PRIC€ LAB

AN INVESTIGATION OF CONSUMERS' CAPABILITIES WITH COMPLEX PRODUCTS

PETE LUNN, MAREK BOHACEK, JASON SOMERVILLE,
ÁINE NÍ CHOISDEALBHA & FÉIDHLIM MCGOWAN

Banc Ceannais na hÉireann
Central Bank of Ireland
Eurosystem

Coimisiún um Iomaíocht agus Cosaint Tomhaltóirí
Competition and Consumer Protection Commission

Commission for Communications Regulation

CER
Commission for Energy Regulation
An Coimisiún um Rialáil Fuinnimh

ESRI

PRICE Lab: An Investigation of Consumers' Capabilities with Complex Products

Pete Lunn, Marek Bohacek, Jason Somerville,
Áine Ní Choisdealbha and Féidhlim McGowan

May 2016

Available to download from www.esri.ie

Whitaker Square, Sir John Rogerson's Quay, Dublin 2

ISBN 978-0-7070-0400-6

The ESRI

The *Economic Research Institute* was founded in Dublin in 1960, with the assistance of a grant from the Ford Foundation of New York. In 1966 the remit of the Institute was expanded to include social research, resulting in the Institute being renamed the *Economic and Social Research Institute* (ESRI). In 2010 the Institute entered into a strategic research alliance with Trinity College Dublin, while retaining its status as an independent research institute.

The ESRI is governed by an independent Council which acts as the board of the Institute with responsibility for guaranteeing its independence and integrity. The Institute's research strategy is determined by the Council in association with the Director and staff. The research agenda seeks to contribute to three overarching and interconnected goals, namely, economic growth, social progress and environmental sustainability. The Institute's research is disseminated through international and national peer reviewed journals and books, in reports and books published directly by the Institute itself and in the Institute's working paper series. Researchers are responsible for the accuracy of their research. All ESRI books and reports are peer reviewed and these publications and the ESRI's working papers can be downloaded from the ESRI website at www.esri.ie

The Institute's research is funded from a variety of sources including: an annual grant-in-aid from the Irish Government; competitive research grants (both Irish and international); support for agreed programmes from government departments/agencies and commissioned research projects from public sector bodies. Sponsorship of the Institute's activities by Irish business and membership subscriptions provide a minor source of additional income.

The Authors

Pete Lunn is a Senior Research Officer at the ESRI and the Principal Investigator of PRICE Lab. Marek Bohacek is a Research Assistant and Áine Ní Choisdealbha a Post-Doctoral Fellow at the ESRI. Jason Somerville and Féidhlim McGowan contributed to this research as part of their studies as research students at Trinity College Dublin.

Acknowledgements

These experiments were conducted as part of PRICE Lab, a research programme funded by the Central Bank of Ireland, Commission for Energy Regulation, Competition and Consumer Protection Commission and the Commission for Communications Regulation. The authors are especially grateful to members of these funding organisations who have contributed to the research programme via PRICE Lab's Steering Group and to others who have engaged with the work of the lab at various presentations and seminars. The experiments have prompted animated discussion, which has proved helpful for experimental designs, interpretation of results and consideration of policy implications. We are also very grateful to Professor Stephen Lea and to Dr Seán Lyons for their contribution both to the steering group and more broadly to the work of PRICE Lab. The work described here has been presented to too many audiences both in Ireland and internationally for each to be thanked here, but we are grateful for much feedback from numerous academic colleagues and policymakers, who have contributed substantially to the programme.

Table of Contents

EXECUTIVE SUMMARY x

PART 1: INTRODUCTION AND METHODOLOGY

SECTION 1 INTRODUCTION 1

1.1. Background 1

1.2 Behavioural Economics 1

1.3 International Research on Biases in Consumer Decision-Making 3

1.4 Policy Responses to Problems of Complex Products 6

1.5 The Logic of PRICE Lab 9

1.6 Research Questions 10

SECTION 2 THE SURPLUS IDENTIFICATION (S-ID) TASK 12

2.1 Making the Surplus Objective 12

2.2 Hyperproducts 14

2.3 Statistical Analyses 16

2.4 Generalisability 19

PART 2: EXPERIMENTAL FINDINGS

SECTION 3 HOW ACCURATELY CAN CONSUMERS RESOLVE A TRADE-OFF? 26

3.1 Introduction 26

3.2 Experiment A: Aims and Methods 27

3.2 Experiment A: Results 30

3.3 Experiment A: Discussion 34

SECTION 4 HOW MANY ATTRIBUTES CAN CONSUMERS COPE WITH? 35

4.1 Introduction 35

4.2 Experiment B: Aims and Methods 35

4.3 Experiment B: Results 37

4.4 Experiment B: Discussion 41

4.5 Experiment C: Aims and Methods 43

4.6 Experiment C: Results 44

4.7 Experiment C: Discussion 46

4.8 Experiment D: Aims and Methods 47
4.9 Experiment D: Results 48
4.10 Discussion 50

SECTION 5 CAN CONSUMERS COPE BETTER WITH CATEGORICAL AND NUMERIC ATTRIBUTES? 53
5.1 Introduction 53
5.2 Experiment E: Aims and Methods 54
5.3 Experiment E: Results 56
5.4 Experiment E: Discussion 58
5.5 Experiment F: Aims and Methods 59
5.6 Experiment F: Results 61
5.7 Discussion 63

SECTION 6 CAN CONSUMERS HANDLE NON-LINEAR ATTRIBUTE RETURNS? 64
6.1 Introduction 64
6.2 Experiment G: Aims and Methods 64
6.3 Experiment G: Results 67
6.4 Experiment H: Aims and Methods 69
6.5 Experiment H: Results 70
6.6 Discussion 73

SECTION 7 DOES INACCURACY GENERALISE TO LARGER RANGES OF PRODUCTS AND MORE FAMILIAR PRODUCTS? 74
7.1 Introduction 74
7.2 Experiment I: Aims and Methods 74
7.3 Experiment I: Results 77
7.4 Experiment I: Discussion 79
7.5 Experiment J: Aims and Methods 79
7.6 Experiment J: Results 83
7.7 Experiment J: Discussion 85
7.8 Discussion 86

PART 3: CONCLUSIONS

SECTION 8 SUMMARY OF FINDINGS 88
8.1 Introduction 88

8.2 Summary of Findings 88

SECTION 9 POLICY IMPLICATIONS 91

9.1 Introduction 91

9.2 From Evidence to Policy 91

9.3 Implications of the Present Findings 94

REFERENCES 98

APPENDIX 101

List of Tables

Table 1 Summary of Experiment Features, Research Questions and Findings 24

Table 2 The value functions used in Experiment H 70

List of Figures

Figure 1 The Three Hyperproducts Used In The Experiments 15

Figure 2 Example S-ID task data 17

Figure 3 Example tasks from Experiment A 29

Figure 4 The 86% just noticeable differences (JNDs) in surplus for four groups of 16 participants and six tasks in Experiment A. 31

Figure 5 Biases across the price range for four groups of 16 participants and six tasks in Experiment A. 33

Figure 6 JNDs for identifying a surplus with one, two, three and four attributes in Experiment B, comparing actual performance with a hypothetical participant who integrates additional attribute information with statistical efficiency. 38

Figure 7 JNDs for the 15 possible combinations of four attributes in Experiment B 39

Figure 8 Bias across the price range by number of attributes in Experiment B. 40

Figure 9 JNDs for identifying a surplus with one, two, three and four attributes in Experiment C, across three consecutive sessions. 45

Figure 10 Bias across the price range by number of attributes in the third session of Experiment C 46

Figure 11 JNDs for increasing numbers of perfectly correlated attributes in Experiment D 49

Figure 12 Bias across the price range by number of attributes in Experiment D 50

Figure 13 Example attributes and products from Experiment E. 55

Figure 14 JNDs in Experiment E for multiple conditions with one visual attribute (blue) or two attributes consisting of different combinsatoins of visual, numeric and categorical attributes (green). None of the differences between the JNDs for the two-attribute conditions are statistically significant. 57

Figure 15 Bias across the price range by number of attributes in Experiment E 58

Figure 16 Example attributes and products from Experiment F 60

Figure 17 JNDs in Experiment F for multiple two-attribute conditions, combining visual (Height, Brick), numeric (Height(n), Age) and two-category (Doors) attributes. 61

Figure 18 Bias across the price range by condition in Experiment F. 62

Figure 19 An example of the varying degrees of diminishing returns employed in Experiment G. 65

Figure 20 Results of the pilot study for Experiment G 67
Figure 21 JNDs for linear and non-linear attribute-price relationships in Experiment G.. 68
Figure 22 Biases across the price range in Experiment G. 68
Figure 23 JNDs for a range of more complex two-attribute non-linear functions in Experiment H. 71
Figure 24 Example screens from the MS-ID task in Experiment I. 76
Figure 25 JNDs and biases for the standard S-ID task in Experiment I across four conditions. 78
Figure 26 Precision in the MS-ID task of Experiment I 79
Figure 27 Onscreen environments for Experiment J. 82
Figure 28 JNDs for surplus identification relative to average market prices for Dublin houses and broadband packages, compared to hyperproducts with the same mathematical relationship between attributes and prices (Experiment J). 84
Figure 29 Biases across the price range in a surplus identification relative to average market prices for Dublin houses and broadband packages, compared to hyperproducts with the same mathematical relationship between attributes and prices (Experiment J). 85

Glossary

Accuracy	The likelihood that a participant correctly identifies a surplus. Accuracy is a combination of precision and bias, measured using the JND and PSE respectively.
Behavioural Economics	The use of insights and methods from experimental psychology (and related disciplines) for economic analysis.
Bias	A reduction in accuracy due to systematic directional error. When judging a surplus, a bias means that the participant consistently overestimates or underestimates the surplus. Note that a participant in a judgement task can be biased but have high precision, because their judgments may be consistently too high or too low yet vary little in relation to one another.
Forced-Choice	An experimental method in which participants are presented with two or more specific response options and have to choose one. That is, participants must commit to an answer – 'don't know' is not permitted. In the standard S-ID task, the participant tries to choose the option that is better value (choosing between the product or price when only one product is presented, or between alternative products when more than one product is presented).
Just Noticeable Difference (JND)	A measure of the precision of judgements. The JND is the amount of surplus required in order for participants to identify said surplus reliably. More technically, the JND is the size of surplus needed for a participant to go from identifying it 50 per cent of the time to 86 per cent of the time.
Linear	A relationship between two variables is linear if a one unit increase in one quantity always produces the same increase in the other quantity. A graph of a linear relationship produces a straight line.
Mandated Disclosure	A form of regulation that requires firms to provide specific product information or to disclose information in a standardised format.
Mandated Simplification	A form of mandated disclosure policy that requires firms to design or adjust their products, product ranges or product descriptions so that consumers can understand and make decisions about the products more easily. Mandated simplification often involves product descriptions that adhere to a standardised format to assist product comparison.
Monotonic	A function is monotonic if an increase in one quantity always increases (or always decreases) the other quantity. For example, where a larger product always results in higher value, or where higher fuel consumption results in lower value, the relationships are monotonic. Monotonic functions contain no turning points.
Multi-Product	A forced-choice task specifically designed to assess the accuracy with which

Surplus Identification Task (MS-ID)	consumers are able to detect (identify) a surplus when choosing among multiple possible products.
Non-linear	A relationship between two variables is non-linear if a one unit increase in one quantity does not always produces the same increase in the other quantity. For example, if the size of a small product is increased it may raise the value by more than if the size of a medium-sized product is increased by the same amount. A graph of a non-linear relationship deviates from a straight line, often because it is curved. A non-linear function may or may not contain a turning point.
Precision	The amount of variability in an estimate. Note that a participant in a judgement task can be imprecise but have no bias, because their judgments can be subject to large variation that is not consistently too high or too low, i.e. they are accurate on average.
Precision-Bias Trade-Off	The idea that the more precise a participant is at identifying a surplus, the more likely their judgments are to be affected by bias across the price range.
Point of Subjective Equality (PSE)	A measure of bias in judgements. The PSE is the amount of surplus at which the participant judges that the surplus is exactly zero, i.e. the product has the exact same value as the price.
Surplus	How much a product is worth product over and above its price or, equivalently, the product's value minus its price. In theory, surplus can be measured in multiple ways: in monetary terms, as a percentage, or (in the present report) as a proportion of the total price range.
Surplus Identification (S-ID) Task	A forced-choice task specifically designed to assess the accuracy with which consumers are able to detect (identify) a surplus. Consumers learn how much the product is worth, which is determined by an objective function set by the experimenters. Their accuracy in identifying the surplus is then tested over multiple experimental trials.
Turning Point	The point at which a function changes from increasing to decreasing or vice-versa. Functions with turning points are hence not monotonic. For example, every increase in a product's size up to a certain point could increase its value, and after this optimum size is reached, every increase in size could decrease the product's value.

Executive Summary

This report describes a series of experiments carried out by PRICE Lab, a research programme at the Economic and Social Research Institute (ESRI) jointly funded by the Central Bank of Ireland, the Commission for Energy Regulation, the Competition and Consumer Protection Commission and the Commission for Communications Regulation. The experiments were conducted with samples of Irish consumers aged 18-70 years and were designed to answer the following general research question: At what point do products become too complex for consumers to choose accurately between the good ones and the bad ones?

BACKGROUND AND METHODS

PRICE Lab represents a departure from traditional methods employed for economic research in Ireland. It belongs to the rapidly expanding area of 'behavioural economics', which is the application of psychological insights to economic analysis. In recent years, behavioural economics has developed novel methods and generated many new findings, especially in relation to the choices made by consumers. These scientific advances have implications both for economics and for policy. They suggest that consumers often do not make decisions in the way that economists have traditionally assumed. The findings show that consumers have limited capacity for attending to and processing information and that they are prone to systematic biases, all of which may lead to disadvantageous choices. In short, consumers may make costly mistakes. Research has indeed documented that in several key consumer markets, including financial services, utilities and telecommunications, many consumers struggle to choose the best products for themselves. It is often argued that these markets involve 'complex' products. The obvious question that arises is whether consumer policy can be used to help them to make better choices when faced with complex products.

Policies are more likely to be successful where they are informed by an accurate understanding of how real consumers make decisions between products. To provide evidence for consumer policy, PRICE Lab has developed a method for measuring the accuracy with which consumers make choices, using techniques adapted from the scientific study of human perception. The method allows researchers to measure how reliably consumers can distinguish a good deal from a bad one. A good deal is defined here as one where the product is more valuable than the price paid. In other words, it offers good value for money or, in the jargon of economics, offers the consumer a 'surplus'. Conversely, a bad deal

offers poor value for money, providing no (or a negative) surplus. PRICE Lab's main experimental method, which we call the 'Surplus Identification' (S-ID) task, allows researchers to measure how accurately consumers can spot a surplus and whether they are prone to systematic biases. Most importantly, the S-ID task can be used to study how the accuracy of consumers' decisions changes as the type of product changes.

For the experiments we report here, samples of consumers arrived at the ESRI one at a time and spent approximately one hour doing the S-ID task with different kinds of products, which were displayed on a computer screen. They had to learn to judge the value of one or more products against prices and were then tested for accuracy. As well as people's intrinsic motivation to do well when their performance on a task like this is tested, we provided an incentive: one in every ten consumers who attended PRICE Lab won a prize, based on their performance. Across a series of these experiments, we were able to test how the accuracy of consumers' decisions was affected by the number and nature of the product's characteristics, or 'attributes', which they had to take into account in order to distinguish good deals from bad ones. In other words, we were able to study what exactly makes for a 'complex' product, in the sense that consumers find it difficult to choose good deals.

FINDINGS

Overall, across all ten experiments described in this report, we found that consumers' judgements of the value of products against prices were surprisingly inaccurate. Even when the product was simple, meaning that it consisted of just one clearly perceptible attribute (e.g. the product was worth more when it was larger), consumers required a surplus of around 16-26 per cent of the total price range in order to be able to judge accurately that a deal was a good one rather than a bad one. Put another way, when most people have to map a characteristic of a product onto a range of prices, they are able to distinguish at best between five and seven levels of value (e.g. five levels might be thought of as equivalent to 'very bad', 'bad', 'average', 'good', 'very good'). Furthermore, we found that judgements of products against prices were not only imprecise, but systematically biased. Consumers generally overestimated what products at the top end of the range were worth and underestimated what products at the bottom end of the range were worth, typically by as much as 10-15 per cent and sometimes more.

We then systematically increased the complexity of the products, first by adding more attributes, so that the consumers had to take into account, two, three, then four different characteristics of the product simultaneously. One product might be good on attribute A, not so good on attribute B and available at just above the

average price; another might be very good on A, middling on B, but relatively expensive. Each time the consumer's task was to judge whether the deal was good or bad. We would then add complexity by introducing attribute C, then attribute D, and so on. Thus, consumers had to negotiate multiple trade-offs.

Performance deteriorated quite rapidly once multiple attributes were in play. Even the best performers could not integrate all of the product information efficiently – they became substantially more likely to make mistakes. Once people had to consider four product characteristics simultaneously, all of which contributed equally to the monetary value of the product, a surplus of more than half the price range was required for them to identify a good deal reliably. This was a fundamental finding of the present experiments: once consumers had to take into account more than two or three different factors simultaneously their ability to distinguish good and bad deals became strikingly imprecise. This finding therefore offered a clear answer to our primary research question: a product might be considered 'complex' once consumers must take into account more than two or three factors simultaneously in order to judge whether a deal is good or bad.

Most of the experiments conducted after we obtained these strong initial findings were designed to test whether consumers could improve on this level of performance, perhaps for certain types of products or with sufficient practice, or whether the performance limits uncovered were likely to apply across many different types of product. An examination of individual differences revealed that some people were significantly better than others at judging good deals from bad ones. However the differences were not large in comparison to the overall effects recorded; everyone tested struggled once there were more than two or three product attributes to contend with. People with high levels of numeracy and educational attainment performed slightly better than those without, but the improvement was small. We also found that both the high level of imprecision and systematic bias were not reduced substantially by giving people substantial practice and opportunities to learn – any improvements were slow and incremental.

A series of experiments was also designed to test whether consumers' capability was different depending on the type of product attribute. In our initial experiments the characteristics of the products were all visual (e.g., size, fineness of texture, etc.). We then performed similar experiments where the relevant product information was supplied as numbers (e.g., percentages, amounts) or in categories (e.g., Type A, Rating D, Brand X), to see whether performance might improve. This question is important, as most financial and contractual information is supplied to consumers in a numeric or categorical form. The results showed clearly that the type of product information did not matter for the level

of imprecision and bias in consumers' decisions – the results were essentially the same whether the product attributes were visual, numeric or categorical. What continued to drive performance was how many characteristics the consumer had to judge simultaneously. Thus, our findings were not the result of people failing to perceive or take in information accurately. Rather, the limiting factor in consumers' capability was how many different factors they had to weigh against each other at the same time.

In most of our experiments the characteristics of the product and its monetary value were related by a one-to-one mapping; each extra unit of an attribute added the same amount of monetary value. In other words, the relationships were all linear. Because other findings in behavioural economics suggest that consumers might struggle more with non-linear relationships, we designed experiments to test them. For example, the monetary value of a product might increase more when the amount of one attribute moves from very low to low, than when it moves from high to very high. We found that this made no difference to either the imprecision or bias in consumers' decisions provided that the relationship was monotonic (i.e. the direction of the relationship was consistent, so that more or less of the attribute always meant more or less monetary value respectively). When the relationship involved a turning point (i.e. more of the attribute meant higher monetary value but only up to a certain point, after which more of the attribute meant less value) consumers' judgements were more imprecise still.

Finally, we tested whether familiarity with the type of product improved performance. In most of the experiments we intentionally used products that were new to the experimental participants. This was done to ensure experimental control and so that we could monitor learning. In the final experiment reported here, we used two familiar products (Dublin houses and residential broadband packages) and tested whether consumers could distinguish good deals from bad deals any better among these familiar products than they could among products that they had never seen before, but which had the same number and type of attributes and price range. We found that consumers' performance was the same for these familiar products as for unfamiliar ones. Again, what primarily determined the amount of imprecision and bias in consumers' judgments was the number of attributes that they had to balance against each other, regardless of whether these were familiar or novel.

POLICY IMPLICATIONS

There is a menu of consumer polices designed to assist consumers in negotiating complex products. A review, including international examples, is given in the main body of the report. The primary aim is often to simplify the consumer's task.

Potential policies, versions of which already exist in various forms and which cover a spectrum of interventionist strength, might include: the provision and endorsement of independent, transparent price comparison websites and other choice engines (e.g. mobile applications, decision software); the provision of high-quality independent consumer advice; 'mandated simplification', whereby regulations stipulate that providers must present product information in a simplified and standardised format specifically determined by regulation; and more strident interventions such as devising and enforcing prescriptive rules and regulations in relation to permissible product descriptions, product features or price structures. The present findings have implications for such policies.

However, while the experimental findings have implications for policy, it needs to be borne in mind that the evidence supplied here is only one factor in determining whether any given intervention in markets is likely to be beneficial. The findings imply that consumers are likely to struggle to choose well in markets with products consisting of multiple important attributes that must all be factored in when making a choice. Interventions that reduce this kind of complexity for consumers may therefore be beneficial, but nothing in the present research addresses the potential costs of such interventions, or how providers are likely to respond to them. The findings are also general in nature and are intended to give insights into consumer choices across markets. There are likely to be additional factors specific to certain markets that need to be considered in any analysis of the costs and benefits of a potential policy change.

Most importantly, the policy implications discussed here are not specific to Ireland or to any particular product market. Furthermore, they should not be read as criticisms of existing regulatory regimes, which already go to some lengths in assisting consumers to deal with complex products. Ireland currently has extensive regulations designed to protect consumers, both in general and in specific markets, descriptions of which can be found in Section 9.1 of the main report.

Nevertheless, the experiments described here do offer relevant guidance for future policy designs. For instance, they imply that while policies that make it easier for consumers to switch providers may be necessary to encourage active consumers, they may not be sufficient, especially in markets where products are complex. In order for consumers to benefit, policies that help them to identify better deals reliably may also be required, given the scale of inaccuracy in consumers' decisions that we record in this report when products have multiple important attributes. Where policies are designed to assist consumer decisions, the present findings imply quite severe limits in relation to the volume of information consumers can simultaneously take into account. Good impartial

consumer advice may limit the volume of information and focus on ensuring that the most important product attributes are recognised by consumers.

The findings also have implications for the role of competition. While consumers may obtain substantial potential benefits from competition, their capabilities when faced with more complex products are likely to reduce such benefits. Pressure from competition requires sufficient numbers of consumers to spot and exploit better value offerings. Given our results, providers with larger market shares may face incentives to increase the complexity of products in an effort to dampen competitive pressure and generate more market power. Where marketing or pricing practices result in prices or attributes with multiple components, our findings imply that consumer choices are likely to become less accurate. Policymakers must of course be careful in determining whether such practices amount to legitimate innovations with potential consumer benefit. Yet there is a genuine danger that spurious complexity can be generated that confuses consumers and protects market power.

The results described here provide backing for the promotion and/or provision by policymakers of high-quality independent choice engines, including but not limited to price comparison sites, especially in circumstances where the number of relevant product attributes is high. A longer discussion of the potential benefits and caveats associated with such policies is contained in the main body of the report.

Mandated simplification policies are gaining in popularity internationally. Examples include limiting the number of tariffs a single energy company can offer or standardising health insurance products, both of which are designed to simplify the comparisons between prices and/or product attributes. The present research has some implications for what might make a good mandate. Consumer decisions are likely to be improved where a mandate brings to the consumer's attention the most important product attributes at the point of decision. The present results offer guidance with respect to how many key attributes consumers are able simultaneously to trade off, with implications for the design of standardised disclosures. While bearing in mind the potential for imposing costs, the results also suggest benefits to compulsory 'meta-attributes' (such as APRs, energy ratings, total costs, etc.), which may help consumers to integrate otherwise separate sources of information.

FUTURE RESEARCH

The experiments described here were designed to produce findings that generalise across multiple product markets. However, in addition to the results

outlined in this report, the work has resulted in new experimental methods that can be applied to more specific consumer policy issues. This is possible because the methods generate experimental measures of the accuracy of consumers' decision-making. As such, they can be adapted to assess the quality of consumers' decisions in relation to specific products, pricing and marketing practices. Work is underway in PRICE Lab that applies these methods to issues in specific markets, including those for personal loans, energy and mobile phones.

Part 1

Introduction and Methodology

Section 1

Introduction

1.1. BACKGROUND

Consumer choice is a fundamental concept in a modern market economy. It is a cornerstone of economic theory, an important branch of experimental psychology and the preoccupation of marketing science. Many important life outcomes are determined by our decisions as consumers: whether we have warm and comfortable homes; how much we are able to save; whether we benefit from the latest technologies; where we live; whether we have adequate income in retirement; what we eat; whether we become ensnared by debt; how well insured we are in the face of calamity; whether we are shocked by an unexpectedly high bill; and so on.

Consumer choice is also central to various policy outcomes of interest. Policymakers often aim to influence consumer decisions and hence to promote specific policy goals through information, incentives and social marketing. Perhaps the most straightforward such policy goal is to help consumers to avoid being overcharged or sold substandard or dangerous goods. Much consumer regulation aims to ensure that minimum standards are met, product descriptions are accurate and that contract terms are upheld. Policy interventions are common also in situations where consumer choices have potential implications not only for the individual concerned but also for wider society. This occurs, for instance, where products are associated with environmental damage, health outcomes or financial risk.

Even where there is a clear and agreed goal for a consumer policy, however, an intervention will only be effective if it influences consumer choices in the desired direction. Thus, to achieve the policy goal, we need an understanding of how consumers decide. Substantial progress has been made in recent years in our understanding of consumer choice, in large part through advances in 'behavioural economics'.

1.2 BEHAVIOURAL ECONOMICS

Traditionally, economists have tended to make and to support strong assumptions about consumer choice. Standard (neoclassical) economics assumes that, at least to a good first approximation, consumers act rationally in pursuit of their own self-interest. The implication is that consumers can, in effect, process an unlimited volume of information and integrate all of the relevant information

accurately when making decisions, provided the decision is sufficiently important to their material interests. These assumptions were not tested empirically until recent times, however, when behavioural economists began to use experimental methods to examine how consumers really make decisions. The findings have highlighted numerous circumstances where consumers deviate substantially from the predictions of standard economic theory. Almost coincidentally with this rise to prominence of behavioural economics, the global financial crisis effectively raised a red flag with respect to the quality of consumer decision-making in financial services markets, not least in Ireland (Nyberg, 2010).

Some specific cases are highlighted in the following sections, together with a brief discussion of their implications for policy. For present purposes, there are two important general points to note. First, consumers' decisions appear to be strongly influenced not only by the quality and prices of goods on offer, but by the context in which the decision is made. Thus, it has become important to study and to understand how and why decision-making varies across contexts. Second, both field and laboratory studies suggest that consumers may have particular difficulties when faced with more complex products. This term, 'complex', is rarely defined, but is generally used to mean circumstances where consumers cannot cope with the volume of important product information or its technical nature. In a number of specific markets, including financial products and contracts for services, evidence suggests that consumers struggle to identify good products from similar but inferior ones.

Central to the progress made by behavioural economics is the use of an alternative scientific method (Shiller, 2005; Lunn, 2012). Behavioural economics is usually defined as the incorporation of psychological insights into economic analysis. Yet it is not only insights that behavioural economists have taken; they have also borrowed methods from experimental psychology. Rather than assume how economic agents behave, behavioural economists observe economic behaviour in scientifically controlled environments. Laboratory experiments, survey experiments and field experiments are used to measure and record how consumer decisions are made.

The present report summarises a series of experimental studies that proceeded in exactly this fashion. As the next chapter explains, the studies adapted methods used for a number of decades in experimental psychology and cognitive neuroscience to study perception and cognition, but which had not previously been used to investigate the capabilities of consumers. The methods allowed experimental investigation, from first principles, of how consumers cope when faced with complex multi-attribute products.

The present report offers a summary for general readers of the findings from ten experiments undertaken in PRICE Lab between 2013 and 2015, together with a discussion of policy implications. Although there are a number of fairly technical scientific elements involved in the experimental designs, the summary descriptions in this report aim to be non-technical, in the sense that the findings should be comprehensible to a general reader who may have no formal training in economics or psychology, but is interested in and familiar with issues surrounding consumer protection and competition policy. Readers interested in a comprehensive scientific description of the experimental designs and statistical analyses are directed to the series of scientific papers listed in the appendix.

1.3 INTERNATIONAL RESEARCH ON BIASES IN CONSUMER DECISION-MAKING

Very many studies in behavioural economics document 'biases' in economic decision-making. The findings typically show that a substantial number of consumers will alter their choices, or even their willingness to make a choice, depending on the context in which the choice is presented. The scientific literature that documents these influences on consumer decisions is now so extensive that space does not permit a thorough review here. Excellent descriptions for general readers are available, in particular, in Kahneman (2011) and Thaler (2015). An comprehensive review by DellaVigna (2009) shows that many biases first identified in laboratory studies are also to be found in field studies conducted in real markets. Rather than rehearse the results and arguments contained in these reviews, which are in any case too numerous for present purposes, this section concentrates on a specific subset of studies that raise issues from a consumer protection perspective. As such, they provide useful background for the material that follows.

At the risk of stating the obvious, from the consumer perspective what matters is that the consumer gets good deals and avoids bad ones. A good deal is one in which the consumer makes a surplus, i.e. the item they purchase is worth more to them than the money they pay for it. Ideally, consumers manage to locate the offering that provides the highest surplus available to them in the market. In the event that consumers are able to choose the best deals accurately, there is no problem to be studied and little concern from a policy perspective.

The accuracy of consumer choice is the primary focus of PRICE Lab. Generally speaking, how accurately consumers locate surpluses has proved difficult to study scientifically, because it is far from obvious how to determine when a decision is a good one or a bad one. Different consumers, obviously, have different preferences. We can observe and record choices; we cannot observe people's preferences. Consequently, we cannot generally measure surpluses or determine

how good a choice is. There is usually no way, therefore, to assess the quality of an individual consumer decision. It is for this reason that the overwhelming majority of studies of biases in consumer decision-making, described in the reviews cited above, do not actually identify poor decisions, but inconsistent ones. Where consumers display strong and systematic inconsistencies in their decisions, choosing product A over product B in one context and vice-versa in another, and where the difference in context is arbitrary or irrelevant to quality and price, it is fairly safe to infer that at least one of the two choices resulted in a loss of surplus, at least relative to the other one. This kind of study is sufficient to identify that decision-making is biased by the context, but usually not sufficient to provide an estimate of the seriousness of the problem, i.e. how much surplus is being lost. That is, it is hard to gauge the level of harm to the consumer.

However there are now a number of investigations that have found a way to estimate the extent of consumer detriment in certain specific markets. The trick is to find products where the good being purchased is effectively the same across different providers, or where the same good is offered in different ways by the same provider. When consumers are presented with multiple versions of essentially the same product and yet opt for the more expensive one, provided we control for things like brand preferences, we can infer that they are missing out on surpluses.

For instance, using data provided by a German internet provider, Lambrecht and Skiera (2006) studied the tariff choices and usage records over five months of approximately 11,000 customers, who chose among just three types of tariff: a flat-rate and two three-part tariffs[1] with different download allowances. They found that the majority of consumers on the three-part tariff with the higher fee would have fared better on one of the other two tariffs. Meanwhile, one-in-five customers on the flat rate would have paid less on another tariff. A smaller proportion on both three-part tariffs would have been better off on a higher fee or flat rate. Overall, the effects were very large. Of the consumers paying too high a flat-rate for their usage, more than half were paying more than double what they could have. In other words, in a choice between just three offerings at the same company, consumers were missing out on substantial surpluses. Similar evidence that substantial proportions of consumers fail to select the lowest cost tariff from among a small range has been recorded in the US mobile phone market (Grubb, 2009; Bar-Gill and Stone, 2009).

[1] A three-part tariff involves a charge for the service, then a fixed fee in return for an allowance (or suite of allowances such as calls, texts and megabytes) of units of the product, then an additional (often much higher) price for any units consumed beyond the allowance.

In financial services, indexed mutual funds offer an opportunity to study how accurately consumers identify surpluses, because the return on the amount invested is tied to the performance of the same financial index regardless of which company offers the product. However, laboratory and field studies show that in choosing which provider to go with, consumers are inclined to underweight the impact of fees and to overweight descriptions of past fund performance, which can be manipulated by simply selecting a beneficial time period (Barber et al., 2005; Choi et al., 2010).

The above examples involve telecommunications and financial products, which have some specific technological and numeric complexities. Domestic electricity markets are less complex. Yet several studies have documented that large numbers of consumers fail to switch to lower cost suppliers (e.g., Giulietti et al., 2005; Brennan, 2007). While brand preferences clearly play a part, Wilson and Waddams Price (2010) found that the majority of a sample of British consumers who switched *specifically to make savings* failed to select the best available deal, while a substantial minority actually increased their bills.

Overall, this evidence suggests that there are at least some markets in which consumers fail to locate the best deals and sometimes opt for genuinely bad ones. That is, consumers regularly miss out on surplus. The specific markets studied have their own idiosyncrasies, but in each case the consumer is required to weigh up the merits of multiple features of the product in order to determine which offering constitutes the best deal. More technically, they must integrate multiple sources of information, including the price, in order to determine how much surplus each product offers. Somewhere in this process, they seem to make mistakes.

One possibility is that consumers suffer from so-called 'inattention' (Choi et al., 2010; Wilson and Waddams Price, 2010; Grubb, 2015), whereby they pay insufficient attention to a product attribute that is important for the overall surplus they ultimately acquire. Wilson and Waddams Price go a little further to suggest that 'many of the choices are consistent with genuine decision error or inattention' (p.665), although they do not define what they mean by 'decision error'. The implication is that the product is too complex for consumers to integrate the necessary information in order to gauge the surplus accurately.

Overall, these studies persuasively show that there are markets in which the challenge is beyond consumers' cognitive capabilities, but they do not provide a sufficiently rich measure of capability that it is possible to generalise about consumers' ability to identify surpluses across the many potential markets,

pricing and marketing practices. In short, the existing international research identifies a problem, but says far less about the scale of the problem or its cause.

1.4 POLICY RESPONSES TO PROBLEMS OF COMPLEX PRODUCTS

The findings of behavioural economics, including those described in the previous section, have spurred policymakers in several markets and countries to consider, and in many cases to implement, regulations to simplify the choices that consumers face. A number of such interventions are discussed and reviewed by Sunstein (2011) and Lunn (2014). The logic is generally straightforward. If the complexity of offerings means that consumers are failing to take account of important aspects of the product, then simplifying the format or volume of information they have to process should make it easier for them to take that key information into account.

There are several possibilities for simplifying the consumers' challenge. One common and simple intervention is for the authorities to promote, endorse or provide price comparison websites. The aim is to support only sites that meet regulatory standards in terms of impartial content, presentation and transparency. This approach is taken in regulated markets such as energy, telecommunications and financial services in many countries, including Ireland. It is extremely difficult to evaluate the consumer benefits of such websites, beyond establishing the volume of traffic that passes through them. From a theoretical point of view, however, it is likely that genuinely independent price comparison sites help to increase consumer surplus. Where the evaluation of products requires consideration of multiple price components or attributes, or where consumers must combine personal usage information with tariff information, an accurate independent price comparison site can point consumers in the right direction or provide a useful check on a potential decision. There may also be benefits arising from the effect such sites have on the incentives facing firms. On the other hand, many consumers are not willing to trust the impartiality or accuracy of price comparison sites, or may not realise the presence or importance of regulatory endorsement. Furthermore, the sites themselves require varying degrees of sophistication to utilise, depending on the nature of the product. Most sites require the consumer to input at least some initial information, for instance about the anticipated level of usage of a service, the subcategory of product, the amount of a loan sought and so on. Consumers may not feel able to provide this information, or may feel that they are being asked to narrow down their options before truly understanding the implications. Nevertheless, where an independent and accurate price comparison is provided, even if imperfectly used, it is likely to insulate those consumers who exploit it against more disadvantageous choices.

A price comparison site is a type of 'choice engine' – a system, usually computerised, to assist in making choices. Other potential choice engines may prove to be of benefit to consumers facing complex products, with opportunities for regulators to promote or endorse their design, promotion and use. These include applications that offer to download service usage histories, then to conduct a price comparison specific to the consumer's personal usage pattern. These kind of choice engines are feasible only where, firstly, it is possible to access data on historical usage patterns held by providers and, secondly, a third party has invested the time and money to develop the necessary software to gather the data and drive the choice engine. In some markets these conditions are met, but in others not. 'Mydata' (or 'midata' in the UK) initiatives are an extension of this idea with greater regulatory input. The aim is to seek industry-wide voluntary agreement or alternatively to mandate the provision of machine readable personal usage data, in order to encourage the development and use of choice engines.

Choice engines may well be of overall benefit to consumers, but they are constrained in what they can achieve. Firstly, the engine will always be a 'black box' for consumers, who must trust the engine to select the best option. Secondly, because using any piece of software requires a degree of knowledge and computer literacy, it is possible that the consumers most likely to adopt and hence to benefit from choice engines are those already most likely to identify the best deals. Lastly, use of a choice engine implies a situation where the consumer has decided to take some time to survey options and to make a choice from among them. That is, the consumer has initiated the activity. Often, however, consumers must make decisions in contexts where providers have initiated the communication. They must decide whether to engage with a door-to-door salesperson apparently able to deliver a substantial saving, whether to respond to a call or email from their existing provider suggesting improved terms or offering an attractive add-on feature, or whether to follow up on an advert displaying a seemingly better package than the one they are on. Of course, the consumer may ignore all these communications, but it is not realistic to consult a choice engine every time the consumer makes this sort of decision. An initial judgement of potential value is required then and there.

An alternative and popular intervention internationally is mandated disclosure aimed at simplification (for multiple examples see Sunstein, 2011, or work on new disclosures at www.consumerfinance.gov). For instance, one long-standing mandated simplification now widely adopted in the developed world is the standardised 'Annual Percentage Rate' (APR) on credit products, which regulations stipulate must appear on certain types of disclosures and marketing material. Mandated simplification has been widely adopted in the US. The Office of Information and Regulatory Affairs (OIRA) at the White House has sought via

executive orders to make regulatory authorities distinguish between 'summary disclosure' and 'full disclosure'. The former is a mandated disclosure that simplifies and standardises product information, which must be offered at the point of sale to assist product comparison. The latter is full product information that needs to be available somewhere, such as on the company website, but is unlikely to be essential to the value sought by most consumers. There is a clear link here to the concept of inattention: mandated summary disclosure is designed to make consumers more likely to pay attention to the most important product attributes. It is possible for regulators to pre-test different mandated disclosures, using laboratory experiments and other methods, to try to establish whether they improve consumer understanding of the product in question. This empirically informed approach has been used extensively by America's Consumer Financial Protection Bureau (CFPB).

There are also more strident possibilities for regulatory responses to consumers' difficulties with product complexity, often specific to the regulated sector. Where legislation allows, regulators can adopt and enforce rules not only with respect to the nature of product information, but also with respect to features of the product or product range. In America, the 2009 Credit Card Accountability Responsibility and Disclosure (CARD) Act simply banned certain types of fees on credit cards, on the grounds that consumers were unlikely to notice them or to assess the likelihood of incurring them accurately when choosing between providers and cards. The UK's energy regulator, OFGEM, has regulated to limit the number of tariffs that providers of gas and electricity can offer, permitting a maximum of four tariffs for each type of meter and payment method. The idea is that once the consumer has chosen the type of service they require, the choice of tariff is simplified.

The logic of mandated simplification is clearly reasonable. Simplification is unlikely to affect those consumers who are already making good choices, but may help those consumers who are struggling to compare complex products to make better choices. However, mandated simplification alone does not guarantee that they will, in fact, make better choices. As of now, there are relatively few empirical evaluations of the success or otherwise of mandated simplification regulations. Lunn (2014) provides a review and discussion of evaluations, concluding that while some regulations appear to have measurably benefitted consumers, others have imposed costs on providers for little or no apparent gain. The CARD Act perhaps stands out as a success, with initial estimates suggesting a consumer gain of US$12 billion per year (Agarwal et al., 2015). Ericson and Starc (2013) provide evidence for successful mandated simplification in the health insurance market also.

One of the difficulties in determining the case for more interventionist regulation of complex products is the width of the gap between evidence and policy. The evidence is sufficient in many cases to show that consumers are failing to identify the best deals, but it is often only suggestive as to exactly why; the psychological mechanisms are generally only sketchily understood. Consequently, evidence of consumer detriment in one context does not necessarily help us to determine where else it might be present. Moreover, it is not clear at what point a product becomes, as it were, too 'complex'. Thus, it may sometimes be the case that a regulation will make the consumer's choice less complex, yet still leave the decision sufficiently difficult that no discernible improvement in outcomes can be established.

1.5 THE LOGIC OF PRICE LAB

Given the international research and policy context described in the previous two sections, PRICE lab set out to develop an alternative empirical approach to the problem of complex products. At its core, the consumer's difficulty is one of accurately perceiving the presence and size of surpluses; it is simply hard in some markets to judge how good a deal is. The policymaker's challenge is to understand the nature of the difficulty well enough to be able to devise interventions that are of sufficient benefit to outweigh any costs they impose. That is, what exactly is it that makes it hard to perceive surpluses accurately? Since the problem is essentially one of perception, we turned to perceptual science (i.e. the scientific study of perceptual systems) for guidance as to how the problem might be tackled.

A key principle of perceptual science is to design laboratory environments that gain complete experimental control over the perceptual inputs, or 'stimuli', that the observer must respond to. Once this has been achieved, the environment can be systematically manipulated and the impact on the observer's responses can be recorded. Applying this logic to consumers' abilities to identify surpluses, this means devising laboratory tasks in which experimental participants must repeatedly assess surpluses while the experimenter manipulates the attributes of the products. This logic underpins the design of the Surplus Identification (S-ID) task, which is described in the following chapter.

The S-ID task is a departure from previous empirical work in this area. It allows researchers to measure the ability to identify surpluses from first principles, beginning with simple products consisting of easily discernible attributes and prices, then increasing the level of complexity systematically. Each time the nature of the attributes is changed, the task generates a quantitative measure of the consumer's capability, allowing the impact of different types of product features to be isolated and assessed. In this way, the technique aims to produce

findings that are not specific to a particular market, but are instead likely to generalise across markets. As the remainder of this report will show, the S-ID task provides general principles about consumer capability across markets and gives some insights into the psychological mechanisms behind variation in consumer capability. Consequently, the findings offer useful guidance for policymakers as to what constitutes a 'complex' product and when consumers are likely to struggle to identify good deals.

1.6 RESEARCH QUESTIONS

The techniques developed in PRICE Lab enable us to address the question of product complexity from first principles. The complexity of a product can be thought of as depending on the number of its attributes, the type of attributes and the nature of the relationship between the attributes and the overall value of the product. Arguably, the most simple product is one which has just a single plainly perceptible attribute, which maps in a linear fashion on to the product's value, i.e. so that unit increases in the single attribute's magnitude translate into equal sized increases in its value. From this starting point, complexity can be gradually increased: by adding additional attributes that also contribute to the product's value, by making the magnitude of the attribute harder to discern, by increasing the number of products in the range, by introducing a non-linear relationship between one or more attributes and the product's value, by allowing the attributes to interact with each other, and so on. Because the S-ID task permits complete experimental control over attributes and prices, each of these potential sources of complexity can be manipulated and tested in isolation, to see what impact it has on consumers' ability to identify surpluses.

The general research question addressed in this report is: How complex does a product have to become before consumers find it hard to identify surpluses? More specifically, we address the following questions:

- How accurately can consumers identify surpluses when asked to judge a simple single attribute product against a price?
- How is accuracy affected when they have to trade off attributes against each other?
- Does accuracy improve or decline when there are more products in the range?
- How many attributes can consumers simultaneously cope with before mistakes become large?
- Is the accuracy of surplus identification affected by whether attributes are correlated with one another?
- Is accuracy affected by making the value of attributes non-linear?

- Is there variation in how accurately consumers can identify surpluses according to the type of attribute, i.e. whether it is visual, described by numbers or put into categories?
- Does accuracy improve when consumers deal with more familiar products?

The experiments presented in the chapters that follow were designed to provide answers to each of these research questions. While some individual findings may be open to interpretation, across the body of the report a fairly clear pattern emerges. The experiments show that consumers are limited in their capabilities. Products do not have to be particularly complex before consumers struggle to spot surpluses accurately. Nevertheless, there are some forms of complexity that have greater negative effects on consumers' decisions than others. Thus, the findings provide a helpful indication of what type of products consumers are likely to find generally more difficult to evaluate.

Following the body of the report, which describes the experimental results in sequence, the final chapter considers the policy implications and directions for future research. The findings provide some principles with respect to consumer capability that can guide policymakers seeking to understand where, when and why consumers are likely to struggle to locate good deals. The methods developed here also offer opportunities to further explore consumers' capability, both across markets and within specific markets, of interest from a consumer policy perspective.

Section 2

The Surplus Identification (S-ID) Task

2.1 MAKING THE SURPLUS OBJECTIVE

Consumer surplus is defined as the benefit the consumer obtains from a purchase over and above the price paid for it. The central problem facing any investigation of how accurately consumers can identify surpluses is that the benefit obtained by any one consumer is subjective and, hence, unobservable to the investigator. In short, different people have different preferences. It is for this reason that the studies cited in the previous chapter focused on situations where consumers appeared to be purchasing effectively the same product at different prices. In such situations, the failure of consumers to perceive surpluses accurately is inferred, given the assumption that consumers do not wish to pay more than necessary for the same product.

The Surplus Identification (S-ID) task takes an alternative approach. The method gains experimental control over consumers' preferences by incentivising them to adopt preferences that are predetermined by the experimenter. The incentive offered is what experimental economists call a 'tournament' incentive, where the participants stand to win prizes if they are among the best performers in the experiment. On a computer screen, participants are shown a product consisting of one or more attributes and a displayed price. The magnitudes of the attributes and the price are controlled by the experimenter. How much the product is worth is a function of the attributes and can be expressed as a monetary value. Participants have to learn, through examples and feedback, how the value of the product relates to the magnitudes of the attributes, in order to then compare it with the displayed price. They then make a series of decisions based on the perceived size of surpluses. Generally, and in most of the experiments reported here, participants are shown a single product with a price displayed on a price tag and they have simply to decide whether the product is worth more or less than the price displayed, i.e. to judge whether there is or is not a surplus. In other cases, participants are shown two or more products, with or without price tags, and they have to decide which option is the best value. The S-ID task always requires the participant to process the available information about attributes and prices, to integrate this information into an assessment of the size of the surplus(es), to decide which of the options is worth more and then to indicate their response by pressing a button on a response box.

Because the attributes and prices are under complete experimental control, the experimenter can vary the difficulty of the task. For instance, in the most simple form of the S-ID task, just one product and one price are presented on a

computer screen. The participant presses a button to indicate whether they think the product is worth more or less than the displayed price, i.e. they have to decide whether the surplus is positive or negative. On some trials the task is fairly easy, because the product is worth much more (or much less) than the price. When the difference is large, the participant presses the correct button almost every time. On other trials the task can be made much harder, with a smaller surplus and, hence, a higher chance that the participant presses the incorrect button. By varying the difficulty of the task in this way, the S-ID task can be used to measure what happens to the accuracy of participants' responses as the size of the surplus varies. Throughout the experiment, the participant's clear incentive is to be as accurate as possible.

Accuracy is not a unitary concept, however. An archer may miss the bull's-eye nine times out of ten because the arrows are sprayed randomly in a circle around the centre of the target. Alternatively, the archer may miss the bull's-eye nine times out of ten because most of the arrows hit too low on the target. In the former case, the accuracy of the archer is damaged by 'imprecision'. Arrows would hit the bull's-eye on average, but they instead end up scattered randomly around it. A measure of the archer's precision would be how far an arrow is, on average, from the centre of the target. In the latter case, where most arrows strike below the bull's-eye, accuracy is reduced by 'bias'. A measure of the archer's downward bias would be obtained by finding the average height of all the arrows fired and recording how far it is below the bull's-eye. Imprecision and bias are thus two different kinds of inaccuracy. The archer might fire all the arrows in a tight circle below the bull's-eye. This would be precise but inaccurate, because of bias. A biased archer may still score higher than an archer who is unbiased but who is less accurate because of imprecision.

Throughout this report, the distinction between precision and bias is important. Consumers may make errors because they cannot distinguish between two options that appear to be of similar value, when in fact one is worth more than the other – they are imprecise. Or a consumer may make errors because they systematically overvalue or undervalue a certain type of product – they are biased. The S-ID task can measure both types of inaccuracy and distinguish between them.

This method of assessing accuracy is adapted directly from experimental studies of perception conducted by psychologists and neuroscientists. The environment is under perfect experimental control, so that all information available to the participant is determined by the experimenter and can be varied from trial to trial. The participant's task on each trial is reduced to determining which of (usually) two alternatives is the correct answer – a 'two-alternative forced choice'

(2AFC) task. They receive feedback after each trial to help them learn. Over multiple trials, the probability of correctly identifying the surplus can be measured precisely. How much this probability changes when different kinds of 'complexity' are introduced can also be measured, while learning can be tracked over a series of trials.

2.2 HYPERPRODUCTS

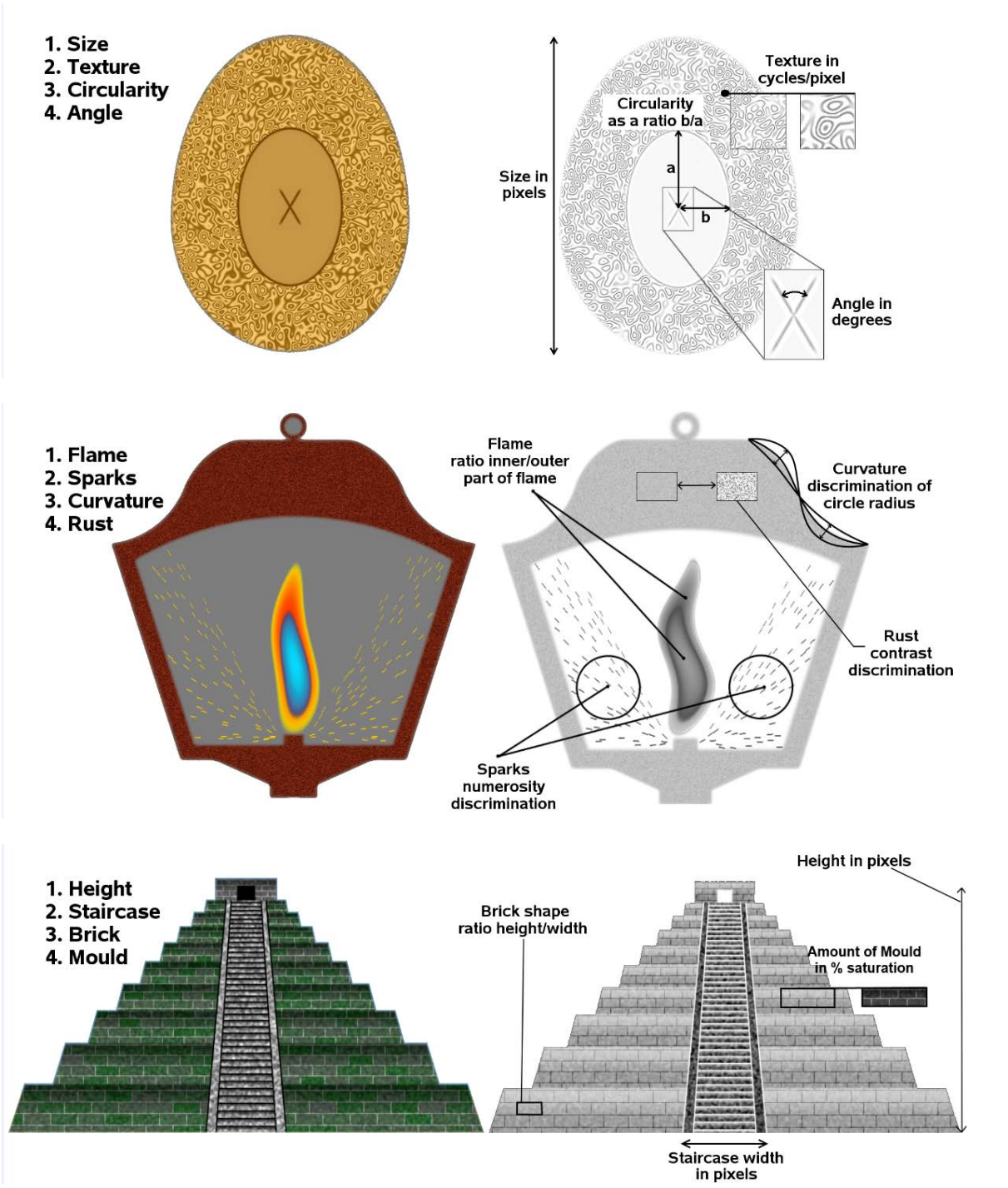
One potential difficulty when trying to impose an objective set of preferences on consumers is that it may require them to override their subjective preferences. That is, the experiment tells them what each product is truly worth, but this may jar with their own idea of what they do and do not like. If judgements were subject to this sort of interference, the result might be to underestimate consumers' abilities.

In order to overcome this problem and to ensure experimental control, most of the experiments described in this report employed a set of hypothetical products which appeared on a computer screen. The idea was that the products would be new to the participant and, consequently, that the participant would be unlikely to have strong initial preferences regarding what made them more or less valuable, making it easier to learn to value the product according to the function determined by the experimenter. Each product had a small number of attributes that consisted of visual features under complete experimental control. The attributes were chosen on the basis of previous studies of visual perception that have shown how the relative magnitudes of these visual features can be discriminated with high accuracy. The aim was to make sure that any inaccuracy in consumers' perceptions of surpluses was due to how well they could integrate the available information, when comparing the attributes with the displayed price, rather than how accurately they could perceive the attributes in the first place.

Figure 1 shows examples of the three computerised products used in the experiments in this report: a golden egg, a Mayan pyramid and a Victorian lantern. These items were chosen for three reasons: (1) they each have intuitive value; (2) it would be highly unlikely that any of our participants had ever had cause to value or to trade one; (3) they were objects for which it was easy to devise and to precisely define many attributes. Each product has up to four possible visual attributes that can be manipulated. In addition, it is possible to assign numeric and categorical attributes. For instance, in the first experiment to be reported here, we used golden eggs that varied in size and surface texture. The larger the egg's size, the more it was worth; the finer the surface texture, the more it was worth. In later experiments we employed a categorical attribute: the eggs could be gold, silver or bronze, with obvious implications for their relative

value. Similarly, we introduced a numeric attribute in the form of a percentage purity, which was written on a stand below the egg. Technically speaking, each presentation of one of these computerised products is uniquely defined in a multidimensional attribute hyperspace. We therefore refer to them as 'hyperproducts'.

FIGURE 1 The Three Hyperproducts Used in the Experiments. The Value of Each Product is a Function of up to Four Precisely Defined Visual Attributes.



For each of the experiments described in the chapters that follow, we outline which attributes mattered and how they related to the overall value of the hyperproduct. For now, the key to understanding how the S-ID task worked is to view it from the perspective of participants. They would arrive in the laboratory and, after reading and signing appropriate consent forms, they would be introduced to one of these new products. We would show them a series of helpful examples of the product, together with prices showing what they were worth. In experiments where more than one attribute had to be taken into account, we would hold each attribute constant and vary the other, to show participants how the individual attributes affected the overall value of the product. This procedure helped the participant to learn the relationship between the attribute magnitudes and the value of the product before the test proper began. After viewing the examples and undertaking some practice trials, participants would then undertake the main task, in which one or more products was shown together with a displayed price and the participant had to decide which was more valuable. Where only one product was on screen in each trial, they would decide whether there was a surplus, or equivalently whether product was worth more or less than the displayed price. Where more than one product was on screen, they would decide which one was the best value at the price shown. After making their response by pressing one of two buttons on a response box, they would be shown feedback in the form of the correct monetary value for the product(s) they had just tried to judge. This provided continual opportunities to learn the relationships between product attributes and prices. Further details are provided in the methods sections accompanying each experiment.

Typically, each experimental participant completed a series of trials arranged into 4-6 experimental runs of between 50-70 trials. They would proceed through these at their own pace – there was no time limit placed on their responses and they could spend as long as they liked observing the feedback. At the end of each run of trials there would be a short break while the experimenter described what was coming next. Approximately half way through the session there would be a longer break during which the participant would be taken to a canteen elsewhere in the building for refreshments. In total, a typical session lasted around an hour, including this break.

2.3 STATISTICAL ANALYSES

The main statistical measures that we employ also make use of techniques adapted from perceptual science. It is common to measure how accurately people can see, hear, feel and so on, using the concept of a 'just noticeable difference' (JND). This is the amount of a signal that a perceptual system needs in order to detect that signal reliably. In our case, a just noticeable difference is the amount of surplus required in order for the surplus to be identified reliably. Referring back to the earlier example of the archer, the JND is equivalent to the

size of the circle one would need to draw for almost all the arrows to be inside the circle. The more precise the archer, the smaller the diameter of the circle; the more precise the consumer, the smaller the amount of surplus that can be detected. The bias is measured by the location of the 'point of subjective equality' (PSE), which is the size of signal at which the participant judges two things to be the same. Adapted to the most simple S-ID task in which the participant has to decide whether the product is worth more or less than the price, this means the point at which the participant perceives there to be no surplus – they decide that the product is worth exactly the price displayed. If a participant is unbiased, then the PSE will lie at exactly zero surplus – the surplus is perceived to be zero when it is in fact zero. If the value of the product is overestimated, the PSE will lie below zero surplus, while if it is underestimated, the PSE will lie above it, just as a biased archer might consistently shoot too high or too low.

FIGURE 2 Example S-ID task data. When there is a large positive surplus, the participant almost always responds that the product is worth more than the price, and vice-versa when there is a large negative surplus. The slope of the curved line fitted to the data is a measure of the precision of surplus identification, while the location of the midpoint of the curve relative to zero surplus is a measure of the bias.

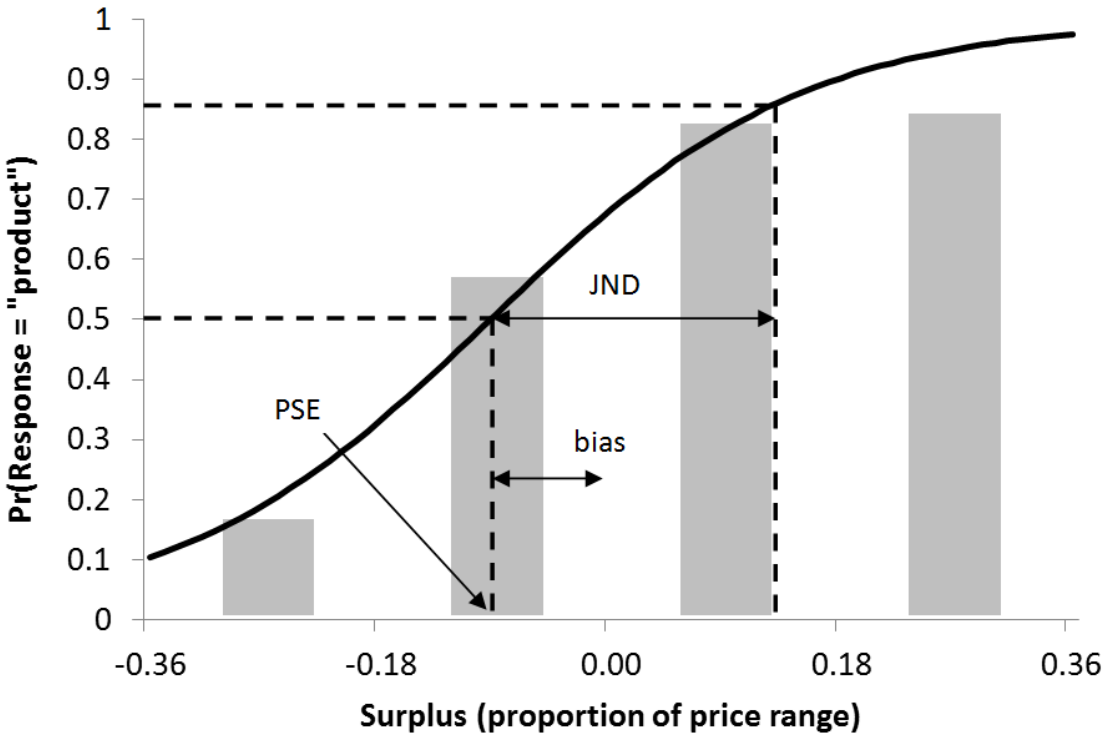


Figure 2 shows example data generated by the S-ID task, designed to illustrate these two key concepts for the analysis to follow. In this experiment the participant had to decide repeatedly whether a golden egg was worth more or less than a price displayed next to it on a tag. The eggs varied in price between

€177 and €423. On the horizontal axis is the true size of the surplus, which is measured as a proportion of the total price range (€246).[2] Thus, the surplus varied from large and positive, when the egg was worth 0.36 of the price range more than the displayed price, to large and negative, where it was worth 0.36 of the price range less than the displayed price. On the vertical axis is the probability that the participant judged the egg to be worth more than the displayed price. When the egg was worth much more than the price, this probability was close to one, indicating the participant almost always perceived the surplus. When the egg was worth much less than the price, the probability was close to zero, indicating that the participant clearly realised there was no surplus.

The JND and bias for this participant in this task are marked in the figure. This participant had a clear bias towards overestimating the value of the product relative to the displayed price. When the egg was actually worth almost one tenth of the price range less than the displayed price (the location of the PSE), the participant had a probability of 0.5 of perceiving the egg to be worth more than the price. This means that the participant perceived zero surplus when in fact there was a negative surplus equivalent to one tenth of the price range. The curve fitted to the data is the best fitting logistic curve.[3] From the slope of this curve, we measure the size of the JND, which is our measure of the precision of surplus identification. The JND is defined as the increase in surplus required to raise the probability of perceiving a surplus from 0.5 to 0.86. In other words, the JND tells us how much surplus is needed for the participant to identify it with a reliability of 86 per cent. This figure of 86 per cent reliability is used as our measure of precision because it is a standard measure used in the study of perception (which in fact corresponds to one standard deviation of a logistic distribution). Returning to Figure 2, this participant required a difference between the value of the product and the displayed price to be just over 0.22 of the price range – a surplus of €54 – in order to identify when there was a surplus with a reliability of 86 per cent.

Throughout this report, we make repeated use of the concepts of the JND and the bias as our main measures of consumers' accuracy of surplus identification. Although at various points we make comparisons across individuals, the analysis primarily concentrates on the average JNDs and biases for the sample of individuals that undertook each experiment, after checking and controlling for

[2] We measure the surplus in this way because it allows comparison across experiments with different products and price ranges.

[3] The logistic is a common function fitted to forced-choice data in both studies of perception and of economic choice. Our results are not sensitive to the precise choice of function or to the estimation procedure employed to determine the best fitting function.

any outliers.[4] We examine how the JND and bias are affected by manipulating the different properties of products listed in the research questions in Section 1.6 above. This allows us to measure how the accuracy of consumers' identification of surpluses varies with the number of product attributes, type and linearity of attributes, correlation between attributes, range of products, familiarity with products and attributes, and so on. In this way, the S-ID task gives us insights into why participants make errors in their judgments of surpluses, whether because of bias or because of a lack of sensitivity to differences between products. These methods inform us, therefore, not just of *when* but *why* consumers fail to spot surpluses. If they are biased, the method can isolate what aspect of the product or product range leads them to overvalue or undervalue the product. If they are imprecise, the method can home in on what aspects of products and prices lead consumers to lack sufficient precision to identify available surpluses reliably.

2.4 GENERALISABILITY

Before embarking on a description of the first results obtained when we undertook the S-ID task with samples of Irish consumers, it is appropriate to consider the likely generalisability of the findings. In other words, to what extent can we extrapolate from these laboratory investigations to consumer behaviour in the real world?

One important point to note in this regard is that unlike many (indeed, probably the large majority of) laboratory studies on consumer decision-making, PRICE Lab does not use samples of students. Instead, it employs samples of consumers. For each of the experiments described in this report, the samples of participants were recruited from the Dublin area by Amárach Research. Each sample was approximately balanced by gender, age (18-70 years) and working status (working, not working). The use of samples of consumers rather than students is of obvious importance for generalising the findings.

As described above, the experiments employed products that are new to participants in order to gain experimental control and, especially, to avoid interference from existing subjective preferences. Yet it might be argued that because the hyperproducts were new to consumers, the findings obtained apply only to unfamiliar products and underestimate accuracy when choosing among more familiar ones. Note, however, that consumers frequently do encounter new products, markets in which product attributes have changed, and markets in

[4] These average estimates are computed from mixed effects logistic regression models that are described in detail in the technical papers from which the current report is drawn. These models allow for individual differences in precision and bias, providing the underlying 'fixed effect', or average for the experimental sample. Readers interested in the details of this technique and the accompanying robustness checks should consult the first paper listed in the appendix, which is available online at www.esri.ie.

which they make a purchase only once in a number of years – far fewer judgments than would be made in one of our experiments. Moreover, in our experiments, feedback was immediate, repeated and exact, whereas feedback on the value of purchased products in the real world may be delayed, infrequent and noisy, making it harder to learn the relationships between attributes and prices. More importantly, this is a concern that we were able to address through the experiments themselves. Experiment J makes the direct comparison between the accuracy of surplus identification with hyperproducts and with two more familiar products. It records almost identical performance in an S-ID task that is matched between the two product types.

Similarly, since many of the experiments reported here reduce the consumer's choice to just two alternatives, it could be argued that this context makes product comparison more difficult. In theory, a broader range of alternatives might help consumers to calibrate the key relationships between attributes and prices better. Again, Experiment I tests explicitly whether increasing the size of the product range improves the ability to trade off attributes and finds that the larger product range in fact makes identifying a surplus more difficult.

A more substantial concern surrounds the fact that the S-ID task imposes preferences upon consumers. That is, the experimenters and not the consumers decide what makes a good product. It is not certain that the psychological mechanisms engaged by the S-ID task are also those employed by consumers when they decide what they like, according to their own subjective attribute weightings. It is logically possible, therefore, that consumers can integrate product information more accurately in the real world, when deciding what *they* prefer, than they can when deciding whether a hyperproduct is worth more or less than a displayed price according to a formula that *we* have devised.

While accepting that further investigation of the relative accuracy of information integration in contexts of objective judgement and subjective choice is needed, there are at least three reasons why we argue that the findings presented here do indeed reflect consumer capability in situations of subjective choice in real markets. First, even where subjective preferences are involved, the information integration required is partially objective. For instance, although people differ in usage patterns and tastes for risk, there are objective elements to combining the attributes of a credit card to determine the overall cost of credit, or the attributes of a mobile phone contract to determine the likely cost for a given pattern of usage. Second, because the S-ID task uses products that are new to participants, it simulates the process by which an individual initially learns what they like. While the value of the hyperproducts is determined by an objective formula, the participants learn by receiving positive or negative feedback on individual decisions, as they do when they learn what they like in the real world. Third, as

results described in the body of this report show, we find patterns of biases in our data that have previously been observed in studies of subjective choices between products. These findings imply that common psychological mechanisms are involved in the S-ID task and in subjective consumer choice.

An alternative objection might be that the time and effort taken by participants in the S-ID task is not representative of the time and effort consumers make when making decisions in the real world, with real consequences. Although participants were incentivised to perform well in the experiments and allowed to take as long as they wanted over responses, the experiment nevertheless required them to make many decisions in the course of a session. Most decisions were taken in a few seconds. Does this mean that our measures of accuracy underestimate capability? We think not. The data for all the experiments displayed a systematic pattern whereby participants took longer over the more difficult decisions. This suggests that they were putting in effort and allocating cognitive resources efficiently across the multiple decisions. Moreover, in Experiment C, we incentivised the participants to take longer and to put additional effort into one final experimental run. In response to this unexpected once-off incentive, they did indeed increase the amount of time spent on their decisions, but with no significant effect on performance. Note also that many consumer decisions are in fact taken very quickly. For instance, even if consumers take a long time to decide between perhaps two or three final options, they are likely to have edited down their decision to this shortlist via much more rapid decision-making processes, during which they may have already discarded options with higher surpluses.

There are some clearer limits to the generalisability of the results presented here, however. Our experimental designs ensured that the decisions of our participants were made by balancing the relative merits of attributes and prices *in their own heads*. There was no opportunity to use formal arithmetic, calculators or other kinds of decision aids. Nor were the decisions subject to conversations with or the opinions of other people. To the extent that these activities are undertaken and found to be beneficial to real-world consumer decisions, they are not accounted for here. On the other hand, our experiments took place in circumstances where the participants were paying full attention to the task at hand. They were not interrupted, not subject to persuasion or marketing and not required to search for the relevant product information, all of which was made easily available.

To conclude, given all of the above, the S-ID task provides a good measure of the limits of consumer capability in circumstances where consumers pay full attention to all relevant product information for at least several seconds, while facing a significant incentive to make a good decision. In our view, this implies that the measures of capability produced apply to a large proportion of real-

world consumer decisions. Furthermore, tightly-controlled lab experiments such as those conducted using the S-ID task have one key benefit from the perspective of generalisability. They are designed to target fundamental psychological processes which are likely to operate across contexts. The experiments described in this report are designed to home in on consumers' capabilities when they must simultaneously juggle different sources of product information. If consumers are only able to reach a certain level of performance in our laboratory tasks, which are designed to make it as easy as possible to execute a psychological process fundamental to identifying surpluses in markets, then it is unlikely that this process will be executed more precisely in real-world contexts. Transient social, subjective and environmental factors, such as getting advice from a friend, researching product features or paying limited attention to the decision at hand, are likely only to feed more or less information into the very psychological mechanism that our experiments isolate. In other words, the limits that we uncover are likely still to apply.

Part 2
Experimental Findings

Summary of Experiments

TABLE 1 Summary of Experiment Features, Research Questions and Findings

Exp.	Feature	Research Questions	Findings
A	Golden eggs with 1-2 visual attributes (e.g. size, texture)	• How accurately can consumers determine which of two products is more valuable: o When they differ on one attribute (e.g. size)? o When they differ on two attributes (e.g. size and texture)? • How does this change when consumers must compare a product to a price instead of to another product?	• Participants found it much harder to tell which of two products was worth most when they had to trade-off two attributes rather than compare just one • It was more difficult to tell whether a product was worth more than a price than to tell whether one product was worth more than another • In some cases, the difference between product and price needed to be **one-third** of the price range for participants to choose reliably • A strong **bias** emerged – products higher up the price range were overvalued while those lower down were undervalued
B	As Experiment A, increasing up to 4 attributes	• How many attributes can a consumer cope with when comparing a product to a price?	• Precision became much worse as more attributes were added o 3 attributes versus price required a difference of almost half of the price range o 4 attributes versus price needed a difference of almost two-thirds of the range • The bias across the price range was reduced as more attributes had to be taken into account, suggesting a **precision-bias trade-off**
C	As Experiment B	• Is comparing multiple attributes to prices easier for highly-educated, numerate participants? • Does performance improve with practice? • Can performance be enhanced if participants are given a financial incentive to improve?	• Highly-educated people performed **slightly better** than the general population • Performance **improved slightly** between the first and second sessions, but not between the second and third sessions • Performance was not improved by the incentive
D	As Experiment B	• If a product has multiple attributes that are correlated (i.e. all signal the same value) does it make it easier to detect if the product is more valuable than the price?	• Precision **improved** somewhat with more attributes when they all signalled the same about whether the product was good or bad • There was a **precision-bias trade-off**: when precision was better the bias across the price range was stronger
E	Eggs with visual, categorical and numeric attributes	• Can consumers weigh attributes against prices more accurately when the attributes are numeric or categorical instead of visual?	• When combining a visual attribute with a numeric or categorical attribute, or when combining a numeric with a categorical attribute, performance was **the same** as when combining two visual attributes

Contd.

TABLE 1 Summary of Experiment Features, Research Questions and Findings *Contd.*

Exp.	Feature	Research Questions	Findings
F	Mayan pyramids, visual, numeric and categorical attributes	• Do the findings apply to a different product? • How is accuracy affected when the same attribute is presented numerically and visually? • Is performance improved when a categorical attribute has just 2 levels? • Is performance improved if category level changes are associated with a % increase in value instead of a fixed € amount?	• Overall performance was **similar** to tasks using the egg • Accuracy was **unaffected** when the same attribute was presented visually and numerically • Performance judging a product with two attributes was **slightly better** when one of the attributes consisted of just two categories than when both were continuous visual attributes • Precision was **the same** regardless of whether the categories involved a fixed or % increase in value
G	Golden eggs, Mayan pyramids, Victorian lanterns, visual attributes	• How is accuracy affected by making the relationship between attributes and prices non-linear (i.e. diminishing returns)?	• Performance was not made worse by non-linear attribute-price relationships and was **slightly improved** for attributes in the case of moderate diminishing returns
H	As experiment G, with additional visual and numeric attributes	• Do consumers struggle more with two-attribute products when one attribute matters more for the value of the product than the other? • Is accuracy reduced by unusual pricing structures, such as increasing or non-monotonic returns?	• Unequal weighting of attributes **did not affect performance** • Whether returns were increasing or diminishing also made **no difference** • Precision was much worse when the price-attribute relationship was non-monotonic, with participants needing differences of **two-thirds or more** of the price range to make the correct choice
I	Golden eggs, Victorian lanterns	• Does performance change when consumers face a wider range of products and must find the one that is worth more than its price?	• As more products were added to the range (from 2 to 4) participants became **substantially less precise**, needing a greater difference in order to identify which product was worth more than its price
J	Mayan pyramids, Dublin houses, broadband packages, Victorian lanterns	• How does performance with new, hypothetical products (used in Experiments A-I) compare to performance with familiar, real-world products and pricing structures?	• Participants' accuracy when judging two real, familiar products (houses and broadband packages) and was **almost identical** to their accuracy when judging unfamiliar hypothetical products, in terms of both precision and bias

Section 3

How Accurately Can Consumers Resolve a Trade-off?

3.1 INTRODUCTION

To identify a surplus requires consumers to negotiate a trade-off. Even for a simple single-attribute product, there is a trade-off between price and quality to be resolved: a higher quality product is only a better buy than a lower quality one if the price difference is not too great. At its heart, a trade-off requires consumers to map one scale on to another. For many products, bigger is better. The size of an object can be measured in a variety of units, but generally this does not include money. The consumer must generate internal representations of size and money and map one on to the other, to decide precisely how much value an extra foot of height, kilogram of weight, or litre of capacity adds. Naturally, this process becomes more difficult when a second product attribute is added. An item at a given price that varies in size and colour requires the consumer to map three different scales on to each other simultaneously, in order to determine the surplus.

Generally speaking, product attributes are, in the language of consumer research, non-alignable. This means that they are incommensurate; there is no obvious or veridical way to map one scale on to another. How do you map a nicer product colour on to a more convenient size, or superior network coverage on to an allowance of text messages? How much must per unit rate for energy drop to match the inconvenience of receiving only an e-bill? How much better must the interest rate on savings be to make it worth accepting an early exit fee? One consequence of this incommensurability is that making a good decision requires more of the decision-maker than simply observing and responding to the information that is immediately in front of them. An assumed or learned relationship between otherwise incommensurate scales is required, based on experience or memory. Thus, comparing attribute magnitudes and prices, by mapping them to a common internal representation of value, requires *absolute* rather than *relative* judgment.

To the best of the authors' knowledge, no previous empirical study has examined how accurately consumers are able to trade off attribute magnitudes against prices, at least in terms of both precision and bias. Yet it is known that humans have limited capacity for making absolute judgments. Going back to seminal work by Miller (1956) and beyond, it is well documented that human observers are generally only able to map perhaps between four and eight unique levels of a perceptual quantity such as length, weight, loudness etc., onto a set of numbered categories. In these 'absolute identification' tasks, there is only modest variation

in accuracy across different types of quantity (Stewart et al., 2005). Even with extensive practice and experience, the upper limit of eight is rarely breached (Dodds et al., 2011). The results of decades of research with 'perceptual categorisation' and 'magnitude estimation' tasks suggest similar limits to the human capacity to make absolute judgements (Ashby and Maddox, 2005; Laming, 1997).

3.2 EXPERIMENT A: AIMS AND METHODS

Aims

The purpose of Experiment A was to produce an initial set of measures regarding how accurately a representative sample of consumers could resolve trade-offs between attributes and prices, and between attributes themselves. The experiment was a simple S-ID task involving a golden egg. The egg could possess one or two attributes, with or without a price. We compared how accurately consumers could decide: (1) which of two eggs was more valuable when they possessed just a single attribute; (2) whether a surplus was present for an egg with a single attribute and a displayed price; (3) which of two eggs was more valuable when two attributes had to be traded off against each other; (4) whether a surplus was present for an egg with two attributes and a displayed price.

Methods

Sixty-four consumers from the Dublin area took part in Experiment A. They each received a fee of €20 for participation. In addition, they were told that one-in-ten participants would win a €50 shopping voucher based on performance in the task. Seven shopping vouchers were posted to the participants who produced the most accurate performance averaged across the experimental runs.

The experiment employed the golden egg hyperproduct and consisted of two types of task. For the 'Egg v. Price' tasks, the participant was presented with an egg and had to decide whether it was worth more or less than a displayed price. For the 'Egg v. Egg' tasks, participants were presented with two eggs and had to decide which was the more valuable. The main attributes used were the size of the egg (measured in pixels from top to bottom) and the fineness of its surface texture (measured by its highest spatial frequency component in cycle/pixel). The value of the egg was a simple linear function of its attribute(s). A certain number of pixels of size and a certain amount of spatial frequency equated to an amount of Euros.

Each participant undertook six experimental runs of 80 decisions. Half the participants completed three Egg v. Egg runs, followed by three Egg v. Price runs, while for the other half this was reversed. On experimental runs 1, 2, 4 and 5, the value of the egg depended on a single attribute (size or texture). On experimental runs 3 and 6, the value of the egg depended on both attributes (size and texture). Before each experimental run, participants were presented with a series of examples of eggs and prices, to help them to learn the relationship between the attribute and the price. The golden eggs had a mean value of €300 and a price range of €180 to €420. Throughout each experimental run, a reminder of the attributes that mattered to the value of the egg was placed at the top of the screen.

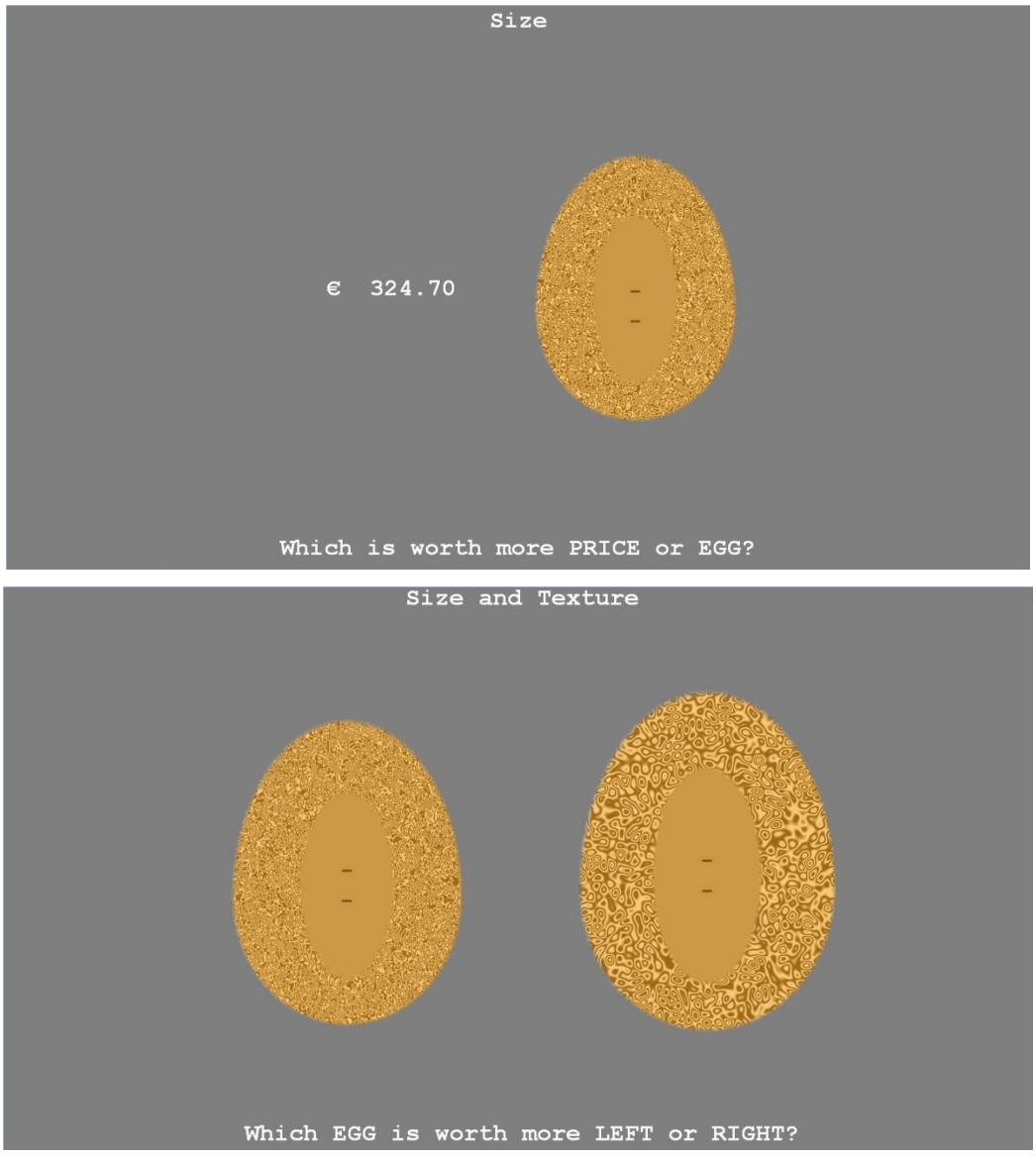
Two example screen-shots are presented in Figure 3. The top display is an Egg v. Price task, where the value of the egg depends only on its size and the participant had to decide whether the egg was worth more or less than the displayed price. After making their decision by pressing the right or left button on the response box, feedback was provided in two forms. First, there was an audible beep if they got the answer wrong. Second, the actual value of the egg was revealed regardless of whether they answered correctly or incorrectly. The bottom display is an Egg v. Egg task with two attributes. On this trial, the egg on the left is smaller but has a finer surface texture. Participants had to learn how to relate a difference in texture to a difference in size, in order to resolve the trade-off and decide which egg was worth more. In this case, the correct answer is the egg on the left, because the scale of the difference in texture trumps the scale of the size difference. Feedback was again given in the form of a beep for an incorrect answer and a '€€€' sign placed next to the more valuable egg.

For each trial, an initial price was selected randomly from a uniform distribution covering the price range. For the Egg v. Price trials, this was the displayed price. For the Egg v Egg trials this initial price determined the value of one of the eggs. A predetermined positive or negative surplus was then added to the display price to set the value of the (other) egg. In the single attribute case, the attribute magnitude was simply set to match this value. In the two attribute case, the magnitude of the two attributes was drawn randomly from the possible set of combinations that matched the value required, subject to the proviso that in the Egg v. Egg trials each egg had to be better on one attribute. That is, the higher value egg could not have a higher magnitude on both attributes – there was always a trade-off.

During each experimental run, an adaptive 'staircase' procedure was employed to determine the surplus for each trial. The run began with large surpluses that were relatively easy to spot. When the participant responded correctly, the size of the surplus for the subsequent presentation was reduced, i.e. the task was made a

little more difficult. Similarly, following an incorrect response, the size of the surplus was increased to make the task a little easier. The size of these adjustments was designed to home in on a level of difficulty at which the participant was able to respond correctly 75-80 per cent of the time. This procedure ensured that there were sufficient incorrect responses to measure the participants' performance accurately, while not demotivating the participant by making the task feel too hard.

FIGURE 3 Example tasks from Experiment A. In the Egg v. Price task (top), the participant decides whether the egg is worth more or less than the price. In the Egg v. Egg task (bottom), the participant decides which of two eggs is more valuable. The egg on the left is smaller but has a finer texture, so participants had to trade off these attributes to decide which was more valuable.



The 64 participants were also split into four groups of 16. This between-subjects aspect of the experiment was designed to test whether the specific attribute or

the extent of the attribute range mattered for performance. For instance, if performance were driven by perceptual limitations, it would be easier to compare an attribute with a larger range against prices. For Group 1, the attributes ranges were 356-708 pixels for size and 0.018-0.142 cycles/pixel for texture. For Group 2 the size attribute was replaced with a different attribute based on the interval between two lines, or 'hallmarks' at the centre of the egg. Pilot studies suggested that despite being easy to discriminate perceptually, this attribute might be harder to map onto prices. The interval varied between 18.6 and 93.4 pixels. For Group 3, the size range was approximately doubled to 177-887 pixels. For Group 4 the texture range was approximately halved to 0.049-0.111 cycles per pixel.

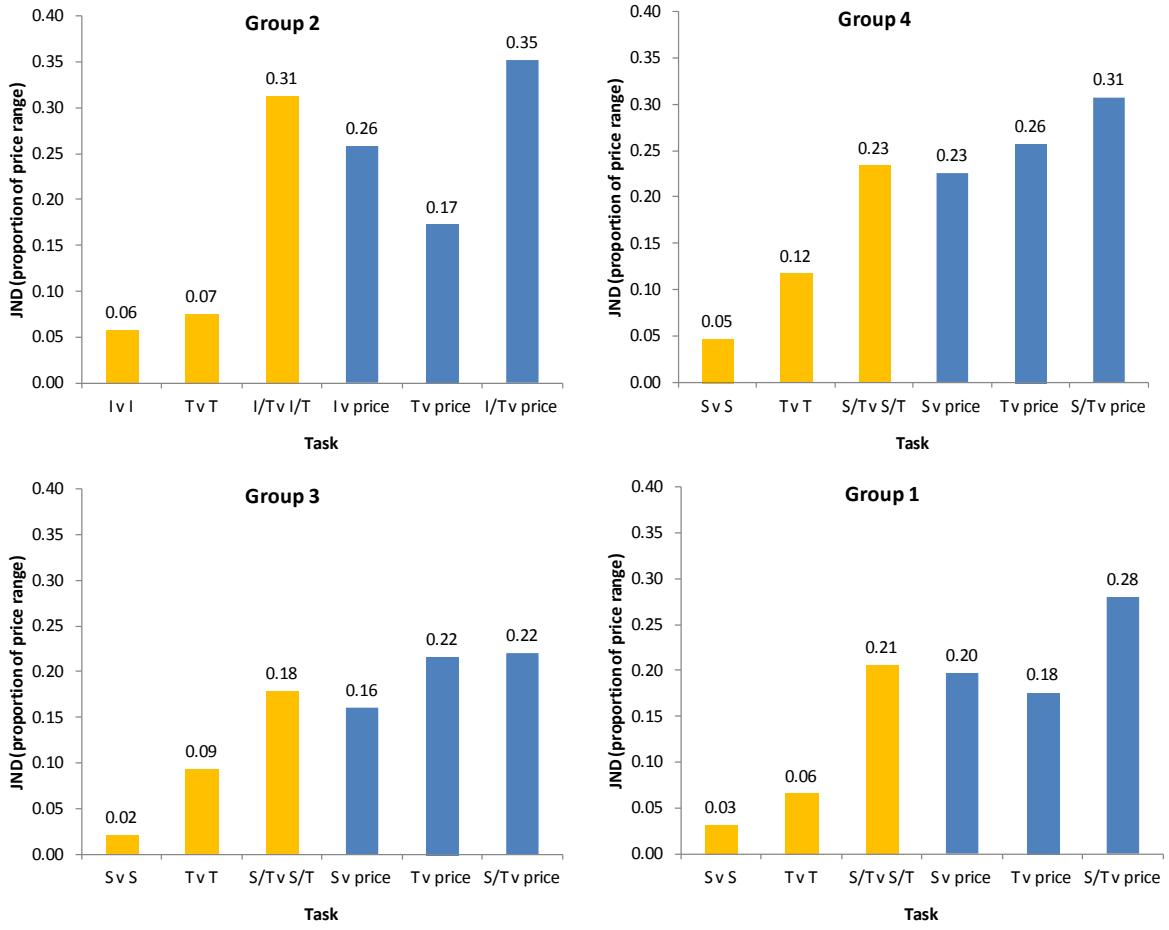
3.3 EXPERIMENT A: RESULTS

Precision

The 86 per cent just noticeable differences (JNDs) for each group and task are presented in Figure 4, where the bars for the Egg v. Egg tasks are coloured gold and for the Egg v. Price tasks coloured blue. A number of results stand out. Firstly, the precision with which participants could perceive differences between attribute magnitudes was very much higher than the precision with which they could match them against prices. Size differences between the eggs were reliably perceived when they were the equivalent of less than 5 per cent of the price range. For texture the variation in JND was 6-12 per cent; for the interval the JND was 7 per cent. However, once an attribute had to be compared with a price, precision decreased markedly. Participants required a minimum of 16 per cent (size, Group 3) and a maximum of 26 per cent (interval, Group 2; texture Group 4). All of the differences in JNDs between single-attribute Egg v. Egg tasks and Egg v. Price tasks were strongly statistically significant.

It is notable that the level of precision when comparing an attribute against a price was largely unrelated to the precision with which it could be compared between two eggs. For example, in both Groups 1 and 2, participants could discriminate which was the more valuable egg better when they differed in size (or interval) than when they differed in texture. Yet both groups were more precise at judging whether an egg conferred a surplus on the basis of texture, especially Group 2. The strong suggestion, therefore, is that precision when spotting a surplus is largely unrelated to the precision with which the attribute itself can be discriminated. In other words, the psychological mechanisms that limit performance are located beyond the perceptual system's representations of perceptual magnitudes; the capacity constraint is not perceptual, but cognitive.

FIGURE 4 The 86% just noticeable differences (JNDs) in surplus for four groups of 16 participants and six tasks in Experiment A. Egg v. Egg tasks in gold; Egg v. Price in blue.



As in the pilot study, for some reason participants found the interval between the central hallmarks harder to compare against prices than either of the other two attributes. One possible reason for participants finding the interval harder to compare with prices is that unlike the surface texture and the size of the egg, it had no obvious absolute benchmarks. The size of the egg could be compared fairly easily with the size of the screen. The egg's surface texture essentially varied between a finest texture that had it been any finer would not have been visible and a coarsest texture that almost ceased to appear as a texture, but became more like a pattern of blobs. In other words, the texture had identifiable end-points. The interval, on the other hand, had nothing obvious that could operate as an absolute benchmark. In support of this account, when the range of the texture was halved for Group 4, thereby making the end-points of the scale less clear as absolute benchmarks, precision when comparing it to prices was significantly poorer.

The second notable result apparent from Figure 4 is that when two attributes had to be traded off against each other in the two-attribute Egg v. Egg task precision

was, broadly speaking, comparable to precision when one attribute had to be compared with a price in the single-attribute Egg v. Price tasks. Overall, there was no significant difference between these tasks across the four groups.

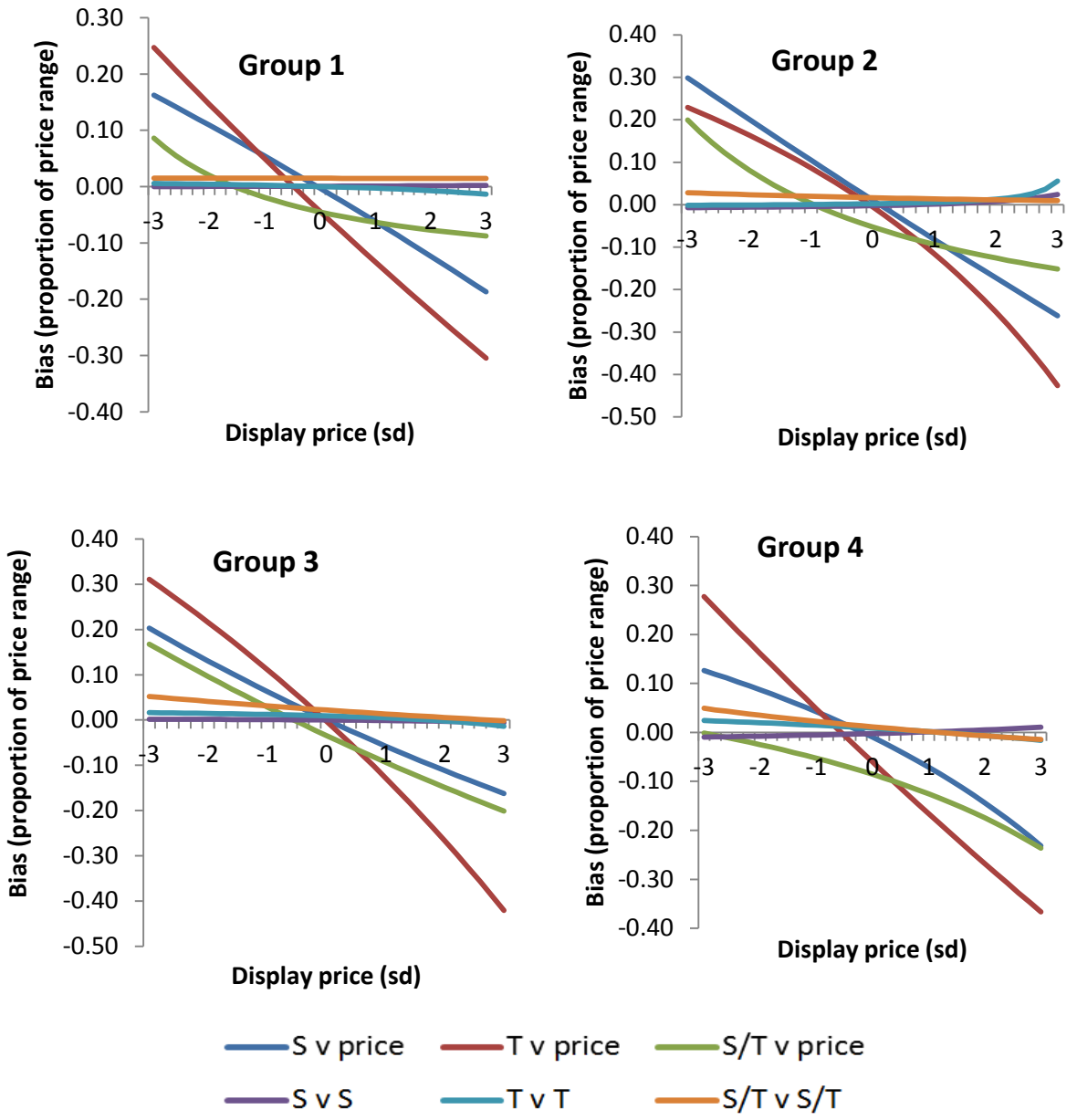
Across all four groups there was a further decline in precision when both attributes had to be taken into account in order to identify a surplus in the Egg v. Price task. Combining these last two results, there is a suggestion from Experiment A that precision when trying to identify a surplus may largely depend on the number of internal scales that must be traded off against each other.

Lastly, we tested for learning by comparing performance early in the experimental run with performance later in the run once the participants had seen many examples and associated feedback. Surprisingly, we recorded no statistically significant effect.

Bias

A bias in the Egg v. Egg tasks would only indicate a preference for right or left, which we did not encounter. In the Egg v. Price tasks, the presence of bias is much more interesting, since it indicates overvaluation or undervaluation of the hyperproduct relative to the price. Overall, this bias was small, notwithstanding a slight tendency for most participants to overvalue the eggs, on average. However, when we examined how performance in the task varied across the price range, we encountered a surprising result, which is shown in Figure 5. The bias (vertical axis) is plotted as a function of the displayed price (horizontal axis), expressed in standard deviations. As in Figure 2, a positive bias indicates undervaluation and a negative bias overvaluation of the product.

FIGURE 5 Biases across the price range for four groups of 16 participants and six tasks in Experiment A. Positive bias indicates undervaluation of the egg relative to the price; negative bias equates to overvaluation (see Figure 2 and accompanying text).



Participants had a strong and consistent bias in the three Egg v. Price tasks. They undervalued eggs towards the bottom end of the price range and overvalued them at the top end. The size of this bias was strikingly large in comparison with the JNDs reported in the previous subsection. At just one standard deviation from the mean price, we estimate that the bias was approximately 5-15 per cent of the price range, while at the edges of the price range it climbed to as high as 30 per cent. Furthermore, the bias was somewhat stronger for the single-attribute tasks than when two attributes had to be taken into account. We again tested this bias to see whether it was subject to any learning across the experimental runs. Yet

we found no statistically significant evidence of the bias diminishing with experience and feedback.

3.4 EXPERIMENT A: DISCUSSION

Experiment A provides an initial indication that the ability of consumers to resolve simple trade-offs between attributes and prices may be surprisingly limited. In an incentivised experiment with plainly perceptible attributes, multiple examples and feedback, surpluses could be reliably identified only when they were of the order of 20 per cent of the entire price range. This lack of precision occurred despite the fact that the attributes themselves could be discriminated much more precisely when two products were placed side by side (down to just 2 per cent in the case of size). Once either an attribute had to be traded off against a price, or two attributes against each other, thus requiring the integration of information from incommensurate scales, surplus identification was imprecise.

Furthermore, in addition to this lack of precision when resolving simple trade-offs, consumers' responses were systematically biased. They undervalued products towards the bottom end of the price range and overvalued them towards the top end. The size of this bias was quite substantial, although it did vary by the number of attributes (one or two).

Neither the imprecision nor the bias we recorded appeared to be subject to any kind of appreciable learning effects, despite the participants completing 80 trials in each condition with feedback. It seems that whatever learning took place occurred during the initial example phase and any subsequent learning was at best too slow to be detected.

The findings of this initial experiment invite further investigation. The suggestion is that precision in identifying surpluses may depend on how many internal scales (for attributes and price) must simultaneously be integrated by the consumer. Experiment A also hints at a potentially interesting relationship between the number of attributes being taken into account and biases in consumers' judgements across the price range. The slope of the relationship between the bias and the price was consistently flatter for two-attribute products than for single-attribute products, implying that it was actually stronger for the single-attribute product.

Section 4

How Many Attributes Can Consumers Cope With?

4.1 INTRODUCTION

The clear suggestion of Experiment A is that whatever psychological mechanism is used to integrate information from incommensurate scales, it is relatively imprecise and subject to systematic bias. The pattern of bias across the price range also may be related to the number of attributes that determine value. While Experiment A establishes a baseline measure of performance and is suggestive about the effects of requiring consumers to take account of multiple information sources simultaneously, the design was limited to products with a maximum of two attributes. Furthermore, accuracy in the two attribute case may even have been overestimated. The two-attribute experimental runs were always preceded by two single-attribute runs that gave the participants extensive opportunities to learn the attribute-price relationship for each of the two attributes separately, before having to integrate information from both simultaneously. In the real world, there are many product attributes that consumers would never have the opportunity to learn to value in isolation like this, before taking them into account in the context of valuing a multi-attribute product.

The three experiments described in this chapter test consumers' capabilities with multi-attribute products more thoroughly. Experiment B increases the number of attributes up to a maximum of four. The results suggest that there are severe limits to the number of attributes that can be integrated into consumers' identification of surpluses and, moreover, that the capacity for learning is modest. Experiment C tests whether this rather striking result continues to hold when a sample of highly educated and numerate individuals is given repeated practice at the S-ID task. Experiment D then tests whether consumers are able to integrate attribute information more accurately when attribute magnitudes are positively correlated.

4.2 EXPERIMENT B: AIMS AND METHODS

Aims

The purpose of Experiment B was to test how increasing the number of attributes affects the accuracy with which they are able to integrate the available attribute and price information in order to identify surpluses. The experiment compared performance for products with one, two, three and four attributes. More specifically, the experiment sought to test whether consumers could integrate an

additional attribute into their decisions with statistical efficiency. In theory, while the addition of an extra attribute should always reduce the decision-maker's precision, it should do so at a predictable diminishing rate if the extra information is integrated efficiently into the decision.[5] There are some human perceptual systems that can integrate additional information about the perceptual scene they are encountering in such an optimal manner, for example when integrating information from vision and touch in order to determine the shape of an object (Ernst and Banks, 2002). A second aim was to determine the relationship between the systematic bias in surplus identification across the price range and the number of attributes.

Methods

The methods employed in Experiment B were identical to those of Experiment A, except in the following respects. Most notably, there were up to four attributes that determined the surplus conferred by the golden egg on any one experimental run. The four attributes used were those depicted in Figure 1: size, texture, circularity (of a central ellipse) and the sharpness of the angle on a central hallmark. In all cases, the value of the egg was a simple linear combination of the attribute magnitudes.

Because the relationship between attributes, prices and surpluses was more complex in this multiple attribute task, participants were given more extensive practice prior to the test runs. They were shown examples where the surplus depended on a first attribute only, followed by 24 practice trials with this attribute. A second attribute was then added. The participant was shown examples of how it affected the value of the egg, and then completed 24 practice trials with two attributes. The third attribute and the fourth were added in the same way, with examples and 24 practice trials in each case. After all this practice, the participant completed four experimental test runs of 80 trials with one, two, three or four attributes in an order that was pseudo-randomised across participants. The sequentially increasing complexity of the product during the practice trials was designed to ensure that participants understood the task, whereas the measurements taken during the test phase were designed to compare accuracy with different numbers of attributes on an equal footing.

Thirty-six participants took part in the experiment. This number was chosen carefully to allow the different conditions to be counterbalanced across

[5] Technically speaking, this calculation assumes that a participant's JND for identifying surpluses when single attributes determine the value of the product is a measure of the variability inherent in the internal mapping from an attribute to a price. If so, then the variability when an additional attribute is added should equal the sum of the variances, provided the additional information is processed efficiently, leading the JND to increase according to a square-root relationship. Further details are to be found in the first technical paper listed in the appendix.

participants. There are 15 possible attribute combinations: four one-, six two-, four three-, and one four-attribute combination (see vertical axis of Figure 7). These subconditions were counterbalanced across the 36 participants such that each combination was tested the same number of times. Throughout each run, a reminder at the top of the screen told participants which attributes must be factored in ('size', 'texture', 'circularity', 'angle').

Pilot experiments were undertaken to determine how accurately each of the attributes could be discriminated when two eggs were presented side by side. The range of each attribute was then set to cover 26 just noticeable differences of attribute magnitude. The price range was €176.50 to €423.50, so that each discriminable difference in magnitude equated to €9.50 in price. For example, we found that people could reliably spot when the angle of the cross on one egg was sharper than the angle on the other when the difference was 1.5 degrees. Thus, the range of possible angles covered 39 degrees and each 1.5 degrees was worth €9.50.

The specific price displayed and egg used on each trial were determined as follows. First, attribute magnitudes were selected randomly from uniform distributions covering their ranges. Second, the corresponding price was calculated to act as the display price. Third, the surplus was added to this price to determine the value the egg should take. Fourth, the relevant attributes were increased or decreased to match the required surplus, with proportions of the increase or decrease assigned across attributes at random. Lastly, the program checked that attribute magnitudes and prices remained within the specified ranges. This included ensuring that for any given displayed price and positive (negative) surplus, the equivalent negative (positive) surplus would also keep the test price within range. Hence, no correlation existed between displayed prices and correct responses – the probability that the test price was higher was always 0.5. If a price or attribute magnitude fell outside the specified range, the programme began the process again.

4.3 EXPERIMENT B: RESULTS

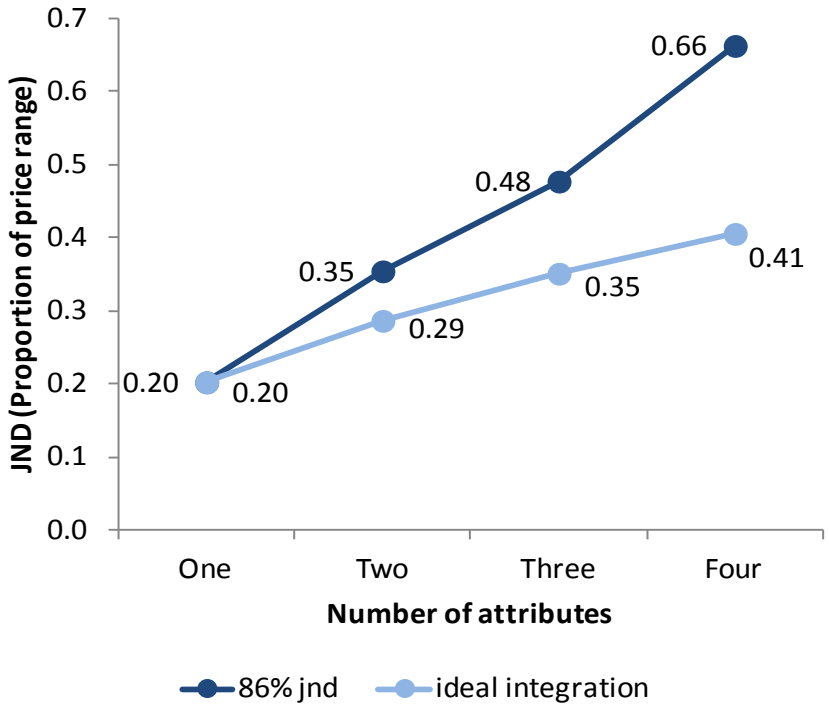
Examination of the 144 experimental test runs revealed six runs in which performance was extremely poor and participants had performed close to chance. The individual participants had generally performed adequately on their other experimental runs and closer examination suggested that the participant had mentally inverted an attribute (e.g., treated a coarse rather than fine texture as more valuable). These six runs were discarded from the analysis.

Precision

The JNDs for the conditions with one, two, three and four attributes are provided in Figure 6. The first thing to notice about this figure is the absolute level of performance for a single attribute, which closely paralleled that obtained in Experiment A. Again, in order to identify it reliably, participants needed a surplus equivalent to approximately 20 per cent of the price range. Put somewhat differently, they were able to distinguish approximately only five different levels of value (e.g., very bad, quite bad, average, quite good, very good), despite the fact that there were 26 discriminable levels of each attribute magnitude when two eggs were placed side by side. Matching attributes to prices is much less precise than comparing them with each other.

When the number of attributes was increased, performance declined steeply (dark blue curve). Each successive difference when an additional attribute was added was strongly statistically significant. Once three or four attributes had to be taken into account simultaneously, most participants required a surplus of half the price range or more in order to identify it reliably. In simple terms, they could just about tell a good product from a bad one. This level of performance was considerably worse than the level that would be predicted by statistically efficient integration of the additional information (light blue curve), given how accurately a single attribute can be compared with a price.

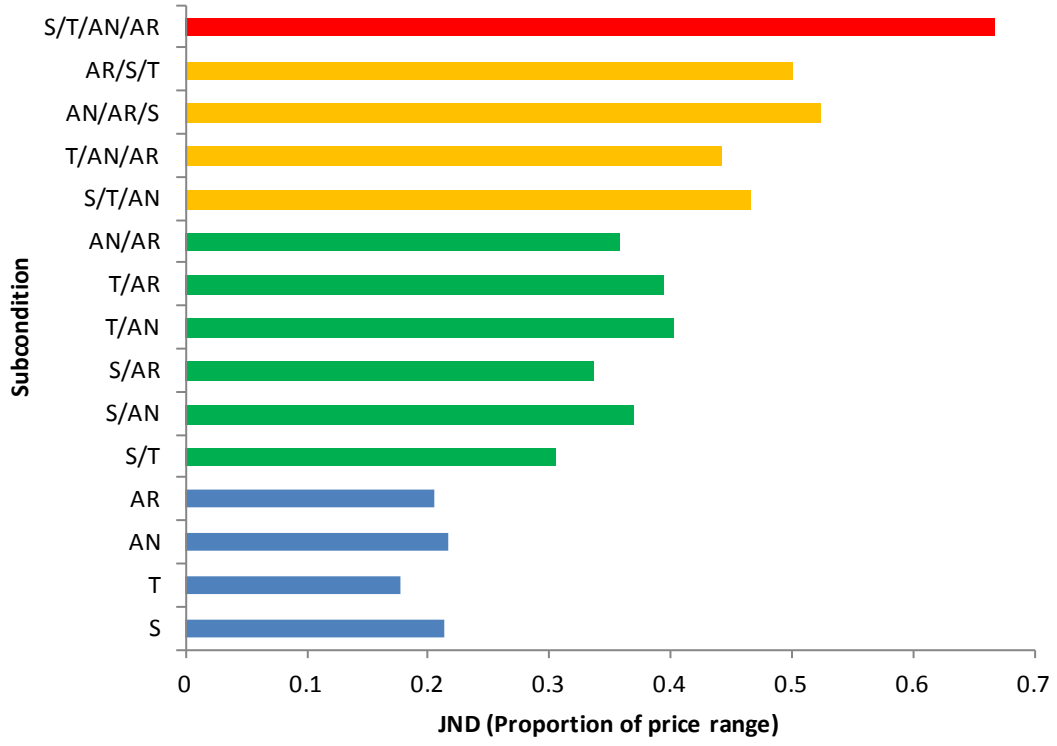
FIGURE 6 JNDs for identifying a surplus with one, two, three and four attributes in Experiment B, comparing actual performance with a hypothetical participant who integrates additional attribute information with statistical efficiency.



As described in the Methods section, all the different possible combinations of the attributes were tested an equal number of times. The JNDs for each of these 15 possible subconditions are shown in Figure 7. There are some small differences between the subconditions with the same number of attributes. For whatever reason, participants found it significantly easier to identify surpluses in the size and texture combination (as used in Experiment A) than in the other two-attribute combinations. Overall, however, the dominant factor behind the precision with which information could be integrated in order to identify a surplus was how many attributes had to be factored into the decision. Surplus identification was more precise in all single-attribute subconditions than in all two-attribute subconditions, more precise in all these two-attribute conditions than in all three-attribute subconditions, and least precise in the four-attribute condition.

Once again, we looked for learning effects across the trials of each experimental run, but found no significant improvements in precision, despite all the feedback supplied.

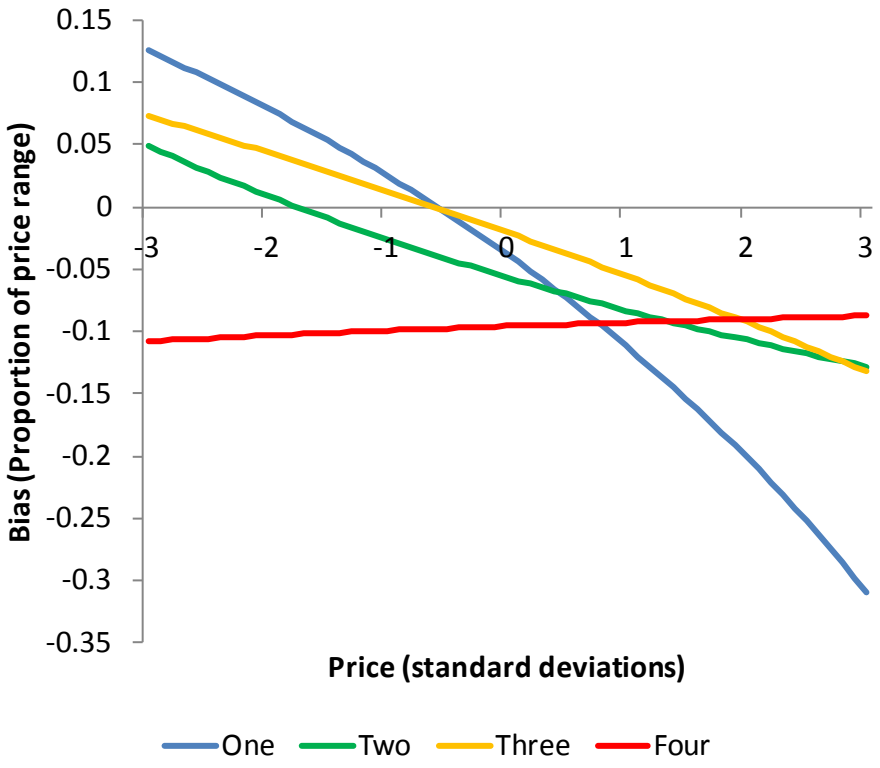
FIGURE 7 JNDs for the 15 possible combinations of four attributes in Experiment B. Although there are small differences in precision between the different combinations, imprecision is overwhelmingly driven by the number of attributes that must be integrated into the decision (S=size; T=texture; AN=angle; AR=aspect ratio (circularity)).



Bias

Figure 8 plots the extent of bias across the price range according to the number of attributes that had to be factored into the decision. As in Experiment A, when a single attribute magnitude had to be integrated with a price to determine the surplus, value was underestimated at the bottom end and overestimated at the top end. This effect was again statistically significant and substantial. However, the slope of this relationship between the bias and the price moderated with the addition of attributes, although overall there remained a tendency to overestimate the value of the egg and hence the surplus.

FIGURE 8 Bias across the price range by number of attributes in Experiment B. Positive bias indicates undervaluation of the egg relative to the price; negative bias implies overvaluation.



Additional tests

The richness of the data generated by this experiment permitted a number of additional tests to be undertaken. In particular, as described in Section 2.4, it is of interest to test for effects that have been observed in previous choice experiments in which preferences were not imposed. If such effects can be found in the data collected in the present experiment, this would support the contention that the psychological mechanisms being investigated in the S-ID task are the same as those involved in subjective consumer choices. From a policy perspective, this is obviously a central contention of the present work.

Meyvis and Janiszewski (2002) report a 'dilution' effect, according to which consumers struggle not to factor in irrelevant attributes when choosing among products. This effect can easily be tested for in Experiment B, since there were experimental runs in which participants had to ignore attributes that previously they needed to factor in. Accordingly, we found a small but statistically significant bias in the direction of products that were better on the irrelevant attributes.

Hauser (2011) describes evidence that consumers' choices are biased towards more familiar attributes. Similarly, Carlson et al. (2006) describe an effect they call 'leader-driven primacy', whereby the order of attribute presentation has an effect on choice, with attributes that are encountered earlier given greater weight in choice. Because, for each participant, there was always one attribute that was the first to be introduced and mattered in all four conditions, any bias towards familiar or initial attributes would materialise in Experiment B as a bias towards this primary attribute. In our data, we again found a small but statistically significant bias towards this attribute.

Chernev (2005) shows that consumers are more likely to choose products that have balanced attributes, i.e. preference is given to options with similar attribute magnitudes rather than those with larger trade-offs between attributes. We tested for this effect in the data from the two-, three- and four-attribute conditions. There was a statistically significant bias in favour of products with smaller trade-offs between attributes. Interestingly, we also found that participants were less precise in their decisions when there were large trade-offs between attributes, although we do not know of any precedent for this result.

4.4 EXPERIMENT B: DISCUSSION

Experiment B produced striking findings. The implication of the variation in JNDs as extra attributes are added is that consumers are likely to struggle to make good product comparisons when they have to take multiple attributes into account at the same time. Once three or four attributes are each important to the product's overall value and have to be factored into the decision, surplus identification is very imprecise, requiring surpluses equivalent to half the price range or more for reliable detection. As in Experiment A, this level of imprecision was found despite employing plainly perceptible attributes for which 26 levels of magnitude could be discriminated when presented alongside each other.

In addition to this level of imprecision, we again found that surplus identification was biased, but the bias was not consistent across products with different numbers of attributes. For a single attribute, the value of products at the lower

end of the range was underestimated, while for those at the top end of the range it was overestimated. But the scale of this effect diminished with the addition of extra attributes, albeit that there was a slight (but significant) bias towards overestimating surpluses generally.

At this stage it is not clear what lies behind this consistent pattern of bias across the price range, but there are two potential explanations for the moderation in the slope of the bias as additional attributes are added. One possibility surrounds attribute averaging. It is well documented that consumers at least sometimes choose among multi-attribute products by averaging attribute magnitudes. This leads to the counterintuitive but replicated empirical finding that an option that is very good on a first attribute and moderately good on a second can be judged as less good overall than an option that is known only to be very good on the first (Troutman and Shanteau, 1976; Weaver et al., 2012). Translating this finding into Experiment B, in which the overall value of the egg was equal to a linear addition of attribute magnitudes, any mechanism that averages attribute magnitudes would, as the number of attributes increases, reduce the value of the better products and increase that of the less valuable ones. This is consistent with the reduction in the bias shown in Figure 8 as the number of attributes increased. It does not, however, explain why there is a bias at all when there is just one attribute in play.

A second possible explanation for the pattern of biases is that there is a relationship between precision and bias, perhaps even a trade-off between them. The basic logic here goes back to observations by Barlow (1961), who pointed out some potential consequences of the fact that neural systems have limited ranges over which they can code responses to external stimuli. In order to distinguish between different stimuli within a range, the system should adapt its coding to disperse responses to those stimuli, to maximise the difference in the neural code arising from likely differences in the real world. Yet, by employing this technique, which will improve the ability to discriminate when one stimulus has greater magnitude than a near neighbour, the coding exaggerates these differences, leading to biased estimates of veridical quantities. Applied to the present experiment, the logic is that the system tunes itself to discriminate differences in the value of the product, but in doing so exaggerates these differences relative to price. If so, the result will be a trade-off between precision and bias.

Overall, the findings of Experiment B are strong in terms of implications for consumer capability. Many products have multiple attributes that matter to the overall value of the product and hence the surpluses on offer. The scale of inaccuracy reported here implies that in such markets consumers will struggle to find best value, or perhaps even simply to locate reasonable value. Given the

strength of these implications, the next experiments sought to ensure that the finding was genuine and robust.

4.5 EXPERIMENT C: AIMS AND METHODS

It is possible that performance in Experiment B might have been influenced by some factors that, from the perspective of generalisability, one might want to rule out. Firstly, there is a possibility of cross-task interference. The requirement for participants to ignore some attributes on some experimental runs but not on others may have impaired performance or hampered learning. Secondly, performance might have been affected by fatigue or some other drain on effort over the one-hour session. The experiment required concentration on a repetitive task, somewhat like a video game. Participants were paid, incentivised, appeared to be engaged competitively in the task (especially from their reaction to errors), and did not vary significantly in performance across the session. However, learning and motivational attrition could have counterbalanced each other. A less demanding session might improve effort and performance. Finally, the task required a level of initial comprehension and numeracy. Since six runs had to be discarded, probably because the participant failed initially to grasp the direction of an attribute-price relationship correctly, it is possible that performance might be affected more generally by uncertainties or misunderstandings that could not so easily be detected in the data. Because each of these factors might mean that the results of Experiment B underestimate consumers' capabilities, Experiment C provided a robustness check.

Aims

Experiment C consisted of a tournament held among research staff at the ESRI, who competed at valuing golden eggs for a prize. The aim was to test how performance was affected when the concerns raised in the previous subsection were ruled out. Because the participants were professionals with high levels of educational attainment and numeracy compared to the general population, the risk of misunderstanding the task was negligible. To eliminate cross-task interference, participants undertook only a single condition. To reduce any influence of fatigue, experimental runs were shorter and the session duration was halved compared to Experiment B. To examine further potential for learning, participants completed three sessions, each separated by more than a week. To assess whether there was a role for additional effort, an incentive manipulation was also employed on the final run of trials.

Methods

Apparatus, products and procedure were as for Experiment B, except that following the initial examples, which were repeated at the start of each session,

participants completed three test runs of 64 trials. Twenty-four participants were pseudo-randomly assigned to a two-, three-, or four-attribute condition. Those in the two- and three-attribute conditions were further pseudo-randomised into one of two subconditions ('size-circularity' or 'texture-angle'; 'size-texture-circularity' or 'size-texture-angle'). Participants were not paid to participate, but were told that a €50 voucher would be awarded to the best performer across all sessions (determined by comparing JNDs with standardised distributions by condition from Experiment B). They expected the third session to be the same, but instead encountered a manipulation. After the first run, when expecting two more runs, participants were told that there would be just one more run, that they should take as long as they needed for each judgment, and that the participant who improved the most relative to previous performance would win a €50 voucher.

4.6 EXPERIMENT C: RESULTS

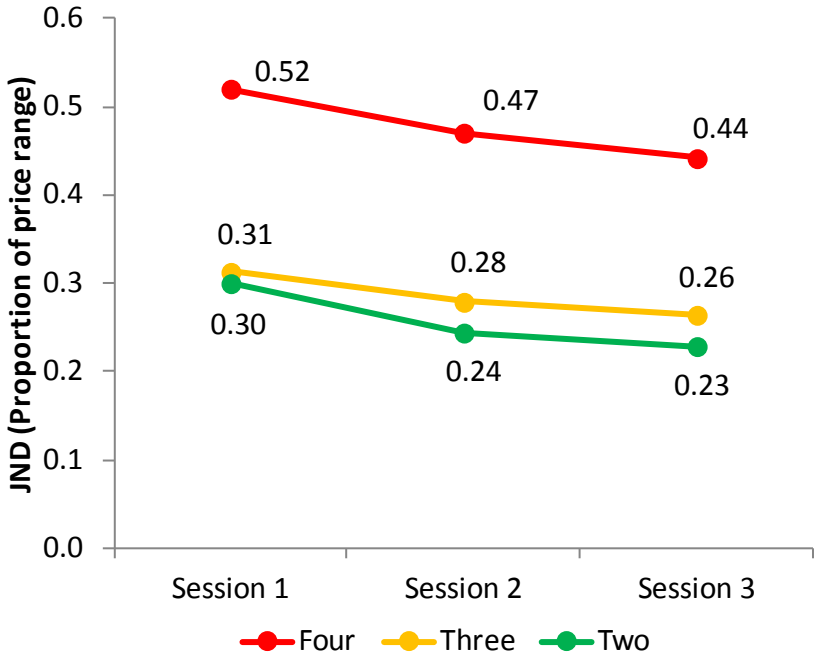
Precision

Figure 9 shows JNDs for the three separate conditions across the three sessions. The use of highly numerate participants, removal of the possibility of cross-task interference and shorter runs and sessions did have a significant impact on the precision of surplus identification. The JNDs in Session 1 were lower than those recorded in Experiment B, especially for the eight participants who were in the three-attribute condition.[6] However, the size of this effect was small. In the first session of Experiment C, the JND of the median participant would place them at the 69th percentile of participants for the equivalent condition in Experiment B. In addition, precision was again lowest when the trade-offs between the attribute magnitudes were larger.

The extent of learning following the first session was consistent across the three conditions. JNDs fell by just under one quarter between the first and third sessions. The difference in JNDs between the first and second sessions was statistically significant but the difference between the second and third sessions was not. This deceleration in the rate of learning is a standard property of what might be considered a 'standard' learning curve that might be observed in many psychological tasks. The limited scale of this learning explains why it was not detected in Experiment B – learning beyond the initial examples in the S-ID task appears to be limited and slow.

[6] This narrowing of the gap between performance in the three- and two-attribute was probably a simple reflection of differences in ability between the two groups of participants assigned to these conditions, since there were only eight participants in each group.

FIGURE 9 JNDs for identifying a surplus with one, two, three and four attributes in Experiment C, across three consecutive sessions.



The incentive manipulation intended to induce greater effort in the final experimental run was effective. Response times in this run increased by half a standard deviation compared to the run that preceded it, implying that participants were taking substantially longer to make their decisions. For 12 of the 24 participants, their precision in this final run improved, while for the other 12 it declined. Increased effort, therefore, had no effect on precision.

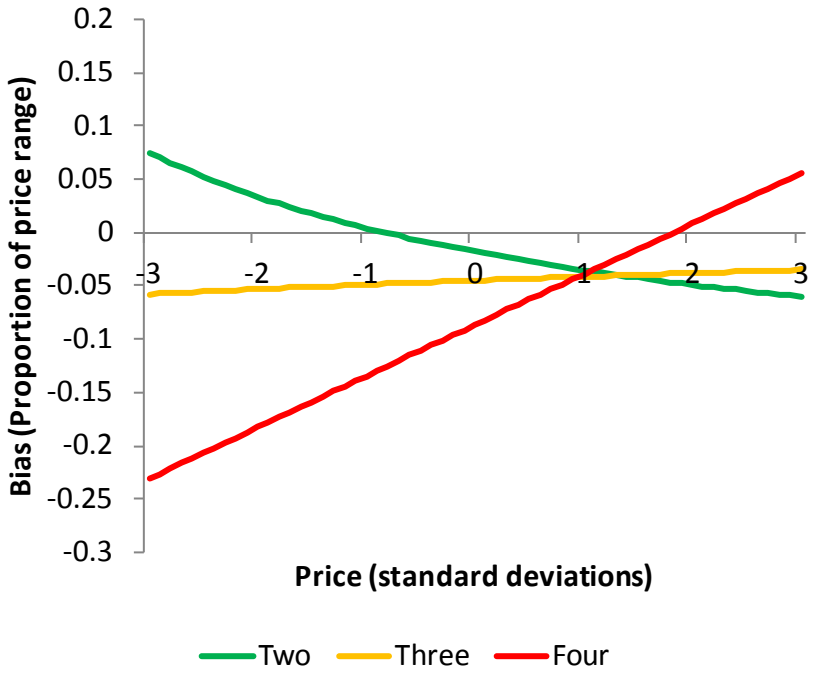
Overall, these results suggest that it is possible to improve only marginally on the low levels of precision seen in Experiment B. Surplus identification remains imprecise, with surpluses equivalent to large proportions of the price range required for reliable detection and limited scope for improvement through learning.

Bias

Throughout Experiment C, the participants displayed a similar pattern of biases to those recorded in Experiment B. Of particular interest was whether they showed any sign of abating over multiple sessions and feedback. Figure 10 plots the bias across the price range in the three conditions during the final session. At this stage, participants had already completed between 384 and 512 trials with feedback. Nevertheless, substantial biases were still recorded, with the slope of the bias across the price range rotating anti-clockwise as the number of attributes increased. Again, there was a slight overall bias in favour of the egg. The only

notable difference between this figure and Figure 8 above is that this rotation is such that the direction of the bias is reversed with four attributes, resulting in overestimation of the value of products at the lower end of the range and underestimation at the higher end. All of these effects were statistically significant.

FIGURE 10 Bias across the price range by number of attributes in the third session of Experiment C. Positive bias indicates undervaluation of the egg relative to the price; negative bias implies overvaluation.



4.7 EXPERIMENT C: DISCUSSION

The results of Experiment C support the primary conclusions of Experiment B. Surplus identification with multi-attribute products was imprecise and subject to persistent biases. Precision improved somewhat with practice, but learning was modest. Even following hundreds of trials of a single S-ID task with feedback, participants with high educational attainment and numeracy could effectively distinguish less than five levels of value when just two attributes were involved and just over two levels once four attributes were in play. Precision was lowest when attribute magnitudes entailed larger trade-offs. Valuations remained subject to systematic biases that varied over the price range, depending on how many attributes were to be factored into the decision. Overall, the results confirmed that lack of effort, engagement or understanding is not behind the limitations to performance in the S-ID task, which instead appear to reflect the limited capacity of the psychological mechanisms that integrate information from incommensurate internal scales.

4.8 EXPERIMENT D: AIMS AND METHODS

In Experiments A to C, the relationships between the attributes and the overall product was additive. In order to keep the value of the products in each of the conditions within the specified price range, it was necessary for the product attributes in these previous experiments to be somewhat negatively correlated with one another. It is possible that the results might have been different had we used positively correlated attributes, as occurs in some markets. For instance, cars tend to have positively correlated attributes: all components of a Mercedes are supposed to be high-quality; all components of a Skoda are designed to be basic yet functional. Thus, in markets such as these, quality may be signalled by very many attributes all pointing towards the same value. Of course, from the consumer's perspective, there can always be one or more attributes that are not entirely in step with the others and which need to be spotted and integrated into assessments of surplus.

Our methods up to this point would have underestimated consumers' capabilities if individuals could integrate information from multiple positively correlated attributes more accurately than they could integrate negatively correlated ones. Experiment D therefore conducted a further robustness check on the results of Experiments B and C, by comparing accuracy in surplus identification across products with perfectly correlated attributes as the number of those attributes increased. We tested this by repeating Experiment B with a fundamental change: the attributes were perfectly correlated. If consumers can indeed integrate correlated attribute information efficiently, performance should improve substantially as attributes are added, because each attribute provides additional information about the same underlying value.

Aims

The primary purpose of Experiment D was to measure precision and bias when increasing numbers of perfectly correlated attributes were added to a product. Again, the number of attributes varied between one and four. As with Experiment B, it is possible to calculate a predicted level for the two-, three- and four-attribute conditions once precision is measured for the single-attribute condition, based on the assumption that the additional information is processed efficiently. A secondary aim was to test our hypothesis that the rotation of the bias across the price range when additional attributes are added (as in Figures 8 and 10 above) is caused by a psychological mechanism that averages attributes. When attributes are uncorrelated and the value of the product depends on adding them together, averaging them will reduce the value of products towards the top of the range and increase the value of those towards the bottom. But with perfectly correlated attributes, this effect should disappear, since the average of the attributes remains the same as more attributes are added. If, instead, the bias is

caused by a precision-bias trade-off, the extent of bias ought to be related to the level of precision.

Methods

The methods were as for Experiment B but with the following modifications. There were 24 participants, who each received a €25 fee for participation. The three best performers stood to win a €50 shopping voucher. This time we employed four different colours of precious eggs: gold, silver, bronze and emerald. The colours changed between runs and were matched to the number of attributes signalling the value of the egg, thereby helping the participant to differentiate between the conditions. The examples and practice trials were the same as for experiment B except that because the task was easier initially to grasp, the number of practice trials was reduced from 24 to 12 for each condition. The number of test trials per condition was 72.

The surplus on each trial was no longer selected via a staircase procedure, but instead via an alternative adaptive method generally referred to as an 'adaptive method of constant stimuli' (AMCS). Each run of 72 trials in fact consisted of six blocks of twelve, although the participant had no knowledge of this. Within each block, the participant was twice presented with each member of a set of three positive and three negative surpluses with a constant separation, i.e. {-5d, -3d, -d, d, 3d, 5d}, where d could be varied between blocks. If a participant responded correctly on more than ten trials in the block, the value of d was reduced to make the task harder. If they responded correctly on only eight or fewer, the value of d was increased.

Thus, as with the staircase method, the difficulty of the task adapted to the participant's performance to aid efficient estimation of their capability. The change from the staircase method to the AMCS was made because the latter presents a combination of easier and harder trials. We reasoned that this might make the task more enjoyable for participants. We did not anticipate that it would have any impact on participants' abilities to identify surpluses.

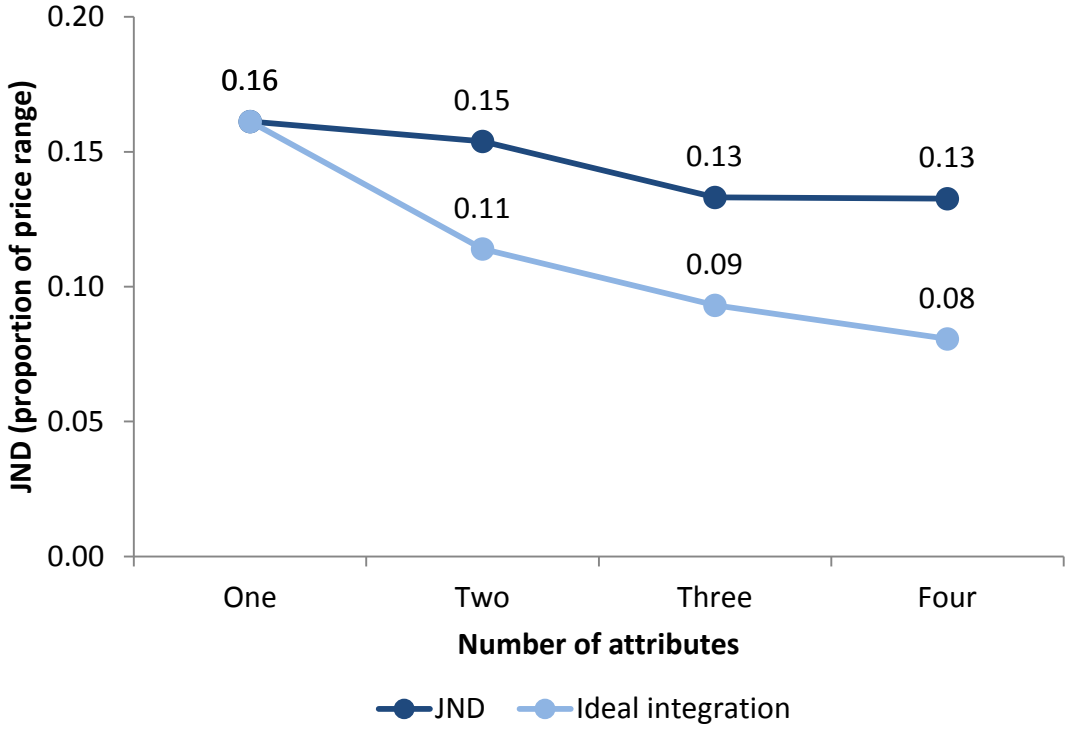
4.9 EXPERIMENT D: RESULTS

Precision

The JNDs by number of attributes are presented in Figure 11. In the single-attribute condition, the level of performance reached by this group of 24 participants is marginally better than that of the 36 participants in Experiment B, at a JND of 0.16 rather than 0.20. One possible explanation for this is that with perfect correlation between attributes, the two-, three- and four-attribute conditions effectively provide additional learning opportunities for mapping the

single attribute to the price, permitting a small amount of additional learning relative to Experiment B.

FIGURE 11 JNDs for increasing numbers of perfectly correlated attributes in Experiment D. The average precision of 24 participants (dark blue line) is compared to performance that would be expected if additional attribute information were integrated with statistical efficiency given the precision in the single attribute condition (light blue line).



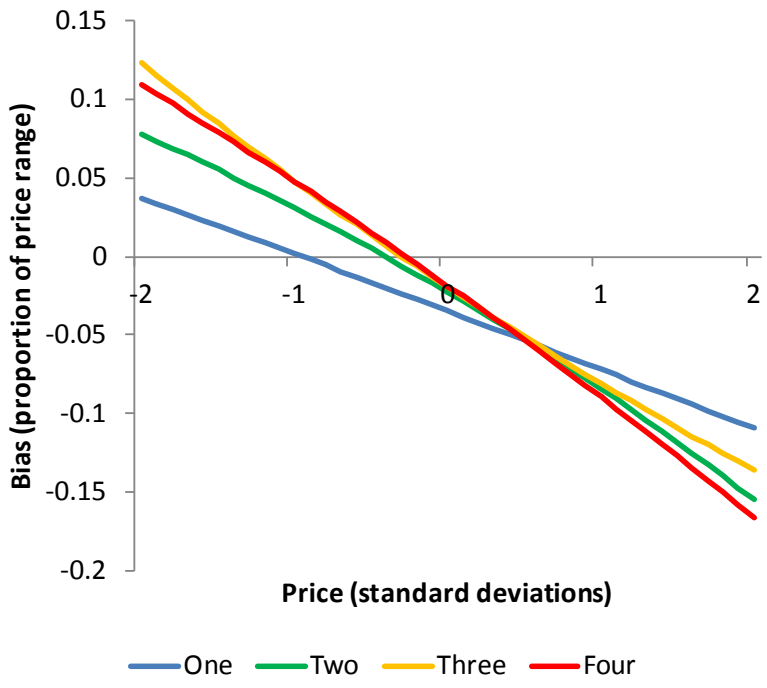
Participants produced a marginal improvement in the JND as the number of perfectly correlated attributes rose from one to four (dark blue line). The differences between the single-attribute and three- and four-attributes conditions were statistically significant. However, precision fell a long way short of efficient statistical integration of the available information (light blue line). Indeed, adding the fourth perfectly correlated signal to the egg's value produced no improvement at all over and above three attributes. Overall, the pattern of precision parallels the situation with the mildly negatively correlated attributes, in that the capacity to process additional attribute information is limited, even with this small number of attributes.

Bias

Figure 12 presents the pattern of bias across the price range. As in previous experiments, there was a slight overall bias towards overestimation of the value of the surplus, but the larger effect was the variation in the extent of the bias over the price range. Again, in the single attribute condition, surplus was

underestimated towards the bottom end of the price range and overestimated towards the top end. The scale of this variation in bias in the single-attribute case was a little smaller than in previous experiments. Most importantly, however, the pattern of biases by number of attributes was radically different. Indeed, it was the opposite of that recorded in Experiments A-C. Once the attributes were perfectly correlated, the slopes of the curves showing the extent of bias across the price range steepened as the number of attributes rose. This effect was statistically significant. Thus, the results are in keeping with the hypothesis of a precision-bias trade-off, such that a mechanism that can discriminate more accurately between levels of value, and hence discriminate surpluses more precisely, induces a degree of bias across the price range.

FIGURE 12 Bias across the price range by number of attributes in Experiment D. The variation in bias strengthens with increased numbers of attributes. Positive bias indicates undervaluation of the egg relative to the price; negative bias implies overvaluation.



4.10 DISCUSSION

In combination, the experiments reported in this chapter imply strong limitations in consumers' ability to integrate information from multi-attribute products in order to identify surpluses. Even when just a single, plainly perceptible attribute must be compared against a price, surpluses need to be 16-26 per cent of the price range in order to be seen with 86 per cent reliability. Across the three experiments, the highest level of precision recorded was for the condition in Experiment D where three or four attributes were all acting as perfectly correlated signals of an identical value. In this case, a surplus of 13 per cent of the

price range was still required. In approximate terms, this is equivalent to being able to discriminate just less than eight levels of value of the product. It is interesting to note that this finding closely parallels previous psychological studies of the accuracy of absolute identification following learning with stimuli drawn from perceptual continua (Dodds et al., 2011). This finding supports the conclusions of Chapter 3 that resolving a trade-off requires absolute rather than relative judgement and that the integration of information from incommensurate scales limits the degree of precision possible in such absolute judgements.

Once additional attributes must be accounted for when judging the surplus, precision declines further. The experiments in this chapter show that when attributes contribute to a product's value in a simple, additive, linear fashion, the extra attribute information cannot be integrated efficiently into decisions. The result is that the precision of surplus identification declines sharply. With four attributes to cope with, individuals can reliably spot a surplus only when the value of the product and the price differ by around half the entire price range. When highly numerate individuals with high educational attainment are given extensive practice, including exposure to examples and feedback that exceeds what would be realistic in most real markets, the level of precision improves, but only marginally. Although we find the extent of imprecision in surplus identification to be somewhat surprising, it is not inconsistent with the body of work in psychology reviewed briefly in Section 3.1 above.

Perhaps more surprising still is the finding that the identification of surpluses is subject to a strong and systematic bias across the price range, which persists despite extensive practice and feedback. Overall, there is a consistent but slight tendency to exaggerate surpluses, but this effect is small in comparison to how it varies over the price range. For single-attribute products, surpluses at the bottom of the range are underestimated while those at the top end are overestimated. This pattern changes with the number of attributes, however. When attributes are additive, the pattern diminishes with additional attributes and can even reverse once there are four. Yet when the attributes are perfectly correlated the bias strengthens when additional attributes are in play. This pattern in the biases supports the view that higher precision in surplus identification may come at the cost of bias, caused by adaptive psychological mechanisms that tune themselves to discriminate between surpluses.

An alternative possibility might be that these patterns indicate the use of (at least) two types of information integration, one that averages attribute magnitudes and another that adds them. Thus, when the experiment is designed such that one of these systems points to the veridical surplus, responses are biased in the direction of the other. That is, when adding gives the right answer,

as in experiments C and D, responses are biased towards averaging; when averaging gives the right answer, as in Experiment D, responses are biased towards adding. While this provides a possible explanation for how the extent of bias changes when attributes are added, it does not explain why there is a strong bias across the price range for a single-attribute product in the first place.

All of the experiments covered in this chapter involved novel products, visual attributes, linear returns to these attributes and judgements of whether single products conferred a surplus at a displayed price. Further experiments are conducted in the following chapters to determine whether changing any of these aspects of the experiments might alter the results.

Section 5

Can Consumers Cope Better with Categorical and Numeric Attributes?

5.1 INTRODUCTION

All of the experiments in the previous chapters involved visual attributes. While many products consist primarily of attributes the magnitudes of which must be judged visually, many products of interest for the present research programme do not. Attributes are often expressed as numbers, such as interest rates, fuel efficiency, service allowances, dimensions, ages, and so on. Attributes are also often expressed as categories, such as quality ratings, model types, colours, or perhaps most simply, brands. The experiments presented in this chapter were designed to assess whether the accuracy of surplus identification is affected by the need to integrate information from numeric and categorical attributes, as opposed to visual ones.

There were good reasons to begin investigation of the accuracy of surplus identification with products consisting of only visual attributes. From a purely scientific standpoint, the use of visual attributes helped to isolate the psychological mechanisms that integrate incommensurate scales. Had categorical or numeric attributes been involved, it may have been possible to exploit alternative mechanisms such as arithmetic rules of thumb ('assume category A is worth €120 more than category B', 'an increase of 3 per cent means a price increase of €10', etc.). Furthermore, if a product consisted of a single numeric or categorical attribute, the method of repeated forced-choice judgements employed here would have failed, because over many observations it would have been possible to remember precise associations between a category or number and a price. Such mechanisms of memory are of course scientifically interesting, but they may not be relevant to the problems associated with complex products that consumers face. Consumers must integrate attribute information on the basis of the more limited, fragmented and less frequent experience and feedback provided in a real market. Once a categorical or numeric attribute must be integrated with at least one other attribute, this concern fades, however, because it is no longer straightforward to learn one-to-one mappings between the attribute and the price. Consequently, the primary method employed in this chapter is to compare performance in two-attribute S-ID tasks when the type of attribute involved is manipulated.

There are reasons to think that surplus identification might be more accurate when attributes are numeric or categorical. It is true that, in Experiment A, no

relationship emerged between the precision of surplus identification based on visual attributes such as size and texture and the precision with which individuals could discriminate relative magnitudes of these attributes when they were presented side by side. This suggested little role for perceptual error in surplus identification with visual attributes. The magnitudes of numeric and categorical attributes can be made perfectly precise. Thus, unless a plainly seen category is somehow misidentified or a number misread, the attribute information is subject to no perceptual error at all. This may make it easier to integrate information from such attributes. Furthermore, because the information is free from perceptual error and the need for extensive perceptual processing, it may be that the presence of numeric and categorical attributes will simply reduce the cognitive load experienced by participants and make the task less cognitively demanding.

5.2 EXPERIMENT E: AIMS AND METHODS

The most basic starting point for investigating the question at hand is to adapt the design of the previous experiments by adding a numeric or categorical attribute to the egg hyperproduct. This approach forms the basis of Experiment E.

Aims

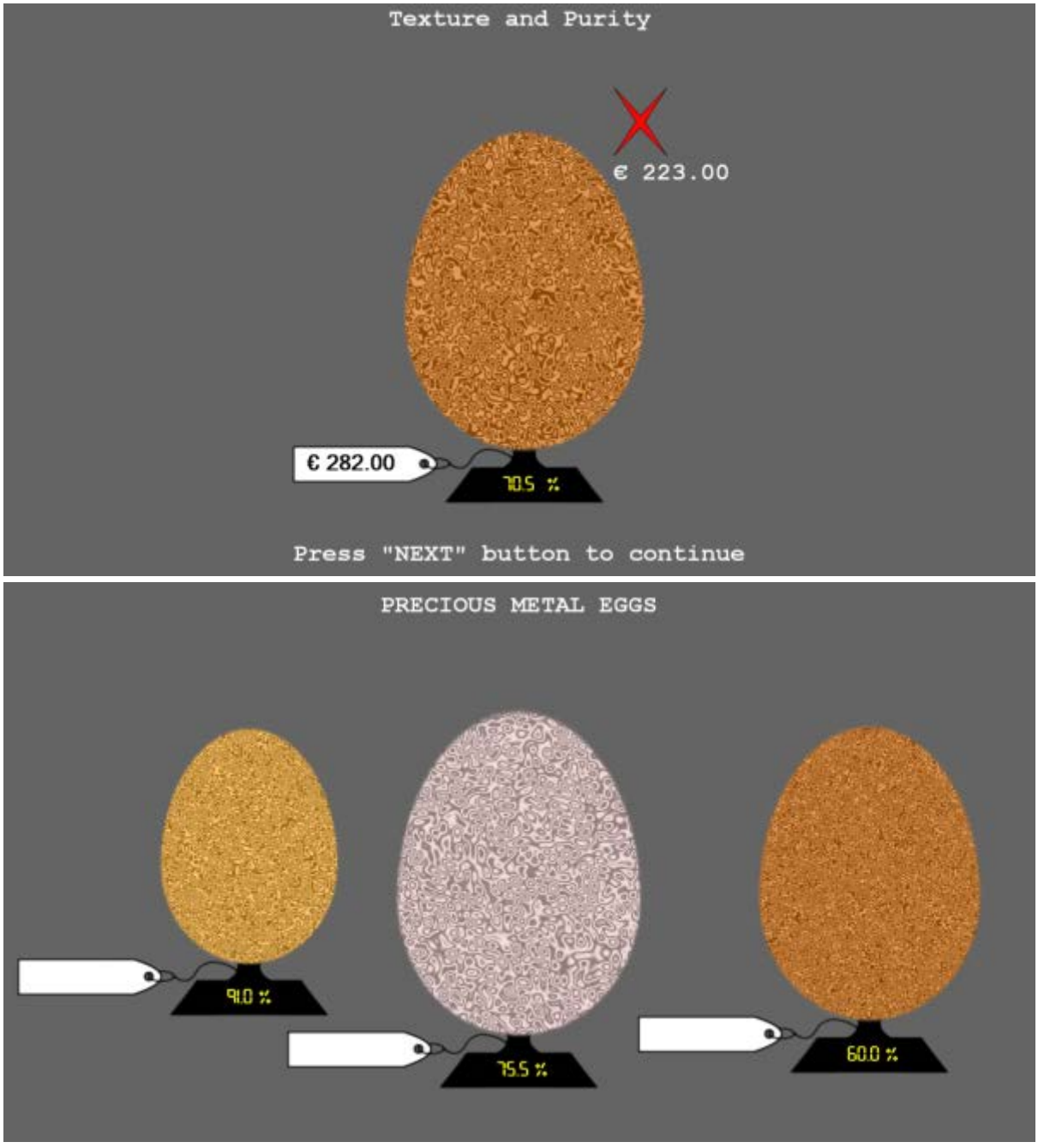
The primary aim of Experiment E was to test whether accuracy in a two-attribute S-ID task is improved by the use of a categorical and/or numeric attribute, rather than a visual one. To this end, the experiment made use of a categorical attribute (gold, silver or bronze) and a numeric one (percentage purity). These were tested alongside the two visual attributes: size and surface texture. In addition to investigating whether the use of categorical and numeric attributes would improve the precision of surplus identification, it was of interest to see what impact, if any, it would have on the bias. One possibility is that the bias recorded in Experiments A-D is unique to visual attributes, perhaps because of non-linearities in the way perceptual attributes are coded by the brain. Experiment E tested this hypothesis.

Methods

The methods were closely similar to previous experiments. The primary difference was the use of two additional attributes. Screen grabs to illustrate these are provided in Figure 13. The eggs could differ not only in size and texture, but in percentage purity, which was written on the little plinth on which the egg stood, or in metal type, which could be gold, silver or bronze (top panel). Percentage purity was loosely based on carat gold, in that it varied from 55 to 96 per cent purity. The ranges of size and texture were as for Experiment A: 356 to 708 pixels and 0.018 to 0.142 cycles/pixel. The price range was set to €177 to

€423. One step change in the category of metal type was worth €60, with an average bronze egg worth €225, an average silver one worth €285 and an average gold one worth €345. Although there was a strong correlation between metal type and the value of the egg, there was no correlation between the metal type and the surplus.

FIGURE 13 Example attributes and products from Experiment E. Eggs could be gold, silver or bronze, with obvious correspondences to value. The stand on which the eggs stood could display a percentage purity for the metal. Eggs could also vary in texture and size.



On any given trial, as in previous experiments, a single egg appeared together with a price tag attached to the plinth. The bottom panel of Figure 13 depicts a

trial at the feedback stage following the response. The egg was bronze with a fairly average texture and a price tag of €282. The participant pressed the right button to indicate that the egg was worth more than the displayed price, when in fact it was worth just €223. Feedback was given via a beep and a red cross when the response was incorrect and a green tick when it was correct. The correct monetary value was always given as shown in the figure.

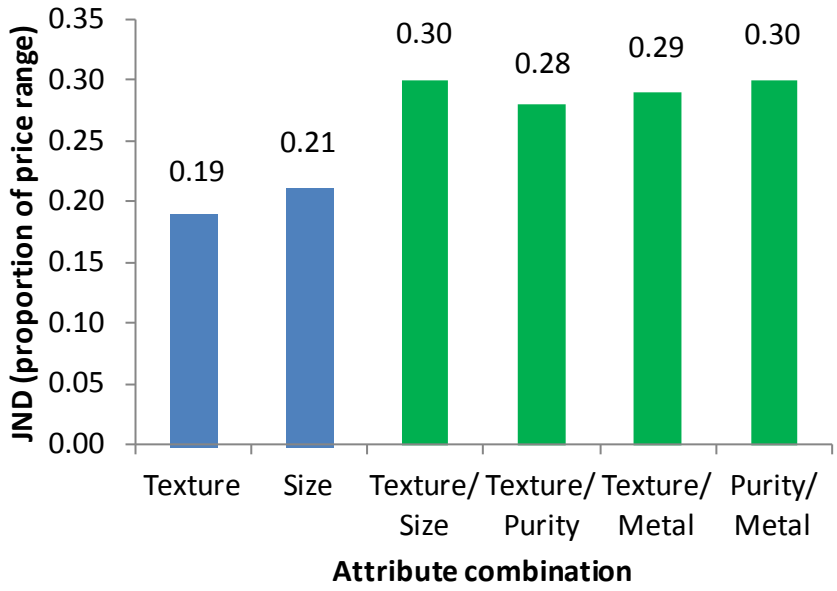
Thirty-six participants undertook six experimental runs, two single-attribute tasks and four two-attribute tasks: size only, texture only, texture and size, texture and purity, texture and metal type, purity and metal type. The two single-attribute conditions always appeared first in a pseudo-randomised order, followed by the size and texture condition, then the other three conditions in a pseudo-randomised order. This order was chosen to assist participants in understanding the tasks. As in previous experiments, each experimental run was preceded by a series of example eggs and a reminder of the relevant attributes remained at the top of the screen throughout the run. The AMCS procedure was preferred to the staircase procedure. Each run was 76 trials long, consisting of four easy practice trials followed by the first of six blocks of twelve trials, as in Experiment D.

5.3 EXPERIMENT E: RESULTS

Precision

JNDs for each of the six conditions are supplied in Figure 14. The JNDs for the single-attribute conditions are in line with previous experiments, at approximately 20 per cent of the price range. Similarly, the JND of 30 per cent for the condition with two visual attributes is in line with previous experiments. Somewhat surprisingly, however, surplus identification for the remaining three combinations of visual, numeric and categorical attributes displayed no statistically significant improvement in precision. This included the condition with no visual attribute at all (Purity/Metal).

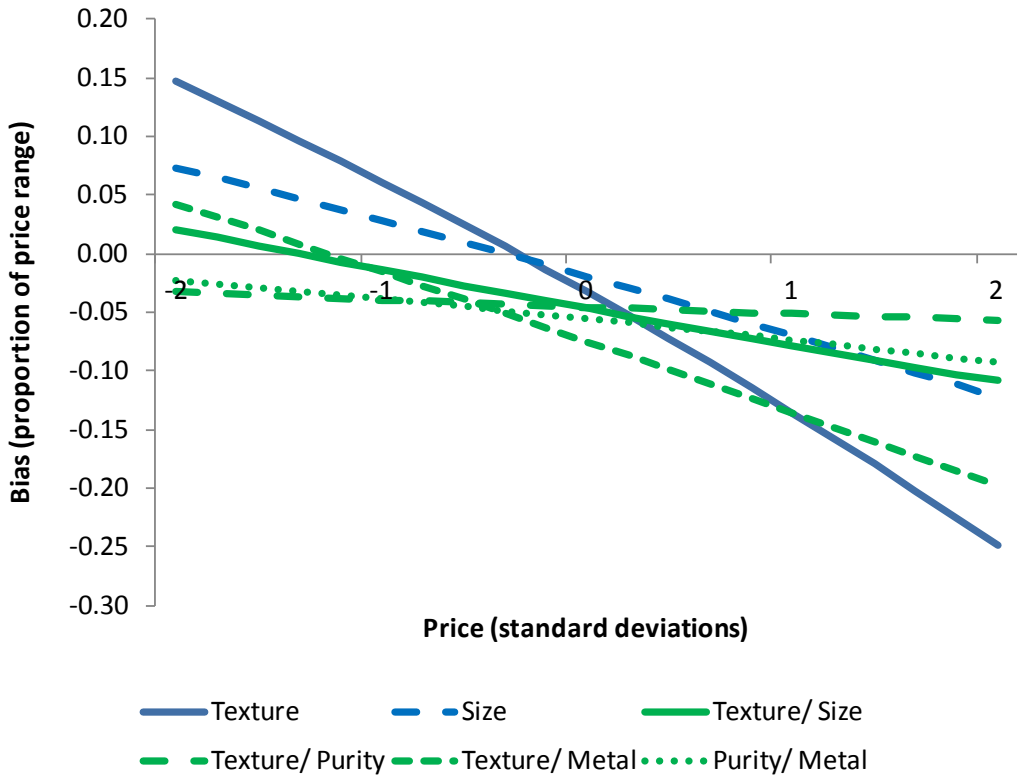
FIGURE 14 JNDs in Experiment E for multiple conditions with one visual attribute (blue) or two attributes consisting of different combinations of visual, numeric and categorical attributes (green). None of the differences between the JNDs for the two-attribute conditions are statistically significant.



Bias

The pattern of biases across the price range is presented in Figure 15. As previously, participants tended to overvalue the product relative to the price, although the larger effect was to exaggerate surpluses at the top of the range and to underestimate them at the bottom. This variation across the price range was again stronger on average for the single-attribute conditions than for the two-attribute conditions. The strongest bias in the two-attribute conditions was for the Texture/Metal type combination. The downward slope of the bias across the price range was statistically significant in the Texture/Size, Texture/Metal and Purity/Metal conditions, and did not significantly differ between them. However, the equivalent slope for the Texture/Purity was significantly flatter than for the Texture/Size condition and not significantly different from zero. Hence, there is some suggestion that the extent of bias may be reduced somewhat by the numeric attribute in this case, although some element of bias existed in all conditions, i.e. all curves departed significantly from a horizontal line at zero bias.

FIGURE 15 Bias across the price range by number of attributes in Experiment E. The variation in bias strengthens with increased numbers of attributes. Positive bias indicates undervaluation of the egg relative to the price; negative bias implies overvaluation.



5.4 EXPERIMENT E: DISCUSSION

The results of Experiment E are somewhat surprising. The complete removal of any perceptual error associated with the attribute magnitudes apparently made no difference to the precision of surplus identification. Furthermore, if categorical and numeric attributes impose less cognitive load than visual attributes, there was no sign of it in the data. Indeed, we expected at least some small improvement in performance, yet did not record one.

The findings of Experiment E are also potentially important, because they suggest that the results of earlier experiments are more likely to generalise to products in key markets. Offerings in telecommunications, financial services, energy, electronics, automobiles and more are frequently described by categorical and numeric attributes. The suggestion of Experiment E is that where offerings in such markets have multiple important attributes, surplus identification is likely to be imprecise and biased.

In one condition with a numeric attribute (Texture/Purity) there was a hint of a reduction in the bias across the price range. However, it is unclear why only this

condition showed an improvement. Moreover, the significant bias in the Purity/ Metal condition implies that the bias across the price range is not confined to visual attributes, and hence not caused by non-linearities unique to perceptual processing.

Overall, the rather stark message about consumer capabilities to emerge from Experiment E warrants additional investigation. Experiment F was designed to do everything possible to locate an improvement in performance associated with numeric and categorical attributes.

5.5 EXPERIMENT F: AIMS AND METHODS

Aims

In simple terms, the purpose of Experiment F was to design an alternative version of Experiment E that might improve relative performance with numeric and categorical attributes. First, we chose a different product with completely different attributes. Second, we removed the single-attribute conditions. Third, we tested whether performance differed when the same attribute was described only visually or both visually and numerically. Fourth, we reduced the categorical attribute to just two categories, in the hope that this would reduce cognitive load. Fifth, we thought it possible that an additive relationship between categories and value might not be as intuitive as a proportional one (i.e. where the category increases the value by a percentage rather than a Euro amount), so we tested both.

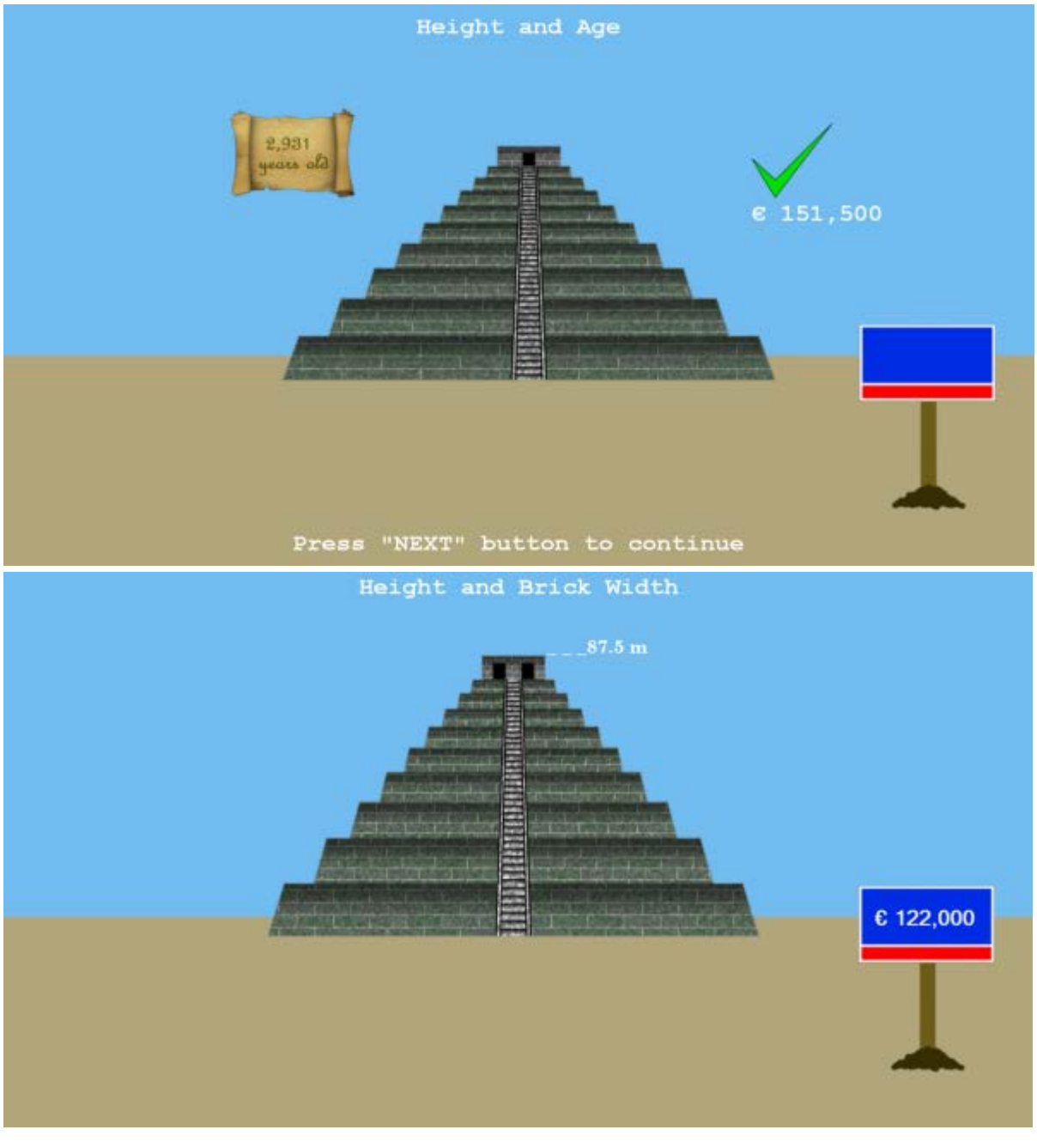
Methods

Methods were as for Experiment E but with the following modifications. Experiment F introduced a new hyperproduct: a Mayan pyramid. The attributes of the pyramid were its height, the shape of its bricks (rectangular aspect ratio), its age and the number of doors at the top (one or two). Taller pyramids, those with more rectangular (less square) bricks, older pyramids and those with two doors rather than one were more valuable. As with the golden eggs, the overall value of the product was determined by the combination of attributes – a two-door pyramid could be worth less than a one-door pyramid which was superior on other attributes, such as age. Height varied between 50 and 90 metres (with a one-to-one mapping with screen pixels); the aspect ratio of the brick varied between one and seven; age varied between 976 and 4,024 years old; the price varied between €39,000 and €193,000.

There were five conditions: Height (visual) and Brick; Height (visual and numeric) and Brick; Height (visual) and Age; Height (visual) and Doors; Age and Doors.

Example screen shots are shown in Figure 16. The top panel shows the Height and Brick condition where the numeric height information was also provided. The displayed price appeared on an estate agent style sign to the right of the pyramid. The bottom panel shows the Height and Age condition at the feedback stage, which was as for Experiment E. Age appeared on a scroll to the left of the pyramid. The five conditions were pseudo-randomised across participants. Because there was no single-attribute condition, in addition to being shown multiple example products at the start of each run, participants were given 12 practice trials before completing 72 test trials.

FIGURE 16 Example attributes and products from Experiment F. Pyramids varied in height, brick shape (aspect ratio), age and the number of doors. Height could be stated numerically as well as visually.



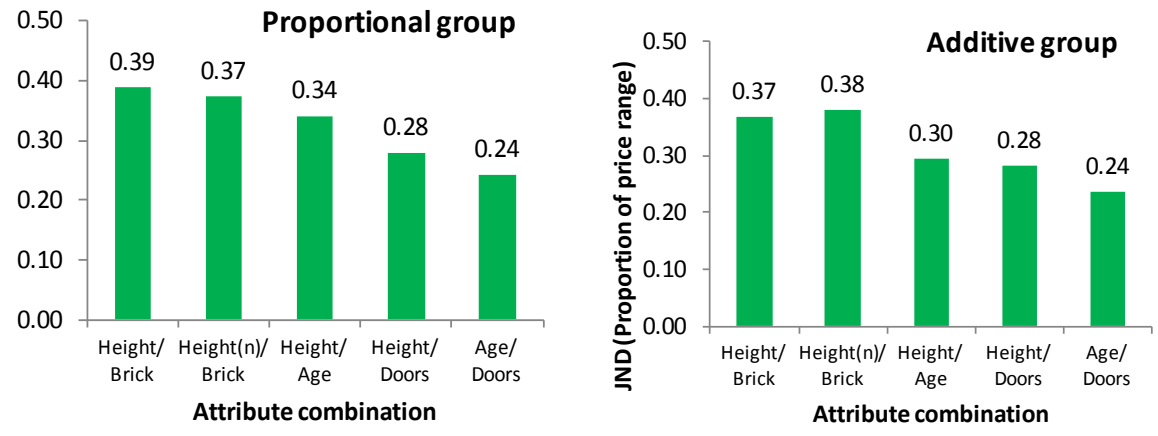
There were 40 participants who were each paid a €20 fee. The four most accurate participants won a €50 shopping voucher. Participants were split into two groups of 20. For the 'additive' group, pyramids with two doors were worth €52,000 more than pyramids with one door. For the proportional group, possessing two doors increased the value of the pyramid by 20 per cent relative to possessing just one door.

5.6 EXPERIMENT F: RESULTS

Precision

Figure 17 provides the JNDs for the two groups and the five conditions. Five aspects of these results require comment. First, participants clearly struggled with the aspect ratio of the brick. The JND of 37-39 per cent for just two attributes was higher than that recorded in any two-attribute condition in previous experiments. Second, adding the numeric height (Height/Brick versus Height(n)/Brick) was of no additional benefit for surplus identification. Third, performance was better in the Height/Age condition than the Height/Brick condition, but only because it was relatively poor in the latter. The JND, measured as a proportion of the price range, was no lower than in the two-attribute conditions in previous experiments. Fourth, the categorical attribute with just two categories (one or two doors) did produce a statistically significant increase in precision relative to the other two-attribute conditions, especially in combination with the numeric attribute. The effect size was, however, small, producing at best a JND somewhere between what would be expected from a single-attribute S-ID task and a two-attribute S-ID task. Fifth, there was no difference in precision between the conditions with additive and proportional categorical attributes.

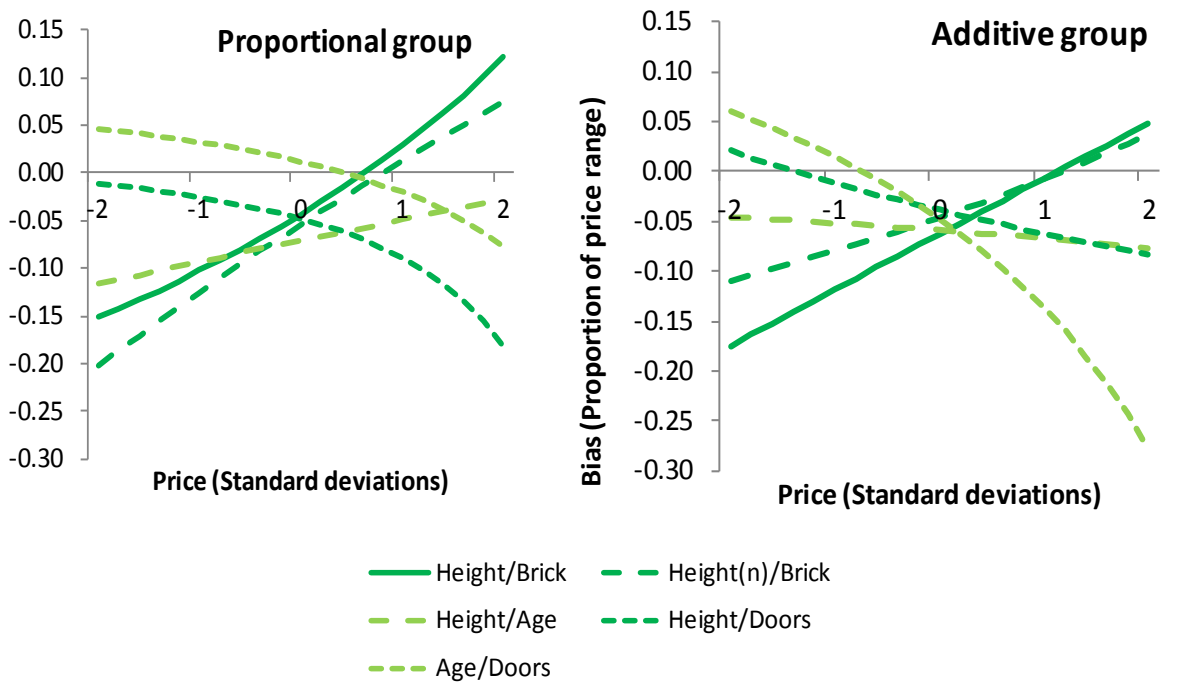
FIGURE 17 JNDs in Experiment F for multiple two-attribute conditions, combining visual (Height, Brick), numeric (Height(n), Age) and two-category (Doors) attributes. For half the participants the two-category attribute was additively related to product value; for half it was proportionately related to product value.



Bias

Figure 18 plots the biases across the price range for the two groups. Again, there were strong and systematic biases, including the same slight overall bias towards exaggerating the surplus. The results were somewhat different from those of previous experiments, however. Most notably, in the conditions where brick shape had to be factored in, the bias had the opposite slope to that seen previously in two-attribute S-ID tasks. Closer investigation suggested that participants had underweighted the brick shape in their responses, leading them to exhibit a form of mean reversion; i.e., their assessment of the value of the pyramid was biased towards the mean value. Adding the numeric value of the height made no significant difference to this bias.

FIGURE 18 Bias across the price range by condition in Experiment F. Positive bias indicates undervaluation of the egg relative to the price; negative bias implies overvaluation.



The remainder of the pattern of biases was not entirely dissimilar to that seen in previous experiments. There appeared to be no systematic effect of numeric or categorical attributes on the scale of the bias. Indeed, the only clear pattern to emerge was, once again, that those conditions that produced the highest levels of precision in surplus identification also produced the strongest bias towards overvaluing products at the top end of the range and undervaluing those closer to the bottom of the range. An additional aspect of the data was consistent with this relationship. Although the condition was identical, the 20 participants in the 'additive' group happened to be marginally (but statistically significantly) more precise in the Height/Age condition (Figure 17). Their bias in this condition also

had a downward slope, whereas the less precise 'proportional' group produced an upwards slope, a difference that was again statistically significant. Lastly, while the two groups did not differ with respect to precision in the conditions where the number of doors had to be taken into account, they did differ in these conditions with respect to the bias. The proportional group was less inclined to be biased towards overvaluation, especially towards the top end of the price range, perhaps because they underestimated the multiplicative impact of the two-door category.

5.7 DISCUSSION

Experiment F succeeded in generating an improvement in the precision of surplus identification with a multi-attribute product when one attribute was reduced to just two categories. However, the size of this effect was small. This result, to a large extent, makes sense. Once there are just two categories for an attribute, the two-attribute S-ID task effectively becomes two separate single-attribute S-ID tasks, with the trials interleaved and the price range for one task shifted upwards, as signalled by the categorical attribute. That is, the number of doors effectively signals which of two different price ranges the continuous attribute should be mapped onto. Performance in this task does not reach the levels of precision typically seen for single-attribute S-ID tasks, suggesting that the need to process this binary signal, and/or to consider two separate ranges, imposes a degree of cognitive load, even if this is less than the load imposed when the magnitude of a second continuous attribute must be integrated into the decision.

Experiment F provides further evidence, meanwhile, that the accuracy of information integration with numeric attributes is no better than with visual ones. While there was a hint in the data of Experiment E that numeric attributes might reduce the degree of bias, this result did not carry through to Experiment F. The strongest bias was in fact observed in a condition with one numeric and one categorical attribute (Age/Doors, 'additive').

Overall, Experiments E and F primarily show that the degree of inaccuracy in consumers' assessments of surpluses generalises to situations where products possess numeric and categorical attributes. This has clear implications for the generalisation of the results of the S-ID task to real markets, suggesting that the cognitive limitations identified are likely to apply in key consumer markets.

Section 6

Can Consumers Handle Non-linear Attribute Returns?

6.1 INTRODUCTION

In all of the experiments described so far, attributes have been related to prices in a simple linear fashion. A given difference in attribute magnitude has always translated into the same price difference, regardless of where it occurred in the price range. However, in many markets there can arise products that are complex not because of the number or type of attributes, but because the attributes are related to price in a highly non-linear way. Non-linear pricing structures are a common feature of economic life, yet some empirical evidence suggests that consumers may have particular difficulty when trying to identify surpluses in the presence of such non-linearities.

For example, with financial products, balancing risk and return is a non-linear trade-off. For some common financial attributes, such as the APR on a credit product, the compounding of interest is a difficult concept to grasp. Consumers typically struggle to compute the total cost of a loan accurately (Lusardi and Mitchell, 2011). When stating willingness to pay for fuel efficiency in the car market, consumers fail to appreciate the non-linear nature of measures such as Miles-Per-Gallon (Larrick and Sol, 2008). Similarly, in the telecommunications market, a substantial number of mobile phone consumers fail to choose the lowest cost tariff for their personal pattern of usage from a limited number of non-linear price plans (Grubb 2009; Lambrecht and Skiera, 2006).

6.2 EXPERIMENT G: AIMS AND METHODS

Aims

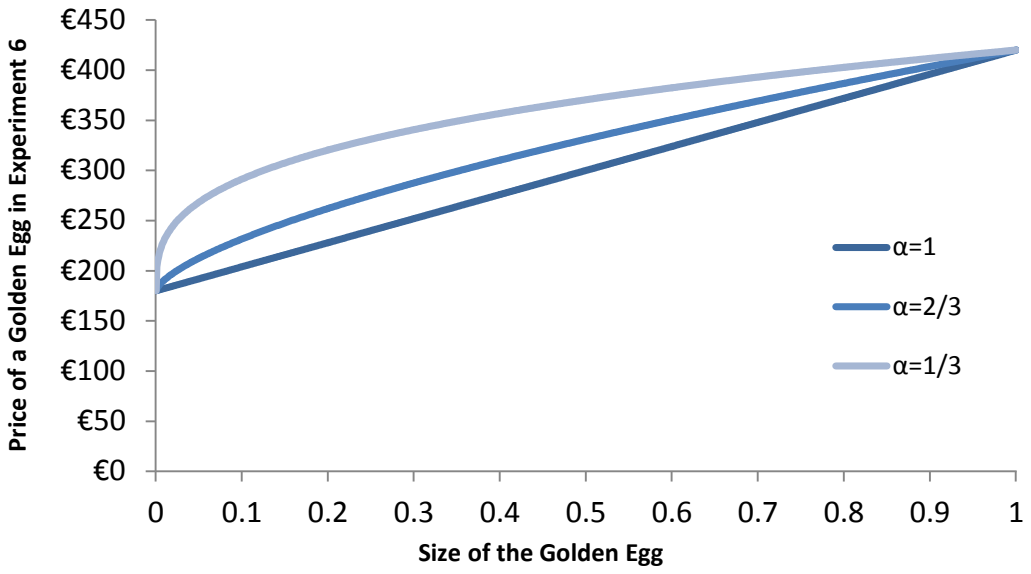
Experiment G aimed to generate baseline measures of accuracy when the value of an attribute was subject to some non-linearity. The simplest test of surplus identification with non-linear returns involves a single-attribute product with an increasing relationship between attribute and price. The non-linear structure we imposed was 'diminishing returns', which implies that the value of an additional unit of the attribute declines as the total amount of that attribute increases. Diminishing returns are common in economic analysis and are an intuitive concept: the value a consumer places on a second sandwich will generally not be as high as the first; the benefit of an extra 1GB of data is large when the current allowance is only 1GB, but largely insignificant when it is already 50GB.

The value of a hyperproduct in Experiment G was determined by the following equation:

$$V = \beta_0 + \beta_1 x_1^{\alpha}$$

where χ_1 was the attribute magnitude, the value of β_0 and β_1 varied according to the price range of the different hyperproducts and the value of α determined the degree of diminishing returns. For instance, Figure 19 shows how the size of a golden egg would affect its overall price for different values of α. In this chart, size has been normalised so that a size of zero equates to the smallest egg in the range and a size of one equates to the largest. The minimum price was €180 (β_0 for the golden egg) and the maximum price was €420 (i.e. β_1= €240). For $\alpha = 1$, the equation reduces to the simple linear case employed in all previous experiments. This is indicated by the dark blue line in Figure 19. Changes in size map onto constant changes in price. When we reduce α to 2/3, moderate diminishing returns kick in, so that the change in price due to an increase in size is greater when the egg is small than when it is large. For $\alpha = 1/3$, this pattern is amplified to produce severe diminishing returns: minor changes in size of the egg have a large effect on price when the egg is small, but very little impact when it is already large.

FIGURE 19 An example of the varying degrees of diminishing returns employed in Experiment G.



Method

The methods employed in Experiment G were identical to those in Experiment B, apart from the following modifications. First and foremost, consumers identified surpluses on three hyperproducts: golden eggs, Mayan pyramids and Victorian lanterns. Although each experimental run required just one attribute to be taken

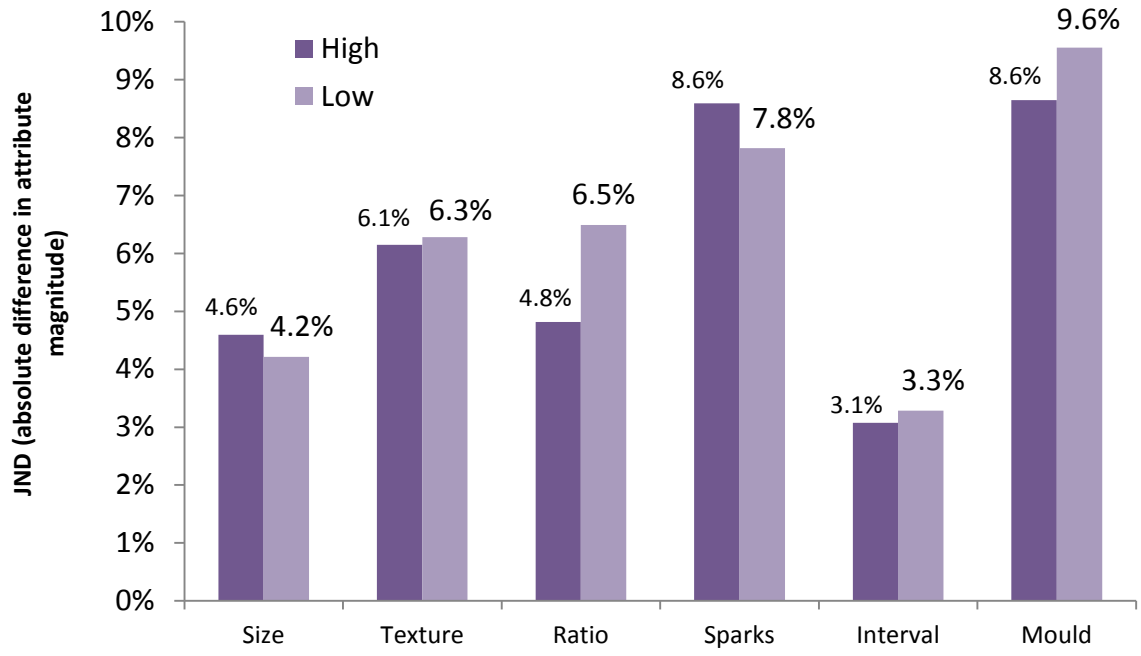
into account, two separate attributes were tested for each hyperproduct, making a total of six in all. For the golden egg, we again employed size and surface texture. For the Mayan pyramid, we used the width of the staircase and the mouldiness (level of green saturation) of the bricks (see Figure 1). For the Victorian lantern, we used the ratio of the inner blue flame to the overall flame and the number of sparks emitted from the base.

There were 36 participants, who each undertook six experimental runs of 72 trials. On each trial the value of the hyperproduct was determined by the equation above. The main experimental manipulation of interest was the strength of the diminishing returns: linear ($\alpha=1$), moderate diminishing returns ($\alpha=2/3$) or severe diminishing returns ($\alpha=1/3$). These conditions were pseudo-randomised across participants and attributes such that each participant completed two runs for each condition, with the proviso that the two attributes of each hyperproduct corresponded to different degrees of diminishing returns. For each hyperproduct, the price range was always the same: €180-€420 for the Golden Egg; €7-€35 for the lantern; €23,000-€172,000 for the pyramid. However, because processing non-linearities accurately might require a greater range of attribute magnitudes than processing linear relationships, we tested two ranges of magnitudes of each of the hyperproducts, one 'high' and one 'low', which differed by a factor of two. This was also pseudo-randomised across conditions, hyperproducts and participants.

Pilot Study

Prior to the main experiment, a pilot study with 26 participants checked the perceptual ability to discriminate the attribute magnitudes when two products were placed side by side. As with experiment A, the aim was to be sure that we were testing the ability to identify surpluses rather than how well people discriminate perceptual magnitudes. Both high and low ranges were used. The results are presented in Figure 20, with the attribute ranges normalised to run from zero to one, to allow comparability across different attributes. All six attributes were discriminated with high accuracy, with differences of less than 10 per cent perceived reliably. There was variability across the attributes, however, with discrimination of the relative magnitudes of some attributes being more than twice that for others.

FIGURE 20 Results of the pilot study for Experiment G. JNDs for discriminating magnitudes of the attributes when two products were placed side by side, showing substantial variation by attribute.

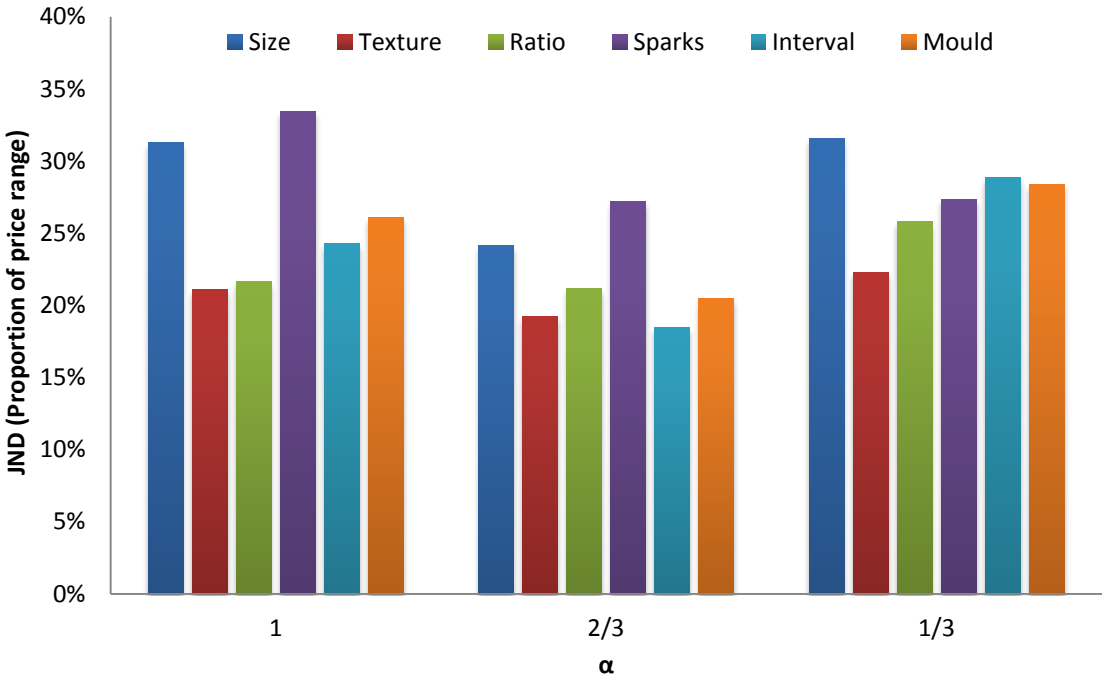


6.3 EXPERIMENT G: RESULTS

Precision

Figure 21 presents the average JND, broken down by attribute, for each of the conditions of diminishing returns. Three findings are of note. The first is the level of absolute performance. Similarly to previous experiments, even with just a single attribute, a surplus of at least 18 per cent of the price range was required for it to be identified reliably. Second, while for some attributes and conditions the JND was considerably higher, this bore little relation to the JNDs for discrimination of magnitudes recorded in the pilot experiment, suggesting that the source of imprecision was post-perceptual, confirming the result we first saw in Experiment A. Third, the non-linearity in the attribute-price relationship had little impact. In fact, there was even a slight advantage for the moderate diminishing returns case ($\alpha=2/3$).

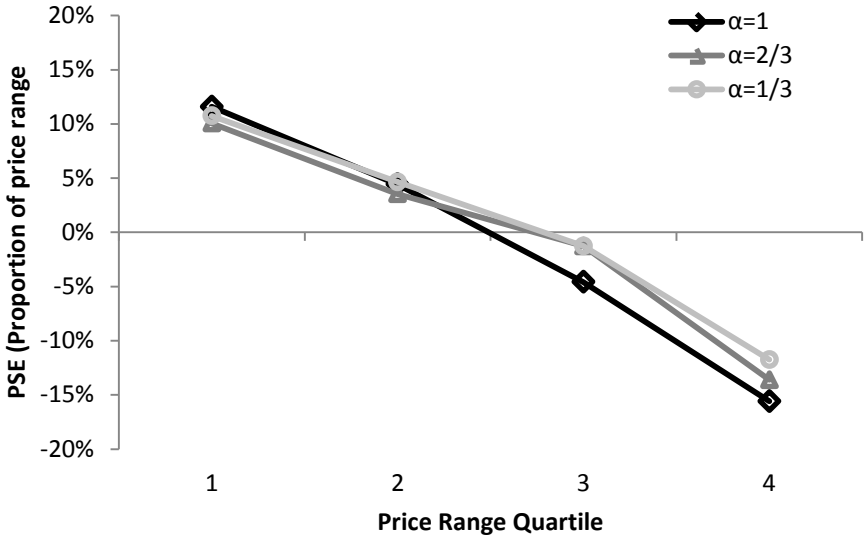
FIGURE 21 JNDs for linear and non-linear attribute-price relationships in Experiment G. Although surplus identification is imprecise, there is no disadvantage associated with non-linearity.



Bias

The pattern of biases across the price range is shown in Figure 22. At this point, the pattern is familiar. There was a slight overall bias, such that surpluses tended to be exaggerated, but the larger effect was that participants underestimated surpluses towards the bottom of the price range and overestimated them towards the top.

FIGURE 22 Biases across the price range in Experiment G. The linear and non-linear cases are indistinguishable.



6.4 EXPERIMENT H: AIMS AND METHODS

Aims

In Experiment G, precision was robust to the introduction of simple non-linear returns. Given enough examples and feedback, consumers can handle these types of non-linearity. In reality, however, the products consumers encounter on a daily basis are more complex. Experiment H increased the complexity of the non-linear returns, first, by incorporating a second attribute and, second, by defining the attribute-price relationship via a range of more complex non-linear functional forms, which instead of simply increasing or decreasing at different rates also included turning points. Attributes that can be both too large and too small are common. Examples of attributes with turning points include portion sizes for food and drink, terms of loans, and engine sizes; consumers often seek a 'happy medium'.

Moving to a two-attribute space also allowed us to test additional hypotheses arising from Experiment G. First, adding a second attribute introduced a new source of complexity in the form of the relative attribute weighting. In previous experiments, all attributes contributed equally to the value of a product. In contrast, Experiment H tested whether performance was affected by different weightings. Second, in the case of the more complex two-attribute products, we hypothesised that learning might not be so complete following the initial examples and might therefore continue throughout the main experiment.

Methods

Methods were as in Experiment G, except for the following modifications. First, six additional attributes were employed, bringing the total number of attributes on each hyperproduct to four. Three more continuous attributes were added. On the golden egg, we used the angle of the hallmark, as in Experiment B. On the Victorian lantern, we employed the 'rustiness' of the metal, defined as the contrast of an orange-brown versus black coloured texture. On the Mayan pyramid, we used the rectangular aspect ratio of the bricks as in Experiment F. In addition, another three attributes, again one for each hyperproduct, were numeric and appeared on a label next to the product. For the golden egg, we displayed the purity in carats on the plinth; for the Victorian lantern, fuel efficiency on a 25-point gradient scale; for the Mayan pyramid, age in years on a scroll. The attribute-price relationship took one of six functional forms (henceforth we refer to these as 'value functions'), which are summarised in Table 2.

Twenty-four participants completed one run for each of the six value functions. On half of the runs, the attribute weights were balanced such that both attributes contributed equally to the overall price. For the other half of the runs, the attributes were unbalanced; one attribute contributed twice as much to the total price than the other. On half of the runs, one of the attributes was numeric, on the other half both were visual.

TABLE 2 The value functions used in Experiment H

Value Function	Description	Equivalence
Linear	Benchmark case; the same change in attribute leads to a constant equivalent change in value	Equivalent to Experiments A-F
Constant returns to scale	Doubling both attributes doubles the price, but individually each attribute exhibits diminishing returns	Equivalent to Experiment G
Increasing returns to scale	Increased complexity via increasing returns to scale, i.e. further improvements in attributes contribute more to the price than previous improvements	New
Perfect complements	The weakest attribute alone determines the price. Attributes must be compared to each other to identify the weakest, which must then be compared to the displayed price	New
Cyclical	One attribute has linear returns, the returns of the other are cyclical, i.e. increasing or decreasing depending on the part of the range	New
Goldilocks	The centre of the attribute range is the ideal, such that the average attribute is 'just right'	New

6.5 EXPERIMENT H: RESULTS

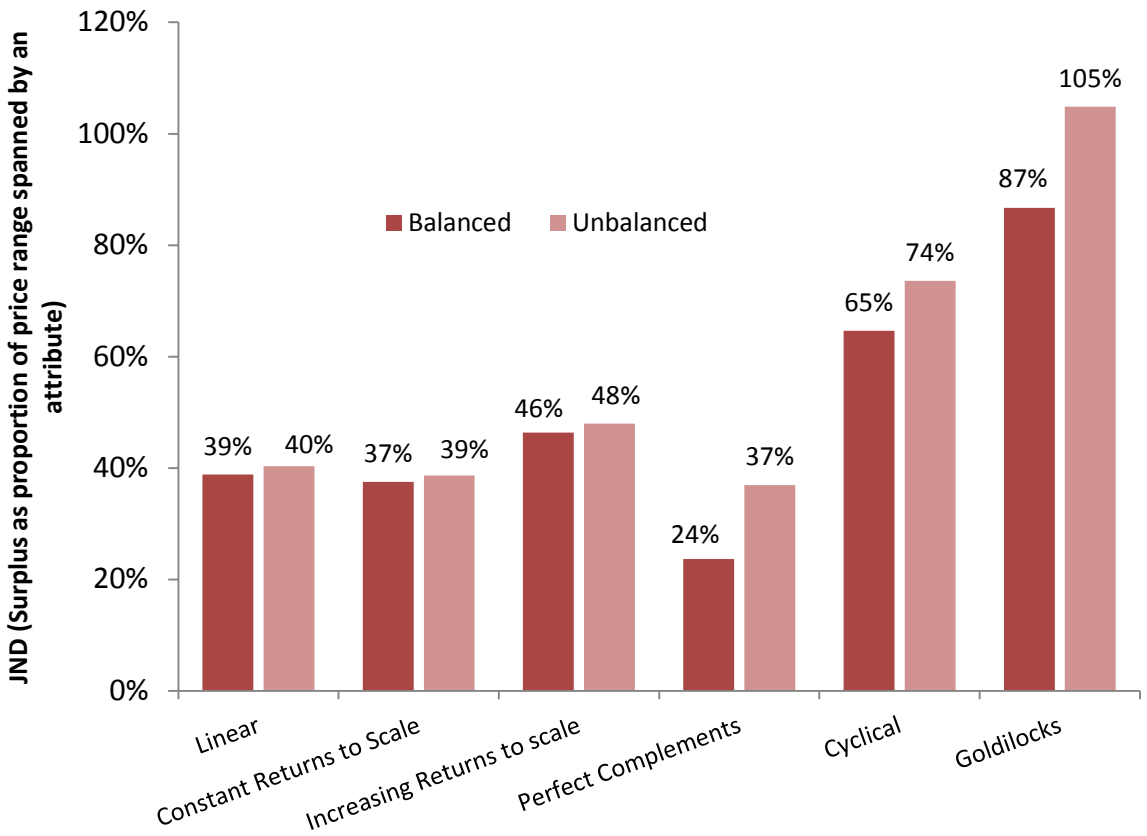
Precision

Figure 23 presents the average JND for each value function, split by whether the attributes were balanced or not.[7] For the simplest value functions (linear and constant returns) reliable identification required a surplus equivalent to one third to one half of an attribute range, compared to one fifth to one quarter for the single attribute case in Experiment G. While absolute performance for the linear case was somewhat below that of Experiment B, adding an attribute increased the surplus required for reliable identification from 24.7 per cent to 38.9 per cent, corroborating the previous finding of inefficient attribute integration (see also Figure 6). Precision was reduced slightly when the returns were increasing in scale, but the difference was not statistically significant. The perfect complements value function with balanced attributes was the only condition to produce JNDs as low as Experiment G. The more complex cyclical and goldilocks value functions resulted in substantially greater imprecision: for the cyclical

[7] In Experiment G, a single attribute range matched the full price range. In Experiment H, to accommodate the more complex non-linear functions, on average, the range of each attribute mapped on to only half the price range. Therefore, for comparison between the two experiments, the JNDs in Figure 23 are measured as a proportion of an attribute range, such that precision when an attribute was mapped to price on its own can be compared directly to precision when a second attribute was simultaneously taken into account.

function, surpluses needed to be more than 30 percentage points greater on average to achieve the same precision as the simple non-linear functions; for the goldilocks value function, surpluses needed to be 57.2 percentage points larger for comparable precision, i.e. more than double.

FIGURE 23 JNDs for a range of more complex two-attribute non-linear functions in Experiment H. Precision for the functions with turning points deteriorated markedly.



Although the complexity of the value function had a strong impact, variation in the relative weights of the attributes did not. Unbalanced attributes significantly disrupted the perfect complements case only. The additional complexity introduced by this value function primarily surrounded the need for participants to make two sequential judgements, first assessing relative attribute magnitudes, then the relationship of the weakest attribute to price. It is likely that the first stage was disrupted by unbalancing the attributes.

Bias

Surplus identification was again biased, with poorer products undervalued and better products overvalued. Despite large variation in the linearity, scale and

relative weight of the attribute returns, this bias was remarkably consistent across all value functions and appeared almost identical to Figure 22 (and is for this reason not repeated here).

Learning

It is entirely possible that the deterioration in precision for complex preferences merely reflects slower learning for difficult non-linear attribute-price associations. However, despite the added complexity of the task, we again found no evidence of improvement beyond the initial exposures; learning was rapid and confined to the practice trials. Imprecise surplus identification seems to be due to a genuine cognitive deficit that is not easily overcome.

Additional tests

Experiment G produced the richest data to date on surplus identification for product versus price comparisons with only one attribute, while Experiment H produced a richer variety of two-attribute decisions than previous experiments. These advantages make possible several additional tests for specific biases that have been commonly observed in subjective choice tasks.

One such bias is the 'attraction effect' (Huber et al., 1982). This refers to a tendency to overvalue a product that dominates another on all attributes. In Experiment H, we looked for a similar effect in comparison to the product that had immediately preceded the current one. For some trials the product was better on both attributes than the previous one, thereby dominating it. Conversely, for some others it was dominated by the previous product. Further analysis revealed that there was indeed an attraction effect: domination of the previous product exaggerated perceived surplus, while being dominated by it diminished perceived surplus.

The notion of 'loss aversion' was first formalised and investigated empirically by Kahneman and Tversky (1979). An extensive literature has explored asymmetries in how humans and animals weight losses and gains in choices. While explanations for these empirical phenomena remain controversial, replicable findings that imply loss aversion in economic choice are numerous. Each successive presentation in the S-ID task entails an increase or decrease in attribute magnitude relative to the previous product. Assuming that the most recently viewed product provides a reference point, loss aversion implies that participants might judge the surplus to be smaller on trials when the attribute magnitude decreases than those when it increases, all else being equal. Experiment G offers an ideal test, as successive presentations varied in a single attribute. We do indeed find evidence for loss aversion: the difference in

magnitude between successive attributes was exaggerated in both directions, but the effect was larger when the attribute magnitude implied lower value than the previous product than when it implied higher value.

6.6 DISCUSSION

Experiments G and H confirm and extend the results of the previous experiments. For both linear and non-linear returns, surplus identification is subject to important capacity constraints even when just one or two directly observable attributes are compared with a price. Following some instructive initial examples, performance for simple non-linear relationships, such as decreasing and increasing returns to scale, is comparable to that of the linear case. Once the complexity of the non-linearity is increased to include turning points, however, precision declines substantially. Moreover, despite the additional scope for learning with these complex products, precision did not improve with repetition, feedback and incentives.

Overall, the results from Experiment G and H imply that for a general class of preferences defined over a broad range of products and attributes, cognitive constraints bind performance. Consequently, surplus identification is prone to large and persistent errors. The generalisability of these empirical results to larger product ranges and to more familiar consumer products is a crucial question to which we turn next.

Section 7

Does Inaccuracy Generalise to Larger Ranges of Products and More Familiar Products?

7.1 INTRODUCTION

In the experiments described thus far, consumers were always faced with a binary choice. In most cases, this was a decision regarding whether or not a product conferred a surplus at a displayed price. Logically, it is possible that the picture might change in a more realistic setting where consumers could compare multiple products.

There are at least two reasons why decisions taken when faced with a product range might be more accurate. First, rather than combine the information from all the attributes into an overall assessment of value for each product, to be compared against its price, consumers might be able to make decisions based on differences between specific attribute magnitudes across the available product range. There is no reason to assume that multi-attribute decisions must be made by calculating the value of each options in isolation (Vlaev et al, 2011). Second, it is possible that the availability of more products in the range will help consumers to calibrate their internal scales for representing attribute magnitudes. Considering a product in isolation, with only memory representations against which to compare it, may produce less accurate representations than considering several products simultaneously.

Alternatively, however, it is possible that increasing the number of products in the range will only serve to increase inaccuracy. This could occur because the greater number of options increases cognitive load while making the decision, or because the availability of more options simply increases the probability of making a mistake when surplus identification is inaccurate.

7.2 EXPERIMENT I: AIMS AND METHODS

Aims

Experiment I sought to test whether surpluses are more accurately identified when products are part of a range. This was done by comparing the standard S-ID task with an adapted version of the task using the same hyperproduct, but in which participants had to decide which of two, three or four products conferred a surplus. Only one product in the range had a surplus and the participant's task was simply to spot which one. Because it is possible that presenting products

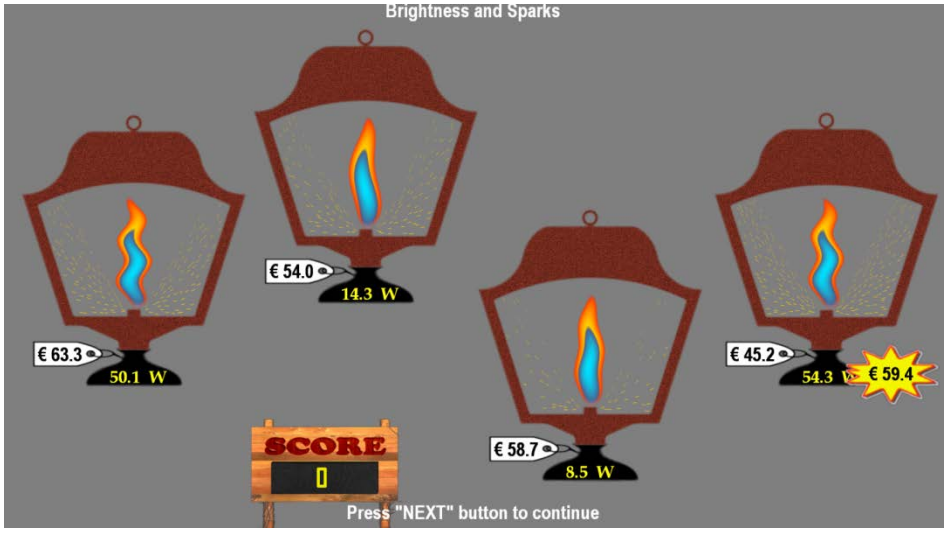
within a range might have different effects for visual and numeric attributes, one of the two hyperproducts had a numeric attribute.

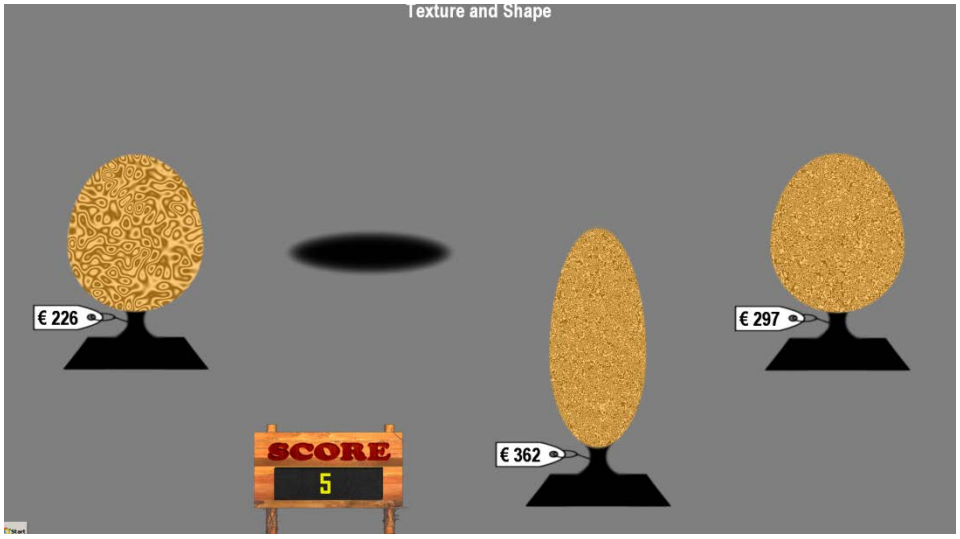
Methods

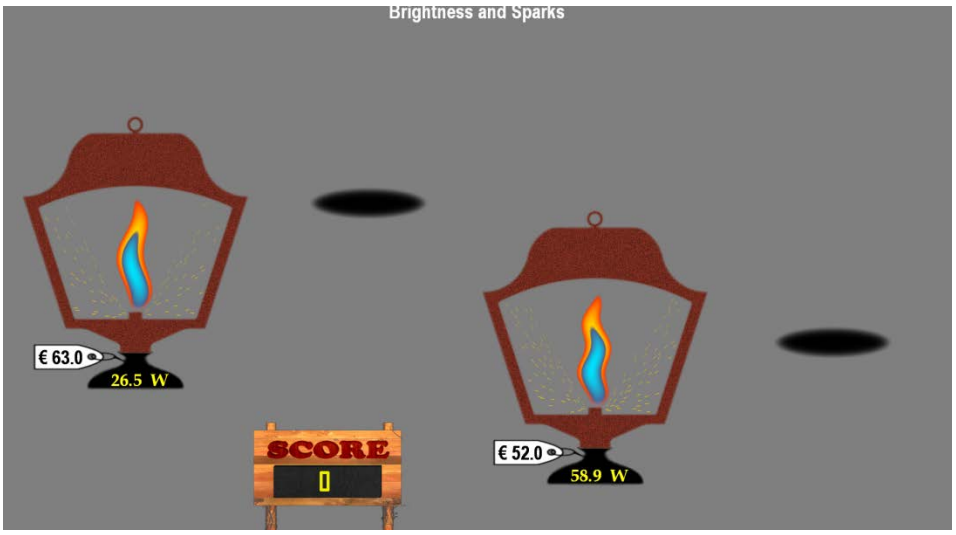
Participants completed four experimental runs. The first was a standard two-attribute S-ID task in which they had to decide whether a hyperproduct was worth more or less than a displayed price. The second was an S-ID task adapted for multiple products and prices, hereafter termed a multi-product S-ID task (MS-ID). In this task they were presented with two, three or four hyperproducts, each at a different displayed price, just as if they were standing before a shop display. Only one of the offerings conveyed a surplus; the value of each of the others was exactly the same as the price (i.e. the surplus was zero). No product in the range was dominated by another (i.e. no product was more expensive and had lower magnitude on both attributes than another). Hence, trade-offs had to be negotiated and the participant responded by pressing one of (up to) four buttons on the response box. The third and fourth experimental runs repeated the sequence with a different hyperproduct.

The hyperproducts used were the golden egg and the Victorian lantern. For the egg, the two visual attributes were the surface texture and overall shape. This latter attribute was similar to the circularity attribute in Experiment B, but the change of shape was applied to the whole egg. The numeric attribute was the percentage purity, as used in Experiment E, which replaced texture on half of the runs. For the Victorian lantern, the two visual attributes were the flame ratio (as used in Experiment G) and the number of sparks. The numeric attribute was the brightness expressed in Watts, which replaced the flame ratio. Two example screen grabs are provided in Figure 24.

FIGURE 24 Example screens from the MS-ID task in Experiment I. Participants had to decide which one of two, three or four hyperproducts conferred the surplus.







Experimental runs for the standard S-ID task were 64 trials. For the MS-ID task, the experimental run was 90 trials long; 30 trials had two products in the range,

30 had three and 30 four. These were randomly interleaved. To facilitate precise comparison, the surpluses were no longer decided by an adaptive method in the MS-ID task, but were at the same predetermined levels for all participants and conditions. Feedback was given via an audible beep and a 'bargain' sticker revealing the true monetary value of the product carrying the surplus (bottom panel). Each time a correct response was recorded a counter at the bottom of the screen added one point to the participant's 'score'.

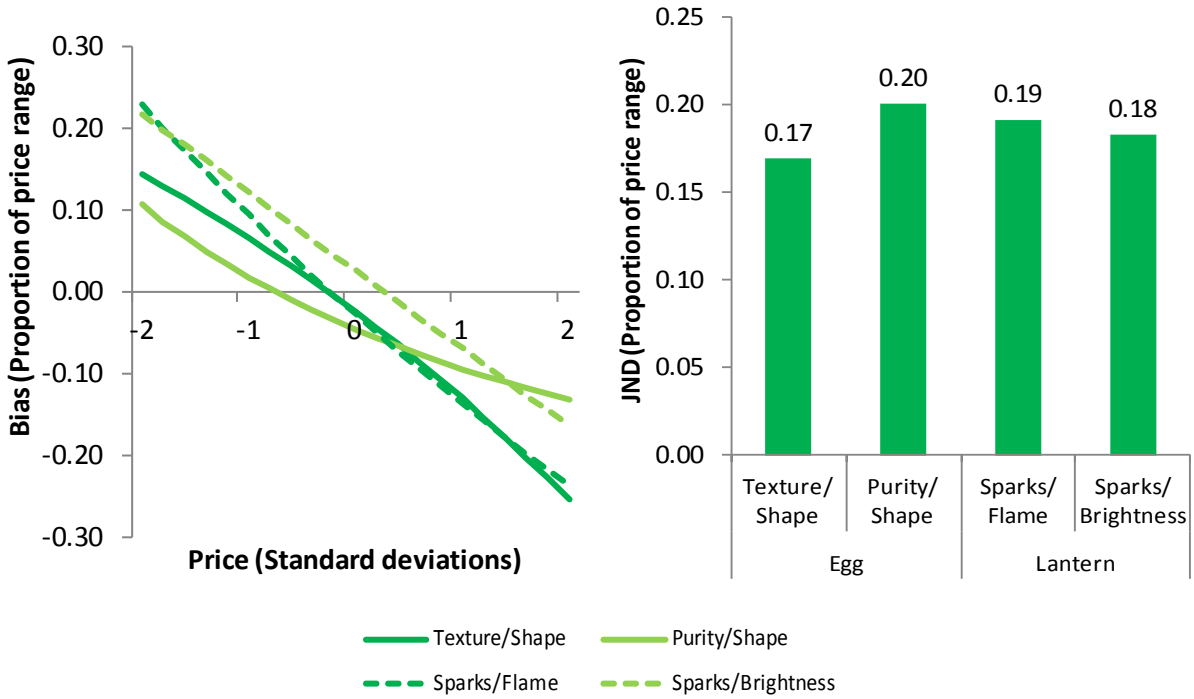
Forty participants took part in the experiment. Each was paid a €25 fee and the four most accurate performers were rewarded with a €50 shopping voucher.

7.3 EXPERIMENT I: RESULTS

Standard S-ID Task

Figure 25 plots the JNDs and biases for the standard S-ID task, in which participants saw only one hyperproduct together with a displayed price and had to determine whether the product was worth more than the price. These results look very similar to those for previous experiments, although the level of precision was the highest we recorded in a two-attribute S-ID task. As previously, we found no difference between performance with visual and numeric attributes. There was also the usual bias across the price range, with surpluses underestimated towards the bottom of the range and overestimated towards the top.

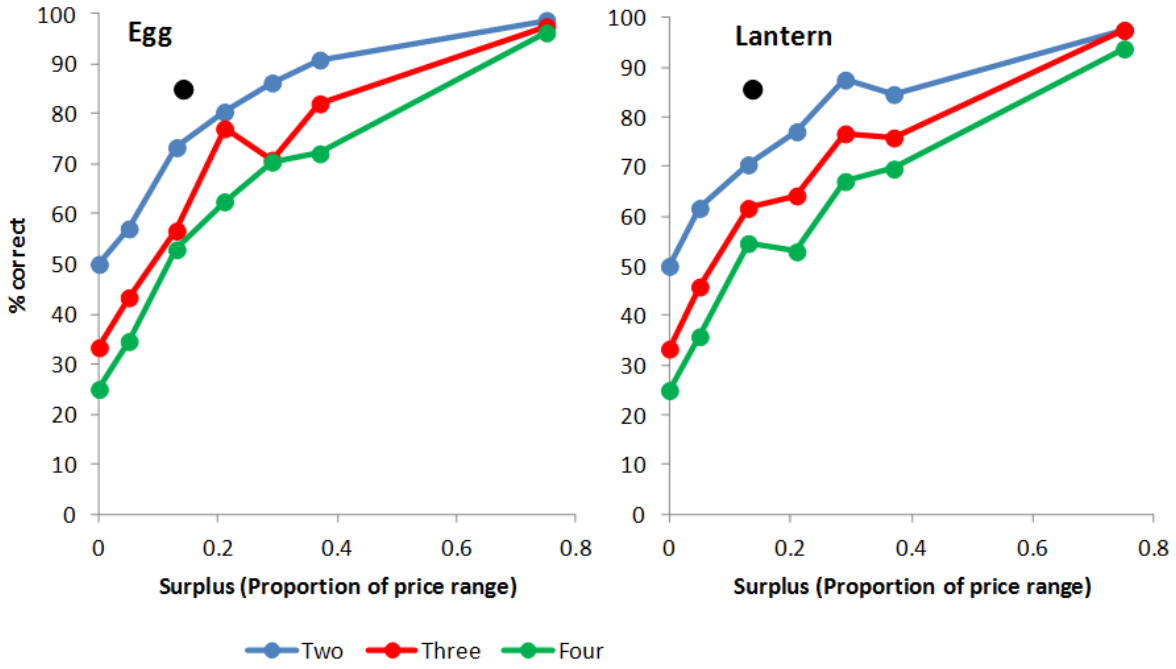
FIGURE 25 JNDs and biases for the standard S-ID task in Experiment I across four conditions.



MS-ID Task

Figure 26 shows the results of the MS-ID task. As the number of products in the range increased, participants found that the chances of detecting the surplus fell substantially and significantly. The 86 per cent correct point from the S-ID task is also shown in Figure 26 for comparison. Thus, while it is possible that there are some benefits to surplus identification from an expanded product range, these do not outweigh the fact that the increased number of products in the range simply means that the task of locating the surplus is more demanding and the probability of making a mistake is higher. Even when the surplus had to be located from among just two products, it needed to be approximately 30 per cent of the price range to be identified with 86 per cent reliability. This was the case despite the fact that the same participants undertook the task, with the same product for which they could reliably detect surpluses of approximately 20 per cent when comparing a single product against a display price (see also Figure 25). When four products were in the range, surpluses of 40 per cent were only located accurately with a probability of 65-75 per cent.

FIGURE 26 Precision in the MS-ID task of Experiment I. The proportion correct is compared when participants had to locate the product with the surplus from among two, three or four products. The 86% JND from the S-ID task is reproduced as a black circle for comparison.



There is no equivalent of the bias in the MS-ID task, because only positive surpluses are presented. Any bias would manifest itself as a preference for a particular screen location.

7.4 EXPERIMENT I: DISCUSSION

Experiment I provides a clear indication that the extent of imprecision that we record in the S-ID task generalises to situations in which there are multiple products. Although surpluses equivalent to approximately 20 per cent of the price range could be identified reliably when the product was compared with a display price, surpluses in excess of 30 per cent were required to locate the surplus from among just two alternative products, much higher when three or four products were in the range.

7.5 EXPERIMENT J: AIMS AND METHODS

All of the experiments undertaken for this report, prior to this final one, involved hyperproducts. We generated all of the hyperproducts on computer screens and consumers had never seen them before entering the lab to participate in one of our experiments. The advantage of using hyperproducts is that they provide excellent scientific control over the levels of attributes, prices and, hence, surpluses. They also minimise any role for participants' subjective preferences,

which is helpful for imposing an objective surplus against which performance can be measured. By employing attributes that we know can be perceived accurately, the hyperproducts also help us to isolate the mechanisms that integrate attribute information, rather than measure the performance of mechanisms that gather it.

However, while much effort went into making hyperproducts that were engaging for participants and designing attributes that were easy to process perceptually, it is possible that accuracy in surplus identification would improve markedly if the product were a familiar one. At a general level, there is some evidence that performance in difficult reasoning tasks can improve when the equivalent reasoning task is placed into a familiar everyday problem, although this result does not always hold (e.g. Griggs and Cox, 1982). More specifically, there are several reasons why one might hypothesise that integrating attribute information to identify surpluses would be easier when products and attributes are more familiar. Firstly, the familiarity of the price range might mean that individuals have a set of approximate reference prices for the product already in memory. Secondly, they may already have some idea of how attributes map on to prices. Thirdly, they may have a pre-existing notion of the appropriate relative weighting of attributes. Lastly, they may simply engage or identify more with the task of identifying a surplus with a familiar and, consequently, more meaningful product than a hyperproduct, despite the incentives on offer in the task.

Aims

Experiment J set out to test whether familiarity assists surplus identification. The aim was to find familiar products for which prices were largely determined by a small number of attributes. After some research, we settled on two familiar products: Dublin houses and broadband packages. We found that the prices of these two products could be modelled statistically with a fairly high degree of accuracy on the basis of just two, three or four attributes. In the case of houses these were: size, postcode, garden size and distance from the city centre. For broadband, they were download speed, platform (cable, DSL, data-card) and brand (Eircom versus the rest).

The aim was to adapt the S-ID task for use with houses and broadband packages. The task facing the participant was to decide whether a given house (or broadband package) was good or bad value at a given asking price (monthly fee). We informed participants that we had a statistical model of all the products and price available on market, such that we knew the average price that any given bundle of attributes would fetch. Their job was to decide whether the price we displayed was good or bad value for that bundle of attributes, relative to the market as a whole, i.e. whether it was lower or higher than the average price for those attributes.

Crucially, we then compared their performance in this task with familiar products to performance in the exact same task with a hyperproduct. In order to accomplish this, we embedded the statistical formula for the asking price of Dublin houses into a Mayan pyramid, and the formula for monthly payments on broadband packages into a Victorian lantern. In this way, the goal was to see whether the accuracy of surplus identification was driven by familiarity with the product, or by the mathematics of how its attributes related to prices.

Methods

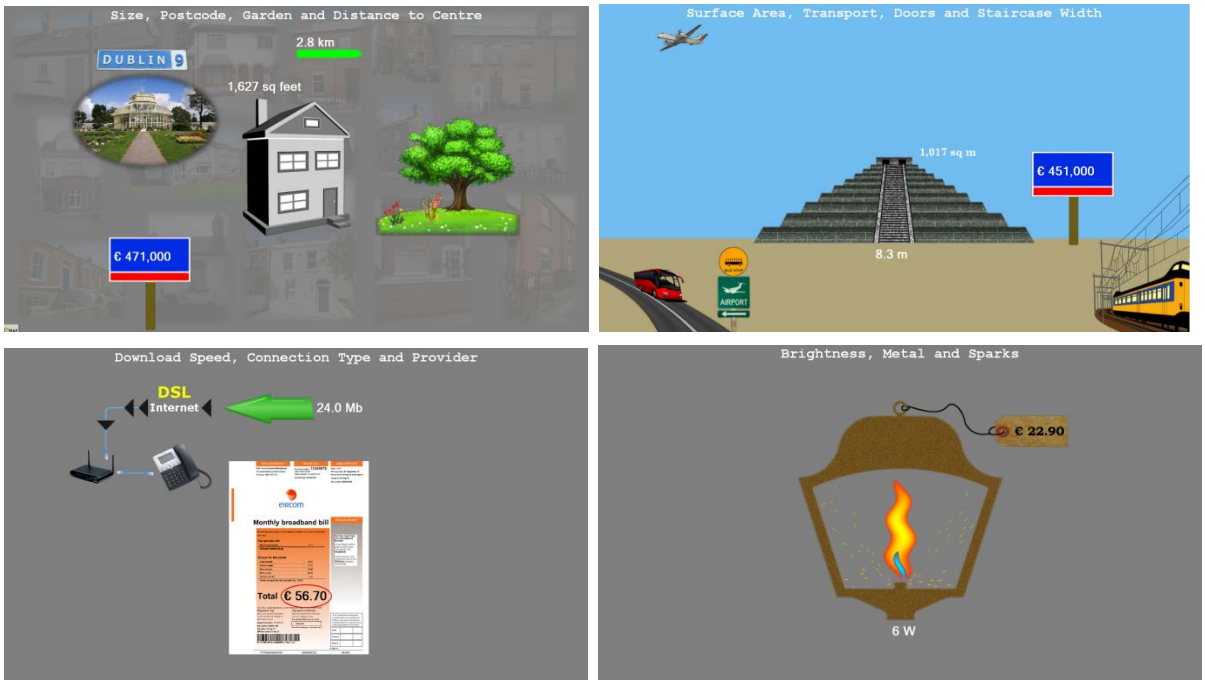
We scraped the data from the listings for all Dublin houses in four central postcodes that were advertised with Dublin's largest online sales and marketing website in July 2014. We used this to compile a dataset listing the asking price, size (in square feet), postcode (D6, D7, D8 and D9), garden size, and distance to the city centre. This allowed us to build a regression model, with the asking price as dependent variable and two (size and postcode) or all four attributes as regressors. The relationship between the size and price was approximately log-linear. A model containing just this variable and the four postcodes as a categorical variable explained more than 70 per cent of the variance of Dublin asking prices. When a three-category variable for garden size and a continuous variable for distance from the city centre were added, this climbed to over 75 per cent.

For broadband we made use of data on available broadband packages supplied by the Commission for Communications Regulation in 2014. A model with a linear and non-linear (squared) term for download speed and a three category variable for platform explained 68 per cent of the variance in monthly fees. Adding a dummy variable for a small premium paid if the package was supplied by Eircom increased this to 70 per cent.

Thus, we had good statistical models of the average prices of both goods for bundles of 2-4 attributes. Screenshots of the presentations of the products are provided in Figure 27. The displayed price was presented on a sign in front of the house and the participant had to decide whether the house was good or bad value at that price, i.e. was the price lower or higher than average for the attributes presented. Similarly, for the broadband packages, the price was displayed on a bill. Performance in these two tasks with real products was then compared to performance with two hyperproducts, the value of which was determined by the exact same formulas. That is, the attributes were different, but their relationship to the price was mathematically identical. The surface area of the Mayan pyramid in square metres was the equivalent of the area of the

house; the transport access (bus only; train only; bus and train; bus train and plane) was the equivalent of the postcode; the number of doors (none, one, two) was the equivalent of the garden size; the width of the staircase was the equivalent of the distance to the city centre. The ratio of the blue part of the lantern flame to the rest of the flame, which was also given numerically as a Wattage, was the equivalent of download speed; the metal type (copper, brass, iron, signalled by colour) was the equivalent of platform; the presence or absence of sparks was the equivalent of Eircom or another brand. Thus, the two sets of two tasks were perfectly matched mathematically. The only thing that differed between the 'real' and 'hyper' conditions was whether the participant was likely to be familiar with the product, its attributes and price range. Moreover, in half the trials with the hyperproducts, we changed the price range by an arbitrary factor while keeping the relative attribute weightings the same. If familiarity with the price range itself was of any benefit, this would be removed by this manipulation.

FIGURE 27 Onscreen environments for Experiment J. The house could have two attributes (size, postcode) or four (plus garden size, distance to city centre), which were perfectly matched to the surface area, transport access, number of doors and staircase width on the pyramid. The broadband packages could have two attributes (speed, platform) or three (plus brand), which were perfectly matched to the flame ratio, colour and presence of sparks on the lantern.



Forty participants took part in the experiment and were paid a fee of €25 for their participation. Over half of them were homeowners and 36 had previously chosen

a broadband package.[8] The four most accurate respondents won a €50 voucher. Each participant either undertook the two-attribute house and pyramid conditions plus the three-attribute broadband and lantern conditions, or the four attribute house and pyramid conditions plus the two attribute broadband and lantern conditions. The order of the four runs was pseudo-randomised. In many previous experiments experimental runs were ordered such that more simple tasks came first, which helped participants initially to understand the nature of task. Because we were interested in removing order effects from our comparison, to get participants accustomed to the task, we started each session with a run of 48 trials using golden eggs that varied in size and texture. This also allowed us to benchmark the performance of the sample against previous samples that had undertaken this condition.

Before each of the four main experimental runs, participants were shown a series of examples of the product and price, followed by 12 practice trials, then 60 test trials. As in previous experiments, feedback was given by way of a beep, green tick or red cross, and the presentation of the true (or average) value of the product presented.

Because the attributes of the products were correlated, e.g. larger houses also tended to have larger gardens, unlike in previous experiments we did not select the attribute magnitudes at random. Rather, we selected a price at random, added or subtracted the appropriate surplus using the same adaptive method as in previous experiments, then selected the actual house from our advertised sample of houses (or actual broadband package from the available packages). The same method was used to select the hyperproducts to be presented. In this way, the correlations between the attributes that existed in the real market were preserved in the experiment.

7.6 EXPERIMENT J: RESULTS

Precision

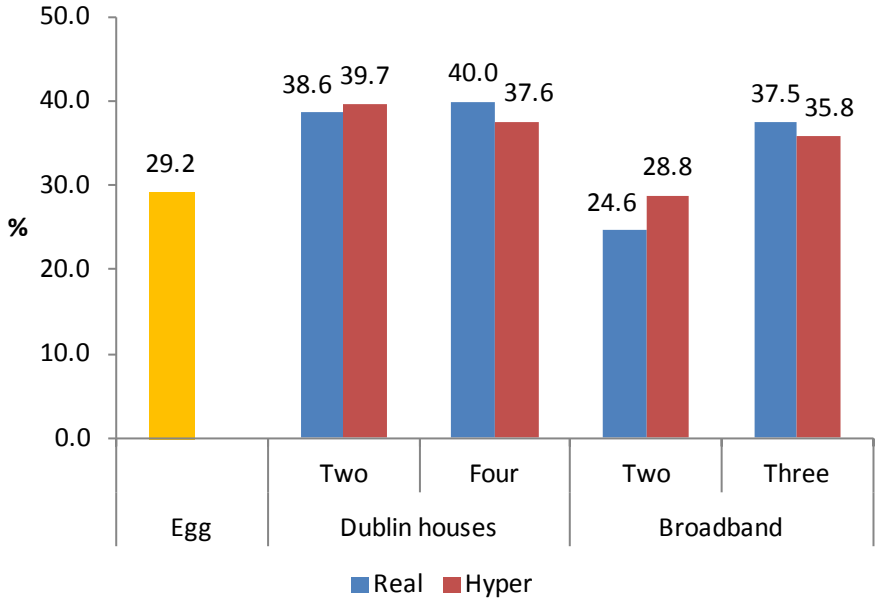
Figure 28 shows the levels of precision achieved by consumers across all conditions. It is important to note that the vertical axis is somewhat different to previous experiments, where the JND is measured as a proportion of the price range. In this case, the price range of both products was highly skewed, making the proportion of the range a poor measure for comparison. Instead, the vertical axis corresponds simply to the surplus expressed as a percentage increase or decrease relative to the displayed price. We recorded no significant difference

[8] Even those who were not homeowners would probably have some familiarity with the Dublin housing market: types of house, relative merits of locations and so on. Although the price range is obviously different, a similar relative weighting of attributes such as size and location would also apply to the rental market.

between the conditions where the hyperproduct shared the price range with the real product and the conditions where it did not, so the data from these are pooled.

The first bar in Figure 28 confirms that the sample of participants in this experiment performed on the initial golden egg task at a similar level to those in most previous experiments (allowing for the different vertical axis). The primary comparisons of interest are the relative JNDs of the real products and hyperproducts in each of the four conditions. Perhaps somewhat surprisingly, we recorded no statistically significant differences in the precision with which participants could identify surpluses with the real, familiar products and the completely unfamiliar hyperproducts.

FIGURE 28 JNDs for surplus identification relative to average market prices for Dublin houses and broadband packages, compared to hyperproducts with the same mathematical relationship between attributes and prices (Experiment J).

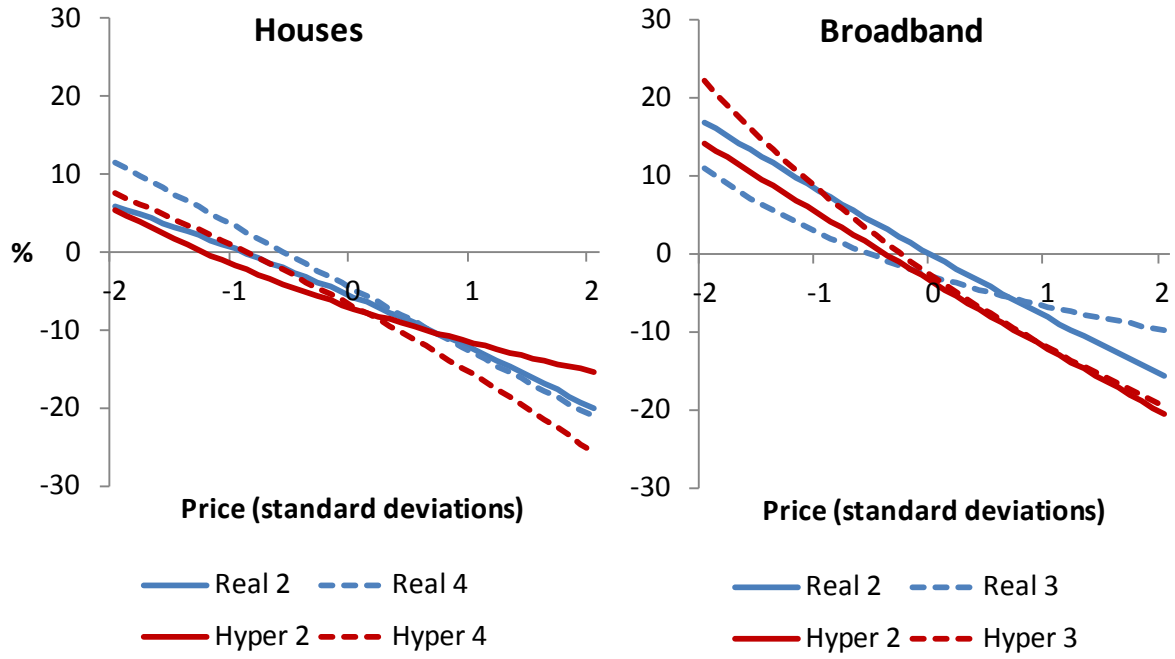


Bias

Figure 29 shows the size of the bias across the price range in each of the four conditions. The pattern is almost indistinguishable and, again, there is no statistically significant difference between the real products and hyperproducts regarding the variation in extent of bias across the price range. The only statistically significant difference between the real products and hyperproducts to be found in this experiment relates to the slight bias towards overestimation of the surplus generally, which was again present here. When the four conditions are pooled and the real and hyper products compared, the overall bias towards overestimation of surplus was significantly larger for the hyperproducts. In

relation to Figure 29, this means that the red curves mostly sit below the blue ones and that on average this difference is significant. As is clear from the chart, however, this effect is very small, especially in comparison to the variation in bias over the price range, which is common to both real products and hyperproducts.

FIGURE 29 Biases across the price range in a surplus identification relative to average market prices for Dublin houses and broadband packages, compared to hyperproducts with the same mathematical relationship between attributes and prices (Experiment J).



7.7 EXPERIMENT J: DISCUSSION

While confident that our experiments with hyperproducts were isolating capabilities associated with the key psychological mechanisms used to integrate product information, we were nevertheless surprised to find essentially no improvement at all in the accuracy of surplus identification for real, familiar products, as opposed to the unfamiliar hyperproducts. The implications of this finding are, firstly, that it is the mathematics of the attribute-price relationship that determine how accurately attributes and prices can be compared and, secondly, that results obtained with hyperproducts are likely to generalise to more familiar multi-attribute products in the real world. Arguably, we should not be surprised by the finding, given the results in relation to learning that have been consistent throughout the experiments described in this report. It seems that initial learning to map attributes to prices is very rapid, requiring just a small number of exposures, and that subsequent learning is modest even following hundreds of trials with perfect feedback. When dealing with these two real world products, consumers are very unlikely to be exposed to such a volume of

accurate feedback, so it is perhaps not so surprising that their familiarity with the products was in fact of little help.

This is not to say, however that there may be no benefit to familiarity. For one thing, it may be that a very large amount of familiarity does result in an improvement in the integration of attribute information to register. It seems highly likely that professional dealers in, say, antiques or livestock, are better at identifying surpluses in their respective markets than non-professional occasional purchasers. Indeed, in the livestock market there is evidence to support the claim, with certain caveats, that some individuals possess genuine expertise in valuation (Phelps and Shateau, 1978). For ordinary consumers, however, a measurable element of expertise may be hard to come by.

One final point is crucial for placing the findings of Experiment J in context. Although the experiment revealed that familiarity conferred no advantage for the integration of attribute information, participants were initially guided via examples and explanation with respect to which attributes they needed to take into account. Familiarity with a market, while not apparently helping to overcome the cognitive constraints that have been investigated here, may nevertheless be very helpful for understanding which attributes of a product to pay most attention to and which to safely ignore.

7.8 DISCUSSION

Experiments I and J investigated very different issues: whether larger product ranges or familiarity with products could improve the accuracy of surplus identification. Yet the two experiments had two things in common. First, they answered these respective research questions with a resounding negative. Surplus identification is, if anything, less accurate when products are in ranges than when they are compared individually against prices. Meanwhile, familiarity with the market does not improve surplus identification, at least once the attributes that must be taken into account are known. Second, both Experiments I and J support the view that the limited ability to integrate information from incommensurate scales is not easily overcome. The implication is that the experiments reported here probe an underlying psychological mechanism with a limited capacity, one that is likely to apply broadly across consumers and markets.

Part 3

Conclusions and Policy Implications

Section 8

Summary of Findings

8.1 INTRODUCTION

This final chapter first presents a summary of the findings of Experiments A to J. It then turns to a discussion of their potential policy implications.

8.2 SUMMARY OF FINDINGS

Rather than repeating the findings of the individual experiments in turn, this section pulls together principles that emerge when looking across Experiments A to J as a whole. These principles are first stated and then supported with reference to the individual experiments.

Lack of precision in surplus identification is due to more than perceptual limitations.

A much larger difference in value between a single attribute and a displayed price was needed for reliable detection of which one was worth more, than the difference in attributes needed when comparing two products (Experiment A). Even when both of two product attributes are numerical or categorical and therefore offer no scope for error in perceptual judgment (Experiment E), the precision of surplus identification (S-ID) does not improve compared to tasks in which both attributes are perceptual. Furthermore, whether the same attribute (height; Experiment F) is depicted visually or numerically, precision does not change. These various results show that mapping an attribute magnitude onto a price is challenging and that limitations in the precision of doing so are not limitations of perceptual discrimination, at least not primarily. Rather, there seem to be cognitive limitations in developing and retaining representations of attribute-price maps and relationships. At no point in any condition across all of the ten experiments was a product mapped on to a price with sufficient precision that the average participant could distinguish as many as eight different levels of value.

The more attributes a product has, the more difficult it is to map these onto a price with precision.

Comparing two products becomes more challenging when they vary on more than one attribute, and precision diminishes further when a two-attribute product has to be compared against a price (Experiment A). Products with three and four attributes that map onto the same price result, unsurprisingly, in further reductions in precision (Experiment B). These reductions are even more extensive

than the changes in performance anticipated from the already limited precision in mapping single-attribute products onto prices.

If product attributes are positively correlated, such that if one attribute is good the others are good too, precision improves as the number of attributes increases (Experiment D). However, in Experiment D precision still lagged behind what would be predicted from single-attribute performance, if individuals could combine additional attribute information efficiently. Furthermore, even if the number of levels of an attribute are limited to preserve cognitive capacity (Experiment F, categorical attribute with two categories), precision in detecting a surplus for a two-attribute product is still worse than precision in surplus identification for a single-attribute product.

These experimental results suggest that there are stark cognitive capacity restrictions, which limit not just how well people can map incommensurate scales such as prices and attributes onto one another, but also how many attributes they can do this with simultaneously, even if those attributes do not conflict with one another or have few possible levels.

Bias when determining whether a product is worth more or less than a given price depends on how good or bad the product is.

Overall, people tend to be biased in how they estimate the value of basic products for which a change in attribute magnitude maps onto a change in price (Experiment A). Generally, they overestimate how much the product is worth, but this bias changes across the price range (Experiments A, B, C, E, F, G, H, I and J). Overestimation of value is greatest for products that are worth the most, whereas for products that are worth the least, underestimation occurs. However, as more attributes are added, the effect of location in the price range diminishes (Experiment B). This could be partly due to a precision-bias trade-off, such that greater precision in detecting a surplus exaggerates its perceived size, while lesser precision reduces this bias. This latter effect is seen in two-attribute tasks in which precision is very low (i.e. the task is difficult; Experiment F). Bias reversal could also be due to participants averaging the value of the attributes. In Experiment D, where the average attribute magnitude was held constant as more attributes were added, the bias did not diminish; overestimation of good and underestimation of bad products was enhanced with increasing numbers of attributes.

Practice, high educational attainment and numeracy, and increased motivation and effort do not overcome limitations in surplus identification.

In Experiment C, a highly educated sample of participants working in a numerate field was given the opportunity to perform the S-ID task over multiple sessions. These participants performed only marginally better than the general public, and were subject to the same patterns of bias and imprecision. Performance improved to a small degree between the first and second but not the second and third sessions, suggesting that the limitations in mapping product attributes onto prices can be slightly attenuated following some experience with the task but that improvement levels off early. Tests for learning in Experiment H also show it to be limited, and indicate that it is not extended even for very difficult tasks. Motivational manipulations do not improve precision either, even if they do enhance effort (Experiment C).

Consumers can perform surplus identification with attribute-price relationships that are non-linear, but only if the direction of the relationship does not change.

Consumers can cope as well with attributes with 'diminishing returns' as they can with attributes that vary linearly with price, at least once they have seen some helpful examples (Experiment G). Bias is identical for linear attribute-price mappings, moderate diminishing returns, and severe diminishing returns. However, once the relationship between product and price becomes non-monotonic (i.e. it changes in direction once or more), precision suffers immensely although bias does not change (Experiment H).

These results can be replicated in studies that conform more closely to standard market structures and that use familiar products.

In most markets there is more than one product available at a time. In theory, accuracy might improve if consumers have to identify which of two, three or four product-price pairings confers a surplus, rather than simply identifying whether a single product is worth more than the price it is shown at. They might combine comparison of attributes within a product with comparisons between products, to assist surplus identification. Yet this possibility is balanced by the fact that with more choices comes more complexity and more opportunity to be incorrect. Experiment I suggests the latter force at least cancels out any benefits of viewing products within a range. Finally, when performance in S-ID tasks with hyperproducts is compared to S-ID tasks where familiar products are compared to prices, there are almost no differences in precision and bias across tasks (Experiment J). The identification of surpluses for multi-attribute products is at best approximate.

Section 9

Policy Implications

9.1 INTRODUCTION

As highlighted in Chapter 1, especially in light of advances in behavioural economics, policymakers in regulated markets are increasingly concerned about consumers' ability to cope with complex multi-attribute products. The results contained in this report give much ballast to this concern. Once a product has more than one or two relevant attributes, where each has a substantial impact on the overall value that the product provides, consumers' choices are likely to be highly approximate. In other words, a product does not have to be particularly complex before its complexity has a clear impact on the accuracy of consumers' decisions and hence on their ability to get the best value from transactions. Some specific consequences of the approximate nature of consumer choice are likely to arise, with implications for consumer policy.

It is important to understand that the policy implications discussed in this chapter are not specific to Ireland or to any particular product market. Furthermore, they certainly should not be read as criticisms of existing regulatory regimes, which already go to some lengths to assist consumers to deal with complex products. Ireland currently has extensive regulations designed to protect consumers. In addition to market-specific regulations, summarised briefly below, all firms must comply with general consumer protection laws that cover, amongst other things, the accuracy of claims about products and of product descriptions. These laws are a mix of Irish and European legislation, such as the Consumer Protection Act 2007 and the EU Consumer Rights Directive, which came into force in 2014. The Competition and Consumer Protection Commission (CCPC) is responsible for the enforcement of these laws.

Key regulations in financial services include those contained in Consumer Credit Act 1995, the Central Bank's Consumer Protection Code (most recently revised in 2015) and the European Communities (Consumer Credit Agreements) Regulations 2010. Amongst other things, the Consumer Protection Code sets out principles as to how regulated entities must disclose all relevant information to consumers and, moreover, includes detailed requirements with respect to the provision of accurate and up-to-date information, the clarity of language and use of plain English, the inclusion in advertising of key information and qualifying criteria, and the completion of a standardised assessment of suitability of a product for a specific consumer. The regulations specific to credit products require that certain standard information be included in all advertising and that this information is

presented by means of a representative example. In addition to these regulations, CCPC provides financial education programmes through schools and workplaces, as well as a general consumer rights and information website, www.consumerhelp.ie, which contains a price comparison facility for certain key financial products.

In energy markets, as a condition of their licence, suppliers must comply with a set of minimum standards contained in the Electricity and Natural Gas Supplier Handbook 2012, which is produced by the Commission for Energy Regulation (CER). This includes requirements for codes of practice covering marketing, billing and disconnection, complaints and vulnerable customers. Many of the minimum standards concern the accuracy and transparency of price information, as well as the presentation and clarity of other useful information for consumers. At the time of writing, CER is undertaking a consultation exercise on changes to the Supplier Handbook. A key aim of the consultation is to find ways to provide energy consumers with the right information to help them to make better choices in the market. CER also accredits two price comparison websites, www.bonkers.ie and www.switcher.ie.

In telecommunications markets, the Commission for Communications Regulation has multiple powers under the Communications Regulation Act of 2002 (amended 2007) to support its statutory obligations. These include promoting the provision of clear information, with particular reference to the transparency of tariffs and of conditions for using services. The Commission provides an interactive price comparison site, www.callcosts.ie, for home phone, mobile, broadband and bundled telecommunications services, which includes information on price and non-price product features such as contract periods, early contract termination penalties, customer service and billing details. The Commission also enforces EU regulations, including the Universal Services Regulations, which govern matters of notification and consent in relation to contractual terms and conditions.

In summary, the aim of the present chapter is in no way to evaluate these regulatory environments, but instead to consider the implications of the experimental findings for the potential development of future consumer protection policies, which will be formed in the light of results from the field of behavioural economics. An international review of such policies appears in Section 1.4. The additional understanding of consumers' capabilities provided by the current findings can contribute to this process.

9.2 FROM EVIDENCE TO POLICY

While policy is clearly likely to be more effective when informed by evidence, there are of course limits to the extent to which one can infer good policy from specific empirical findings. For instance, while the empirical evidence supplied by the present report may be used (in conjunction with other recent findings in behavioural economics) to support regulatory policies that are somewhat more interventionist than has been the norm in recent decades, some caution is clearly warranted. While the experimental findings do suggest potential *benefits* to more prescriptive rules in relation to how firms can and cannot market and promote certain pricing structures and product attributes, none of the evidence supplied here deals with the potential *costs* of any regulations imposed, nor with the issue of how firms will respond to a regulatory change. Both factors deserve attention in a full assessment of the costs and benefits associated with any regulatory changes.

Furthermore, the findings reported here are of a general nature, in the sense that they uncover principles of consumer capability that are likely to apply across products and markets. There may be specific markets in which the determinants of consumer choice are somewhat idiosyncratic. The empirical methods employed in the present work suggest much scope for experimentation, whether in the laboratory or in the field, that is more specific to the product or market concerned. The repeated forced-choice methods developed here can be used to examine consumer capability not only by testing capability with specific products of interest, but also by testing the likely impact on decisions of a specific regulation. For instance, the methods can be adapted to test the impact on consumer decisions of particular warnings and information mandates, or to measure the likely impact of consumer advice. They can also be used to test the effect on consumers of specific pricing or marketing techniques that firms might introduce to a market by providing quantitative measures of the likely impact of such techniques on the accuracy of consumer choice.

Lastly, as is generally the case with research programmes, the findings raise questions as well as answering them. Given the limits to the integration of product information revealed here, one key area for future research opened up by the present findings is how consumers actually cope when overwhelmed by product information. How do consumers go about ‘editing’ a decision down to a manageable size? Are there ways in which it is possible to assist them in this task? The methods developed here have the capability to address these types of research questions, but until we have the answers, the evidence base for policy remains incomplete. Nevertheless, while recognising the need to conduct further research that addresses both these questions and others that may be specific to sectors, products or potential interventions, the experiments reported here

provide new empirical findings and it is worth considering some specific potential implications.

9.3 IMPLICATIONS OF THE PRESENT FINDINGS

While, for good reason, policymakers have spent much regulatory effort on reducing barriers to switching for consumers, on its own this may be insufficient to prompt consumers to make the effort to switch. Reducing and removing barriers makes sense, since they can deter consumers from switching to gain better value. Yet the results of the present report also suggest that another barrier to consumer activity in markets with multi-attribute products is likely to be consumers' ability to identify gains reliably, and perhaps their own perceptions of their ability to do so.

In this context, the provision of additional consumer information has the potential to be a double-edged sword, depending on the volume and complexity of information that consumers are already trying to cope with. Given the limited number of attributes that consumers are able to factor into decisions simultaneously, the provision of additional information may in some contexts hinder as much as help. The key question may not be whether consumers have sufficient information, or even whether they make an 'informed choice', but which information ultimately makes it into the decision-making process. This implies a potentially strong role for independent and accurate consumer advice, which can be designed to highlight the key attributes of complex products that cannot be ignored without risking negative consequences. Currently, most regulators try to assist consumers through the provision of impartial advice, through websites, leaflets, publicity campaigns and so on. With limited scope for the simultaneous inclusion of multiple sources of information in consumer decisions, good advice may be that which limits the volume of information provided and instead aims to ensure that the most important product attributes are made prominent and ultimately drive the decision.

In modern market economies, reliance is placed on competition to ensure that consumers obtain value. As well as incentivising firms to offer good quality and to keep prices down, competition is important because of its effects on innovation, customer service and, most straightforwardly, choice itself. However, the present results imply some limits to what competition can achieve where products are complex. The pressure that competition exerts on quality and price is likely to be directly related to the extent to which consumers are able, or indeed unable, to identify surpluses and switch to obtain them. Furthermore, there is a danger that competition incentivises some firms, perhaps especially those defending high market shares, to act in ways that increase the complexity of the product. This need not involve any lack of disclosure or unclear presentation of information,

because the present results show that merely the addition of a product attribute of potential concern to consumers has the capacity to reduce the accuracy of choices. Of course, where new attributes reflect genuine innovations that add value for consumers, they exemplify the long-term benefits of competition, as firms improve products to attract customers. But where the added value is minimal or absent, the present findings suggest a cost in terms of increased complexity. This evidence should be considered in conjunction with other recent advances in 'behavioural industrial organisation' that suggest ways in which firms can artificially generate market power, despite apparent competition (e.g. Gabaix and Laibson, 2006; or see Grubb, 2015, for review). Of course, the impact of competition will vary between specific markets and between the short and long term, depending on such things as product differentiation and the pace of change within the industry. The present contribution nevertheless implies limits to the amount of downward pressure on prices and upward pressure on quality likely to be exerted by consumer activity when products have multiple important attributes.

In showing how products that possess more than just one or two attributes can greatly reduce the accuracy of consumer choices, the experiments suggest that the list of products that might be considered 'complex' is perhaps surprisingly long. This raises issues of prioritisation. The suggestion is that there may be much scope for assisting and improving consumers' decision-making in many markets. If so, then it makes sense from a consumer welfare perspective to target markets and products where the potential for consumer detriment is greatest. For instance, the largest transactions that households make surround homes, cars and pensions. However, there are also markets such as energy and telecommunications in which, while detriment to any one individual is likely to be relatively low, the total cost spread across consumers may be very high. Hence, special attention might also be paid to those markets where a very large proportion of the population are customers and where specific pricing or marketing practices are suspected of adding to the complexity of the consumers' task. The present results imply that what might appear to be small changes to products and product descriptions, such as innovations or marketing based on single additional attributes, have the capacity to reduce the accuracy of consumer decisions and hence to generate small amounts of additional revenue from very large numbers of consumers. It is important to understand that no deception or mis-selling need be involved for this to occur.

It is perhaps in this latter context that the present results give the most clear indication of concrete regulations that may be of assistance to consumers. The findings show straightforwardly that the simple number of factors that must simultaneously be taken into account when making choices among products has a substantial impact on the quality of choices. Given such evidence, some

marketing and pricing practices that may appear to be of little or no obvious benefit will be detrimental to consumers to the extent that they increase the number of attributes that consumers try to take into account. These include situations where firms avoid expressing prices as total costs and instead employ multiple price components (non-linear price structures, drip pricing, partitioned pricing etc.). They also include situations where attributes of products are split into multiple components without a clear rationale for the consumer. Furthermore, given the limited capacity for integrating product information that is uncovered here, the marketing of irrelevant product attributes may be more damaging to consumer interests than it first appears. The issue may not be so much whether an irrelevant attribute draws the consumer towards a specific offering, but whether it crowds out attributes likely to have a greater impact on consumer outcomes, given limited capacity to simultaneously process information. The results therefore support a tough regulatory line on increases in choice complexity that appear to be of no plausible benefit, whether by splitting prices, splitting attributes or promoting attributes of apparently no or little value. Regulators must of course be careful, since consumers are the ultimate arbiters of what product attributes matter, but the present results suggest that this sort of complexity can confuse consumers and generate a form of market power, without any deceptive practice or unfaithful communication of information to the consumer.

The insight that the number of factors that consumers can simultaneously consider is small also suggests potential ways in which consumers might be assisted in their choices. The results provide backing to the efforts made by policymakers to promote, support, and potentially provide price comparison sites and other 'choice engines' (see Chapter 1). The evidence supplied by the experiments in this report suggests the potential for quite substantial losses (and opportunity costs) to consumers from inaccuracy in identifying surpluses. Independent price comparison sites made available or endorsed by regulators are likely to be of considerable benefit especially when multiple product attributes must be taken into account. As explained in Chapter 1, while not guaranteeing that the consumer will locate best value, choice engines make a good choice more likely and reduce the chances of making a strongly disadvantageous choice.

As described in Chapter 1, mandated simplification policies are increasing in popularity and the present research offers up some lessons for what may make a good mandated simplification. Consumer decisions are likely to be improved if the mandated simplification makes it more probable that the most relevant attributes feature in the decision-making process. The present results offer guidance with respect to how many key attributes consumers are able simultaneously to trade off, with implications for the design of standardised disclosures. The findings also lend support to the use of mandated 'meta-

attributes', similar to the APR on credit products, where multiple attributes are reduced by regulatory mandate to single comparable attributes. There may be more scope for regulators to support useful meta-attributes in other markets, perhaps especially where it is possible to express service contracts, which often contain separate price components some of which are time limited, as total costs paid over the contract. This may be particularly important in those markets where products are increasingly 'bundled'. The present findings suggest that, while it may or may not be the case that bundles tend to offer better overall value, for example through price savings, they are almost certain to be more difficult for consumers to compare, since by definition they increase the number of product attributes in play. In keeping with this implication, some recent evidence suggests that consumers who opt for bundles in telecommunications then become less likely to switch provider (Burnett, 2014). Note that this does not mean that bundles are bad for consumers, as competition based on bundles may ultimately be good for consumers, but it does imply that consumers will struggle to compare bundled products accurately where the bundling requires a larger number of product attributes to be taken into account simultaneously.

Finally, the S-ID task is a novel contribution to consumer research and one that has the potential to address other questions of interest for understanding consumer choice and providing evidence for consumer policy. For instance, the method can be adapted to examine willingness to switch, or how consumers factor in risk, or how they deal with dynamic attributes that change over time. By gaining experimental control over the size of surpluses, the respective impacts on consumer capability can be measured with some accuracy.

References

Agarwal, S., S. Chomsisengphet, N. Mahoney and J. Stroebel (2015). 'Regulating Consumer Financial Products: Evidence from Credit Cards'. *Quarterly Journal of Economics*, 130, 111-164.

Ashby, G.F. and W.T. Maddox (2005). 'Human Category Learning'. *Annual Review of Psychology*, 56, 149-78.

Barber, B.M., T. Odean and L. Zheng (2005). 'Out of sight, out of mind: The effects of expenses on mutual fund flows'. *The Journal of Business*, 78, 2095-2120.

Bar-Gill, O. and R. Stone (2009). 'Mobile Misperceptions'. *Harvard Journal of Law & Technology*, 23, 49-118.

Barlow, H. (1961). 'Possible principles underlying the transformation of sensory messages'. In Rosenblith, W. (ed.), *Sensory Communication*, Chapter 13, pp. 217-234.

Brennan, T. (2007). 'Consumer Preference Not to Choose: Methodological and Policy Implications'. *Energy Policy*, 35, 1616-1627.

Burnett, T. (2014). The Impact of Service Bundling on Consumer Switching Behaviour: Evidence from UK Communication Markets. Working Paper, Centre for Market and Public Organisation.

Carlson, K.A., M.G. Meloy and J.E. Russo (2006). 'Leader-driven Primacy: Using Attribute Order to Affect Consumer Choice'. *Journal of Consumer Research*, 32, 513-18.

Chernev, A. (2005). 'Context Effects without a Context: Attribute Balance as a Reason for Choice'. *Journal of Consumer Research*, 32, 213-23.

Choi, J.J., D. Laibson and B.C. Madrian (2010). 'Why does the law of one price fail? An experiment on index mutual funds'. *Review of Financial Studies*, 23, 1405-1432.

DellaVigna, S. (2009). 'Psychology and Economics: Evidence from the Field'. *Journal of Economic Literature*, 47, 315-372.

Dodds, P., C. Donkin, S.D. Brown and A. Heathcote (2011). 'Increasing capacity: Practice effects in absolute identification'. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 477-492.

Ericson, K.M. and A. Starc (2013). 'How Product Standardization Affects Choice: Evidence from the Massachusetts Health Insurance Exchange.' NBER Working Paper 19527.

Ernst, M.O. and M.S. Banks (2002). 'Humans integrate visual and haptic information in a statistically optimal fashion'. *Nature*, 415, 429-433.

Gabaix, X. and D. Laibson (2006). 'Shrouded Attributes and Information Suppression in Competitive Markets'. *Quarterly Journal of Economics*, 121, 505-540.

Giulietti, M., C. Waddams Price and M. Waterson (2005). 'Consumer choice and competition policy: a study of UK energy markets'. *The Economic Journal*, 115, 949-968.

Griggs, R.A. and J.R. Cox (1982). 'The elusive thematic-materials effect in Wason's selection task'. *British Journal of Psychology*, 73, 407-420.

Grubb, M.D. (2009). 'Selling to Overconfident Consumers'. *American Economic Review*, 99, 1770-1807.

Grubb, M.D. (2015). 'Behavioral Consumers in Industrial Organization'. Working Paper, Boston College.

Hauser, J. (2011). 'A Marketing Science Perspective on Recognition-Based Heuristics (and the Fast-and-Frugal Paradigm)'. *Judgment and Decision Making*, 6, 396-408.

Huber, J., J.W. Payne and C. Puto (1982). 'Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis'. *Journal of Consumer Research*, 9, 90-98.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin.

Kahneman, D., and A. Tversky (1979). 'Prospect theory: An analysis of decision under risk'. *Econometrica*, 47, 263-291.

Lambrecht, A. and B. Skiera (2006). 'Paying Too Much and Being Happy About it: Existence, Causes and Consequences of Tariff-Choice Biases'. *Journal of Marketing Research*, 43, 212-223.

Laming, D.R. J. (1997). *The Measurement of Sensation.* Oxford University Press.

Larrick, R.P. and J.B. Soll (2008). 'The mpg illusion'. *Scienc*e, 320(5883):1593.

Lunn, P.D. (2012). 'Behavioural Economics and Policy making: Learning from the Early Adopters'. *Economic and Social Review*, 43, 423-449.

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

Lusardi, A. and O.S. Mitchell (2011). *Financial literacy and planning: Implications for retirement wellbeing*. Technical report, National Bureau of Economic Research.

Meyvis, T. and C. Janiszewski (2002). 'Consumers' Beliefs About Product Benefits: The Effect of Obviously Irrelevant Product Information'. *Journal of Consumer Research,* 28, 618-35.

Miller, G.A. (1956). 'The Magical Number Seven, plus or Minus Two: Some Limits on Our Capacity for Processing Information.' *Psychological Review*, 63, 81–97.

Nyberg, P. (2010). *Misjudging Risk: Causes of the Systemic Banking Crisis in Ireland*.

Phelps, R.H. and J. Shanteau (1978). 'Livestock judges: How much information can an expert use?' *Organizational Behavior and Human Performance*, 21, 209-219.

Shiller, R.J. (2005). *Behavioral Economics and Institutional Innovation*. Cowles Foundation Discussion paper No. 1499, Yale University.

Stewart, N., G.D.A. Brown and N. Chater (2005). 'Absolute Identification by Relative Judgment'. *Psychological Review*, 112, 881-911.

Sunstein, C. (2011). 'Empirically informed regulation'. *University of Chicago Law Review*, 78, 1349-1429.

Thaler, R.J. (2015). *Misbehaving: The Making of Behavioural Economics*. Allen Lane.

Troutman, M.C. and J. Shanteau (1976). 'Do Consumers Evaluate Products by Adding or Averaging Attribute Information?' *Journal of Consumer Research*, 3, 101-06.

Vlaev, I., N. Chater, N. Stewart and G.D. Brown (2011). 'Does the Brain Calculate Value?' *Trends in Cognitive Sciences*, 15, 546-54.

Weaver, K., S.M. Garcia and N. Schwarz (2012). 'The Presenter's Paradox'. *Journal of Consumer Research*, 39, 445-60.

Wilson, C.M. and C. Waddams Price (2010). 'Do Consumers Switch to the Best Supplier?' *Oxford Economics Papers*, 62, 647-668.

Appendix

PRICE Lab Scientific Papers

Lunn, P.D., and J. Somerville (2015). Surplus Identification with Non-linear Returns. ESRI Working Paper No. 522. www.esri.ie/publications/surplus-identification-with-non-linear-returns.

Lunn, P.D., M. Bohacek and F. McGowan (2016). The Surplus Identification Task and Limits to Multi-Attribute Consumer Choice. ESRI Working Paper, forthcoming.

Lunn, P.D., M. Bohacek and A. Ní Choisdealbha (2016). How Accurately can Humans Resolve Trade-Offs. Working Paper, forthcoming.

Lunn, P.D, M. Bohacek, J. Somerville and F. McGowan (2016). Consumers' Ability to Integrate Visual, Numeric and Categorical Attribute Information. Working Paper, forthcoming.

ISBN 978-0-7070-0400-6

The Economic and Social Research Institute, Whitaker Square, Sir John Rogerson's Quay
Telephone +353 1 8632000 **Fax** +353 1 8632100 **Email** admin@esri.ie **Web** www.esri.ie