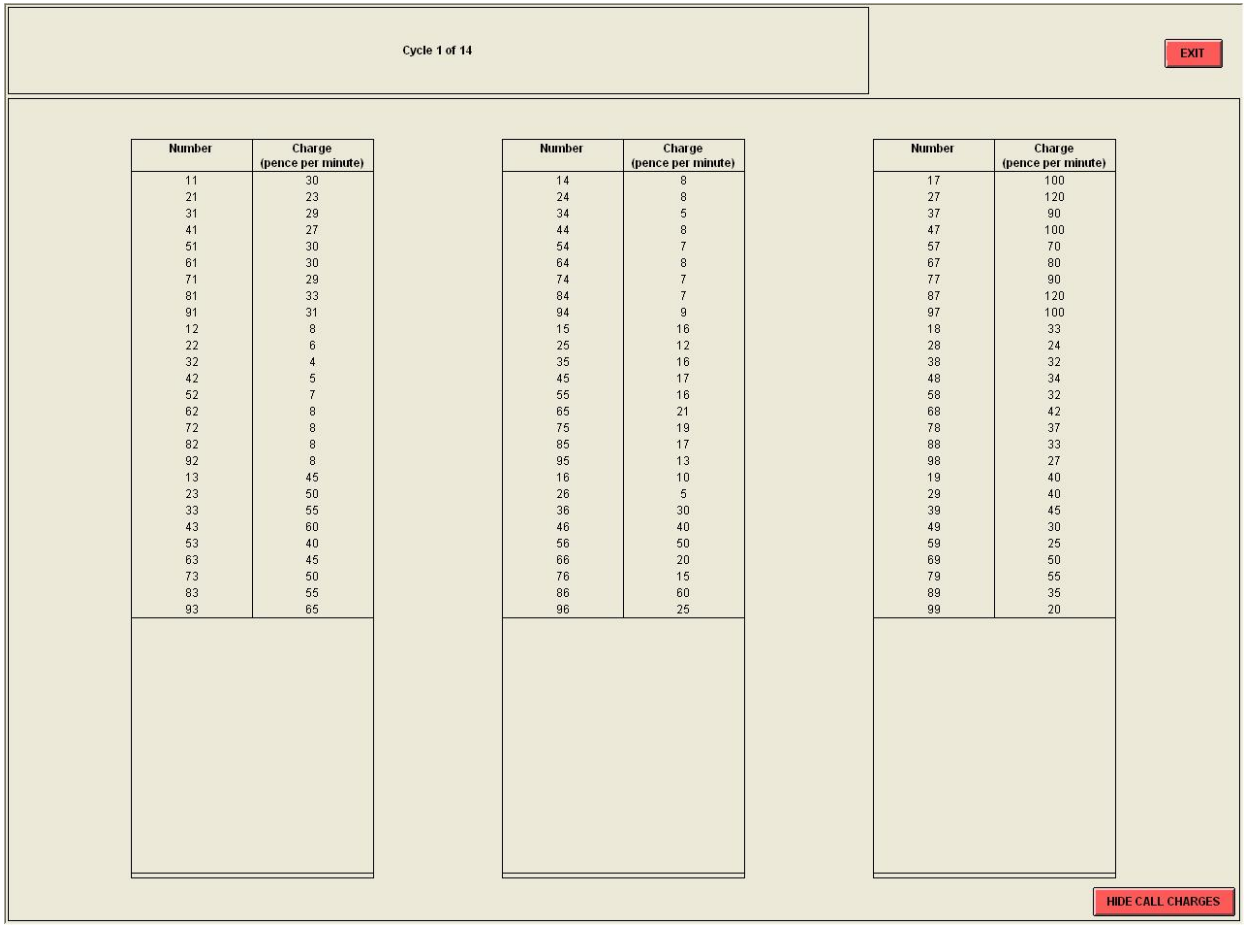
There are a couple of noteworthy details about the design of the intervention treatments. Regarding the pre-call announcements, we have introduced small real-time delays between the announcement and the actual start of the call – mirroring the time that would pass in real life listening to the announcement. Similarly, we have built in some time delays when subjects decide to opt-out

or opt-in again, reflecting the idea that such changes in status would require some time and effort.

Regarding the bill information, one should note that while the numbers are displayed systematically on the bill, the order does not reflect the grouping of numbers by task. On a bill numbers are simply ordered numerically, while task groupings are rather different. That is, the last digit is kept constant for numbers that solve the same task. See Figure 2 for a screen shot of a bill with price information.

Finally, we should perhaps add the observation that interventions (a) and (b) are only different for selection tasks while the pre-call announcements for the binary tasks where there is only one number are identical.

Figure 2: The telephone bill



Number	Charge (pence per minute)
11	30
21	23
31	29
41	27
51	30
61	30
71	29
81	33
91	31
12	8
22	6
32	4
42	5
52	7
62	8
72	8
82	8
92	8
13	45
23	50
33	55
43	60
53	40
63	45
73	50
83	55
93	65

Number	Charge (pence per minute)
14	8
24	8
34	5
44	8
54	7
64	8
74	7
84	7
94	9
15	16
25	12
35	16
45	17
55	16
65	21
75	19
85	17
95	13
16	10
26	5
36	30
46	40
56	50
66	20
76	15
86	60
96	25

Number	Charge (pence per minute)
17	100
27	120
37	90
47	100
57	70
67	80
77	90
87	120
97	100
18	33
28	24
38	32
48	34
58	32
68	42
78	37
88	33
98	27
19	40
29	40
39	45
49	30
59	25
69	50
79	55
89	35
99	20

4.3 Procedures

For each treatment we recruited between 20 and 30 subjects from the UCL-ELSE subject pool which consists of all kinds of UCL students who have registered their interest to take part in economic experiments. In total, we observed 211 participants in the main experiment. (In the pilot phase that preceded the main experiments we observed another 60 subjects. As we made several small adjustments after the pilots we do not include these subjects in our data analysis.) [21]

The experiments were fully computerized using Fischbacher's (2007) z-tree. Instructions were, however, distributed on paper. During the experiment subjects were sitting in isolated cubicles ensuring they could only focus on their own tasks. Subjects were neither allowed to talk, nor were they allowed to take any notes. However, they were invited to ask clarifying questions. For that they would have to raise their hand and the experimenter would come to them to answer the question privately.

After the experiment, subjects were immediately and privately paid in cash. The average duration of an experimental session was 150 minutes and the average payment was £23.65.

4.4 Data Analysis and Results

4.4.1 Aggregate Performance

The main results of this experiment are presented in Table 2 to Table 5, which show for both, the landline and the mobile treatments, how subjects performed in the baseline and under the four interventions. Table 2 and Table 3 show average pay of subjects – which is the relevant overall consumer welfare measure. Table 4 and Table 5 show subjects' pay excluding the search costs they incur. One could argue that in real life search costs might be impossible to measure such that any realistic approach to measuring consumer welfare in the field would probably exclude them.

All tables show overall performance over all 14 cycles as well as performance for the first and the second half (cycles 1-7 and cycles 8-14) separately.

[21] The biggest change was that one of the original treatments was dropped and replaced by the pre-call announcements with maximum call charge information. There were also several minor changes such as changing the colour of the search button in the short codes treatment from red to green to signify the change of the environment and the drastically lower search costs.

The welfare (pay) is presented in relative terms by comparing the outcomes achieved by the subjects in the experiment to what an omniscient consumer (who would always call the cheapest number if it pays to make the call) would have earned in expectation.[22] Some of the entries are negative signifying losses. For example, in the 1st half of the landline baseline, the welfare is –99.1% which means that subjects made as large a loss as the omniscient consumer would have made in gains. (Never making a phone call would have ensured 0%.)

This performance measure indicates that subjects do not find the environment easy to deal with. However, there is tremendous improvement in performance over time. In the 2nd half of the landline baseline subjects are able to cut their losses by 75%. This is similar in the mobile environment where welfare increases from –94.5% to –16.9%. We will later analyse this learning process in more detail.

There is similar learning within the intervention phases where performance is invariably much better in the second halves than in the first. Again, this is true for both, landline and mobile environments.

One of the clearest results of this experiment concerns the effect of the interventions relative to the baseline. All interventions increase consumer welfare significantly. The quantitative effects are huge. Even the worst performing intervention in the 1st half of the landline environment reduces losses by 50% (–47.8% in the 1st half of Price List compared to –99.1% for the baseline).

[22] Omniscience here simply means that a consumer would have all the price information but would still not know the actual realizations of call lengths. We compute then the profit such an omniscient consumer would make on average provided he always chooses optimally, ie, call the cheapest number in the selection tasks and only call the profitable numbers in the binary tasks.

Table 2: Performance in the landline environment				
Landline	**Number of participants**	**Total Welfare**	**Welfare 1st half**	**Welfare 2nd half**
Baseline	113	–62.5%	–99.1%	–25.8%
PCA-exact	30	25.4%	6.7%	44.0%
Short codes	30	12.4%	4.1%	20.7%
Price list	25	–16.1%	–47.8%	15.5%
PCA-max	28	–13.9%	–43.1%	15.4%

Table 3: Performance in the mobile phone environment				
Mobile	**Number of participants**	**Total Welfare**	**Welfare 1st half**	**Welfare 2nd half**
Baseline	98	–55.7%	–94.5%	–16.9%
PCA-exact	26	27.4%	16.2%	38.6%
Short codes	26	13.8%	0.9%	29.5%
Price list	26	11.8%	–11.7%	35.3%
PCA-max	20	–9.4%	–33.0%	14.1%

Table 4: Performance in the landline environment (excluding search costs)				
Mobile	**Number of participants**	**Total Welfare**	**Welfare 1st half**	**Welfare 2nd half**
Baseline	113	-23.9%	-32.1%	-15.7%
PCA-exact	30	32.8%	21.4%	44.1%
Short codes	30	17.2%	13.0%	21.4%
Price list	25	-4.7%	-25.0%	15.5%
PCA-max	28	3.4%	-9.6%	16.5%

Table 5: Performance in the mobile environment (excluding search costs)				
Mobile	**Number of participants**	**Total Welfare**	**Welfare 1st half**	**Welfare 2nd half**
Baseline	98	-10.7%	-15.5%	-6.0%
PCA-exact	26	29.7%	20.7%	38.7%
Short codes	26	17.7%	5.0%	30.4%
Price list	26	23.5%	11.4%	35.5%
PCA-max	20	10.3%	4.0%	16.7%

The better interventions, PCA-exact and Short Codes, even imply positive payoffs during the 1st halves. This is true for both, landline and mobile.

In the long run, all interventions lead to substantial positive payoffs (although none achieves what omniscience would do, which is perhaps to be expected).

Comparing the different interventions, it is striking that PCA-exact comes out on top in both environments. This is true for first halves, second halves, and overall performance. The advantage of PCA-exact over the other interventions is particularly pronounced in the 1st halves when the interventions are introduced and, in the case of the landline environment,

also in the long run where subjects achieve the best results of the entire experiment. The difference to the next best intervention, the Short Codes treatment, where consumers achieve only half of the average payoff of PCA-exact in the 2nd half is statistically significant with a p-value of 0.07 (two-tailed Mann-Whitney U test). This difference stems mainly from subjects in PCA-exact making significantly cheaper phone calls than subjects in the Short Codes treatment.

The differences between PCA-exact and the two weaker interventions, PCA-max and Price List, are not only substantial but also highly significant throughout with p-value of .00 for the 1st half in the landline environment and p-values of .07 and .08 in the 2nd half (two-tailed Mann-Whitney U tests). In the mobile environment the picture is slightly murkier with statistical significance only detected in case of the weakest intervention, PCA-max.

It is interesting to examine in a little more detail through which channels the interventions work. Further below, we will see that excessive search activity in the baseline is one important ingredient for the bad overall performance of subjects and all interventions radically dispense with this.[23] However, it is important to note that the interventions also lead to better quality decisions when it comes to making telephone calls. This can easily be seen in Table 2 and Table 3, but in order to demonstrate this in a little more detail we decompose the overall performance measure further into all its different components and compare them to what the omniscient subject would have achieved. For the baseline, we find that subjects not only pay too much for the phone calls, they also make too many calls. The baseline landline premia are 16% above the premia the omniscient subject would earn, 19% in the mobile. In addition, the baseline phone bills are 64% (landline) and 67% (mobile) above the optimal bill. This generates premia to charges ratios of 70% (landline) and 71% (mobile).

All four interventions are similar in that they reduce the number of phone calls made and the overspend on the phone bills. In fact, under the interventions subjects do pretty much earn the same total premia that optimal behaviour would generate. The deviations are just between +/-9%. In addition, telephone bills are also substantially reduced but remain at above optimal levels. In the worst cases (the price list intervention in the mobile phone environment) average bills that are still up to 40% higher than the optimal bill. However, in the most successful intervention (PCA-exact in the

[23] This is largely an artefact of the experimental design. Even though search is extremely costly it is essentially easy as it just requires the click of a button. Moreover, there are (always in experiments) so-called "demand effects" that essentially describe that every button that can be clicked will be clicked. Subjects in experiments have always the desire to explore the environment fully and generally prefer action over inaction. When we extrapolate our results to the real world (that is, when we discuss our results external validity) we will consider this, of course.

landline) subjects' total telephone bills are just 13% above the optimal bill: The premia to charges ratios are around 80% (+/-5%).

The overall picture is that all interventions are desirable in the short and in the long run. They improve overall pay but they also improve performance if we take out search costs as can be seen in Table 4 and Table 5. If the only improvement came from lower search costs such improvements would a) be hard to measure in the field and b) probably be much smaller as we observe rather intense search activity in the baseline. So it is crucial to observe that the interventions – by reducing search costs - substantially improve the quality of choices. However, there is a clear winner, PCA-exact. PCA-exact never does worse than its main rival, Short Codes, and often outperforms it. Both, Price Lists and PCA-max do clearly worse than the better two interventions, in particular in the short run. The weakest intervention is PCA-max. It does particularly poor in the 2nd half of the mobile environment where it is even significantly worse than the Price Lists treatment.

4.4.2 A Closer Look at Learning

In order to understand how consumers make use of presently available information we analyse the baseline treatment for both, landline and mobile environment, in more detail. In particular, we are trying to address the following questions. (a) Do consumers actively search out pricing information? (b) Do consumers make different decisions depending on whether they know the cost of a call? (c) Is the currently available pricing information useful to consumers? (d) Do consumers learn from 'bill shock' (i.e. unexpectedly high charges observed in the bill)?

We shall answer these questions step by step.

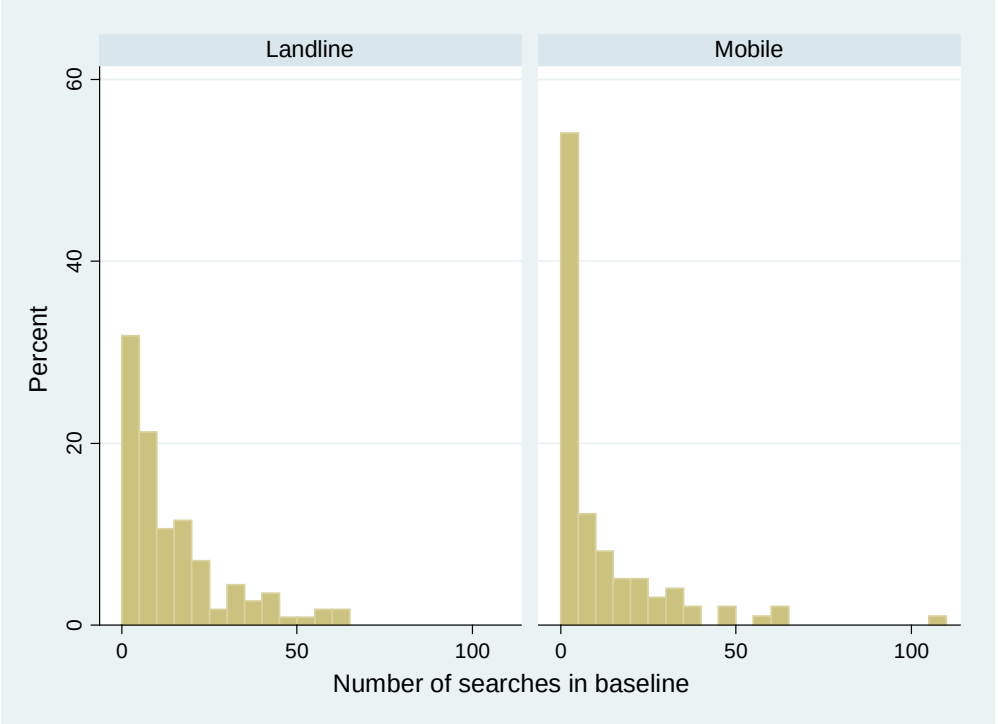
Do consumers actively search out pricing information?

To answer this question we simply plot the distribution of total number of searches over subjects for the two baseline environments. These are shown in Figure 2 below. The figure is to be read in the following way. The total number of searches is shown on the x-axis in multiples of 5. The y-axis shows how many subjects (as a percentage) have conducted that many searches. For example, in the landline treatment a little over 30% of subjects conducted less than 5 searches. The equivalent percentage for the mobile treatment is a little over 50%.

The figure reveals that there is substantial search activity. The median subject conducts 8 searches in the landline and 3 in the mobile. There is also a small minority of subjects conducting vast amounts of search. With more than 50 or even 100 searches these are subjects who, in all likelihood, did not have a particular good understanding of their environment.

Finally, the figure also reveals that there is substantially less search activity in the mobile environment (this is significant with p = .00, Mann –Whitney).

Figure 3: Distribution of total number of searches



Do consumers make different decisions depending on whether they know the cost of a call?

In order to understand how search activity affects the choice of numbers dialled we plot in Figure 3 the average call charges (per minute) that subjects incur as a function of the number of searches they conduct. We concentrate for this on the first cycle of selection tasks from the two baselines. (In later rounds search activity is confounded with memory and as such we cannot accurately isolate the impact of search and information collection on choice.)

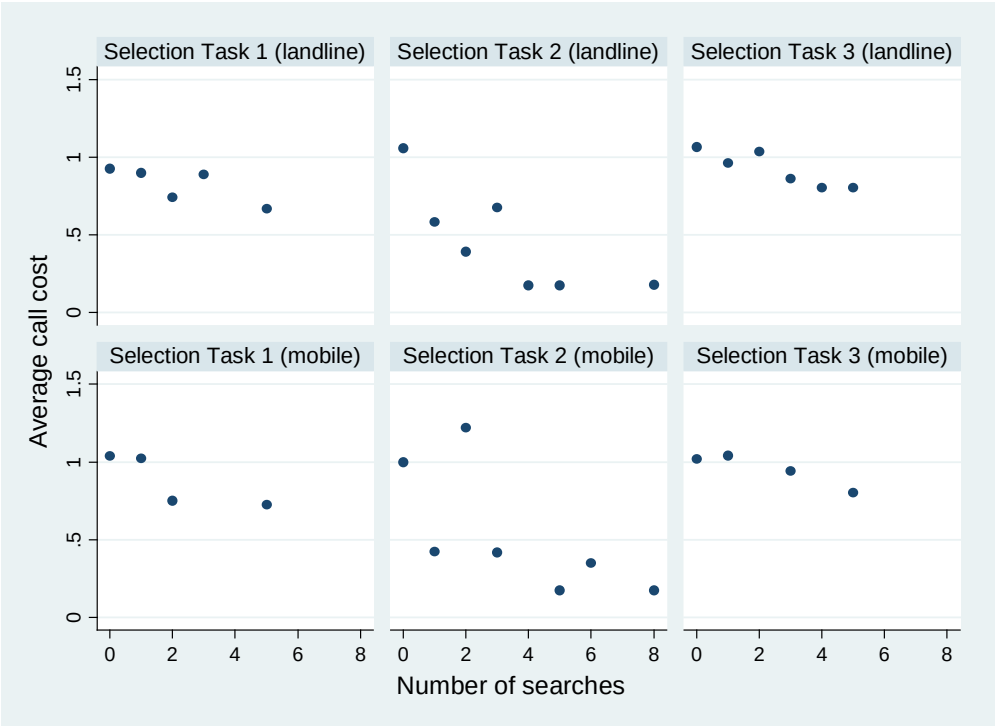
The graphs are normalized in that a level of one corresponds to the call charge that would be incurred if one were to choose at random. In other words what the subjects would do if they did not use any price information in their choices of what number to dial. Hence, observations below 1 signify

that, on average, subjects do better than random because they actively use the price information they searched.

The figure shows clearly that search matters. Incurred calling charges are generally falling in the number of searches and they do so more dramatically when there is a large variance between call prices (see the middle panel). This reduction in charges is highly significant (with a p-value of .00 obtained through an OLS estimate). However, one can also see that the marginal benefit of additional searches falls quickly. For example, there is very little difference in incurred charges for 5 or more searches.

So, the answer is a very clear yes. Subjects make use of the search information. And if they search more, they call cheaper numbers.

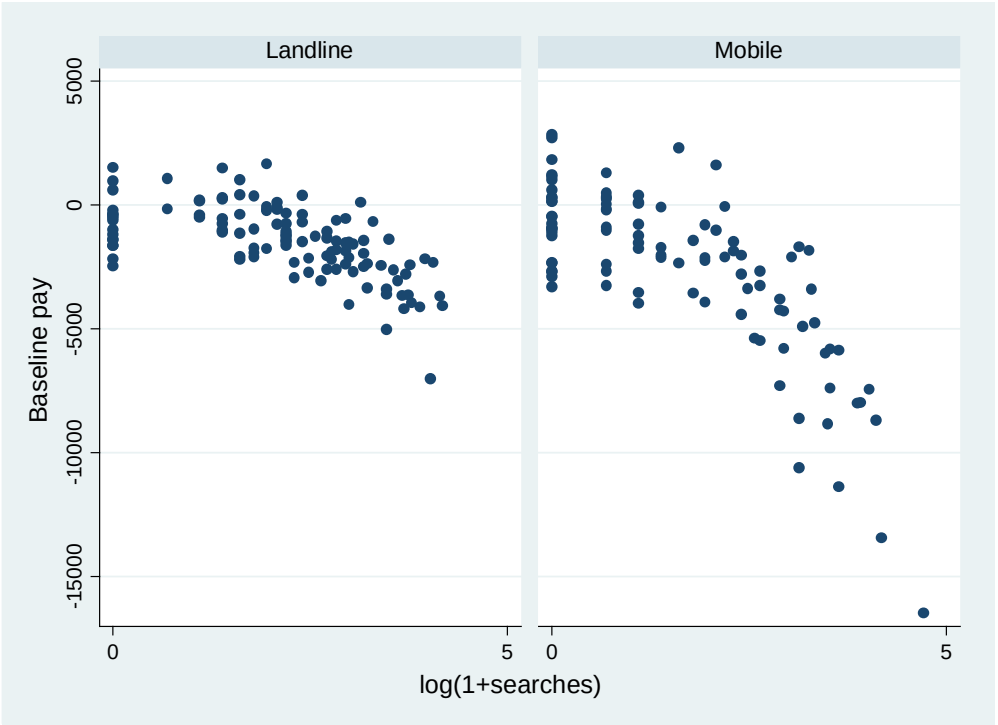
Figure 4: Does searching help consumers make better choices when making phone calls?



Is the currently available pricing information useful to consumers?

This question has two aspects. One aspect we have already tackled in the previous paragraph. Price information is useful in the sense that it helps consumers to reduce incurred call charges. The larger question is, however, given that search is costly, does it pay to search? To answer this question we show in Figure 4 for the entire baseline experiment, for both the landline and the mobile environment, a scatter plot with total number of searches (we take the log to space out the values close to zero) on the x-axis and total pay (in pence) on the y-axis.

Figure 5: Does it pay to search for price information?



There is a stark effect. Overall, subjects in the experiment simply search too much. Total pay is falling in the number of searches conducted. This is also shown in a regression analysis which confirms that the effect is highly significant (p-value .00).

For those subjects who perform small numbers of searches, the picture is different though. As the figure already reveals performance seems to be flat

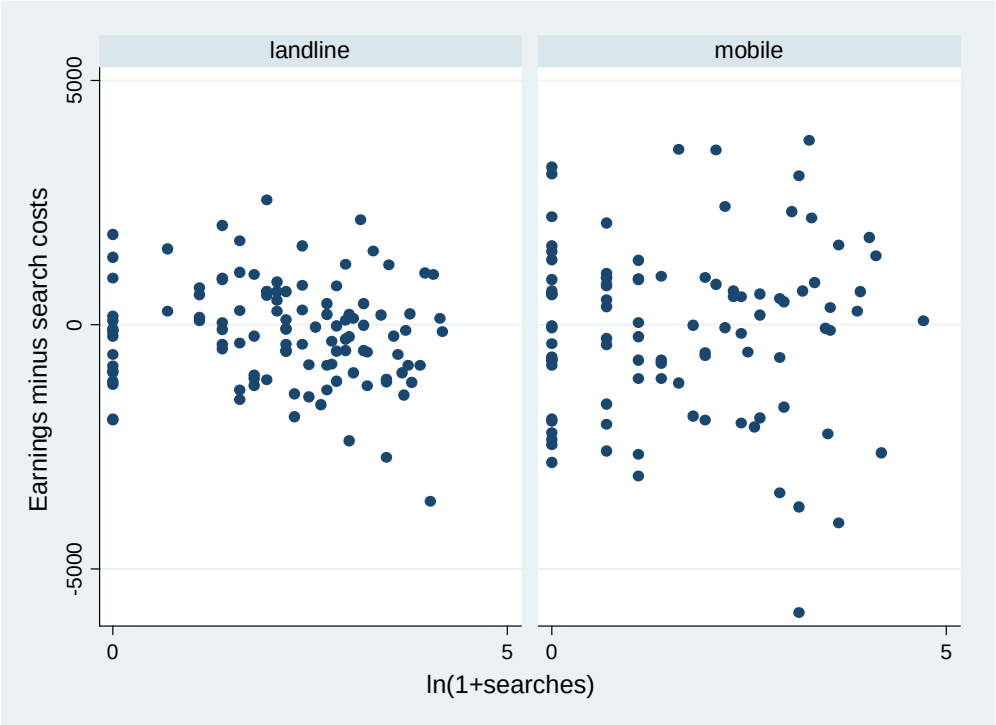
initially and then starts to fall off quite rapidly from around 10 searches onwards. This is confirmed in regressions. For example, for the landline there is a weak positive effect of number of searches for those who do not conduct more than 7 searches. As there is substantial variance in performance this is, however, not significant.

Of course, as search is very costly, one wonders whether the bad effect of excessive search is simply due to the high search costs that subjects incur. For a better judgement on this, we show in Figure 5 how search activity correlates with pay minus search costs. One could expect that subjects who search more over the entire experiment have better price information and are, hence, smarter when it comes to making calls. We have just seen that this holds for the first task cycle. But does it also hold if we look at the complete baseline?

The figure reveals it does not. Subjects who search more are not better when it comes to making the phone calls. Excessive search is simply waste.

This has an interesting implication for how subjects learn. As we have shown in the previous subsection subjects do initially learn from price search. But not in the long run. In the long run it appears that subjects rather learn from their bills, a theme we will revisit in the next subsection.

Figure 6: Do participants that search more in the long-run have better price information?



The detrimental effect of excessive search on overall performance in this experiment requires some discussion, in particular with a view to the problem of external validity.

In an experiment like this it is important to note that we are bound to overestimate search activity due to what the experimental literature calls demand effects. Demand effects refer to a private expectation of how one should behave driven by the environment you are placed in.

In this experiment because the screen has a button which is clearly visible, subjects may press this button, and undertake more search than if search information is difficult to find. This is much less the case in real life where price information is hidden in the depths of the internet and where there is consequently less search activity. In other words, the experiment overestimates the detrimental effects of excessive search. However, this should not bias any of our main results, as all pricing strategies are tested in the same environment; we discuss this further in section 4.5.

Do consumers learn from bill shock?

There are two ways of learning about the true price of a call, price search and reading the bill. The idea of bill shock is that consumers may, after being charged very high prices on some calls, decide to search more or simply to try out alternative numbers. In order to understand the role of the bill we analyse how the likelihood to call the same number again depends on the charge incurred in the previous period. We get the cleanest results for analysing this question by focussing on behaviour in the second baseline cycle as a function of the bill from the first cycle.

Specifically, we run probit regressions where we examine how the likelihood of choosing to call on a binary task in the second cycle depends on experience in the first cycle. We regress the choice to call again only on call length, which allows us to pool landline and mobile data.[24]

There is a substantial, highly significant, negative effect of call duration on the likelihood to call again. Estimating the marginal effects we find that each extra minute of call length decreases the likelihood of calling again by 1.9%. Similar findings persist throughout the baseline (that is, seeing how decisions about whether to call in cycle n+1 depend significantly on call length in cycle n).

A similar analysis on a task-by-task basis yields more interesting information: for the tasks with lower premia/call costs, the effects are lower. This documents an important psychological effect. Bill shock is not simply driven by the price of a call as measured by the charge per minute, but is a function of the total costs. If the total costs are large, this attracts subjects' attention and leads to a change in behaviour. In other words, bill shock matters and is driven by the total size of the bill.[25]

Whenever learning takes place it is important to check whether it is still ongoing towards the very end of the experiment. Our data suggest that this is not the case. Learning converges in the second half of the experiment.

[24] Details of the statistical analysis are in the annex.

[25] This raises the question whether learning is suboptimal. Of course, a fully rational subject would simply focus on the call charge per minute. However, if attention is a finite resource that needs to be allocated by subjects, then it may well be optimal to focus on large amounts as this helps to optimize learning in expensive tasks where the payoffs from learning are higher than in cheaper tasks.

4.4.3 Further Observations

The role of cognitive ability

In an environment as complex as the one studied here one might wonder to what extent performance is driven by cognitive skills. For that purpose we control for IQ in a post-experimental questionnaire. We find that IQ does matter for both, absolute performance and learning.

IQ is significantly correlated with overall performance in the mobile environment. This holds for both, the baseline and the interventions. There is no such effect for the overall performance in the landline. This difference stems from the relatively higher search costs in the mobile environment. With substantially higher search costs there is a higher premium for memorizing information once obtained.

We also find that IQ speeds up learning within a fixed environment. Improvements from first halves to second halves are generally greater for those with high IQ (significantly so in the landline baseline and the mobile interventions).

A more complex picture emerges when we examine how IQ interacts with the interventions. If we just focus on the immediate effects of the interventions we find again significant correlations between IQ and improvements in performance. However, if we examine the improvement from the 2nd half of the baseline to the 2nd half of the interventions we find that this improvement is negatively correlated with IQ. The explanation for this is that high-IQ subjects learn quickly how to make the best of the intervention while low-IQ subjects need more time for that but eventually catch up in the second half. This also explains why the overall improvements from baseline to intervention do not correlate with IQ. In the long run, everybody benefits from these interventions.

University students may be considered to have better skills than the population as whole. While we cannot compare the IQ distribution of the students used in this study to that of the population as a whole, what we can say is that across this 'higher' IQ level those that are at the lower end of the spectrum also benefit from the interventions. In order to assess if this holds across the general population an experiment with consumers drawn form different IQ levels would need to be used.

The landline-mobile split

While we have highlighted some differences between the landline and the mobile environment above the overall picture is that the differences are slight. This is good news in several ways. First of all, it indicates that data we have obtained are quite robust. Despite much higher incentives in the mobile

treatment performance is not really better. In the baseline, the steeper incentives in the mobile environment are, of course, offset by the higher search costs but this is no longer true for most of the interventions. Take for example PCA-exact where costly search becomes superfluous. Here we detect no differences at all in the relative long-run performance between the two environments. (In fact, subjects in the landline do even slightly better on average than in the mobile although this difference is not significant.) Such robustness of behaviour in response to doubling the incentives is a good indicator for external validity.

There is, however, one interesting difference between the landline and mobile environments that are revealed by a comparison of Table 2 and Table 3. Price lists do much better in the mobile environment than in the landline environment. This difference is in line with what economic theory would predict for agents who have finite memory. The more costly search is the bigger are the returns to memorize information. This is what we observe here. In the mobile environment subjects memorize the information from price lists more effectively.

Second - and this is equally important – we find that the interventions that work well for the landline environment also work well for the mobile phone environment. This suggests that there is no need to tailor interventions for the two environments which would, of course, greatly simplify any such intervention.

The role of opt-outs in the PCA interventions

As we have seen above the best performing intervention is PCA-exact. Just as with PCA-max we have designed this intervention to allow opt-out. Opting out has two consequences in the experiment. On the minus side, subjects no longer get the free price information once they start a call. On the plus side they save a little time. We find that subjects trade off these aspects very efficiently. Initially, they use the PCAs as a cheap substitute for search (and systematically terminate calls they find too expensive). But very soon, usually from the 4th cycle onwards most subjects decide that they need the PCAs no longer and opt out. Broadly speaking, subjects fall into two categories. Those that never opt-in again and those that realize that, perhaps, they have opted out too early, opt in again to gather some more information and then opt out for good.

Thus, the opt-out/opt-in option is used very effectively and subjects are able to avoid the downside of the intervention almost completely while making full use of the information provided initially. It is noteworthy in that context that we have started off all subjects as opted in. Given the efficient use of the freely provided information this appears to be a desirable feature.

Post-experimental questionnaire

The post experiment questionnaire consisted of two elements, a personality test which looked at subjects' personality traits and a feedback questionnaire where they could tell us about their approach to the experiment and provide comments.

The personality test consisted of a 15-item ("big five") questionnaire tried and tested by Schupp and Gerlitz (DIW German Institute for Economic Research, 2005). This measured five personality traits – openness, conscientiousness, extraversion, agreeableness and neuroticism.[26] We found no significant effects for any of these on total earnings or any other measure.

The feedback questionnaire indicated that the vast majority of subjects understood the instructions. Further, subjects were able to comment on the strategies that they used. There were a very large number of strategies ranging from the very simple ("guess work" or "call at random") to more thorough ("Work out which calls did not make much money or negative money and then not press them at all. Work through all the possible numbers and work out what the call rate per minute was for each number.") and from risk averse ("I tended not to call the numbers which had high charges") to risk loving ("Gamble on the larger premium tasks at the expense of the less rewarding ones").

Subjects were offered the chance to comment on any aspect of the experiment. Typical comments included:

- "I found waiting for the phone call to finish incredibly frustrating!"
- "I normally avoid numbers with special charges, so I would not call if not extremely necessary. I avoid companies offering this numbers as their main contact number."
- "Very interesting but felt like I was losing so much money just like when I have to call with pay as you go phone."

We asked the following question to find out whether there was widespread knowledge of the true cost of calling an 0870 number. "Suppose you had a query about a NatWest student credit card. You would need to dial 0870 333 9091. How much do you think this would cost from a BT landline during a weekday morning in pence per minute (assuming calls to this number are outside your bundle of free calls)?" Subjects typically overestimated the true cost, the median response being 25p (true cost approximately 6p).

[26] Gerlitz, JY, Schupp, J (2005) Zur Erhebung der Big-Five-basierten Personlichkeitsmerkmale im SOEP, DIW Research Notes 4.

Subjects were also asked whether they found the experiment frustrating or stimulating. The split was about 50:50.

4.5 Extrapolating the experimental results to the field

The extent to which one can extrapolate from laboratory data depends largely on how robust the data are within the laboratory. Clearly, if effects are very fragile in the lab in the sense that they depend a lot on precise parameter values it is everybody's guess on whether or not these effects could be replicated outside the laboratory. Luckily, the patterns observed in this experiment appear very robust. There are no changes in behaviour in response to steeper monetary incentives. The ranking of the interventions is stable. IQ matters in a consistent manner, etc.

Perhaps the most robust finding is the division between the two better and the worse performing interventions. In case of PCA-max the bad performance is, of course, in line with what standard economic theory or, in fact, common sense would predict. There is much less information that is injected in this intervention so it would be odd to believe that it can do as well as the other interventions that present much clearer information.

If anything, the experiment will even overestimate the effect of PCA-max as, in our specific case, PCA-max does provide the precise information in the binary tasks. Outside the laboratory such clear cut cases may not exist which would surely dampen the performance of PCA-max further.

Similarly, we have reason to believe that we rather overestimate the effect of the Price List intervention. Subjects can benefit from this intervention only if the carefully study the price list and manage to locate and then remember the relevant information. This is comparatively easy in our experiment. The price lists are not particularly long. It is comparatively easy to locate the relevant numbers. And because there are not that many numbers it is much easier to remember those that one needs to call. This is further aided by the repetitiveness of the 14 cycles and the fact that the 14 cycles are crammed into not much more than an hour. All these features aid memory. Consequently, we feel confident to say that the comparatively poor performance of the Price List intervention will translate in relatively poor performance in real life.

As regards the two more successful interventions let us start with a word of caution. Generally, speaking laboratory experiments tend to overestimate the quality of choice behaviour as compared to the general population. This is, as it appears, mainly due to the fact that we sample a different part of the IQ distribution. So while it is comparatively easy to predict that things that do not work well in the lab also won't work well in real life, it is more difficult to predict that things that do work well in the lab will also work well in real life.

This said there are many reasons to believe that the two interventions that perform best in our study would also work well in the field. First, we find that all subjects benefit in the long run from the interventions, including those at the lower end of the IQ distribution. Second, there are many reasons to believe that subjects in our baseline know much more about the cost of telephone calls than real-life consumers do. This is simply because a) our environment really encourages good performance as subjects are always eager to make as much money as they can and b) our environment is, in the end, much less complicated than real life with just 9 different tasks that are repeated 14 times. Consequently, we probably *underestimate* the effects of the two interventions that make precise price information readily available when it is needed (that is not when the bill arrives but when the consumer wants to make a phone call).

Moreover, when it comes to considering the interventions' effects on different parts of the general population it should be obvious that those interventions that require little cognitive effort will do comparatively better. In the laboratory experiment the use of a student sample means we oversample high IQs and under sample low IQs. Thus, we would expect that interventions that require higher cognitive abilities (clever search, more memorizing) would do relatively worse in the general population and those that do not require particular cognitive skills would do relatively better. Again, this speaks in favour of the two better interventions, in particular, in favour of PCA-exact where the relevant information is served on a silver platter and hard to ignore even if one is cognitively more challenged.

On balance we think we can say with confidence that the dichotomy with better interventions (PCA-exact and Short Codes) on the one hand and worse interventions (PCA-max and Price List) on the other would have external validity. If anything, the experiment is biased to underestimate the difference between these pairs of interventions. Given the potential bias that stems from oversampling smart subjects we would probably expect PCA-exact to come out as the clear winner in the field.

In the experiment we can also quantify these differences and we can quantify the improvement from the baseline to the intervention. While it would, of course, be nice if we could somehow translate these quantitative measures into predictions of the real-life welfare gains expected under the different interventions, we fear that this is simply not possible. This has to be seen as a clear limitation of the experimental lab approach. Any such quantification would require an increase in the number of participants such that statistical inferences about the general population could be drawn. However, as previously stated, increasing the sample size will still have limitations. We would really need a field experiment to get to the true welfare relevant costs and benefits.

It is worth noting that the experiment in many ways tests 'perfect' versions of each policy intervention: they work as intended, without any technical hitches; consumers are made well aware of each intervention and they are also easy to use, for example, pre call announcements can be quickly and easily turned off at any point. Such factors, while important for policymaking, are clearly outside the scope of experiments. Nevertheless, the experiment yields valuable insights which can be factored into the analysis.

The experiment also assumes that firms will not change their behaviour in an attempt to undermine any policy intervention. As discussed below, this could in principal be incorporated, with some subjects taking the role of firms and making various decisions about the level and presentation of prices. However, this would significantly increase the complexity and cost of the experiment and the gains would be uncertain.

4.6 Other designs

This experiment could have looked much different and we need to ask the question to what extent the things we have learned depend on the specificities of our design. We might also want to consider what else we might have learned had we taken a different route.

Here we focus on some of the design options that we did discuss in the process but that we ultimately rejected.

Handsets and real phone numbers. In order to make the experiment look and feel more real we could have given subjects actual handsets and we could have given them real (longer) phone numbers. We doubt that any of this would have mattered. Subjects in this sort of experiment are typically very good at figuring out what is at the core of the task. A handset would neither have guided this process nor would it have detracted from it. We believe it would not have made any difference. Real (longer) phone numbers probably would have made a difference as longer numbers are harder to remember and hence the associated price information is harder to remember. That would have made the baseline even more difficult and would have further accentuated the benefits from short codes and PCA-exact.

Real search rather than monetary costs. As standard practice in experimental economics we opted to give subjects a simple easy-to-click search button and mirror the complicatedness of search through high monetary search costs. As we have discussed above this has led to rather excessive search. Could this have been avoided if subjects had to perform an actual search, for example, through browsing through some lists available to them through a series of clicks? This is hard to say as, in general, demand effects would also be expected to play a role here. Subjects would, surely, again have explored all the features the experimenter has embedded in the design. One undesirable

consequence of this would have been a potentially big increase in the duration of a task which, given that the experiment was already quite long, would have forced us either to reduce the number of tasks or the number of cycles. As we have seen there was considerable improvement from the first to the second half (with learning converging towards the end) so we might have stopped too early before learning had converged. This would have been rather undesirable. On the other hand, real search would have been perhaps more sensitive to different cognitive skills which would might have further accentuated the role of IQ that we observe in our performance measures.

Quality differences. In our selection tasks all telephone numbers are equally good in the sense that they all completely solve the tasks and all have call durations chosen from identical distributions. Hence, all learning can completely focus on price. This greatly simplifies the environment. If call durations would have been sampled from different distributions subjects would have been faced with a much more complicated task. This would have been particularly interesting if in some cases numbers that are expensive by the minute actually outperform cheaper numbers simply because call durations are on average shorter. This could have led to some suboptimal behaviour in response to easily available price information. While this would still be an interesting question to explore it is generally the right thing to start with simpler designs and move from there to more elaborate environments.

The inclusion of firms: In experiments it is possible to include the interaction between consumer decisions and firm decisions. Therefore it would be possible to include firms in the experiment implemented for Ofcom. Firms could be included either as simulated agents that operate to maximise their own production (or objective) function. Or alternatively, human participants could take on the role of firms making decisions in the experiment to maximise their earnings. Depending on the specific design of the experiment, one could observe how firms pricing practices change (or not) when faced with different information set-ups for call price, and in response to the consumers' behaviour.

5 Criticisms of experiments in economics

The two main (related) criticisms of experiments are the following:

- Representativeness of the subject pool or participants in the experiment
- Extrapolation beyond the laboratory

We can consider these in turn.

Subject pool representativeness: Traditionally economists have used university students in economic experiments. The reason for this is that experiments have traditionally been undertaken by research institutes within universities, and as such, access to students was easy. Perhaps more importantly, however, the use of students minimises the unobservable and uncontrollable private perceptions, experience and information that may influence behaviour and therefore the observations from the experiment. However, when experiments are used for policy development, policy-designers and relevant stakeholders can raise concerns that the observations from student subject pools cannot be translated to the field because students do not have experience in the field. The counter to this argument is that economic principles of behaviour that underpin policy are general principles, and therefore, if the principles are robust behaviour should not change across different participant groups.

The issue of subject pool representativeness is under debate, and the published literature, both from psychology and economics, provides differing outcomes both inter and intra discipline. We briefly review this cross-discipline research below.

Cognitive psychologists have been concerned with the effect of context and experience on individual performance for many decades. For example, Johnson-Laird, 1983, argued that if subjects have a 'mental mode' – experience, knowledge or training – in a related model or set of rules, then they are more likely to have insight into a new task. Similarly, Salomon and Perkins, 1989, explore the impact on behaviour in new contexts and situations of previous experience in similar contexts; Salomon and Perkins find that earlier related experience does impact positively upon performance in related new tasks, and this is termed 'transfer' in the psychology literature. In the experimental economics literature, there is a growing body of research that investigates the effect of context (using fictitious goods or the real terms of the goods e.g. phone calls, mortgage contracts, water) and experience on performance within different experimental settings. Cooper and Kagel, 2003,

find that the use of context in experiments with student subjects can increase the speed of subjects' learning such that context can be a substitute for experience. Harrison and List, 2003, investigate the impact of experience in naturally occurring markets (the sports card market) on the performance (falling subject to the "winners curse", that is, the common tendency to overbid) in a one shot auction. They observe that field subjects with experience as dealers in the naturally occurring market make lower bids than subjects without experience (non dealers) in both artefactual experimental treatments and framed field experimental treatments: They conclude that 'transfer' does occur. Other researchers have explored the effect of demographics. An interesting example is provided by Carter and Irons, 1991. Carter and Irons find that economics students in ultimatum games tend to offer less (as the proposer) and accept less (as the responder) as compared to students from other faculties. [27] Carpenter, Burks and Verhoogen, 2005, also observe demographic differences across different student populations in the ultimatum game.

What these findings imply is that that the fundamental underlying behaviour, if it is robust, given the incentives tested in the experiment , will hold across all types of subject pool. However, there may be some difference in the magnitude of outcomes if the subjects have had prior experience, or if the cognitive ability of subjects is different.

Comparing across student and field subject pools; Carpenter et al., 2005, find that employees of a warehouse are more likely to settle upon an equal distribution of benefits in the ultimatum game than student subjects. Potters and Winden, 2000, find, in their artefactual experimental treatments, that professional lobbyists perform closer to the signalling game theoretic prediction as compared to student subjects. Fehr and List, 2004, also observe that field participants, in this case CEOs, perform closer to the efficient equilibrium in trust games as compared to student subjects. Cummings et al. 2004, use both students and farmers in conventional lab and framed field experiments to explore alternative auction designs for agricultural irrigation reductions. They observed 'behavioural regularities' across subject pools and experiment type. These observations helped to instil confidence in the state regulators that auctions can be used effectively in the field. Ward et al., 2006, compare the performance of student and irrigator (farmer) subject pools in a framed field experiment designed to investigate co-operative behaviour in the management of a common pool resource (water in a river). Ward et al. compare across subject pools the impact of providing different information sets about over-extraction of water (extraction above the social optimum

[27] The ultimatum game is a game in which two players interact to decide how to divide a sum of money that is given to them. The first player proposes how to divide the sum between the two players, and the second player can either accept or reject this proposal. If the second player rejects, neither player receives anything. If the second player accepts, the money is split according to the proposal. The game is played only once so that reciprocation is not an issue.

target). While they find behaviour across the two subject pools is generally the same, they do observe (some) difference in subjects' response to public disclosure of information about individual behaviour. Namely, student subjects did not adjust behaviour toward the social optimum when information about individual behaviour was revealed. On the other hand, irrigator subjects did change behaviour when this individual information was publicly revealed. Ward et al. suggest that social connections, group norms, reputation and reciprocity may be influencing behaviour in the irrigator sessions. [28] In a similar applied application, Duke, 2007, conducted experiments using students and farmers to test the design of water markets and water pollution policies. Duke observed some difference between the behaviour of students and farmers, but the difference was statistically weak and was only present when negative incentives were introduced (such as penalties or taxes) not when positive incentives were used (rewards or subsidies). This difference could therefore be explained by the psychological and economic concept of "loss aversion" as opposed to actual differences in behaviour across subjects.[29]Another example is Johnson et al. (2003) in which observations from the real world, in regard to opt-in and opt-out framing for insurance contracts, was the same as the behaviours in the laboratory.

Overall the topic of subject pool choice remains under debate and continued research in the experimental economics field. However one guiding principle should be kept in mind: If a policy does not work in simple setting using students who bring no prior perceptions about behaviour to the experiment, then it is unlikely to work in a more complex field setting with multiple stakeholders all holding prior perceptions about how the policy should operate.

The use of students is good at identifying what will not work in the field, i.e. if students fall prey to behavioural biases then it is very likely that the general population will also fall prey. However, if students perform well, then it is not as easy to say that the general population will also perform as well.

Extrapolation beyond the laboratory: The main debate about the use of economic experiments for policy testing and design is external validity.

[28] Observations by Casari and Plott, 2000, may (also) be helpful in explaining this behaviour. Casari and Plott, in their conventional experiments, observed that spite and altruistic behaviour can explain (improved) contributions to a common pool resource. The Casari and Plott work, builds upon Walker, Gardner and Ostrom, 1990, who find that cooperation in public good games is below (even) the expected self interested outcomes. These papers are very interesting, as they help to explain why historical self governing communities (such as indigenous communities) can often effectively manage common pool resources.

[29] Loss aversion is a well known concept in psychology and economics, Kahneman et al., 1990. In the experiments conducted by Duke, as more realism was introduced including the use of real water consumers in the experiments, the loss aversion from the negative incentive became stronger while the reward motivation from the positive incentive remained the same across the experiments.

External validity relates to the argument as to whether the observations from experiments can be transferred to the real world.

Some researchers criticise conventional laboratory experiments because they are simplified and all other influences on behaviour are removed from the experiment such that only the specific feature of interest is tested, holding all else constant. This is very similar to economic theory where in order to make a problem tractable, *ceteris paribus* (all else the same) is invoked and the feature(s) of interest is solved in the theoretical framework holding all other factors constant. Like economic theory, conventional experiments build upon one another.

Economic experiments are also similar to physical science experiments, because economic experiments change one variable at time and build a sequence of experiments to understand how behaviours (be it chemical, biological, psychological or economic) are expected to occur in the field. Similarly, economic experiments are like engineering tests that isolate, for example, one part of a bridge, skyscraper or aeroplane and test this individual part before adding more layers of complexity and interactions between individual components.[30]

In this light, the traditional response to the criticism of extrapolation is embodied by the concept of parallelism where the experimenter tackles the criticism through conducting new experiments that investigate the precise nature of the criticism. For example, if the critique is that certain experiments do not have external validity because in real markets there are many more participants, the researcher can carry out new experiments doubling the number of market participants to find out whether this is a real issue or not.

Artefactual, framed and natural field experiments trade-off some control and introduce increasing real world features in order to increase external validity. Often conventional laboratory experiments are followed by one or more of these (more real world) experiment types - particularly in order to improve policy-making and as a useful tool for engaging stakeholders. An example of using a sequence of experiments to increase external validity is presented below.

As part of the Australian Governments' initiative to investigate the feasibility of using flexible incentives to manage natural resource management outcomes, a series of economic experiments were conducted to pre-test the design and performance of the proposed incentives. The experiments began with conventional laboratory experiments using university students and stylised goods called 'one', 'two' and 'three'. The university students were

[30] This is a strength of experiments, it forces the policy-maker to isolate each incentive that may influence decision-making, and to test these incentives in a robust way such that causality can be isolated: exactly what is impacting upon behaviour and when.

playing the role of farmers, but were not told this, and the goods one, two and three represented environmental outcomes such as reductions in pollution entering a stream, but again the students were not told this. However, actual field parameters were used. The motivation for using a conventional experiment with no framing was to quickly, and relatively cheaply, test if the proposed incentives would actually work removing all complexity from the field and ensuring no moral or emotional beliefs (surrounding environmental protection) impacted upon outcomes. If the incentives failed in the conventional laboratory experiment, then there was little reason to expect them to operate in the field and as such the incentives could be abandoned (Duke, Cason and Gangadharan 2008).

Following the successful performance of the incentives in the conventional experiment for reasons supported by the general principles of economic theory (i.e. the incentives worked for the same reasons as that postulated by theory), the experiment was then conducted using farming, namely using the terms "water", "pollution" and "government policy intervention to manage the water pollution". The university students were also told that they were playing the role of farmers, and were given maps that showed them where their farm was (these were real maps of a well known region called the Murray River). Again the incentives performed well for the reasons expected by theory (Duke and Gangadharan, 2008).

The final step was then to conduct experiments using farmers and representatives from water companies. When the market players were used, their performance was the same as the students when they were faced with a positive incentive for good behaviour, but when the market players were faced with a negative incentive – a penalty for bad behaviour – the market players were less willing to pay the tax, and put more effort into avoiding the negative incentive as compared to the students (Duke, 2007). The difference between the students and the market players could be driven by the fact that the market players wanted to signal to government, that taxes on pollution from their production activities would not be a popular outcome within the region.

A skilled experimental economist will be able to suggest what type of experiment may be most useful for the remedy question posed, and what existing experimental evidence there is in the refereed literature that can inform the current question. For example, if we consider internet shopping and want to test different remedies to encourage self-regulation by sellers, then we would build upon existing conventional experiments including the work by Bohnet *et al*. (2005), who found that different information revelation procedures in the market have different impacts on the trustworthiness of sellers in internet markets.

Two additional considerations are learning in the laboratory versus learning in the field, and self selection bias of participants.

Learning in the laboratory: Learning in real life depends much on the frequency with which problems are encountered. Some types of decisions are made with high frequency (grocery shopping), others with extremely low frequency (choosing a pension plan). In the laboratory one can provide as much opportunity for learning as one wishes. In some cases, one might be interested in mirroring the high-frequency scenario and there are laboratory studies where subjects have to make the same type of choice several hundreds times. In other cases, one might be interested in how people decide if they encounter a new problem and encounter it only once, again this can be implemented in the laboratory.

Of course, learning in the laboratory is always compressed in time through the limits imposed by having subjects typically for not much more than a couple of hours. However, if one is interested in very experienced behaviour one can bring subjects back a few days or weeks later. See for example, Kagel and Richard (2001).

Self-selection of participants: The issue of self-selection bias is the same in economic experiments as it is in other quantitative methods such as surveys. For example, if one uses people who have signed up to be on a research panel (as many market research firms do), then these people have pre-selected themselves as willing participants. If the internet is used then the participants will need to have access and knowledge of using the web. If an experiment is conducted in a laboratory at a university then people will need to be able to travel to the location. How subjects are selected, and how and where the experiment is implemented, depends on budget and time constraints.

If there is reason to believe that there is some feature of the target population that is special then the sample should be selected to include people with these features. The same principles for sample selection apply to experiments as they do to other testing methods using humans such as quantitative and qualitative surveys and focus groups.

6 Conclusions and recommendations

The experimental laboratory is an artificial construct. The processes within the laboratory are, however, very real. The laboratory uses people who participate for real profits and follow real rules to make these profits (Plott, 1982). Laboratory processes are simple compared to complicated real world processes. They must be simple in order for the policy designer to isolate and control the economic incentives that influence behaviour. Laboratory processes that are too complicated will fail to identify what agents are responding to (lose control over incentives).

In order for experiments to provide meaningful insights for policy design, the experimental design must be close enough to the real world. As mentioned previously, this refers to the concept of parallelism, 'the extent to which the environment and institutions in the experimental design characterise in a meaningful way the complicated and changing real world that is relevant for those aspects of the agent behaviour under study' (Cummings, McKee and Taylor, 2001). This means when designing an experiment for policy the experimenter must de-construct the real world: The environment and incentives must be simplified and those that substantially influence the behaviour of interest are identified and induced in the laboratory. The outcomes must be interpreted carefully and the implications of the outcomes must be confined within the experimental design. This means, the experiment can only explain behavioural processes, and the resulting outcomes that were controlled for and induced in the laboratory. This no easy process and requires both formal training and on the job (or in the lab) experience.

It becomes clear then, that not everything that influences decisions in the real world can be included in the laboratory. How then can we be confident that an experimental laboratory process explains enough of the real world that we can rely on the outcomes? Through careful design and interpretation, and by replication. Repeating the experiment by the same researcher across different environments, as has been done with this experiment for Ofcom where the information set-ups have been tested across a landline and mobile phone environment. Repeating across different researchers, different subject pools, and by gradually changing the experimental design to more closely represent the complexity of the real world and observing if behaviour persists as complexity is increased in a systematic and a controlled way.

Through these methods – just as in the physical sciences - experimentalists can minimise the chance that there is some special aspect of the experiment which they are inducing but not controlling.

6.1 Designing experiments for public policy

If policy-makers use economic experiments for policy design, what are some of the key questions that may arise?

First, one may consider how abstract (or not) the environment should be. The key here is to capture the features of the real world field that are important for the behaviour of interest, but to remove features that are not of direct importance. This is an important judgement because the greater the complexity of the experiment the more difficulty the policy-maker will have in identifying what is actually driving behaviour (causality). Likewise, if the important features are removed from the environment, then the explanatory power of the experiment will be reduced.

A second, and related question, is what types of subjects should be used in the experiment. As mentioned previously, student subjects have been historically used but field participants, drawn from outside the student population, are also used. Likewise, experience of the subjects (either student or field) in the same or a similar environment may be of importance.

When making these decisions the following considerations should be taken into account.

- Financial Budget: As a rule of thumb lab experiments with students are much cheaper. However, it is always worthwhile to think about possibilities for field experiments. Sometimes one can come up with very creative, elegant and cheap field experiments.

- Time Budget: Lab experiments are almost invariably faster to set up and conduct.

- Specificity: Is the problem of a general nature or is there reason to believe that all elements of the problem in its natural occurrence in the field, matter for behaviour? If the former, a lab experiment will be appropriate, if the latter too much might be lost in the laboratory and a field experiment may be considered.

- Precise quantitative measurements: The less "real" an experimental environment (that is the more different it is from its real-world counterpart that one is interested in) the more difficult it is to translate measurements taken in the experiment into precise quantitative predictions for the field. The lab is always good in detecting relative differences between treatments or interventions but it is typically hard to extrapolate absolute measurements. For example, a laboratory experiment might be able to say under which conditions consumers will get the best prices in markets for homogenous goods, but it would obviously be difficult to say what the absolute price level

would be in the real world --- trivially, this would depend on the precise nature of the good.

Therefore, if the objective is to test the fundamental underlying principles upon which expected behaviour is based then a controlled laboratory experiment may be preferred. For example, a controlled laboratory experiment would be a robust and cost effective way of testing if consumer decision-making is affected by endowment or default decisions. Controlled laboratory experiments are good at testing if an effect is present, and comparing the relative magnitude of effects.

Introducing more complexity from the "real world" field may be desirable if one believes there are specific features of the real world that are fundamental to expected behaviour. For example, if we wanted to test if different types of consumers behave differently when faced with exactly the same incentives, then one may use field participants drawn from across different parts of the population. In this instance the effects, and relative magnitudes of effects, can be assessed across the different types of participants. There will, however, be unobservable (private) characteristics which the experimenter does not know and cannot control for that will have impact on behaviour. In this sense causality is reduced.

Similarly, there may be features of the goods and services that are deemed important. For example, in situations of philanthropic giving, the type of cause and how the money is expected to be used may be important to behaviour and in this instance more realism may be introduced. Likewise, the use of framing (context) may be important for the presentation of the results. If one believes it will be difficult to present the observations if too much abstraction is used, then providing some context may be sensible.

If the objective is to assess the aggregate impacts in the real world field, then more realism and large sample sizes are required. For example, in this experiment implemented for Ofcom, if we wanted to assess the magnitude of the effects of different information set-ups across the general population then we could have increased the sample size to an extent such that statistical inferences can be drawn (as is necessary with quantitative surveys, for example).

Therefore, there is no rule for which type of experiment may be best used. The goal rather is to select a design which provides the best opportunity to learn something useful and to answer the questions that motivate the investigation or research.

6.2 When not to use experiments

Controlled laboratory experiments with student subject pools are typically not vey good at measuring the magnitude of impacts in the field. The

strengths of controlled laboratory experiments are in isolating cause and effect between incentives and behaviours, and in measuring the relative impact of incentives.

In order to get an estimate for the real-world magnitude of effects observed in a laboratory one would have to calibrate the experiment very carefully. If the true parameters that characterize the relevant real-world decision environment are known, the lab environment can be scaled in a way that would allow some inference about the magnitude of real-world effects. But it may be preferable to actually test these inferences in a (limited) field experiment where one has direct access to the true costs and benefits that market participants experience.

If the objective is to collect field data on the prevalence of particular behaviours, then economic experiments would not be used because economic experiments do not collect actual field data. In this instance an alternative method such as quantitative surveys may be used.

If the objective is to collect beliefs and perceptions, then again economic experiments may not be best. Economic experiments test and observe behaviour driven by economic reward and penalties (monetary pay-offs), and therefore qualitative beliefs and perceptions are not the main objective of the experiments. However, economic experiments are often combined with qualitative surveys to gain further insight into participants' beliefs and perceptions.

Experiments present the policy-maker with a new method that allows the observation of actual human behaviour in a controlled setting such that cause and effect can be isolated, and relative impacts observed. It allows policy-makers to test the underlying behavioural model to see if in fact consumers and firms behave as the framework predicts. The method allows rapid, and relatively cheap, comparisons of interventions such that unexpected outcomes can be identified early on in the process and undesirable outcomes mitigated. Experiments open the box on the economic agent, and test if the complicated human does operate as economics predicts.

Annex 1 References

Abeler, J. And Marklein, F., (2008) "Fungibility, Labels, and Consumption", IZA Discussion Paper No. 3500.

Atlas.V., Cameron.L., Chaudhuri.A., Erkal.N. and Gangadharan.L., (2006) Gender and Corruption: Insights from an experimental analysis, University of Melbourne Research Paper No. 974, October.

"*Experiments and Competition Policy*" (2009) eds., Jeroen Hinloopen and Hans-Theo Norman, Cambridge University Press.

Alatas.V., Cameron.L., Chaudhuri.A., Erkal.N., and Gangadharan.L. 2006. "Subject pool effects in a corruption experiment: a comparison of Indonesian public servants and Indonesian students", University of Melbourne Working Paper.

Baye.M., Gatti.R., Morgan.J., Kattuman.P., (2006) Did the Euro Foster Online Price Competition? Evidence from an International Price Comparison Site, Economic Inquiry, vol. 44(2)

Brown.J.,, Hossain.T., and Morgan.J. 2007. Shrouded Attributes and Information Suppression: Evidence from the Field. Working Paper. November.

Bohnet.I., Harmgart.H., Huck.S. and JR Tyran (2005) Learning Trust, Journal of the European Economic Association, vol. 3(2-3), pp. 322 – 329

Carpenter.J.P., Burks.S. and Verhoogen.E. 2005. "Comparing students to workers: The effects of social framing on behaviour in distribution games", In Field Experiments in Economics, Research in Experimental Economics vol. 10, (eds.) J.P. Carpenter, G.W. Harrison and J.A.List, Elsevier, Oxford, UK.

Carpenter.J.P., Harrison.G.W, List.J,A. 2005. "Field experiments in economics: an introduction", In Field Experiments in Economics, (eds.) Carpenter, Harrison and List, Research in Experimental Economics vol. 10, Elseiver, Oxford UK.

Carpenter.J.P., Harrison.G.W. and List.J.A. (2005) Field experiments in Economics and Introduction, in *Field Experiments in Economics*, eds. Carpenter, Harrison, List, Elsevier.

Carter..J., and Irons.M. 1991. "Are economists different, and if so why?", *Journal of Economic Perspectives*, 5(2): 171 – 177

Casari.M and Plott.C.R. 2000. "Keeping and eye on your neighbours: agents monitoring and sanctioning one another in a common pool resource

environment", California Institute of Technology, Social Science Working Paper 1072, October 2000.

Cason, T.N, L. Gangadharan and Duke.C. 2003. a. "A Laboratory Study of Auctions for Reducing Non-point Source Pollution", *Journal of Environmental Economics and Management*, 46: 446 - 71.

Cason, T.N., L. Gangadharan and Duke.C. 2003. b. "Market Power in Tradable Permit Markets: A laboratory test bed for Emissions Trading in Port Phillip Bay, Victoria", *Ecological Economics*, 46: 469 – 91.

Cason.T. and Friedman.D., (2003) Buyer search and price dispersion: a laboratory study, *Journal of Economic Theory*

Cason.T., Friedman.J., and Milam.G. (2003) Bargaining versus Posted Prices in Customer Markets, *International Journal of Industrial Organisation*, 21(2), 223 – 251.

Chamberlin. E.H. (1948) An experimental imperfect market, Journal of Political Economy, 56(April), 95-108

Cooper.D.J., an Kagel.J.H. 2003. "The impact of meaningful context on strategic play in signalling games", *Journal of Economic Behaviour and Organisation*, 50:311 – 337.

Competition Commission and Office of Fair Trading, 2009, *Road Testing Consumer Remedies,* a report by London Economics

Cummings.R.G., Holt.C.A., and Laury.S.K. 2004. "Using laboratory experiments for policy making: an example from the Georgia Irrigation Reduction Auction", *Journal of Policy Analysis and Management*, 23(2):341 – 352.

Davis.D. and Holt.C., (1996) Consumer search costs and market performance, Economic Inquiry, vol.34 (1)

Dufwenberg.M., Gneezy.U., Goeree.J.K., and Nagel .R. 2007. Price floors and competition, Economic Theory, 33, 211-224.

Duke.C.(2008), Using prices to manage environmental externalities evidence from a field experiment, PhD Dissertation, University of Melbourne Australia.

Fehr.E., and List.J.A. 2004. "The hidden costs and returns of incentives – trust and trustworthiness among CEOs", *Journal of the European Economic Association*, 2(5):743 – 771.

Güth. W., Schmittberger.R., and Schwarze.B. (1982) An experimental analysis of ultimatum bargaining, Journal of Economic Behaviour and Organisation 3(4), 367-388.

Fehr.E. and Schmidt.K. (1999) A theory of fairness, competition and cooperation, The Quarterly Journal of Economics, (August), 817-868.

Forsythe.R., Palfrey.T. and Plott.C.R. 1982. "Asset valuation in an experimental market", *Econometrica,* 50(3): 537 – 567.

Harrison.G.W. and List.J.A. 2003. "Naturally occurring markets and exogenous laboratory experiments: A case of the winner's curse", Working Paper 03-14, Department of Economics, University of Central Florida.

Harrison.G.W. and List.J.A. 2004. "Field Experiments", *Journal of Economic Literature*, 43(4):1013 – 1059.

Hoffman.E., McCabe.K., Shachat.K. and Smith.V., 1994, "Preferences, property rights and anonymity in bargaining games", *Games and Economic Behaviour*, (7):346 – 380.

Huck.S., Lunser.G., and Tyran.J.R., 2008. Prising and trust, ELSE working paper (http://eprints.ucl.ac.uk/14453/).

Huck, S. And Rasul, I. 2007. Comparing charitable fundraising schemes: Evidence from a natural field experiment, ELSE Discussion Paper.

Johnson-Laird.P.N. 1983. Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness, Cambridge MA, Harvard University Press.

Kagel.J., and Richard.J-F., 2001. Super-experienced bidders in first-price common-value auctions: Rules of thumb, Nash equilibrium bidding and the winners curse, The Review of Economics and Statistics, 83(3): 408-419.

Kahneman.D. Knetsch.J.L and Thaler. .R.H. 1990. "Experimental tests of the Endowment Effect and the Coase Theorem", *Journal of Political Economy*, 98(6):1325-48.

Karlan, D. and Zinman, J. Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment. Econometrica, *forthcoming*

Lacko.J., and Pappalardo.J. (2004) The effect of mortgage broker compensation disclosures in consumers and competition: A controlled experiment, Federal Trade Commission.

Loewenstein.G. 1999 "Experimental Economics from the vantage point of Behavioural Economics", *The Economic Journal*, 109(Feb.):F25 – F34.

Lynch.J. and Ariely.D. (2000) Wine online: Search cost affect, competition on price, quality and distribution, *Marketing Science*, vol.19(1)

McMillan.J., (1994) Selling Spectrum Rights, *The Journal of Economic Perspectives*, vol. 8(3)

Murphy.J.J. and Stranlund.J.K. 2006. "Direct and Market Effects of Enforcing Emissions Trading Programs: An Experimental Analysis", *Journal of Economic Behaviour and Organisation*, 61(2): 217 – 233.

Murphy.J.J. Stranlund.J.K. 2007. "A laboratory investigation of compliance behaviour under tradable emissions rights: Implications for targeted enforcement", *Journal of Environmental Economics and Management*, 53(2): 196 – 212.

National Action Plan for Salinity and Water Quality, 2005, "Cap and Trade for Salinity: Property Rights and Private Abatement, a Laboratory Experiment Market", Final Report, December, http://www.napswq.gov.au/publications/books/mbi/pubs/round1-project10.pdf

Noussair.C., Plott.C. and Riezman.R. 2003. "Production, Trade and Exchange Rates in Large Experimental Economics", *mimeo,* California Institute of Technology, Pasadena, California.

Plott.C. (1982) Industrial Organisation Theory and Experimental Economics, *Journal of Economic Literature*, vol. 20, pp. 1485 – 1527

Potters.J. and van Winden.F. 2000. "Professionals and students in a lobbying experiment: professional rules of conduct and subject surrogacy", *Journal of Economic behaviour and Organisation*, 43: 499 – 522.

RG Cummings, M McKee, LO Taylor (2001) "To whisper in the ears of princes: laboratory economic experiments and environmental policy", In Frontiers of Environmental Economics. Edward Elgar Publishing

Roth. A., (2002), The economist as engineer: Game theory, experimentation and design economics, *Econometrica*, vol. 70(4)

Salomon.G. and Perkins.D.N. 1989. "Rocky Roads to Transfer: Rethinking Mechanisms of a Neglected Phenomenon", *Educational Psychologist*, 24(2): 113 – 142.

Smith.V. (1962) An experimental study of competitive market behaviour, *Journal of Political Economy*, 70(2) , 111-137.

Vickrey.W., (1961), Counter speculation, auctions and competitive sealed tenders, *The Journal of Finance*, vol. 16.

Walker.J.R. Gardner.R. Ostrom.E. 1990. "Rent dissipation in a limited access common pool resource: experimental evidence", *Journal of Environmental Economics and Management*, 19: 203 – 211.

Annex 2 Experiment instructions

In this section, we have firstly the Experimental Instructions for the baseline (entitled General Experimental Instructions & Phase II) and then experimental instructions for each of the four interventions (entitled Phase II). The four interventions are coded:

- PCAE: precall announcements with exact information
- PCAM: precall announcements with maximum price information
- SC: short codes
- INFO: price list

Instructions are only given for the landline treatment. The instructions for the mobile treatments are exactly the same with the exception that premia and call costs are doubled (where stated) and search costs are trebled.

In the experiment, all subjects would receive the baseline instructions, do the baseline and then receive one of the four interventions instructions.

General Experimental Instructions

Welcome to our experiment! In the course of this experiment you can earn a substantial amount of money. The precise amount will depend on your choices and some luck. We kindly ask you to remain silent throughout the entire experiment. Do not attempt to communicate with your neighbours and do not try to look at their screens. If you have any questions, please, raise your hand and we will come and answer it in private.

Notice that, in contrast to some other experiments, we do not allow you to take any notes during the experiment.

If you violate these general rules on behaviour, we will not be able to pay you.

This experimental session will consist of several on-screen stages:

1. **Experimental Phase I** for which instructions are below
2. **Experimental Phase II** for which instructions will be distributed at the end on Experimental Phase I
3. **Multiple choice quiz**
4. **Questionnaire about yourself**
5. **Feedback questionnaire about the experiment & final payment**

Your payment for this session will consist of the amounts you earn in Stages 1, 2, 3 together with the £5 show up fee. We will pay you in cash. You will need to sign a receipt, which we will supply.

We certainly should finish within the time allocated for the experiment.

Experimental Instructions: Phase I

In the course of this experimental phase you have to complete different "tasks." Each task is completed by "making a telephone call." There are altogether nine different tasks, called task 1, task 2, …, task 9. This phase of the experiment will consist of 14 task cycles where in each cycle you will have to do each task exactly once (in order). Once the 14 cycles are completed we move on to Phase II of the experiment. More information about this will follow then.

Your payment in this phase of the experiment will determined by premia (premia is another word for prizes or rewards) that we will pay to you for the completion of each task. Each task carries a specific premium and we will inform you of that premium prior to the task (they are also listed on page 3 of these instructions).

In addition, we will pay you 10p for every minute that you stay under an hour for completing all the 14 task cycles. From your total earnings (all task premia plus payment for staying under an hour) we will subtract "call charges" and "search costs" that you incur during the task completion. Let us now explain in more detail.

In order to complete a task you have to "call a telephone number". So, think of these tasks as making a dinner reservation, or a cinema booking; calling your bank for information on your account or calling a plumber to get your boiler or heating in your house fixed.

There are two different types of tasks. **SELECTION** tasks where you can choose among different telephone numbers, but you must successfully complete the task. **BINARY** tasks where there is only one telephone number and your choice is between calling or not calling it (if you don't call, you don't successfully complete the task).

Different telephone numbers (all of which are two digit numbers) will have different prices (the charge per minute). However, each telephone number has a fixed price that will not change during this phase of the experiment.

In contrast to that, *call durations* vary from task to task and may also vary from cycle to cycle. Specifically, each task is characterized by the minimum amount of minutes it takes to complete the task and the maximum number of minutes it will take to complete the task – these numbers are independent of the telephone number chosen. But these minimum and maximum times are fixed for each task and will remain constant throughout the experiment.

The actual call time will vary each time you pick a number. Specifically, the computer will draw a random duration somewhere between minimum and maximum each time you pick a number. All possible durations between minimum and maximum (in multiples of 30 seconds) are equally likely. You are not informed of the minimum and maximum call durations.

If you want to find out prices for different telephone numbers, you can carry out price searches by clicking on the "search" button next to each number. **Each price search will be charged at the flat rate of 80p** and will inform you about the price *per minute* for the chosen number. The results of these price searches will not be stored. However, if you forget the price of a number and want to search again we will only charge you half the search costs (40p). That is: every time you click a 'search' button, it will cost you 80p of your earnings (or 40p if you have clicked that button in a previous cycle).

In order to complete a given task, you have to press the "Call" button next to the number that you want to call. Once you press the call button, a timer will appear that will illustrate the length of the call. Call durations increase in steps of 30 seconds and every step will take about 2 seconds of real time. You can at any point in time press the "Hang up" button in order to terminate the call. The call will then only be charged up to that point in time. However, you will not successfully complete that task and you will not receive the task premium. If you hang up, you will have the opportunity to redial. In the case of a SELECTION task, you could instead choose to call another number. In the BINARY task you could instead choose not to complete the task by clicking "Don't call".

The total call charge results from the price per minute and the duration of the call (as simulated through the time, in multiples of 30 seconds).

Once you have completed task 1, you will move on to task 2. For each task, there are different telephone numbers. So there is no number that you can call for two different tasks. Once task 2 is completed, you move to task 3, and so on.

Once a cycle (of the 9 tasks) is completed you will receive a "phone bill" for that cycle. The bill will list all numbers you have called and will show the total charge for each number. You will also be shown the total of your premiums and search costs for the cycle. You will then move on to the next cycle.

Once you have completed all 14 cycles we will compute your total payment for Phase I of the experiment as the sum of the payments from each cycle. Hence, the total payment is computed as follows.

We will add up all the premia you have collected for completed tasks. (Remember if you decide not make a call or terminated a call, you do not complete the task.) In addition, we will take the number of minutes it took you to complete all the cycles. For each minute that you remained below an hour you will receive 10p on top of the task premia. So, for example, if you completed all the tasks within 45 minutes you will earn an extra £1.50.

From that we will subtract two amounts: (i) your total phone charges, i.e, the sum of all 14 telephone bills; and (ii) the sum of all search costs you have incurred during the 14 cycles.

For example, if you completed all the tasks in 45 minutes and the sum of your task premia is £95 and the sum of all of your phone bills is £70 and you did 5 new price searches (5 * 80p), your total payment would be £95 + £1.50 - £70 - £4 = £19.50.

At the end of the Phase, you will need to wait until all subjects have finished. Then we will distribute the instructions for Phase II. If you finish before

others, please sit quietly. You may read any materials you have brought, but we ask you not to use the computers nor mobile devices.

The following table lists the 9 tasks that you will complete in order together with their types and the premium for the task.

Task Number	Task Type	Task Premium (pence)
1	SELECTION	15
2	BINARY	60
3	BINARY	120
4	BINARY	60
5	BINARY	260
6	SELECTION	60
7	BINARY	120
8	BINARY	260
9	SELECTION	90

Experimental Instructions: Phase II [PCAE]

Phase II of the experiment is structured just like the Phase I. There are still nine tasks in a cycle and 14 cycles to complete. There are, however, new premia and numbers for each task.

All rules will stay the same with just one exception: when you press the call button, you will see the price for the number you have chosen on the screen (this is called the "precall announcement"). This price will remain on the screen for a short while before you are connected to the number and everything works like before (the call proceeds). You can choose to cancel before the call is connected if you wish. If you do so, there will be no charge for the call.

If you don't want the price announcement anymore you can press the "change precall setting" button.. Should you change your mind and wish to have the price announcements back, you need only press the "change precall setting" button again. Changing your precall setting takes a few seconds.

Task Number	Task Type	Task Premium (pence)
1	BINARY	120
2	SELECTION	15
3	SELECTION	90
4	BINARY	60
5	BINARY	60
6	SELECTION	60
7	BINARY	260
8	BINARY	120
9	BINARY	260

At the end of the Phase, you will need to wait until all subjects have finished. Then instructions will appear on the screen for the quiz and following that, the questionnaire and final payments.

If you finish before others, please sit quietly. You may read any materials you have brought, but we ask you not to use the computers nor mobile devices.

The code to start Phase II is 19796

Experimental Instructions: Phase II [PCAM]

Phase II of the experiment is structured just like Phase I. There are still nine tasks in a cycle and 14 cycles to complete. There are, however, new premia and numbers for each task. All rules will stay the same with just one exception: when you press the call button, you will see the maximum per minute price for a number on this task on the screen (this is called the "precall announcement").

This price will remain on the screen for a short while before you are connected to the number and everything works like before (the call proceeds). You can choose to cancel before the call is connected if you wish. If you do so, there will be no charge for the call.

If you don't want the price announcement anymore you can press the "change precall setting" button.. Should you change your mind and wish to have the price announcements back, you need only press the "change precall setting" button again. Changing your precall setting takes a few seconds.

Task Number	Task Type	Task Premium (pence)
1	BINARY	120
2	SELECTION	15
3	SELECTION	90
4	BINARY	60
5	BINARY	60
6	SELECTION	60
7	BINARY	260
8	BINARY	120
9	BINARY	260

At the end of the Phase, you will need to wait until all subjects have finished. Then instructions will appear on the screen for the quiz and following that, the questionnaire and final payments.

If you finish before others, please sit quietly. You may read any materials you have brought, but we ask you not to use the computers nor mobile devices.

The code to start Phase II is 19796

Experimental Instructions: Phase II [INFO]

Phase II of the experiment is structured just like Phase I. There are still nine tasks in a cycle and 14 cycles to complete. There are, however, new premia and numbers for each task.

All rules will stay the same with just one exception: at the bottom of the screen where your phone bill is displayed, there is a button labelled "show call charges". If you press this, you will see a screen which will contain a long list with telephone numbers (including all those that you might want to choose, but also others) and next to each number its price will be shown. To go back to the screen containing your phone bill, click the "hide call charges" button.

Task Number	Task Type	Task Premium (pence)
1	BINARY	120
2	SELECTION	15
3	SELECTION	90
4	BINARY	60
5	BINARY	60
6	SELECTION	60
7	BINARY	260
8	BINARY	120
9	BINARY	260

At the end of the Phase, you will need to wait until all subjects have finished. Then instructions will appear on the screen for the quiz and following that, the questionnaire and final payments.

If you finish before others, please sit quietly. You may read any materials you have brought, but we ask you not to use the computers nor mobile devices.

The code to start Phase II is 19796

Experimental Instructions: Phase II [SC]

Phase II of the experiment is structured just like Phase I. There are still nine tasks in a cycle and 14 cycles to complete. There are, however, new premia and numbers for each task.

All rules will stay the same with just one exception: the search cost for finding out the price of a call is now reduced to 5p. This 5p search cost applies both for the first search and for all later searches for a given number.

Task Number	Task Type	Task Premium (pence)
1	BINARY	120
2	SELECTION	15
3	SELECTION	90
4	BINARY	60
5	BINARY	60
6	SELECTION	60
7	BINARY	260
8	BINARY	120
9	BINARY	260

At the end of the Phase, you will need to wait until all subjects have finished. Then instructions will appear on the screen for the quiz and following that, the questionnaire and final payments.

If you finish before others, please sit quietly. You may read any materials you have brought, but we ask you not to use the computers nor mobile devices.

The code to start Phase II is 19796

Annex 3 Parameter Choices

In this section, we provide tables detailing the specific parameter choice for the experiment.

The table below lists the parameter choices for the tasks of the experiment (recall that the same 9 tasks were performed in the same order constituting one 'cycle' and the cycle was repeated 14 times). The same set of choices was available in the baseline and the intervention, but not in the same order.

The same set of parameters was used for landline and mobile treatments. The premia and per minute call charges were doubled for the mobile

The types of task presented are selection (SEL) where subjects had to choose one number from many and call it and binary (BIN) where subjects only had one number and the choice was whether to call or not.

Premiums are given in pence and call costs are given in pence per minute. Minimum and maximum call lengths are given in minutes (the actual call time was rounded up to the nearest 30 seconds). In the selection tasks, the numbers were ordered randomly, so without searching subjects couldn't identify which number was cheapest.

The optimal action is based on the expected payoff.

Parameter choices								
Task Type	**Premium (pence)**	**Prices of available numbers (pence per minute)**	**Min. call length (minutes)**	**Max. call length (minutes)**	**Optimal action**	**Optimal Payoff**	**Order in baseline**	**Order in intervention**
SEL	15	4, 5, 6, 7, 8	1	3	Call 4p number	7p	1	2
SEL	90	40, 45, 50, 55, 60	1	3	Call 40p number	10p	9	3
SEL	60	5, 10, 15, 20, 30, 40, 50, 60	1	3	Call 5p number	50p	6	6
BIN	60	8	0	20	Don't call	0p	2	4
BIN	60	16	2	3	Call	20p	4	5
BIN	120	33	2	4	Call	21p	7	8
BIN	120	30	0	10	Don't call	0p	3	1
BIN	260	100	2	4	Don't call	0p	5	7
BIN	260	40	0	10	Call	60p	8	9

The following table lists the search costs for the experiment in pence. These are the costs subjects incurred every time they did a price search. They pay the 'first' cost the first time they search on a specific number. Every time they subsequently search again for the cost of the same number (the cost of calling a given number does not change in the experiment), they pay the 'repeat search cost'.

Search costs				
	Landline		**Mobile**	
	First search	**Repeat search**	**First Search**	**Repeat Search**
Baseline	80	40	240	120

Search costs				
	Landline		**Mobile**	
	First search	**Repeat search**	**First Search**	**Repeat Search**
PCAE	80	40	240	120
PCAM	80	40	240	120
SC	5	5	5	5
INFO	80	80	240	120

Annex 4 Summary of subject performance

The table below provides a summary of the number of subjects in each treatment and the average aggregate performance.

Relative Premia and call costs are given in comparison to an omniscient subject.

The first line of the table indicates that there were 113 subjects in the Landline Baseline who on average gained 17% more in premia and paid 64% more in call costs than the omniscient subject. The average subject in this treatment also made 14.1 price searches.

	Treatment	Subjects in treatment	Relative Premia	Relative call costs	Average searches
Landline	**Baseline**	113	117%	164%	14.1
	PCAE	30	96%	114%	2.6
	PCAM	28	104%	136%	5.7
	SC	30	102%	130%	22.6
	INFO	25	91%	122%	3.6
Mobile	**Baseline**	98	120%	167%	11.1
	PCAE	26	100%	124%	0.5
	PCAM	20	98%	129%	4.4
	SC	26	97%	125%	36.0
	INFO	26	109%	139%	2.3

A4.1 Summary of calls

The table below gives the distribution of calls made in the baseline, successful and terminated, by cycle (average per subject). Subjects quickly learn to call less (recall that it is optimal to make 6 successful and 0 terminated calls), but there is some persistence in the early termination of calls.

Cycle	Successful calls	Terminated calls
1	7.9	0.6
2	7.2	0.8
3	7.0	0.8
4	6.9	0.8
5	6.7	0.6
6	6.5	0.7
7	6.5	0.7
8	6.4	0.7
9	6.4	0.7
10	6.4	0.6
11	6.3	0.6
12	6.2	0.6
13	6.2	0.6
14	6.3	0.5

The following table shows the distribution of calls made in the baseline, successful and terminated by task – that is, we tabulate the average number of calls per subject over the course of 14 cycles for each task. Hence, for the selection tasks the average would be 14 as subjects had to successfully complete these tasks. We also tabulate the premium for the task and maximum call time.

Task	Optimal Action	Premium	Max length	Successful Calls	Terminated Calls
SEL	Call	15	3	14.0	0.5
BIN	Don't call	60	20	5.7	3.5
BIN	Don't call	120	10	7.5	1.9
BIN	Call	60	3	9.3	0.3
BIN	Don't call	260	4	8.1	0.7
SEL	Call	60	3	14.0	0.2
BIN	Call	120	4	10.7	0.4
BIN	Call	260	10	9.6	1.5
SEL	Call	90	3	14.0	0.5

6.3 Regressions I: pay against searches

	baseline pay	baseline pay (<= 7 searches)
logsearches	-1,217.86***	73.52
	(111.50)	(203.38)
Constant	308.42	-617.08**
	(256.24)	(250.47)
Observations	211	46
R-squared	0.36	0.00

Standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%

The first column of the table shows the results of a regression of pay in the baseline against (log of) number of searches done. There is a strong negative link.

The second column shows the results when this regression is restricted to subjects doing fewer than 7 searches. There is a positive link, but it is not significant.

6.4 Regressions II: total pay against aptitude

	Landline	Mobile
	Total pay	Total pay
Aptitude	114.01	432.45**
	(86.92)	(207.90)
Constant	-2,505.36***	-6,196.41***
	(870.72)	(2,052.11)
Observations	113	98
R-squared	0.02	0.04

Standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%

This table reports the results of a regression of *Total pay* (in the experiment) against *Aptitude* (measured on a scale of 0 to 12). This regression indicates that there is a strong correlation between total pay and aptitude, especially for the mobile treatment where a 1 mark increase in aptitude (out of 12) was associated with an increase in earnings of £4.32 (landline £1.14)

6.5 Regressions III: learning in baseline

	Landline	Mobile
	Pay difference	Pay difference
Aptitude	72.32*	69.20
	(38.36)	(115.43)
Constant	160.39	1,160.71
	(384.28)	(1,139.33)
Observations	113	98
R-squared	0.03	0.00

Standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%

This table reports the results of a regression of *Pay difference* (between the first half and second half of the baseline) against *Aptitude,* separately for landline and mobile treatments. Improvements from the first half to second half is associated with high aptitude.

6.6 Regressions IV: learning in interventions

	Landline	Mobile
	Pay difference	Pay difference
Aptitude	-47.24	-150.03*
	(30.45)	(75.67)
Constant	963.14***	2,294.50***
	(305.02)	(746.91)
Observations	113	98
R-squared	0.02	0.04

Standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%

This table reports the results of a regression of *Pay difference* (between the first half and second half of the intervention) against *Aptitude,* separately for landline and mobile treatments. Improvements from the first half to second half is associated with low aptitude. This indicates that the high aptitude subjects quickly learn how to best use the interventions whereas the lower aptitude subjects take more time.

6.7 Regressions V: effect of search on call cost

	Cost
Searches	-0.09***
	(0.01)
Constant	1.00***
	(0.02)
Observations	633
R-squared	0.12

Standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%

This table reports the results of a regression of *Relative Call Cost* (call cost relative to the expected call cost of phoning a number at random) against *Searches* for all subjects in the first cycle of the baseline (i.e. before they had any chance to learn which numbers were cheapest).

The results show that for each search done, there was an associated call cost reduction of 9%. This link is very strong (significant at 1% level). Individual regressions by task and mobile/landline are equally strong and show the same thing.

6.8 Regressions VI: bill shock

	Cycle 1	Cycle 12
	Repeat	Repeat
Call length	-0.02***	-0.01***
	(0.00)	(0.00)
Observations	1028	668

Standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%

This table reports the results of a probit regression. The dependent variable *Repeat* is the probability of a subject successfully completing a binary task in the following cycle and the independent variable *Call length* is the call length of that task in the current cycle. The two columns report the results where the current cycle is the first cycle or the penultimate cycle.

The results indicate that, for the first cycle, for every minute that a call lasts, the probability decreases by 2% (recall that some tasks may take a maximum of 20 minutes to complete). This is indicative of bill shock.

To see that the bill shock effect persists, we get an equally significant although smaller effect in the penultimate cycle even though subjects should have a good idea of the distribution of call lengths by this point.

11-15 Betterton Street
London WC2H 9BP
Tel: +44 20 7866 8185
Fax: +44 20 7866 8186
Email: info@londecon.co.uk

London | Brussels | Dublin | Paris | Budapest | Valletta