EXPLORING THE ELICITATION OF VALUE

LIFE, HEALTH AND CONSUMER GOODS

by
Danae Arroyos Calvera

Thesis submitted to the University of Warwick
for the degree of
Doctor of Philosophy

Warwick Business School

September 2018

Table of Contents

List of Figures

List of Tables

Acknowledgements

Writing this thesis is probably one of the hardest things I have ever done. I would like to use this space to express my gratitude to the people that have lightened the weight and have, in one way or another, furthered the development of this piece work.

First and foremost, I would like to thank my supervisors, Graham Loomes and Andrea Isoni. Thank you for giving the me chance of pursuing a PhD. Thank you for your guidance. Working with you has taught me many valuable lessons.

Second, I am grateful to Rebecca McDonald for her selfless support. Your informal supervision has made the ride fun, inspiring and empowering; and every time we work together I am reminded of how much I love doing research. I cannot tell you how much I *value* all this.

Third, I want to acknowledge the contribution of Madre, Padre y Eva. My education has always been their priority, and for that I am forever indebted to them. *No habría escrito la tesis, ni sería la mitad de la persona que soy hoy sin vuestro apoyo, consejo y las alas que siempre me habéis dado*.

Last but not least, I need to thank my friends: the people who make good times better and stick around through thick and thin. I feel immensely fortunate that our paths crossed. I would not have been able to produce a PhD thesis without your caring, generous, and enlightening friendship. Leaving my thesis aside, you make me strive to grow as a person, and for that I am thankful.

Declaration

Name and Surname: Danae Arroyos Calvera

Warwick ID: 1363605

Degree: Doctor of Philosophy

Title of the Thesis: Exploring the elicitation of value: life, health, and consumer goods

Department: Warwick Business School

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. I declare that this thesis is the result of my own work. This thesis has not been submitted in any previous application for any degree.

A former version of Chapter 2 was presented at the 2016 Foundations of Utility and Risk (FUR) conference in Coventry (UK), the 2016 Judgement and Decision Making Meeting for Early-Career Researchers (JDMx) in Basel (Switzerland), and the 2016 Behavioural Science PhD Research Exchange in Coventry (UK). My supervisors, as well as Rebecca McDonald, provided feedback at all stages of the development of this chapter.

A former version of Chapter 3 was presented at the 2017 Subjective Probability, Utility and Decision Making (SPUDM) conference in Haifa (Israel), the 2018 FUR conference in York (UK), and the 2018 Economic Science Association (ESA) conference in Berlin (Germany). My supervisors and Rebecca McDonald provided feedback at all stages of the development of this chapter.

A former version of Chapter 4 was presented at the 2018 Summer Health Economists Study Group (HESG) meeting in Bristol (UK). Graham Loomes, Rebecca McDonald, and Judith Covey were involved in the design of the study, but the majority of the study preparation, data analysis and write-up are the result of my own work.

Signature of Student:

Name of Student: DANAE ARROYOS CALVERA

Abstract

In this thesis, we set out to explore some of the issues that arise when we elicit the societal value of life and health. Many governments use societal preferences to inform their policy and budget allocation decisions. However, there is plenty of evidence contesting the reliability and validity of the elicited preference estimates. A possible driver for this may be the fact that the economic theory underlying elicitation methods and value metrics lacks descriptive validity. First, expected utility theory (EUT) may not provide a valid description of people's behaviour under risk. Second, the account that people exclusively care about maximising their own utility, disregarding that of others, may not be accurate. Using the wrong value estimates could lead to sub-optimally allocating enormous amounts of money, with the consequent loss of welfare; hence the relevance of this research.

The first part of this thesis focuses around violations of EUT, in particular the reduction of compound lotteries axiom (RCLA) and the independence axiom. In Chapter 2, the disparities between the utility estimates elicited directly and indirectly using the standard gamble method are studied. They could be caused by a failure of the RCLA, but also by the context in which the estimates were elicited. We found that after having removed the effect of affect and the lack of incentives, and when choice complexity was the lowest, the disparities, and hence the violation of the RCLA, persisted. This suggests that context may play a role, but it is not sufficient to explain the disparities. In Chapter 3, the common ratio effect was examined as the complexity of choices was manipulated by having either money or objects as outcomes. As prior studies did, we found the independence axiom was violated with money, but we did not find this with objects. The higher level of consistency with objects should not be taken as evidence of the validity of EUT; rather, it is likely to stem from participants resorting to heuristics (such as "always choose the prospect with the best payoff") when choices are too complex and hence difficult to make. Both of these chapters provide evidence that decisions may be domain-specific; this should be taken into account when testing models of decision making in the lab.

The second part of the thesis focuses on social preferences in the context of fatality and physical risk reductions. In Chapter 4, the effect of self-interest on the efficiency-equity trade-off was studied. Efficiency was found to be the most important concern, closely followed by equity. Self-interest significantly modified this trade-off, but the magnitude of this effect was not large enough to overturn the general preference for efficiency over equity. On the other hand, stakes (whether the risk under consideration was fatal or non-fatal) did not change these preferences.

Chapter 1. Introduction

1.1. Economic Evaluation of Life and Health

In many countries, citizens trust their governments to collect taxes and use them to provide public healthcare, transport infrastructure and many more services. One of the outcomes of the provision of these goods is the reduction of the population's physical (i.e., illness or injury) and fatality risks, thanks to improved healthcare and additional transport safety. Utilitarian economists, who measure the success and failure of policies by the change in utility they achieve, prescribe that the provision of public sector goods and services should be such that the population's welfare is maximised. This maximisation is subject to the limited availability of resources, and hence governments need to prioritise some interventions or policies over others.

Two tools that inform these trade-offs are cost-benefit analysis (CBA) (Williams, 1974) and cost-effectiveness analysis (CEA; see Neumann, Sanders, Russell, Siegel, and Ganiats, 2016). As their names indicate, in these analyses the cost of the resources that would be used up is weighted against the welfare increase that the population would accrue. The difference between these methods is the units in which the benefit or increase in welfare is measured. In CBA, the benefits are measured in monetary units. Benefit metrics used for CBA include the value of statistical life (VSL), which has been shown to be equivalent to the population marginal rate of substitution between money and risk of death by Jones-Lee (1989) and has been widely used (see meta-analyses of studies using it by Viscusi and Aldy (2003) and De Blaeij, Florax, Rietveld, and Verhoef (2003)). Another example is the value of a life year, which was first used by Johannesson and Johansson (1996, 1997). In CEA, on the other hand, the effect of interventions is accounted for in health benefit units such as quality-adjusted life years (Drummond, Sculpher, Claxton, Stoddart & Torrance, 2015) and disability-adjusted life years (Jamison, Mosley, Measham & Bobadilla, 1993; Murray, 1996). The advantage of CBA over CEA is that making the trade-off between costs and benefits is straightforward, as both are measured in the same units. However, as we will explain later in this section, translating health benefits into monetary units may be challenging.

Traditionally, the method used for both health and safety valuation was CBA (Neumann, Goldie & Weinstein, 2000) but currently healthcare interventions are mostly appraised using CEA (see the 2013 National Institute for Health and Care Excellence guidelines) while CBA prevails in transport safety (see the 2018 guidelines by the Department for Transport). In the 90s, the

benefits of healthcare and transport safety were estimated using the *human capital approach* (see Johannesson and Jönsson, 1991). This method equated the benefits of health improvements to the earnings that would otherwise be lost – find an early application of the method in Mushkin (1978). While the benefit metric was easy to compute, this method was criticised because it lacked an economic welfare theory foundation (as it ignored people's preferences) and it implied that leisure and retirement had no value (Schelling, 1968; Mishan, 1971). These critics also highlighted that people do not spend all the money they earn on reducing their risk of illness, injury or death, but rather they make trade-offs between their money and their risks. Hence, considering the amount of money people would be willing to pay to obtain a given risk reduction (or willing to accept to be compensated for the increased risk) would be a more appropriate way of measuring the value of health improvements to them. This approach is grounded on the economic principle of consumer sovereignty, which states that consumers are the best judges of what is best for them (read more in Knox (1960)). Thus, in order to maximise welfare, public budgets should be allocated so as to satisfy their preferences.

Depending on how these preferences are estimated, we distinguish between *revealed* and *stated* preferences. Revealed preferences are inferred from decisions that members of the relevant population make in a related market. One way of capturing these is through hedonic wage studies[1] (Rosen, 1974), where the estimates of the trade-offs that people make between money and physical risk are obtained by looking at the wages of workers that are exposed to varying degrees of physical risk in their workplace. In stated preference studies, a representative sample of the relevant population is directly asked about their preferences, using interviews and/or questionnaires.

The strength of revealed preference methods is that the values obtained are inferred from real decisions instead of hypothetical ones. Their downside is that they make strong assumptions about how much people know about risks involved in different situations and ignore the impact of factors such as collective bargaining or monopolistic power on wages. Moreover, it may not be adequate to extrapolate the findings from a given revealed preference study to people who

[1] Not all revealed-preference studies rely on wages. Other studies used to estimate metrics such as the VSL have used demand for consumer-safety products, and residential-property values instead – see reviews in Viscusi and Aldy (2003) and Blomquist (2004). However, Cropper, Hammitt and Robinson (2011, p. 317) explain that these are usually considered inferior: "Researchers often argue that these studies are less suitable for valuation than are hedonic wage or stated-preference studies because of difficulties in estimating actual or perceived risks, the need to make assumptions about key factors such as time costs (in some product studies), whether cancers are likely to be fatal (in some hedonic property value studies), and other factors".

are not facing those decisions. Stated preference methods do give hypothetical answers and the way in which participants perceive and respond to questions can sometimes be problematic. For example, answers suffer from embedding, which makes respondents declare similar willingness to pay for a subset and for the whole benefit (Kahneman & Knetsch, 1992; Hausman, 2012). But they provide answers that are tailored to the scenario of interest from a representative sample of the population.

Stated preference techniques for CEA can be classified in two categories, depending on the response method – see Drummond et al. (2015) for a comprehensive review of these methods. Under scaling, the most popular method is the visual analogue scale or VAS – Torrance, Feeny and Furlong (2001) provide a critical appraisal of this method. Under choice, the most commonly used are the standard gamble or SG (Torrance, 1986; Bennett & Torrance, 1996), the time trade-off or TTO (Torrance, Thomas & Sackett, 1972), and the person trade-off or PTO (Patrick et al., 1973; Nord, 1992). In these, respondents are asked to make trade-offs between different health states with different levels of risk (SG), durations (TTO), and numbers of people affected (PTO).

The most relevant elicitation technique to obtain monetary estimates of the benefits is contingent valuation (CV), also known as the willingness-to-pay (WTP) approach. Respondents are asked how much money they would be willing to pay for a physical risk reduction or health improvement (or willing to accept as compensation for taking the risk or being in ill health), contingent on a market for the benefit existing. It was first used by Davis (1963) and further developed in the area of transport safety by economists such as Jones-Lee (e.g., 1976, 1985), but it has also been used in the area of healthcare – see reviews by Smith (2003), McIntosh, Clarke, Frew & Louviere (2010), and Donaldson, Mason & Shackley (2012).

In the UK, the National Institute for Health and Care Excellence (2013) and the Department for Transport (2018) endorse the use of stated preference methods to estimate the value of physical and fatality risk reductions. However, using these methods to estimate societal preferences presents several challenges. The wish to gain a better understanding of these challenges and the possible policy-relevant implications that may come with this motivates the research in this thesis. Drawing on this question, we will also learn more about judgement and decision making under risk, and the nature of people's preferences.

1.2. Challenges

It is of paramount importance to obtain estimates of the value of physical risk reductions that accurately reflect society's concerns. Otherwise, public budgets of millions of pounds could be sub-optimally allocated, resulting in an over-provision of some publicly provided goods and services in detriment of others that would have achieved a higher welfare increase. However, a substantial body of empirical evidence casts doubt on the reliability of the value estimates elicited with the methods outlined in the previous section, and their validity as accurate expressions of people's preferences.

The methods outlined above offer different ways of eliciting the value of health improvements or life extensions to society. According to the principle of procedural invariance (Kahneman & Tversky, 1984), the estimates obtained should be the same regardless of the method used to elicit them, but that is not what is found. Jones-Lee, Chilton, Metcalf & Nielsen (1995) used both the CV and SG methods to elicit the value of several non-fatal health states. They then computed the ratios between each of the states and a worse one, to find out how bad they were relative to each other. They found that the SG ratios were smaller than those from CV, implying that the SG elicited less severe judgements of the health states than CV did. Along the same lines, Dolan, Gudex, Kind and Williams (1996) found differences between the estimates from the SG and the TTO method. They elicited health state value estimates that could range from 0 (death) to 1 (best health), and found that the median values were generally at least 5% higher when the TTO method was used, compared to the SG. Bleichrodt, Pinto and Abellán-Perpiñán (2003) found the same, although a few other studies found the opposite (e.g., Rutten-van Molken, Bakker, van Doorslaer & Van Der Linden, 1995; Stiggelbout et al., 1994; Bennett & Torrance, 1996). This is only a sample of the studies that compare estimates elicited with different methods and find that there are disparities, but the evidence does not end there – for example, see a comparison of SG and VAS by Llewellyn-Thomas et al. (1984) or Torrance et al. (1996); one of CV vs. TTO by Gyrd-Hansen, Kjær and Nielsen (2012); and a comparison of SG, TTO, and CV, and other scaling methods by Frober and Kane (1989).

These findings question the appropriateness of using these methods to inform public policy, but their alternatives (e.g., using the effect of health on happiness to determine how to increase the population's welfare (Dolan & White, 2007; Dolan, Layard, & Metcalfe, 2011)) are not exempt from problems (Dolan, 2011). That is why understanding better the reasons behind the problems with the elicitation methods outlined above is very important.

Out of all the methods, the most worrying evidence comes from the two that have the strongest theoretical background: the CV method (Jones-Lee, 1989) and the SG method (Brazier, 2005; Drummond et al., 2015). They are both grounded on expected utility theory (EUT), which as Schoemaker (1982), Camerer (1995) and many others since have argued, may lack descriptive validity (i.e., accuracy when describing how people *do* make decisions, rather than how they *ought to*).

The deviations from EUT predictions found when using the CV are numerous. Theory predicts that the preferences people state should only be affected by changes in the elements they are trading off. Depending on the elicitation method, these could be the severity of the health impairment, the magnitude of the risk reduction, duration, or the number of affected people. However, value estimates have been found to hinge on irrelevant factors (such as the wording and the order in which the questions were asked) while being insensitive to changes in the factors that matter (mentioned above) (e.g., Kartman, Stålhammar, & Johannesson, 1996; Beattie et al., 1998; Carthy et al., 1999; Abellán-Perpiñán, Martínez-Pérez, Méndez-Martínez, Pinto-Prades, & Sánchez- Martínez, 2011).

Less is known about the failure of EUT in the SG method. This is the focus of the first part of this research. Value estimates can be elicited using one-stage or two-stage SGs (a full description of these is provided in the next chapter). EUT's reduction of compound lotteries axiom (RCLA) postulates that these two variants of the SG should elicit the same estimates. However, this is not what has been found empirically (Llewellyn-Thomas et al., 1982; Rutten-van Molke et al., 1995; Bleichrodt, 2001; Oliver, 2003). In Chapter 2, we aim to find out whether the disparities found within the SG method are due to the RCLA failure or whether context is responsible. we use *context* to refer to the affective nature of the good being valued, the lack of realism due to the absence of incentives, and the complexity of the choice (which is given by the ease with which the dimensions of the goods can be compared).

Others have previously looked at *choice complexity* and its effect on decision making. Although choice complexity has no unanimous definition, each author's own conceptualisation always encompasses at least some of the following features. Namely, the number of questions in the task (e.g., Dave, Eckel, Johnson & Rojas, 2010; Shiloh, Koren, & Zakay, 2001), the number of alternatives in the choice (e.g., Payne, Bettman, & Johnson, 1988; Shiloh, Koren, & Zakay, 2001), the number of attributes or dimensions in each alternative (e.g, Greifeneder, Scheibehenne, & Kleber, 2010; Danthurebandara, Yu, & Vandebroek, 2011), the levels and the range of the attribute levels (e.g., Caussade, de Dios Ortúzar, Rizzi, & Hensher, 2005; Dellaert, Brazell, &

Louviere, 1999), the correlation between these attributes (in the cases the attributes can be expressed in numerical terms, e.g., price, weight, battery life, etc. in the case of mobile phones) (e.g., Mazzotta & Opaluch, 1995; DeShazo & Fermo, 2002), and the similarity of the alternatives in utility terms (e.g., Swait & Adamowicz, 2001a; Dellaert, Donkers, & Soest, 2002). In an attempt to provide an overall measure of complexity, Swait and Adamowicz (2001b) develop an 'entropy' metric. This measure comprises the impact of the number of alternatives and attributes, the correlation between attributes, and how close in utility terms the alternatives are.

The higher the number of choices, alternatives, etc., the harder it is for people to make decisions (Timmermans & Vlek, 1992; Swait & Adamowicz, 2001a). Higher choice complexity has been found to lead to more noisy answers (Dave et al., 2010), higher response errors (Danthurebandara, Yu, & Vandebroek, 2001), and changes in choice strategy (e.g., Olshavsky, 1979; Timmermans, 1993; Payne, Bettman & Johnson, 1988, 1993; Ball, 1997; Swait & Adamowicz, 2001a; Dellaert, Brazell & Louviere, 2012). Most importantly, previous research has found a relationship between choice complexity and consistency. The work by Jacoby, Speller, and Kohn (1974), Jacoby, Speller, and Berning (1974), and Keller and Staelin (1987) links consistency with the amount of information that participants need to process, finding that the higher choice complexity is, the less consistency there is. DeShazo and Fermo (2002) also manipulate choice complexity, and using a random utility model they are able to quantify the effect of choice complexity on utility estimates, and hence its effect on consistency. Caussade, de Dios Ortúzar, Rizzi and Hensher (2005) find similar results. Arentze, Borgers, Timmermans and DelMistro (2003) do not find a change in the error variance when they manipulate the number of alternatives, but they do when they manipulated the number of dimensions per alternative.

The above-mentioned research on choice complexity has been done using a lot of different goods, all within the same domain. That is, it tells us more about how choice complexity affects decisions between a few mobile phones that vary in dimensions such as price, battery life, and memory size (Fasolo, Carmeci, & Misuraca, 2009). Other goods used are monetary lotteries (Payne et al., 1988), university majors (Shiloh et al., 2001), jobs (Keller & Staelin, 1987), apartments and stereo receivers (Olshavsky, 1979), landfill locations (Mazzota & Opaluch, 16995), and transport routes (Caussade et al., 2005), just to name a few. However, the existing literature does not shed light on how complexity brought by having to choose between goods in different domains (e.g., a stereo and a mobile phone, which have dimensions such as price in

common, but do not share a lot of their other characteristics), would affect consistency. The current research contributes to reduce this gap in the literature.

In Chapter 3, we turn to another well-known deviation from EUT: the common ratio effect, a preference shift that occurs when the probabilities of the better outcomes in two prospects are scaled down by the same factor. It was used by Kahneman and Tversky (1979) to show that people's actual decisions deviate from EUT predictions. In this chapter, the premise is that people's everyday decisions are not quite as simple as those that they make in the laboratory, where they are asked to make trade-offs between two uni-dimensional scales (money and probability). The goal of this chapter is to further explore the role that choice complexity brought by divergence in the outcome attributes may play in decision making. Neither EUT nor any of its most prominent alternatives (e.g., Kahneman and Tversky's (cumulative) prospect theory (1979; 1992)) predict choice complexity to have an effect on decisions. It is crucial to find out more about this since a lot of the existing models of decision making are being developed and tested using evidence from simplified choices between monetary payoffs. More importantly, these models are then being generalised to decision making with all outcomes without checking that behave is essentially the same in the new environment.

The second part of this research shifts away from the value elicitation methods to focus on the metrics that capture value. We want these metrics to be a good representation of people's preferences. Yet, we rely on standard economic theory that assumes that people are exclusively self-interested, (expected) utility maximising agents, while there is a great deal of empirical evidence suggesting that people also have *social* preferences (e.g., Fehr & Schmidt, 2001). This evidence implies that people may not only care about their own risk reduction, but they would also be willing to give up some of their own resources in order to make other individuals better off. It also suggests that people may not only be concerned about maximising output, in this case the overall risk reduction (*efficiency*), but they may prefer to have the benefit evenly distributed across the population (*equity*), even when it comes at the expense of efficiency. If the metrics do not account for this, they may fail to provide an accurate representation of the population's preferences. The question addressed in Chapter 4 is how the balance of self-interest and altruism affects the efficiency-equity trade-off. The findings from this study should give us a better idea of how value metrics should best account for the populations' preferences.

In the remaining of this thesis, the three above-mentioned chapters follow. The thesis then concludes with a general discussion chapter.

Chapter 2.
Can Context Explain the Disparities between the Direct and Indirect Standard Gamble Estimates?
Testing the Reduction of Compound Lotteries Axiom with Objects and Money

Abstract

The standard gamble (SG) method is widely used to elicit societal preferences about health. With the SG method, the elicitation can be done directly (in a single stage) or indirectly (two-stages). According to expected utility theory's (EUT) reduction of compound lotteries axiom (RCLA), the estimates elicited with these two procedures should be the same. However, when valuing health, the estimates elicited indirectly have been found to be significantly higher than the direct estimates (e.g., Oliver, 2003). This challenges the appropriateness of using these estimates to inform policy, and calls for a greater understanding the disparities found between the estimates. We aim to find out whether these disparities should be fully attributed to the failure of RCLA or whether the preference elicitation context is also playing a role. We use *context* to refer to the abstract and affective nature of the good being valued, the lack of incentives, and the complexity of the choice. We manipulated choice complexity by asking questions involving outcomes (with the same probability and subjective value) that were either money or consumer goods. We found that when complexity was the lowest, the disparities still persisted. This indicates that context alone is not sufficient to explain the disparities. While the number of consistent estimates was approximately the same for money and objects, we found that the direction of the disparities changed depending on whether the objects shared some dimensions or not. These findings call for caution when using the SG method, and provide evidence that subjective value is not the only characteristic of the outcomes that influences decisions; their complexity matters too.

Keywords: reduction of compound lotteries axiom, standard gamble, context

2.1. Introduction

Most economists and government bodies such as the UK Treasury (2018) and the US Office of Management and Budget (2003[2]) agree that the allocation of budgets to public healthcare and transport safety should be done according to the affected population's preferences (Carthy et al., 1999). These preferences are used to estimate the societal value of the benefits that would be generated by the policies or interventions. One of the methods used to elicit people's preferences, risk preferences in particular, is the standard gamble (SG). This method has been favoured over others because it reflects the risk involved in decisions involving health, and because of its theoretical grounding on expected utility theory (EUT; Oliver, 2003).

The axioms of EUT imply that we can locate the utility of any outcome *x* on a utility scale between two defined boundary outcomes. The lower bound corresponds to the utility of an outcome less preferred than *x*, and the upper bound to that of a more preferred outcome. In its simplest form, which we refer to as the *direct* SG scenario, outcome *x* is compared to a lottery that offers the upper bound outcome with probability *p*, and the lower bound outcome with probability 1-*p*. Normalising the utilities of the bounds to 1 and 0, the utility of *x*, for an expected utility maximiser, is given by *p*. When used to value health, the SG method allows us to find out how a given non-fatal health state compares to full health and death. The SG scenario offers two options: the certainty of being in the non-fatal health state being valued, or a chance of full health (otherwise death). The probability of ending up in full health that makes respondents indifferent between the certain and the risky options is the utility of the non-fatal health state relative to that of death and full health, which are set to 0 and 1.

If EUT's reduction of compound lotteries axiom (RCLA) holds, then we can elicit the utility of *x* *indirectly* through two or more linked SG scenarios. We introduce a new reference outcome *y*, whose utility lies between that of *x* and the lower bound outcome. In the first gamble, we elicit the utility of outcome *x* relative to that of the original upper bound and *y*; in the second gamble, the utility of outcome *y* is elicited relative to the original bounds. Combining these estimates reveals the utility of *x* relative to the original bounds. Importantly, if the RCLA holds, the direct and indirect utility estimates for outcome *x* should not differ.

We heavily rely on the equivalence between the direct and indirect utility estimates for *x* when valuing health. Direct elicitation can be problematic when the valued health state lies very close

[2] This Circular was still in place in 2017 – see Perkins and Carey (2017), the latest available edition of the "Cost-benefit analysis and financial regulator rulemaking" report for the US congress.

to one of the extremes; this is the case, for example, when valuing mild health conditions. In such cases, the difference between the utilities of the upper bound (i.e., full health) and the health state may be minimal, making people unable to discriminate well enough between large indifference probabilities and 1. Using different bounds allows us to increase the relative distance between the bound and the outcome being valued, increasing the sensitivity of the elicitation procedure. Following the indirect elicitation example above, the new reference outcome *y* could be a severe but non-fatal health state. Considering the mild health state in relation to full health and outcome *y*, rather than death, would make the difference between the mild condition and full health more visible. This, in turn, would make respondents' estimates of the utility of the mild condition distinct from 1.

In actual health valuation studies, the direct and indirect utility estimates have been found to differ systematically, in violation of the RCLA. When replacing the original lower bound (e.g., death) by a new one (e.g., a less severe illness or injury), the indirect estimates have been found to be significantly higher than the direct ones (Llewellyn-Thomas et al., 1982; Rutten-van Molke and et al., 1995; Bleichrodt, 2001; Oliver, 2003).

We conceive of two possible explanations for the disparities found in the health literature. The first is that EUT, and particularly the RCLA, do not accurately describe individual decision making. The RCLA requires a linear treatment of probabilities, hence one way this axiom could fail is if people treated probabilities non-linearly. Bleichrodt (2001) and Oliver (2003) explored whether SG choices would be better described by different specifications of cumulative prospect theory (CPT; Tversky & Kahneman, 1992). Both Bleichrodt and Oliver found that applying an inverse-S shaped probability weighting function, of the kind proposed in CPT, did not result in better descriptive accuracy compared to EUT. Oliver also investigated whether different specifications of CPT (probability transformation and loss aversion weighting, probability transformation only, loss aversion weighting only, and the transformation of small and large probabilities only) would do better than EUT. Out of those, the one specification that resulted in improvements in descriptive validity compared to EUT was the addition of loss aversion only.

Alternatively, the disparities may be induced by some of the features of the design, or context, of the SG method. Some of these are inherent characteristics of health. Health is affect-rich, i.e., considering (hypothetical) health states may prompt affective reactions that other outcomes may not prompt, and these reactions may push respondents to react in extreme ways. Beattie et al. (1998) had suggested that dealing with death could be too emotional, and Carthy et al. (1999), in order to make decisions "more manageable", asked participants about milder

conditions, implying that the more severe conditions are, the harder it is for people to make decisions. Moreover, choices in health valuation studies are very complex, in that they require respondents to make trade-offs between items that vary along many dimensions that are hard to compare to each other. Last, answers in health valuation studies cannot be incentivised, and this may raise questions about the quality of respondents' answers[3].

Our goal is to examine whether the disparities between the direct and indirect SG utility estimates are caused by the elicitation context or the lack of validity of the RCLA. We simplify the elicitation context as much as we can to give the RCLA a good chance of holding. First, in order to rule out the impact of affect, we steer away from health and use simpler outcomes. In our study, half of the gamble scenarios have everyday objects as outcomes, and in the other half the outcomes are the money amounts that each participant stated were equivalent in value to the objects. The money amounts allow us to explore the validity of the RCLA when choices are the least complex. Here, the complexity of a choice is defined depending on its outcomes. The lower the fraction of the dimensions that can be used to compare the outcomes to each other, the more complex the decision is. For instance, questions involving money amounts are the least complex because the outcomes have one dimension only, and it is shared. By comparing the money utility estimates with those elicited using objects, we can observe the effect of increasing choice complexity while the subjective value of the outcomes remains constant. Using consumer goods and money also frees our results from the possible time discounting confound (see Gafni (1994) for a more detailed discussion) as, unlike health, our outcomes materialise with no delay. Another advantage of using these goods is that we can fully incentivise responses.

The high stakes in health valuation make understanding these disparities crucial. EUT and the RCLA provide the theoretical grounding for the chaining in the SG method and hence the estimates elicited with it. If this theory is not descriptively valid, then there is no reason to believe that the estimates derived from it accurately reflect the population's preferences. Public

[3] The provision of incentives is crucial in experimental economics. Smith (1982) formalised this as the *nonsatiation* precept, which implies that in order to achieve a controlled experimental environment where respondents truthfully reveal their preferences, respondents' rewards should be directly linked to their decisions during the study. In studies where participants need to choose between lotteries with monetary payoffs, participants shall be paid an amount of money according to their choices at the end of the session. However, when the choices respondents are making involve physical and fatality risks it is much harder, and probably unethical, to make these outcomes materialise.

policy decisions based on such estimates could result in millions of pounds sub-optimally allocated to the wrong treatments or safety measures (e.g., Carthy et al., 1999).

We found substantial disparities between the direct and indirect estimates. That is, under the most favourable conditions we could set up, the RCLA was often not descriptively valid. With regards to our complexity manipulation, we found that even though participants had stated that the objects and the money amounts were equally valuable to them, whether objects or money were used in the elicitation systematically influenced the utility estimates. Looking at the direct estimates (elicited with a single gamble scenario), we found that participants were less risk averse with objects than they were in gambles with their counterpart money equivalents. Looking at the comparison between the direct and the indirect estimates, we found that whereas the amount of disparities was approximately the same for all payoffs, for objects the distribution of the disparities between the estimates was skewed in opposite directions depending on how complex the choice between them was (i.e., depending on the fraction of comparable dimensions they had).

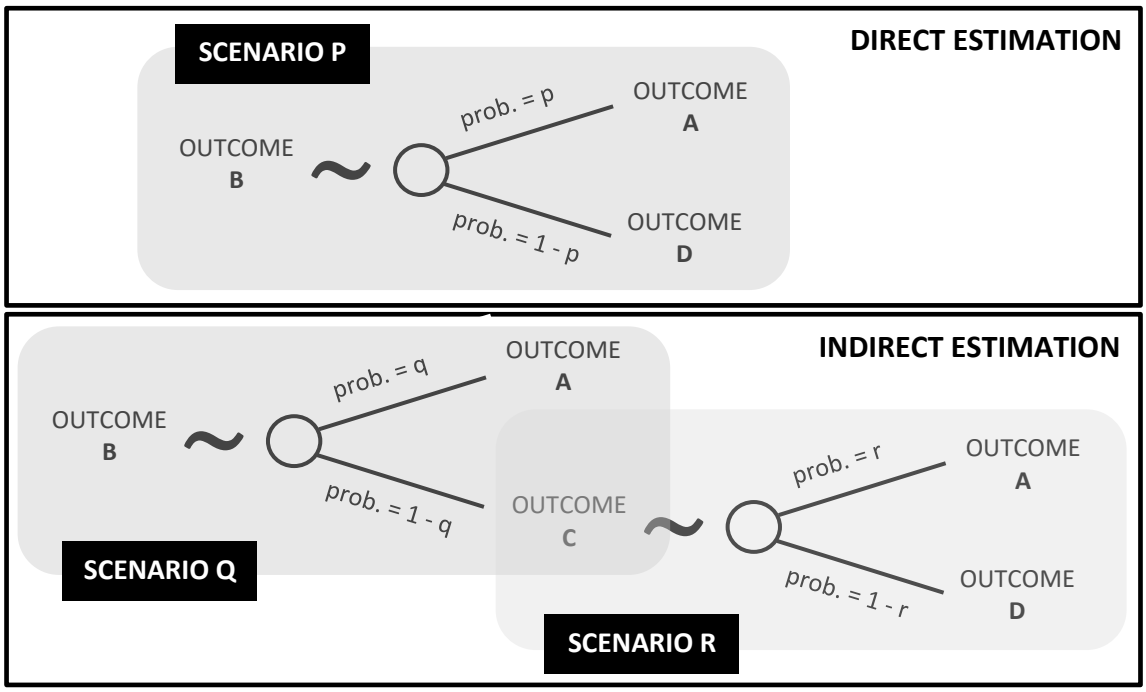
The rest of the chapter is organised as follows. Section 2 includes a summary of the theory underpinning the standard gamble, and previous empirical work on the consistency between the direct and indirect utility estimates. Section 3 explains the design of our study. Section 4 examines our results. And the chapter concludes with the discussion of our results, in section 5.

2.2. Theoretical Background

2.2.1. Consistency Requirement in the Standard Gamble Method

EUT is a general theory that applies to consequences of any kind. Consider four outcomes A, B, C and D such that u(A) > u(B) > u(C) > u(D) where u(.) is a standard Von Neumann and Morgenstern utility function (1944). In order to elicit the utility of the intermediate outcome B directly and indirectly, respondents would see three SG scenarios (P, Q, and R) such as those in Figure 1. Throughout this paper, we use 'SG scenario' to refer to a two-prospect choice. Consider gamble scenario P. The *sure* prospect offers a given outcome for certain. The *risky* prospect is a two-outcome lottery: one of the outcomes is better than the one given for certain and is offered with probability *p*; the other outcome is worse than the certain one, and is associated with probability *1-p*. For each of the SG scenarios, we elicit respondents' *indifference probability* (p, q and r): the chance of receiving the best outcome that they would require to be equally well off, or indifferent, by taking the sure or the risky prospects.

Figure 1. SG scenarios P, Q and R



According to EUT, the utility of a prospect corresponds to the expectation of the utility of its payoffs. When respondents are indifferent between the two prospects of an SG scenario, the utilities of the intermediate outcomes in each of the scenarios are given by the following equations.

$$Scenario\ P: u(B)_P = p \cdot u(A) + (1-p) \cdot u(D) \tag{1}$$

$$Scenario\ Q: u(B)_Q = q \cdot u(A) + (1-q) \cdot u(C) \tag{2}$$

$$Scenario\ R: u(C)_R = r \cdot u(A) + (1-r) \cdot u(D) \tag{3}$$

The utility of intermediate outcome B can be elicited directly using scenario P: an estimate of this utility $u(B)_P$ is given by equation (1). The RCLA implies that an indirect estimate of the utility of B ($u(B)_{PQ}$) can be obtained by combining the estimates from scenarios Q and R as follows. After we normalise the utility of the best outcome (A) to 1, and that of the worst outcome (D) to 0, equations (1) to (3) become the following.

$$Scenario\ P: u(B)_P = p \tag{4}$$

$$Scenario\ Q: u(B)_Q = q + (1-q) \cdot u(C)_R \tag{5}$$

$$Scenario\ R: u(C)_R = r \tag{6}$$

Equations (5) and (6) can be further combined into equation (7), which captures the estimate of the utility of the intermediate outcome B in terms of the indifference probabilities of scenarios Q and R.

$$u(B)_{QR} = q + (1 - q) \cdot r \tag{7}$$

If EUT and the RCLA held, we would expect scenarios P, and Q and R to give consistent estimates of the utility of the intermediate outcome B. That is, the utility of outcome B given by gamble P ($u(B)_P$, equation (4)) and that given by scenarios Q and R ($u(B)_{QR}$, equation (7)) should be the same.

If the direct and indirect elicitation procedures result in the same estimate, then equations (4) to (6) can be combined into one to express the *consistency* requirement in terms of the indifference probabilities of the three gambles.

$$p = q + r - q \cdot r \tag{8}$$

As depicted in Figure 1, the indirect estimation in this paper is done by chaining through the failure outcome (the worst one in the gamble scenario): scenario R is used to elicit the utility of the worst outcome of scenario Q, rather than the success outcome (the best one in scenario Q). When valuing health, chaining through different outcomes (either failure or success) has systematically influenced whether or not the consistency requirement is met (e.g., Bleichrodt, 2001; Llewellyn-Thomas et al., 1982; Rutten-van Molken et al., 1995). Namely, chaining through the success outcome results in consistent estimates, whereas chaining through failure results in the indirect estimates being larger. Because the goal of this study is to understand further these disparities, we chose the methodology that has been found to produce them.

2.2.2. Failures of the Consistency Requirement

It is not only in the health and safety valuation context that the RCLA has been found to fail. There is a substantial body of research that has refuted the axiom in the lab. Studies generally present the *reduced* and *compound* forms of a choice (equivalent to our direct and indirect elicitation procedures) and observe whether, as should happen if the RCLA held, participants do choose the same prospect irrespective of the version of the choice that has been presented.

There is some evidence from non-monetary domains: Ebert and van de Kuilen (2016) found the violation of the axiom using waiting time as the outcome, and Deck and Schlesinger (2016), using pleasant and unpleasant jelly beans. Most of the studies, though, test the validity of the axiom using monetary outcomes. For example, Bar-Hillel (1973) asked participants to choose between the reduced and compound forms of the choice to investigate whether conjunctive and disjunctive events were judged differently. She found that the RCLA failed to describe behaviour. So did Kahneman and Tversky (1979), who then explained the violation in terms of respondents

‘isolating’ or ignoring the first stage of the compound lottery. Keller (1985a) found that the violation was susceptible to the presentation mode. They tested the axiom representing the choices with drawings of marbles in tubes, bar graphs, written description, and matrices where cell sizes were proportional to the probability. She found a percentage of violations that ranged from 32% (with the marbles) to 45% (with the matrices). Bernasconi and Loomes (1992), who were studying two-stage Ellsberg-type (1961) of problems, also found that the axiom was violated. Bernasconi and Bernhofer (2017) also observed the violation in the laboratory with monetary consequences, but with a tax evasion framing instead of the usual neutral one.

On the other hand, Harrison, Martin-Correa, and Swarthout (2015) found that the violation took place when one out of all the questions were incentivised, but it disappeared when respondents answered one question only which was incentivised for sure. Using the same procedure as the studies above (i.e., having participants choose between prospects in both the reduced and compound forms of the choice), Hajimoladarvish (2018) did not find evidence against the validity of the axiom either.

It is worth noting that the task in this study and the one used to elicit health state utilities are not choice tasks, but rather (probability) equivalence tasks. It is well-established that choice and equivalence tasks may elicit different preferences. The classic example of this is the preference reversal between the p-bet and the $-bet, first reported by Slovic and Lichtenstein (1968), Lichtenstein and Slovic (1971), and Lindman (1971). Despite both gambles having similar expected value, in a choice exercise respondents would choose the p-bet (high probability, low payoff), while in an equivalence task they would state a higher certainty equivalent for the $-bet (low probability, high payoff). Butler and Loomes (2007) find the opposite preference reversal when, instead of using certainty equivalents, the preferences stated in the choice task are compared to those inferred from respondents’ probability equivalents, which is what we elicit with the SG method.

We found three studies that test the validity of the axiom using a certainty equivalence task rather than a choice task: Miao and Zhong (2012), Abdellaoui, Klibanoff, and Placido (2015), and Hajimoladarvish (2018). All three found evidence against the validity of the RCLA using monetary outcomes. To our best knowledge, there is only one study that used a probability equivalence task *and* non-health outcomes: Spencer (1998), where the outcomes were money amounts. Chilton and Spencer (2001) used Spencer’s data to explore the consistency between the SG utility estimates elicited directly and indirectly. Looking at mean estimates, they found that RCLA

held in half of their comparisons. When the RCLA failed, contrary to what happens with health, they found the direct utility estimates were higher than the indirect ones.

Chilton and Spencer's results provide valuable insights regarding the validity of the RCLA in a setup close to that in health valuation studies, but they do not answer our question regarding the effect of context on the disparities found when using the SG method. With their design, it is not possible to disentangle which of the two factors that differentiated Spencer's study was behind the occurrence and the direction of the disparities found. Differences may have been found because they used money instead of health outcomes, so choices were less complex. But it may have also been that they chained through the success outcome, rather than the failure one, as previous studies that had found the disparities did. Further, at no point were the utilities of the money amounts and health directly compared, hence the differences in disparities across studies could be due to differences in the utilities of the outcomes, rather than because they were health-related or not. Last, their answers were not incentive compatible (Spencer, personal communication, July 20, 2018). While this could help to make health and money valuation studies more comparable, it prevents the study from shedding light on the role of (the lack of) incentives in the disparities found.

2.2.3. Possible Reasons for the Failures

We hypothesise that other than the lack of descriptive validity of the RCLA, the disparities found may be caused by the context of valuation studies. One of these contextual factors is the nature of the good being valued.

First, health is affect-rich, so considering it may trigger a strong emotional response. In this study, we elicit preferences over goods that are not affect-laden: our outcomes A, B and C are everyday objects and money (the remaining outcome, D, is to receive nothing). Most people will be familiar with money and the objects in our experiment, which should not carry strong emotional responses, and this allows us to rule out the effect of affect.

Second, decisions involving health states are complex: health varies along many dimensions that are difficult to compare to each other, and it is very abstract; it may be difficult to conceptualise for people, especially if they have never experienced the given health states before. We link choice complexity to the ease with which consequence dimensions can be compared. More specifically, the higher the fraction of *comparable* dimensions, the less complex, in relative

terms, the choice will be[4]. For instance, in a choice between £5 and £10, there is only one dimension to consider: the amount of money. Because the only dimension is comparable across consequences, this choice will be relatively simple. In contrast, consider choosing between a coffee machine and a pair of jeans. Each of these items has lots of dimensions, and only a very small fraction of them are comparable. Hence, this decision would be much more complex, and harder to make.

From lower to higher complexity, we ask participants to make choices between lotteries that involve (1) the money amounts that each participant had stated were equivalent in value to the objects (they have the 'amount' dimension only, and it is shared), (2) *similar* objects (which have some obvious comparable dimension but differ in the rest of them, i.e., alarm clocks with different additional features), and (3) *dissimilar* objects (which do not have obvious comparable dimensions, e.g., a toaster and a pair of headphones).

Manipulating choice complexity allows us to test whether this is one the contextual factors that may affect consistency. Given EUT's prediction that people make decisions on the basis of the utility or subjective value of the outcomes but not their comparability to each other, consistency should be the same regardless the complexity of the choice. However, we hypothesise that the smaller the fraction of shared dimensions, the more difficult it will be to make choices, and the less consistency there will be. To our best knowledge, no study to date has investigated the effect of shared or divergent dimensions (which we characterise as choice complexity) on consistency. But difficulty induced by an increasing number of alternatives in a choice, dimensions, correlation among these dimensions, etc. has been found to have a negative effect on consistency (e.g., DeShazo and Fermo, 2002; Arentze et al., 2003; Caussade et al., 2005).

Using consumer goods and money also enables us to capture the effect of another factor that could be causing the disparities, the lack of incentives. In theory, incentive compatibility ensures that it is in respondents' best interest to provide truthful answers to the questions we pose to them. However, in many of the previous studies using the SG method respondents were valuing health, so for obvious ethical reasons, their answers could not be incentivised. The reason for

[4] Note that this definition should also apply to health. For example, a choice between two health states defined in terms of the same dimensions (e.g., EQ-5D (Brooks, 1996) dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) should be *relatively* less complex than one where one health state is defined with the EQ-5D instrument, and the other is defined with another instrument which differs in some of the dimensions (e.g., SF-6D (Brazier, Roberts & Deverill, 2002), where the dimensions are physical functioning, role limitations, social functioning, pain, mental health and vitality).

the disparate answers found when valuing health could hence be that respondents were not devoting enough effort to accessing their true preferences.

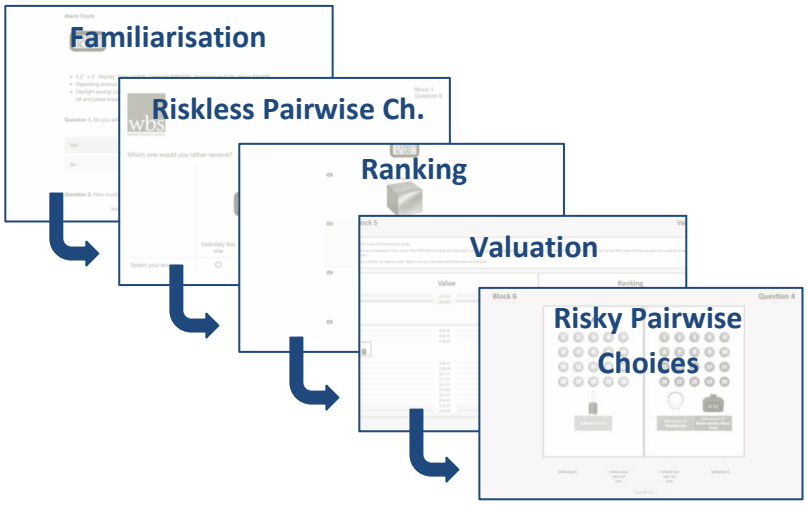
2.3. Method

2.3.1. Design and Procedure

In order to test the consistency between the direct and indirect utility estimates, and between the estimates elicited with objects and money amounts, we need respondents' monetary valuations of the objects and their indifference probabilities for the gamble scenarios. In order to elicit these, we conducted the following study at the University of Warwick.

Participants completed five different tasks on a computer (see Figure 2), all to do with objects and money amounts: familiarisation, riskless pairwise choice, ranking, valuation, and risky pairwise choice. The first four tasks were carefully designed to elicit respondents' valuations of the objects. They gave respondents several chances to reflect upon their preferences before they gave a monetary valuation that was as precise as possible. The fifth and last task is the one we used to obtain participants' indifference probabilities.

Figure 2. Study Flow



Familiarisation

The 10 objects[5] involved in the decisions respondents made throughout the study (Figure 3) were on display at the lab and participants were encouraged to have a look at them before the experimental session started. Within these objects, there were three chosen to have at least

[5] These objects were chosen out of a pool of 30 objects based on their desirability (as reported by respondents in a pilot study; see more details about the pilot studies in Appendix A).

one dimension in common ('similar objects' set; an alarm clock, an alarm clock with a radio, and an alarm clock with a radio and a Bluetooth speaker). Three other objects were chosen so that they were dissimilar to each other in terms of their dimensions. These were an air mattress, a clicker (used in presentations to navigate the slides and point at the screen), and a toaster, and we refer to them as 'dissimilar objects'. The remaining objects (a bottle, a headphones set, a computer mouse, and a suitcase) were included to add variety in this study, and for answering questions that are not relevant in this study.

Figure 3. Consumer Goods in this Study

Alarm Clock | Radio Alarm Clock | Radio-speaker Alarm Clock | Air Mattress | Clicker (for presentation)

Toaster | Bottle | Headphones | Mouse | Suitcase

The study started with the object familiarisation task: participants saw each object on their screen together with a brief description. At any point during the study, participants could click on the image of the objects to see this information again. They also had this information on printed cards on their desks. To encourage them to think about each object, we also asked them whether they owned such object, and how much they would like to receive them (on a 7-point scale from "not at all" to "very much").

Riskless Pairwise Choices

Then, participants made pairwise choices that involved no risk. This task was included in the study after findings from a pilot study suggested it would be helpful to participants in order to introspect about their preferences with regards to the objects (see more details about this pilot study in Appendix A.1). In this task, respondents had to indicate which item, out of two, they would prefer to receive. There were 75 of these questions: 45 were choices between every

possible combination of two out of the ten objects, and 30 were choices between one object and a money amount (either £5.25, £8.75, or £19.50). The money amounts were included in this task to ease respondents into the valuation task, where they had to think of their money equivalents for the objects. The three money amounts were carefully picked so that participants did not take them as anchors and that influenced their valuations.

Respondents had the incentive to make each of these choices carefully because at the end of the study, one of these choices could be randomly selected, in which case they would receive the item they said they preferred.

Ranking

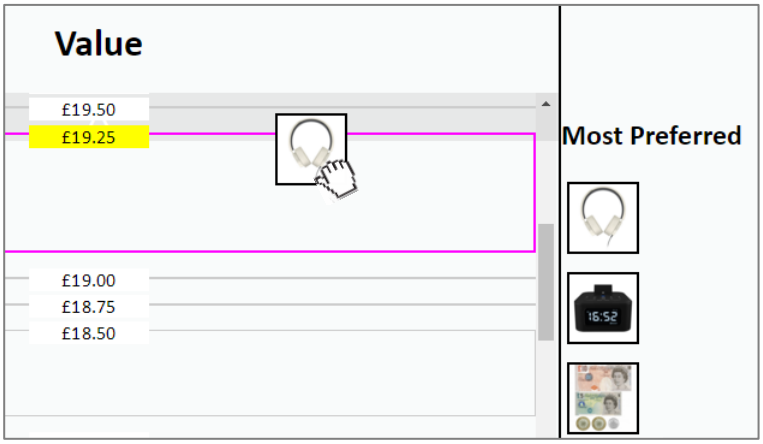
Based on each respondent's answers in the riskless pairwise choice questions, a provisional ranking of preference was inferred, and respondents were invited to adjust it to better represent their preferences if necessary. They had an incentive to do so, because if the ranking task was chosen to be played out for real at the end of the study, two of the items would be selected at random, and respondents would receive the one they had ranked higher.

Valuation

Once the ranking was submitted, the valuation task started, and participants could see their ranking on the screen again. Participants had to drag and drop each item (the 10 objects and 3 money amounts) to the cell in a ladder with the amount of money they considered equivalent to the item[6]. For example, if according to one respondent, the headphones were as valuable to them as £19.25, that respondent would drag the picture of the headphones to the cell with the "£19.25" label (as shown in Figure 4). The money amounts in the ladder started at £0 and increased in £0.25 intervals, but one could scroll up indefinitely.

[6] The instructions in the valuation task were the following: "We want to know what amount of money you regard as equivalent to the items you just ranked. That is, we would like you to tell us what amount of money YOU PERSONALLY regard as being exactly as desirable TO YOU as the item. Please try to forget about what it might cost in the shops or online. Just think: if you had to choose between this item and some money, what amount of money would make you feel you could spin a coin and be equally happy to take the item or else the money?" That is, the objects' monetary equivalents correspond to the equivalent gain.

Figure 4. Detail of the Valuation Interface



The elicited monetary valuations (or *money equivalents*) were used to construct the gamble scenarios. First, they allowed us to find out about the relative ordering of the objects within the similar and dissimilar objects sets; this determined which object would appear in the gamble scenarios as outcome A, B and C[7]. Second, they were used to substitute the objects in the money gambles.

In the valuation task, the incentive scheme was based on Becker-DeGroot-Marschak's (1964) procedure: one of the items was randomly picked, and its valuation according to the participant was compared to a random money amount[8]. If the randomly-generated money amount was higher than the stated valuation, the participant received the random sum of money rather than the object; whereas if the money amount was less than the stated value, the participant received the object. If the two were exactly the same, then a toss of a coin would determine whether the participant would receive the money or the object.

Risky Pairwise Choices

Last, participants completed 100 risky pairwise choices, where they had to choose the prospect they preferred out of two. Not all of these questions were relevant for this study. The choices of interest can be organised in 4 groups, each of which comprise gamble scenarios P, Q and R with a different set of outcomes. The questions in the first group involved the similar objects. In

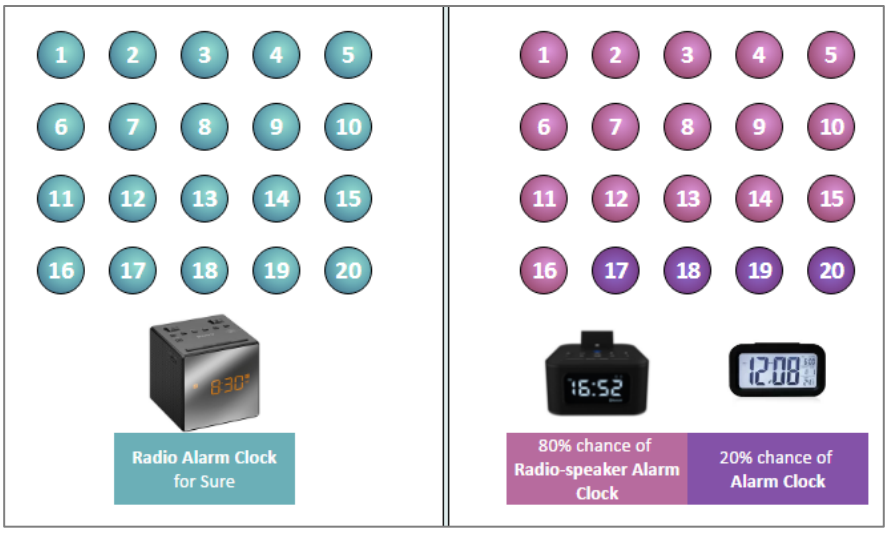
[7] When someone had given the same valuation to two or more objects within one of the objects set, the tie was solved at random. For 79% of participants, there was never a tie, and the remaining 21% had a tie in one of the sets (9 participants), or two ties within one of the sets (3 participants). This does not affect our ability to run the planned analyses, as we are testing consistency.

[8] This random money amount was drawn by the computer from a uniform distribution with minimum £0, maximum £31.50, with £0.25 intervals. The maximum amount of money, £31.50, was the highest median out of all the valuations in the first pilot study. Respondents were not told about the distribution, as this has been found to influence answers under a Becker-DeGroot-Marschak incentivisation procedure (Horowitz, 2006).

the second group, the payoffs were each participant's monetary valuations of the similar objects. The third group had the dissimilar objects as outcomes, and the fourth, their money equivalents. In addition to the objects and their monetary equivalents, scenarios P and R featured an additional payoff that was intended to be the least preferred one, receiving nothing.

Figure 5 provides an example of SG scenario Q for a participant whose relative ordering of the similar objects was radio-speaker alarm clock ≻ radio alarm clock ≻ alarm clock, making the radio-speaker alarm clock payoff A, the radio alarm clock payoff B, and the alarm clock payoff C. In this example, the sure prospect offers the radio alarm clock for certain, and the risky prospect offers an 80% chance of receiving the radio-speaker alarm clock, and a 20% chance of receiving the alarm clock. In the choices, probabilities were presented numerically, but also in terms of occurrence with coloured balls (each representing a 5% chance), as it has been suggested that presenting probabilities in terms of frequencies makes it easier for people to process them (e.g., Viscusi, Magat & Huber, 1991; Desaigues & Rabl, 1995). The side on which each option was presented, and the colours used to represent each probability were randomised.

Figure 5. Example of Risky Pairwise Choice with Similar Objects

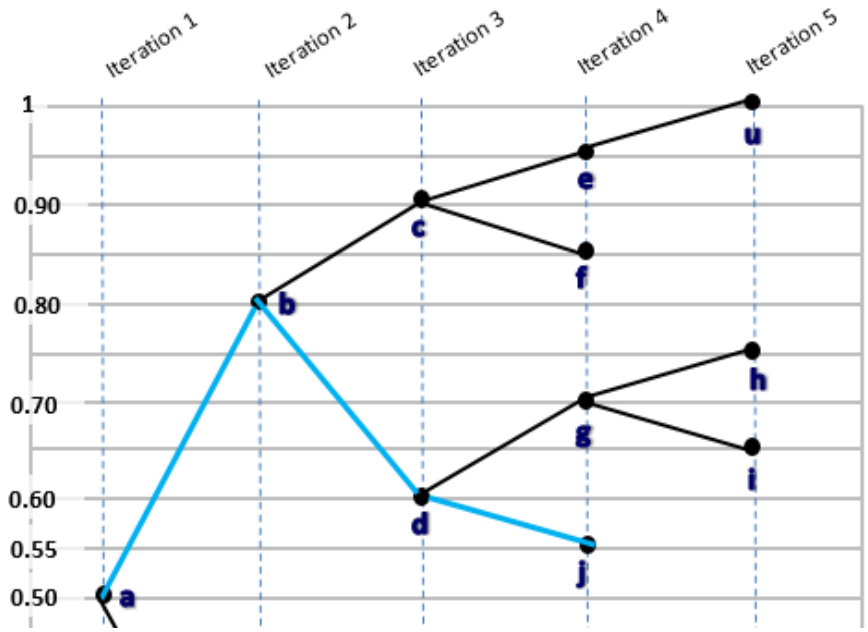


The indifference probability for each SG scenario is the probability (*p*, *q* or *r*) of receiving the best payoff that makes respondents indifferent between the two prospects. These probabilities were used in order to compute the intermediate item's utility elicited directly and indirectly.

Indifference probabilities were elicited using an iterative procedure. The advantage of using iteration trees (like the one partially shown in Figure 6) is that they minimise the number of

questions required to reach a reasonably fine-grained indifference probability[9]. Our iterations began from one out of three starting probabilities (0.35, 0.50, and 0.65) determined at random[10]. Questions went on until the indifference probability range was narrowed down to 5 percentage points, and then the middle point of the range is taken as the indifference probability.

Figure 6. Example of an Iteration to Indifference with Starting Probability = 0.50



An iteration starting from probability 0.50 could look like the example highlighted in Figure 6. The figure shows the upper half of a probability tree, which is a graphical representation of all the possible iteration paths from the starting probability up to indifference (see the full probability trees we used for the three starting probabilities in Appendix A.3). The letters are used to label *nodes*: each unique risky pairwise choice that participants can face on their iteration to their indifference probability in each of the gamble scenarios. The probabilities on the vertical axis of the figure correspond to the chances of receiving the best outcome in the choice in each node.

[9] This comes at the expense of errors having possibly large repercussions. Once the respondent goes into one of the branches, it is not possible for them to go back and reach the other section of the tree. For example, choosing the sure prospect when one would actually prefer the risky one in node a (iteration 1) would lead to not being able to reach indifference probabilities ranging from 0 to 0.50. We do not consider our results were greatly affected by this, if at all, because according to our attention check questions participants were paying attention and that should translate into few response errors (see Appendix A.5 for more details about attention checks).

[10] Details about the randomisation procedure for the risky pairwise choices can be found in Appendix A.4.

The iteration consists in determining the probabilities shown in one question depending on the respondent's choice in the previous one. If the respondent chose the sure prospect, as the respondent in the example would have done in node a, that means that the probability of receiving the best payoff was not high enough for them to be indifferent between the prospects. In the next question, the probability would be increased, going to node b. If, on the other hand, the respondent chose the risky prospect, as in node b, that means that the probability of receiving the high payoff is high enough; in this case, the next question would have a lower probability (node d).

Throughout the iteration, the probability in node a, 0.50, becomes the floor of the indifference range, and the probability in node j, 0.55, the ceiling. At that point, the distance between the floor and the ceiling is 5 percentage points, so the iteration ends. We take the middle point of that range, 0.575 in this case, as the indifference probability for that gamble scenario. In this example, the iteration to indifference took four questions, but it could take up to five questions, and that is why different participants could have a different total number of risky pairwise choices of interest, ranging from 48 to 60.

In order to have a fixed number of questions (100) so that the incentive scheme was the same for everyone, we included filler questions. They were interspersed among the choices of interest, and were also used to conceal the iteration[11], to add variety in the choices to keep respondents engaged, and to test other hypotheses elsewhere[12].

Respondents had incentives to reflect about their preferences and give truthful answers because at the end of the study one of these questions could be randomly selected to be played out for real. The question selected could be like the one in Figure 5. If the participant had chosen the sure prospect, then they would receive the payoff offered by it (in this example, a radio alarm clock). If they had chosen the risky prospect, the reward would be decided by taking a numbered ball out of an opaque bag. In this case, if the number on the ball was between 1 and 16 inclusive, the respondent would receive the radio-speaker alarm clock; whereas if the number was between 17 and 20, the payoff would be the alarm clock.

[11] A non-transparent iterative procedure was chosen over a transparent one because it has been shown that they elicit answers that are more consistent across elicitation techniques (particularly choice and matching). This has been found in the health domain (Pinto-Prades, Sánchez-Martínez, Abellán-Perpiñán & Martínez-Pérez, 2018) and others (e.g., hypothetical job offers as in the 1999 work by Fischer and colleagues), and money lotteries (Loomes & Pogrebna, 2016)).

[12] Some of these are reported in Chapter 3 of this thesis. Additional information about these questions can be found in Appendix A.4.

The risky pairwise choice task was the last one of the study. Once it was over, each participant's payment (in addition to a £3 show-up fee) could be determined. This would be either a money amount (which could go up to £50), one of the objects, or nothing, according to each respondent's answer in a randomly selected decision. We divided the tasks in the study in 9 blocks of approximately the same duration: 3 blocks of 25 riskless pairwise choices each, 1 block for the ranking task, 1 block for the valuation task, and 4 blocks of 25 risky pairwise choices each. The computer randomly chose 1 out of the 9 blocks, and respondents were paid out according to the incentive scheme of the selected task. Participants had been informed about the incentive structure before doing each task and completed test questions with feedback about it.

2.3.2. Participants

A total of 74 participants completed this study but some of them had to be excluded from the analysis. There were four exclusion criteria. First, a computer error during data collection affected 4 participants, who did not see all the questions or saw some of them from two to four times, so their answers were not usable. Second, 7 participants valued some objects at £0, and this made some of the questions impossible to analyse[13]. Third, 5 participants valued some objects above £50 and saw a pre-set money equivalent gamble rather than one with their actual money equivalents (see Appendix A.4); that means that their objects and money gamble scenarios were not comparable. The reason why we had the pre-set money gamble is that while respondents' valuations could go to infinity, our budget could not. One of the choices each respondent made during the study was incentivised, and so we had to cap the reward that they could receive at £50[14]. Last, some of the filler gambles were designed to spot participants who were not paying attention, but we did not have to exclude any participants because of this (see more details about the attention checks in Appendix A.5).

After excluding these participants, our sample of 74 was reduced to 58. Participants were university students, 21.9 years old on average, and 39 (67%) of them were female. They were invited to the study through the University of Warwick SONA online recruitment system, and came to the laboratory for an experimental session that lasted for an hour.

[13] For instance, when there is a single object within the set valued at £0, then that object must be the worst one (C), given that we infer their ranking (A≻B≻C) from their valuation. If the worst object's value equals £0, then gamble scenarios P and Q are the same, and gamble scenario R is a choice between nothing vs. a chance of nothing and a chance of the most preferred object.

[14] Pilot studies study indicated that most respondents' valuations were below £50.

2.4. Results

2.4.1. Valuation and Gambles Summary

The following tables show the summary statistics of the money equivalents of the objects (Table 1), and the cases in which each of the objects in the similar and dissimilar objects sets were ranked as the best, middle, and worst payoffs (Table 2).

The first table offers reassurance about the success of the task in eliciting accurate valuations. While the most expensive objects are clustered in the top half of the table, and the cheapest ones are in the bottom, objects are not ordered by market price within the two halves (see what the objects prices were in Amazon when the object pool was piloted in Table 18 in Appendix A.1). In addition to market price considerations, valuations were sensitive (in the direction that one would expect) to the self-reported measures about the objects we collected in the familiarisation task. Summary statistics of these (participants' desire to receive the objects, and whether they already owned such objects) can be found in Table 24 and Table 25 in Appendix A.6. First, valuations were positively correlated with the desire to receive the objects (ρ = 0.5418, p-value < .01). Second, with regards to owning them already, valuations were significantly lower for those objects that participants owned than for those they did not own ($mean_{own}$ = £9.78, SD_{own} = £8.29, $mean_{not\_own}$ = £12.56, SD_{not_own} = £11.37, Wilcoxon rank-sum two-sided test p-value < .01).

The second table lends credibility to our complexity manipulation: in the case of the similar objects, participants could generally distinguish the increasing features and placed additional value on them; with dissimilar objects, on the other hand, there was greater heterogeneity in people's valuations, which might indicate that establishing an order of preference was not as straightforward as it was with the similar objects.

Table 1. Object Valuation (n = 58)

Good	VALUATION			Object Set
	Median	Mean	SD	
Suitcase	£16.13	£18.57	£16.83	
Radio-speaker (RS) Alarm Clock	£16.00	£19.59	£17.88	Similar
Headphones	£15.00	£18.44	£16.67	
Air Mattress	£10.00	£15.20	£20.00	Dissimilar
Radio Alarm Clock	£9.75	£13.22	£12.23	Similar
Toaster	£7.50	£10.71	£10.94	Dissimilar
Mouse	£6.63	£8.24	£8.63	
Alarm Clock	£5.50	£6.80	£7.85	Similar
Bottle	£5.00	£6.06	£7.47	
Clicker	£4.63	£6.63	£8.66	Dissimilar

Table 2. Object Ranking (n = 58)

SIMILAR OBJECTS			
Best	**Middle**	**Worst**	**Cases**
RS Alarm Clock	Radio Alarm Clock	Alarm Clock	49
RS Alarm Clock	Alarm Clock	Radio Alarm Clock	0
Radio Alarm Clock	RS Alarm Clock	Alarm Clock	9
Radio Alarm Clock	Alarm Clock	RS Alarm Clock	0
Alarm Clock	Radio Alarm Clock	RS Alarm Clock	0
Alarm Clock	RS Alarm Clock	Radio Alarm Clock	0
DISSIMILAR OBJECTS			
Best	**Middle**	**Worst**	**Cases**
Air Mattress	Toaster	Clicker	20
Air Mattress	Clicker	Toaster	11
Toaster	Air Mattress	Clicker	18
Toaster	Clicker	Air Mattress	1
Clicker	Toaster	Air Mattress	3
Clicker	Air Mattress	Toaster	5

2.4.2. Consistency

According to EUT, the utility of a prospect is given by the expectation of the utility of its payoffs. That is, subjective value is the only dimension of the payoff that should matter when making decisions. It also follows from the RCLA, and hence EUT, that the utility estimates obtained directly (equation (4)) and indirectly (equation (7)) should be the same. Thus, we should find consistency between the utility estimates elicited with different consequences (e.g., money vs. objects) for as long as they are equivalent in subjective value; and consistency between the estimates that are elicited directly or indirectly.

In this study, in order to consider two estimates as consistent we did not require them to be identical (i.e., we did not require both sides of equation (8) to be exactly the same). The utilities were estimated by asking respondents to choose between prospects. These prospects offered payoffs associated with probabilities that ranged from 0 to 1, and varied in steps of 0.05. We accommodate a slight lack of sensitivity to probability changes and trembling hand errors[15] by allowing for a 0.05 discrepancy in the estimates. This means that, for example, we would consider an estimate implied from an indifference probability between 0.35 and 0.40, and another estimate implied from a probability between 0.40 and 0.45 as consistent with each other.

Before our main analysis, we ran an OLS regression (clustering errors by respondent) to check that the elements of the gambles design we randomized did not have an effect on the elicited indifference probabilities. These factors were the probability from which the iteration started, the colours of the balls that were used to represent probabilities, and whether the risky alternative was shown on the right- or the left-hand side. None of these had a significant effect on indifference probabilities, so the data did not need to be analysed separately on the basis of these factors in our subsequent analysis. The regression output can be found in Appendix A.6.

We opted for a within-subjects analysis. Previous valuation studies were aiming to estimate the societal value of health or physical risk reductions, and so they looked at the means and medians of the utility estimates. In this study, however, we are testing the validity of the RCLA in different domains, and aggregating the data could hide individual choice patterns that we care about.

We start the analysis looking at the effect of the choice complexity manipulation by comparing the estimates elicited with objects and money. Next, we test the validity of the RCLA by comparing the estimates elicited directly and indirectly for money, similar and dissimilar objects.

Consistency across types of goods: objects vs. money

According to EUT, the comparability of the outcomes should not make a difference in choice; the only dimension of the outcomes that enters people decisions is their utility. If that were the case, we should not find any difference between the object and the money estimates.

We classified the estimates from each of the three gamble scenarios into three groups: one in which the estimates elicited with money were smaller than the ones elicited with the equivalent

[15] These refer to mistakes when answering the questions, which in this case could place participants in a higher/lower branch of the iteration tree than they intended.

object, one where they were consistent, and one for when the money estimate was higher than the object one. Table 3 shows the percentage and the frequency of the cases (out of 58) that fell in each of the groups.

Table 3. Objects vs. Money Equivalents Estimates (n = 58)

		CASES			Chi-squared test p-value
Estimate	**Type of Good**	**u(£) < u(Obj)**	**u(£) ≈ u(Obj)**	**u(£) > u(Obj)**	
$u(B)_P$	**Similar**	14% (8)	36% (21)	50% (29)	.7522
	Dissimilar	19% (11)	34% (20)	47% (27)	
$u(B)_Q$	**Similar**	19% (11)	24% (14)	57% (33)	.0304
	Dissimilar	38% (22)	28% (16)	34% (20)	
$u(C)_R$	**Similar**	24% (14)	22% (13)	53% (31)	.9725
	Dissimilar	24% (14)	21% (12)	55% (32)	

The difference between the money and objects estimates followed a pattern: the largest group for all gambles was almost always the one where the utilities elicited with money amounts were higher than their object counterparts. The exception to this was the $u(B)_Q$ estimate for the dissimilar objects, which we assume to be the hardest to make, as dissimilar objects are the most complex good in our study, and gamble scenario Q involved a choice between three objects (rather than two objects and 'nothing', like in the other two scenarios). This means that, generally, respondents were more averse to risk when the payoffs were money than when they were either of the two object types; that is, they required a higher probability of success (i.e., chance of receiving the best outcome) in order to be indifferent between the sure and the risky prospects. Wilcoxon signed-rank tests confirmed that the difference between the money and the objects utilities was statistically significant (i.e., rejected the null hypothesis of equal distributions between the objects and the money estimates) in all cases but the exception mentioned above (p-values < .05, see all the p-values, together with histograms of these differences in Appendix A.6).

With the exception of $u(B)_Q$, we found no difference in the patterns of objects and money estimates depending on whether the objects were similar or dissimilar to each other. Chi-squared tests (see right column of Table 3) confirmed this at a 5% significance level.

Our complexity manipulation provides evidence against the descriptive validity of EUT: our participants' decisions were not exclusively influenced by the value of the gamble outcomes, but their nature mattered too. Even when respondents had stated that the given objects and money amounts were equivalent in utility terms, they were more risk averse when decisions

involved the money amounts than when they involved the objects. That is, apart from the relative utilities of its outcomes, the SG captured respondents' risk attitude towards them. There was a difference between money and objects, but within the objects we did not find a systematic effect of complexity on the estimates; whether the objects shared a fraction of their dimensions (similar objects) or not (dissimilar objects) generally did not make a difference.

Consistency across types of estimates: direct vs. indirect

We then turned to the consistency between the direct and indirect utility estimates. We pooled the money estimates together, as they were not different to each other in terms of complexity. The relation between respondents' direct and indirect estimates falls in three groups, like the ones used in our previous analysis. Table 4 summarises the percentage and the number of cases in each category. We found consistency in about a third of the cases with each type of good (the percentage ranged from 31% to 38%). Consistency was roughly the same with money (the least complex good) and with dissimilar objects (the most complex one).

Table 4. Comparison of u(B): Direct ($u(B)_P$) vs. Indirect ($u(B)_{QR}$) Estimates

	n	CASES		
		$u(B)_P < u(B)_{QR}$	$u(B)_P \approx u(B)_{QR}$	$u(B)_P > u(B)_{QR}$
Money	116	29% (34)	37% (43)	34% (39)
Similar Objects	58	21% (12)	31% (18)	48% (28)
Dissimilar Objects	58	40% (23)	38% (22)	22% (13)

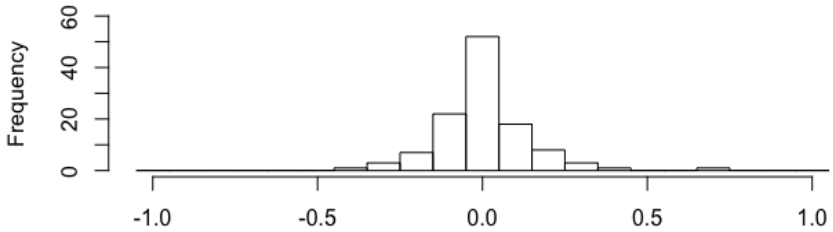
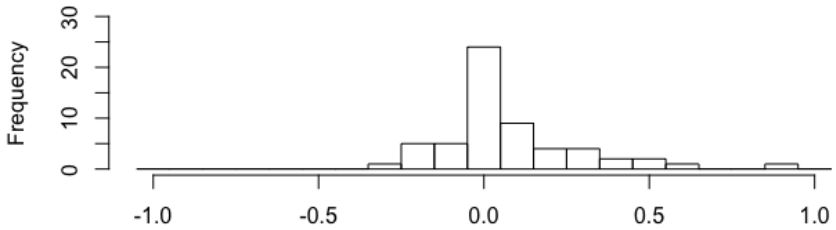
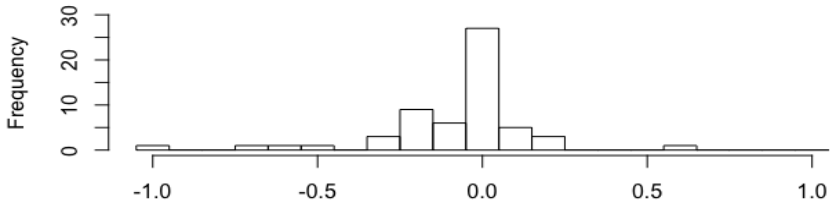
The gamble scenarios used to elicit the utility estimates did not only differ in whether they were chained or not. As we hypothesised, the complexity of the gamble outcomes may have had a systematic effect on consistency. Other factors such as the value of the outcomes, their relative ranking (i.e., which object was ranked as best, middle, and worst), and the probability at which the iteration to indifference started were not kept constant across gamble scenarios either. Hence, the disparities we found between the direct and indirect estimates may have been driven by the failure of the RCLA, but also by any of the previously mentioned factors. In the following sub-sections, we explored whether there was any systematic relation between these factors and the (in)consistency cases found.

Effect of the type of good

We had hypothesised that the smaller the fraction of dimensions the outcomes shared, the less consistency we would find. This implied that we would find the highest consistency with money, followed by similar objects, and that the least consistency would take place with dissimilar objects.

The percentage of consistency cases was about the same with money, similar and dissimilar objects (31-38%), but there were differences in the inconsistent cases across the goods. With our least complex good, money, the inconsistent cases were approximately evenly split between the two inconsistency groups. Wilcoxon signed-rank tests (see p-values in Table 5) confirmed that the distributions of the direct and indirect money estimates were not significantly different with this good. Just like Chilton and Spencer did when they asked their respondents to estimate the utility of money, we found a few more cases where the direct estimate was higher than the indirect estimate, but this asymmetry was much less pronounced than the ones we found with objects. With objects, we found a statistically significant asymmetry in the distribution of the cases in the two categories other than consistency (see histograms in Table 5). In the case of the dissimilar objects, the largest group was the one where the indirect estimates were higher than the direct estimates. That is, the indirect estimates portrayed the worst object (or its money equivalent) as being worse than indicated by the direct estimate. With the similar objects, the opposite was true.

Table 5. Differences in u(B) graphically: Direct - Indirect Estimates

	$u(B)_{direct} - u(B)_{indirect}$	**Wilcoxon signed-rank p-value**
Money (n = 116)		.5631
Similar Objects (n = 58)		.0303
Dissimilar Objects (n = 58)		.0125

The disparities between the direct and indirect estimates elicited with money were symmetrically distributed around zero. This pattern could correspond to that of consistent

estimates surrounded by random noise. However, in the case of objects, the distribution of the disparities is skewed: to the right in the case of similar objects, and to the left in the case of dissimilar objects.

Effect of other factors

It may be the case that the patterns we observed were not only driven by the complexity of the goods, but also by the differences in their value. We looked at value differences for the consistent and inconsistent estimates (see Appendix A.6 for full analysis). We found that there were no significant differences in the value range (the difference between the monetary equivalents of the best and the worst objects), or the money equivalent of the best possible payoff for the consistent and the inconsistent cases. This indicates that the differences in the subjective value of the payoffs could not have had an effect on the (in)consistency patterns found.

Last, we looked at the two other factors that varied across the gamble scenarios: the starting probabilities (0.35, 0.50, and 0.65), which had been randomly assigned to each set, and the relative ranking of the objects, which would determine in which of the prospects each payoff would appear. We had found the starting probabilities were not significantly related to the indifference probabilities, but they might still have had some effect on consistency. We ran Chi-squared tests to check whether there was a statistically significant relation between consistency and the starting probabilities of the three gamble scenarios in each group. We could not reject the null hypothesis of no relation (all p-values >0.10). We also used Chi-squared tests to see whether consistency was associated to the different possible combinations of objects (the ones shown in Table 2), and the relationship was not statistically significant for the similar objects (p-value = .503) or the dissimilar objects (p-value = .122).

2.5. Discussion

We set out to examine whether disparities between the direct and indirect SG utility estimates are caused by the elicitation context or the lack of validity of the RCLA. We modified the context to try to minimise its effect on consistency and observe whether the RCLA held. We shifted away from the affective and abstract nature of health, and the effect of the lack of incentives on consistency; and we manipulated the complexity of the choices.

If the RCLA held, and our hypothesis were true, then having removed the effect of affect, and providing incentives in the lowest complexity treatment (i.e., when valuing monetary amounts), should have led to consistency. We performed a case-by-case analysis and found consistency

rates that ranged from 31% to 38% across the three types of goods. With the simplest possible context we could set up (i.e., money amounts valuation), more than 6 in 10 cases were not consistent. Having removed the effect of some contextual factors, we found no effect of the factors that varied across gamble scenarios, which suggest that the reason why the disparities occur is the failure of RCLA, rather than context.

Had we not found systematic disparities between the direct and indirect SG estimates, our findings would have provided grounds to tentatively attribute the disparities previously found in health to contextual factors. In this case, further work would be needed to identify the role that each of the factors that we removed or manipulated may play in reducing the disparities. This could be done with the current study design by bringing back the elements we removed one by one. First, we would run the same study without incentives, and then, without incentives and with health rather than consumer goods. This output would then help to improve the design of health valuation studies in order to minimise the disparities. Our findings point in the opposite direction, to the lack of descriptive validity of the RCLA. We leave it for others to judge whether using an instrument with an overall rate of inconsistencies of 64% across goods is appropriate for policy purposes.

Turning to our complexity manipulation, its goal was two-fold. First, we used it to test the EUT prediction that the only dimension of outcomes that enters people's decisions is subjective value (or utility). We asked questions involving prospects that were identical in terms of probabilities and the subjective value of the payoffs, but varied in the comparability of the latter: in the more complex case, the outcomes were objects, and in the less complex case, they were money amounts. If the EUT premise held, then we should expect to elicit the same estimates. We looked at the difference in estimates for each of the three gamble scenarios individually. We found that in most cases the estimates elicited with the monetary equivalents of the objects were higher than those elicited with the objects. That is, participants were more risk averse with money than with objects, and this was reflected in the utility estimates. This might be because of the complexity of the choices, or because of their nature (e.g., money is fungible, but objects are not); further research is needed to address this question.

Second, we explored the effect of choice complexity on consistency between the direct and indirect SG utilities. We had hypothesised that the smaller the fraction of shared dimensions among the outcomes, the less consistency there would be. We found that this was not the case: the rates of consistency in the money and objects gambles were very close, and so were those from the two object groups, the similar and dissimilar objects. What did differ for different types

of outcomes was the pattern of the inconsistent cases: while the inconsistent cases with money were approximately evenly split between the two inconsistency groups (one where the direct estimates were higher than the indirect estimates, and the other way around), we found different patterns for the two types of objects. We replicated the asymmetry that is usually found in studies with comparable designs to ours when valuing health (the indirect estimate is higher than the direct) with the dissimilar objects and found an asymmetry in the opposite direction for similar objects. This further supports our complexity manipulation, as the patterns we found with the most complex choices in our study (i.e., those with dissimilar objects) were the same as those in choices involving health.

In sum, our findings contribute to the existing literature on the descriptive validity of EUT and the RCLA, the SG method, and decision making across domains. To our best knowledge, this is the first study where the descriptive validity of the RCLA is refuted with a (probability) equivalence task, in a non-monetary domain, under incentive compatibility. We also uncovered an asymmetry in the direction of the disparities between the utility estimates elicited directly and indirectly depending on whether the objects shared some of their attributes (similar objects) or they did not (dissimilar objects). Further exploring the reasons behind this pattern may shed light on how the fraction of shared dimensions among the outcomes affects people's decisions, and whether it has a systematic and robust effect.

In addition to this, our results show that using monetary outcomes as a proxy for other types of outcomes could lead to the wrong inferences, as outcome dimensions other than subjective value, and their comparability thanks to having (or not) dimensions in common, also appear to influence people's decisions. This is a particularly relevant finding for experimental economists who, on the basis of our findings, should not automatically assume that the predictions of models developed and tested with one good will also hold with other types of goods. We found this was not the case for risk attitudes, when respondents considered money and consumer goods.

Chapter 3.
Exploring the Role of Choice Complexity through the Common Ratio Effect in Monetary and Non-Monetary Domains

Abstract

The common ratio effect (Allais, 1953; Kahneman & Tversky, 1979) is a violation of expected utility theory (EUT) widely replicated with monetary consequences. However, many of our day-to-day decisions are more complex than the trade-offs between two unidimensional scales (money and probability) participants make in the lab, i.e., they may involve outcomes whose dimensions or attributes are not shared. Neither EUT nor any of its most prominent alternatives predict choice complexity to have an effect on decisions, but it is crucial to find out more about it since a lot of decision making models are developed using evidence from choices between monetary payoffs, but generalised to decision making with all outcomes. We aim to find out more about the role of choice complexity on decision making. Our design allows us to compare (incentivised) choice patterns in questions that differ in complexity: the outcomes were either money or consumer goods, which had identical probabilities and equivalent subjective value. We replicated the common ratio effect with money. With objects, however, we did not find it: the deviations from consistency went in the CRE direction, but the asymmetry between them was not statistically significant. The greater consistency found with objects may be caused, rather than by the descriptive validity of EUT, by participants resorting to heuristics (such as 'always pick the prospect that offers the best payoff') when decisions are too complex. These results suggest that preferences and attitudes may be domain-specific, and so that using money to proxy decisions with other outcomes may lead to wrong inferences.

Keywords: common ratio effect, choice complexity, consumption goods

3.1. Introduction

Expected utility theory (EUT) is an influential theory of decision making under risk that postulates that the expected utility of a prospect is given by the utility of its payoffs, weighted by their respective probabilities. Its axiomatic base was developed by von Neumann and Morgenstern (1944), and while these axioms are normatively appealing (Machina, 1982), there is a large body of research that questions their descriptive validity. This evidence comes from fields such as psychology and behavioural economics, which, in the light of evidence contrary to EUT, have provided alternative descriptive models of risky choice (see Starmer, 2000; Sudgen, 2004; Fox, Erner & Walters, 2015).

One of the most famous violations of EUT is the common ratio effect (CRE), which was put forward by Allais (1953) and popularised by Kahneman and Tversky (1979) as evidence that actual behaviour deviates systematically from the implications of EUT. The best-known form of the phenomenon is illustrated in Figure 7. The CRE involves two questions, *scaled-up* and *scaled-down*, with two prospects each: S and R, and S' and R' respectively. In these prospects, there are two possible consequences: c_l, which has lower utility, and c_h, with higher utility, such that $c_h \succ c_l \succ 0$. The consequences are the same in both questions, and so is the ratio of the probabilities of receiving these consequences. This common probability ratio is achieved by scaling down the probabilities of the better outcomes in the scaled-up question by the same factor (in this example, by 4).

Figure 7. Common Ratio Effect

Scaled-up question:	S = (c_l, 1)*	R = (c_h, 0.80; 0, 0.20)
Scaled-down question:	S' = (c_l, 0.25; 0, 0.75)	R' = (c_h, 0.20; 0, 0.80)*

Note. The stars indicate which prospect is chosen more often.

According to EUT, in the absence of error (or noise), the choices in the scaled-up question should be the same as those in the scaled-down question. That is, individuals who choose prospect S (R) in the scaled-up question should choose S' (R') in the scaled-down question. However, it is often found that in the scaled-up question, the safe prospect S is preferred to the risky one R; but in the scaled-down question, when the probabilities of the nonzero outcomes are scaled down by the same factor, the riskier prospect R' is preferred to the safer one S'. This choice

pattern is robust and, if made by the same individual, it cannot be explained by a deterministic set of EUT preferences[16].

The CRE is probably the most frequently replicated violation of EUT (Loomes & Sudgen, 1998). In most cases, replications are done following the experimental economics tradition of using money as outcomes. In the most popular variation of the problem, which we refer to as the *classic* CRE, the consequences are money amounts with a ¾ value ratio (e.g., 3,000 and 4,000 Israeli shekels, or 30 and 40 US dollars or British pounds). Von Neumann and Morgenstern argued for using monetary consequences as a way of simplifying decisions when testing models, assuming that money is "divisible, and substitutable, freely transferable and identical […] with whatever 'satisfaction' or 'utility' is desired by each participant" (1944, p. 8). However, this may not suffice when pursuing a broader descriptive account of behaviour. People's everyday decisions involve complex, multi-faceted consequences (rather than simpler, unidimensional ones as money). Therefore, models built on evidence solely from choices involving monetary amounts may fail to capture the nuances of actual decision making.

EUT predictions for money and for other consequences are not different. When laying down the concept of numerical utility, Von Neumann and Morgenstern (1944) prompted the reader to imagine prospects where the consequences were "any two objects or […] events" (p. 17), implying that their EUT axioms should hold regardless of whether the outcomes were money amounts, objects, or events. After them, Savage (1954) used omelettes as examples of consequences, and Kahneman and Tversky (1979, p. 264) stated that "the domain of the [EUT] utility function is not limited to any particular class of consequences". However, EUT assumes that decisions are only influenced by the utility or subjective value of the consequences and the probabilities associated with them. Hence, it does not allow for the complexity that different consequences bring into choices to have any effect.

In this study, we aim to explore the role that choice complexity may play in decision making. *Choice complexity* is defined, as in the previous chapter of this thesis (see page 16), on the basis of how easy it is to compare the consequences' dimensions. The higher the fraction of

[16] The choice of prospect S over R implies that u(cl) / u(ch) > 0.80, whereas the choice of prospect R' over S' implies the opposite, that u(cl) / u(ch) < 0.80. Given that the consequences and the ratio of probabilities are the same across questions, these choices constitute a violation of the independence axiom. In its generic form, this axiom implies that for prospects S, R, X, if prospect S is preferred to prospect R, then any probability mixing of prospect S with another prospect X should also be preferred to the same probability mixing with prospect R. In the case of the CRE, the prospects in the scaled-down question originate from mixing those in the scaled-up question with prospect X, which is a 0.75 probability of receiving nothing. This results in prospects S' = (0.75, 0; 0.25, S) and R' = (0.75, 0; 0.25, R).

comparable dimensions, the less complex, in relative terms, the choice will be. Following this definition, choices between money amounts are the simpler ones, while choices that involve two different objects (e.g., a coffee machine and a pair of jeans) are more complex, and hence more difficult to make.

Our design allows us to match the utility of each consumer good with that of a money amount, for every participant. To our best knowledge, this is the first fully incentive compatible study to do so. Using the CRE paradigm with monetary consequences as a benchmark, we manipulate choice complexity by asking participants to make choices between prospects with consumer goods as payoffs. In order to isolate this effect, we keep the probabilities of receiving each consequence constant, and compare the choice patterns with objects and money amounts that participants deemed equivalent in utility terms.

We replicated the CRE when the consequences were monetary, but not when they were objects. This evidence contradicts the implicit assumption of EUT formulation that choice complexity does not influence decisions. The absence of the CRE with objects suggests that, with these goods, participants did behave as expected utility maximisers. Alternatively, this pattern could arise if respondents resorted to heuristics (such as "choose the prospect with the better outcome") when it was difficult to compare the consequences. The use of these heuristics when choices were complex may have led to greater consistency with objects compared to money, and yet not be evidence in favour of EUT.

In section 2, we discuss the relevant literature. In section 3, we present the design of our study, and in section 4, our results. We conclude with a discussion of our findings in section 5.

3.2. Theoretical Background

The CRE fostered the development of a large number of alternative models that account for the observed preference switch. The most prominent of these, and currently the dominant behavioural models (Fox, Erner & Walters, 2015), are prospect theory (Kahneman & Tversky, 1979) and its later formulation, cumulative prospect theory (CPT) (Tversky & Kahneman, 1992). In CPT, the utility function is replaced by a value function that encompasses all types of consequences, and cumulative probabilities non-linearly weighed replace the objective ones. This theory predicts systematic violations of the independence axiom and hence can explain the CRE.

Neither EUT nor CPT make different predictions depending on the complexity of the choice. The only aspects that these theories predict to influence decisions are the subjective value (or utility) of the consequences and their (possibly weighted) probabilities. Irrespective of the complexity of the decision (i.e., the comparability of the consequences), EUT cannot account for the CRE. On the other hand, CPT does account for the CRE. In either case, as long as the consequences are equivalent in terms of utility, and given that the probabilities are the same and are weighted in the same order, different choice complexity should not result in different choice patterns.

Although CRE replications with monetary outcomes are abundant, to the best of our knowledge there are only two studies that have tested this effect with more complex consequences than money. In both cases, the CRE was replicated. First, Kahneman and Tversky (1979) report hypothetical choices between a 50% chance of winning a 3-week tour of England, France and Italy, versus a 1-week tour of England for sure, and the scaled-down version of the question where probabilities of nonzero outcomes were reduced by a factor of 10. The switch from the safe to the risky prospects was also replicated then, with 78% of participants choosing the safe option in the scaled-up question, and 67% of participants choosing the risky option in the scaled-down question.

Second, Robin Keller (1985b) tested the CRE using three different ways of presenting the probabilities: written statements, matrices where cell size was proportional to probabilities, and graphically with pictures of marble tubes. She found the effect for all three presentation modes. She used a Porsche car as the better payoff of the pair, and a convertible Volkswagen as the worse one; both of them were hypothetical rewards. The probabilities she used were not the classic ones. She used 4 different probability pairs and combined them into three sets of scaled-up and -down questions. She used 100% and 75% for the scaled-up question in the first two sets. For the scaled-down questions, she reduced these by a factor of 25 (to 4% and 3%) in the first set, and a factor of 1.25 (to 80% and 60%) in the second set. In her third set, she used the 80% and 60% chances in the scaled-up question, and scaled them down by a factor of 20 (to 4 and 3%).

These two studies tested the CRE with non-monetary consequences. Despite the choices they asked participants to make were more complex than those that involve monetary consequences only, these decisions were far from many of those that people typically make in the following two respects. First, they were not incentivised, but decisions people make in their daily lives have real consequences. Second, they involved consequences that were relatively straightforward to compare to each other: it is reasonable to assume that people would

consider the longer holiday with more destinations better than the shorter, 1-destination trip; and that a Porsche would generally be preferred to a Volkswagen. But many of the decisions that people make involve difficult trade-offs.

With this study, we expand the limited existing empirical work by tackling these two issues. First, we move from hypothetical to incentive compatible questions. Second, we ask questions that involve trade-offs between familiar consumer goods. We are able to isolate the effect of choice complexity by comparing the choices participants make in the money and the objects questions, while keeping everything else (i.e., the probabilities and the subjective value of the consequences) constant. To our best knowledge, this is the first fully incentive compatible study to do so. Within the objects, we further manipulate complexity. The so-called *similar objects* in our study have some comparable features, and hence we presume that people will regard the objects with additional features as better. The *dissimilar* objects, on the other hand, do not have comparable features and hence establishing a preference ordering should be more difficult; this is closer to what may happen with choices outside of the laboratory.

We aim to find out whether the CRE persists with consumer goods. This will allow us to test the descriptive validity of EUT using non-monetary consequences. Bust most importantly, it will let us explore the differences in choice patterns that arise as we move from decisions involving monetary consequences to more complex non-monetary ones. For the same purpose, we also compare choice patterns between similar and dissimilar objects.

EUT and CPT imply that the mechanisms behind simpler and more complex decisions are the same, and so that we should expect that whatever choice pattern is found for monetary outcomes should also be found in more complex choices, as long as the subjective value of the consequences and their (possibly weighted) probabilities are the same. Intuitively, deciding between outcomes that cannot be compared using the same measuring stick seems more difficult, so we put the hypothesis of no difference to test.

3.3. Method

3.3.1. Design and Procedure[17]

In order to explore the role of outcome complexity, we elicit respondents' monetary equivalents of the objects. Then, we compare their choices in CRE-type questions when the consequences are objects, and when they are the money amount each respondent indicated is equivalent to the objects in terms of subjective value.

The study was conducted at the University of Warwick. First, we familiarised participants with the objects in the study. Second, we elicited the monetary equivalents (or *valuations*) for each object by means of three tasks: riskless pairwise choices, a ranking, and a valuation grid. We used each participant's valuations to decide which objects to include in their CRE questions, and to substitute those objects when the consequences were money. Last, we asked the CRE questions.

Familiarisation

There were 10 objects in this study (see Figure 8). They were on display at the laboratory and participants were encouraged to have a look at them before the session started. Within these objects, there were 3 that shared some attributes ('similar objects'): three alarm clocks, one of which also had a radio, and a third one that had a radio and Bluetooth speakers. Within the similar goods, there were dominance relationships that should be straightforward to identify for most participants. For example, we would expect them to regard the radio alarm clock as more valuable than the alarm clock, given that it had the same alarm feature but also a radio. The remaining objects did not share any feature with each other or the alarm clocks. Hence, we expected it to be harder to establish an order of preference where the objects clearly dominated one another.

[17] The data used in this chapter come from the same experiment used to collect the data for the previous chapter. A detailed explanation of the design and procedure for this study was included here to fulfil the requirement that each of the chapters should resemble a stand-alone paper, but readers who have covered Chapter 1 may find most of the information redundant.

Figure 8. Consumer Goods in this Study

The study started with the object familiarization task, where participants saw a screen per object with a picture of it and a brief description. The information about the characteristics of the objects was available for the participants throughout the study: they had printed cards with it, and they could also see it on the screen by clicking on the image of the object at any time. To ensure that participants spent some time considering each object, we asked them whether they already owned such an object, and how much they would like to receive it (on a 7-point scale going from “not at all” to “very much”).

Elicitation of the Monetary Equivalents

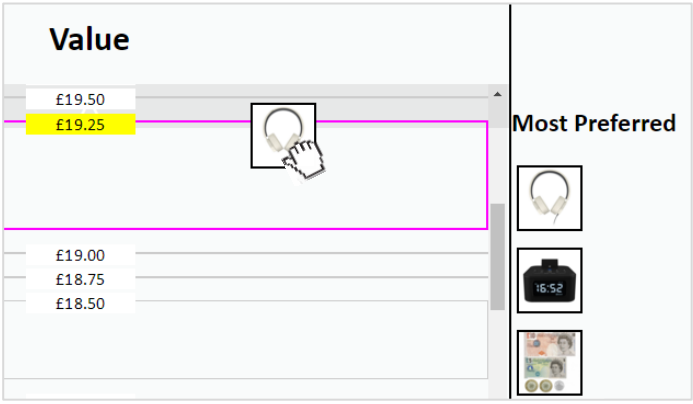
The monetary equivalents were elicited using three tasks, each of which gave participants further chances to reflect about their preferences. This was crucial for our study, as we needed participants’ valuations to be as accurate as possible. First, in the riskless pairwise choice task, participants were asked to make 75 choices between all the possible pairwise combinations of the 10 objects with each other and with each of three money amounts (£5.25, £8.75, £19.50). These money amounts were included to prompt respondents to start thinking about trade-offs between money and the objects, as they would have to do on a later task. Participants had incentives to make each of these choices carefully because at the end of the study, one of them could be randomly selected to be paid out, in which case participants would receive the item they said they preferred.

Based on their answers to the pairwise choice questions, a provisional ranking of preference was inferred, and participants were invited to adjust it to better represent their preferences if necessary. Once again, they had incentives to do this accurately, because if this task was randomly chosen to be the basis of their payment, two items in the ranking (out of the 10 objects and the 3 money amounts) would be randomly selected and they would receive the one that they ranked higher.

Once the ranking was submitted, they moved on to the valuation grid task. Their ranking appeared on the screen again, and this time participants had to drag and drop each item to the cell in a ladder with the money amount they considered equivalent to the item. The money

amounts in the ladder started at £0 and increased in £0.25 intervals, but the ladder had no endpoint. For example, if according to one participant the headphones had the same value as £19.25, that participant would drag the headphones picture to the cell with the "£19.25" label (as you can see in Figure 9).

Figure 9. Detail of the Valuation Interface

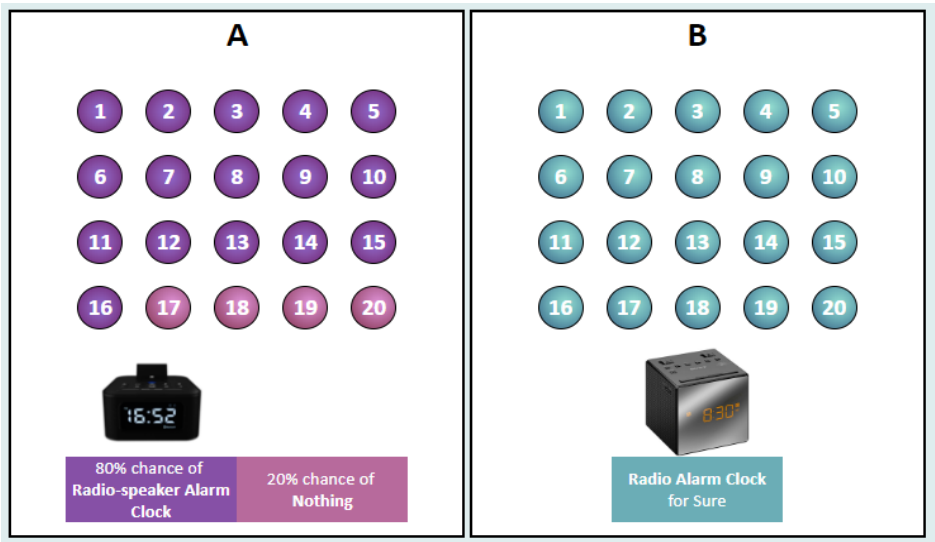


These monetary equivalents were used to decide which 4 out of the 10 objects would appear on the CRE questions.

CRE Questions

Last, participants answered 10 CRE questions (see an example in Figure 10). They were embedded in a larger set of questions that were part of a study with different goals[18], and appeared in random order.

Figure 10. Interface Sample of Scaled-up Question



[18] Find more details about this study in Chapter 2 of this thesis.

We use 'set' to jointly refer to the scaled-up and scaled-down questions that feature a given pair of consequences. There were 5 different payoff pairs (2 featuring objects, and 3 with money), and each pair was shown in two questions that differed in their probabilities. Table 6 summarises the pairs of consequences in the five sets. The payoffs of the classic CRE set were the usual £30 and £40, whose value ratio is ¾. The second set featured similar objects (two alarm clocks). The payoffs in the third set were the money amounts that each participant stated were equivalent in value to the similar objects in the second set. The fourth set featured two dissimilar objects (any two objects from the pool as long as they were not both alarm clocks), and the fifth set, their money equivalents. The pairs of objects were selected so that they were all as close as possible to the ¾ value ratio in order to make the sets as comparable as possible across consequence domains. Therefore, participants were liable to see different objects than one another, and also different monetary payoffs (other than for the classic CRE, where everybody saw £30 and £40).

Table 6. Consequence Pairs

SET	PAYOFFS
1. Classic Common Ratio Effect	£30 and £40
2. Similar Objects	2 out of: - Alarm Clock - Radio Alarm Clock - Radio-speaker Alarm Clock
3. Money Equivalent (ME) of Similar Objects	Monetary Equivalents of the similar objects
4. Dissimilar Objects	2 out of: - Alarm Clock - Toaster - Radio Alarm Clock - Air mattress - Radio-speaker Alarm Clock - Mouse - Headphones - Clicker - Suitcase - Bottle but not two of the alarm clocks
5. Money Equivalent of Dissimilar Objects	Monetary Equivalents of the dissimilar objects

3.3.2. Participants

Respondents in this study were students recruited through the University of Warwick SONA participant recruitment system. In exchange for their participation, they received a show-up fee of £3 and an additional payoff. This payoff was either a money amount (which could go up to £50), one of the objects, or nothing, according to each respondent's answer in a randomly selected decision. 67 people answered all the CRE questions, but 7 of them were not included

in the analysis because their similar or dissimilar objects ratio closest to the target one (¾) was 1. Such ratio implied that they had given both objects in the pair the same value, and hence the questions they faced were dominance questions (i.e., choices between prospects offering consequences with the exact same subjective value, one of which had a larger probability associated to it). The remaining 60 participants were 21.75 years old on average (SD = 2.61), and 40 (67%) of them were females.

3.4. Results

3.4.1. Question Payoffs

Table 7 shows the object pairs that appeared in the CRE questions. The objects in the columns correspond to the consequence with the higher value (c_h), and those in the rows correspond to the consequence with the lower value (c_l). Each cell contains the count of the times that the pair of objects was shown. The light blue cells correspond to the similar objects.

Table 7. Pairs of Objects Displayed

c_l \ c_h										
		1	1	1		1	1	1	1	3
	1		2	20	10	2		1		
	3						1	2	2	
		1			23	5		1	1	
			1	6		3		1	1	1
	1				1		1	2		
		1	1	2		2		2		
				1	1					1
		1	2			1		1		
		2	1	2			1	2		

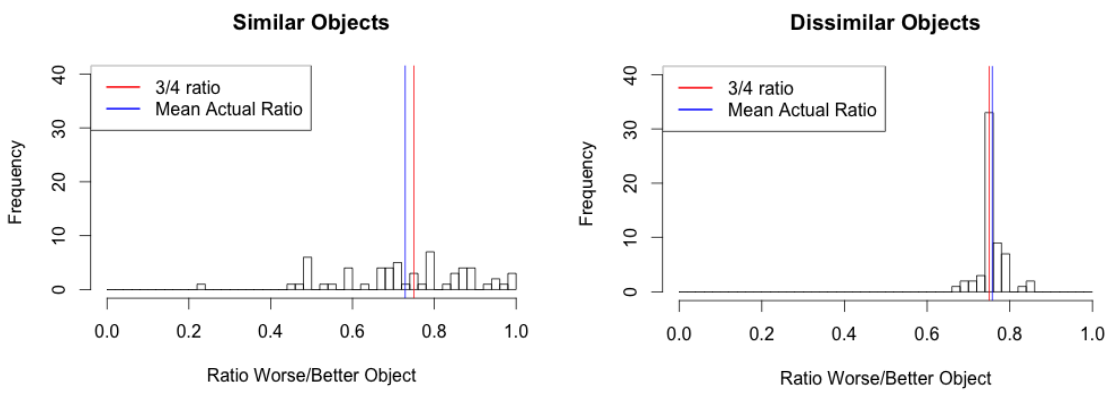
For the similar objects pair there was not much room for variation as there were only 3 objects to choose from. There were 6 possible combinations that featured these 3 objects either as c_l or c_h (e.g., the alarm clock as c_l and the radio alarm clock as c_h; or the opposite, the alarm clock as c_h and the radio alarm clock as c_l). The two most common combinations were the radio-speaker alarm clock as the better object, paired with the radio alarm clock (38% of the pairs shown), and the radio alarm clock as the better object, paired with the alarm clock (33%). This

confirms our assumptions that participants would be able to establish an ordering of preference, giving a higher value to the alarm clocks with additional features. For most participants, the radio-speaker alarm clock was more valuable than the other two clocks, and the radio alarm clock was better than the simple alarm clock.

In the dissimilar objects pair there was much more diversity. Only about half of the possible combinations of objects were displayed, but that adds up to 45 different pairs. The most common combinations were shown 5 times (6% of the time) and had the headphones as the better consequence, and the radio alarm clock as the worse consequence; or 3 times (3.5% of the time) and featured the headphones and the radio-speaker alarm clock, the toaster and the mouse, and the mouse and the bottle.

Regarding the value ratios, we were aiming for a 0.75 (¾) ratio, and the mean of the actual ratios for the similar and dissimilar objects was 0.73 and 0.76 respectively. However, because there was a limited pool of objects to choose from, the actual payoff ratios for every pair of objects were not always close to ¾ (see their distributions in Figure 11). Deviations from the ¾ ratio are more extreme for the similar objects.

Figure 11. Distribution of the Actual Ratios: Similar vs. Dissimilar Objects (n=60)

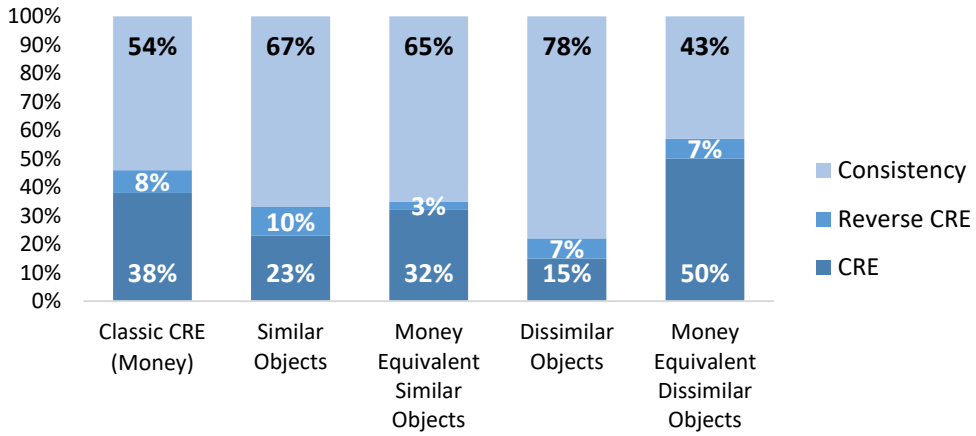


3.4.2. Choice Patterns

First, we looked at the effect of our complexity manipulation by comparing the choices people made in the object sets to those they made when the consequences were money amounts. Within the object sets, we looked at whether choice patterns differed depending on whether the objects have some dimensions in common (similar objects) or rather their dimensions were distinctive (dissimilar objects). Second, we tested whether the classic CRE with monetary payoffs of £40 and £30 was replicated; and whether the effect would persist with consumer goods.

We start the section with an overview of the choice patterns. EUT predicts consistency in people's choices in the scaled-up and the scaled-down questions. That is, either an S-S' or a R-R' choice pattern. But just like in many other studies, a substantial share of our participants made different choices in the two questions. There are two types of reversals: the CRE (S-R'), in which participants choose the safe prospect in the scaled-up question, and the risky prospect in the scaled-down question; and the reverse CRE (R-S'), with the opposite choice pattern. See Figure 12 below for a summary of the choice patterns observed.

Figure 12. Percentage of Choice Patterns per Set



We found more preference switches in the money sets than in their object counterparts; that is, there was more consistency with objects than with money. Leaving consistency cases aside, the direction of the preference switches was the same for both goods: there were more CRE-type reversals than the opposite type for both money and objects. However, the asymmetry between these two types of switches was smaller in the object sets compared to the money sets.

A more detailed account of respondents' choice patterns in the scaled-up and scaled-down questions in each set is provided in Table 8. Choices in the scaled-up questions (S and R prospects) are shown in the rows, and choices in the scaled-down questions (S' and R' prospects), in the columns. The intersections of the rows and the columns show how many participants followed each of the four choice patterns in a given set. Two of these patterns correspond to consistent preferences (S-S' and R-R'), and the other two, to a preference switch – either the CRE or the reverse CRE. The numbers in bold on the diagonal indicate the number of people whose choices deviate from consistency, and the cells highlighted in grey correspond to the classic CRE choice pattern. The numbers in brackets next to the set name correspond to the number of the two types of deviations from consistency: first, the CRE, and second, the

reverse CRE. We used McNemar's Chi-squared tests[19] to assess whether these deviations were random[20] or systematic.

Table 8. Detail of Choice Patterns per Set

1. Classic Common Ratio Effect (23:5)

		Scaled-down		
		S'	R'	TOTAL
Scaled-up	S	27	**23**	50
	R	**5**	5	10
	TOTAL	32	28	60

McNemar's Chi Square Test: p-value < .01

2. Similar Objects (14:6)

		Scaled-down		
		S'	R'	TOTAL
Scaled-up	S	13	**14**	27
	R	**6**	27	33
	TOTAL	19	41	60

McNemar's Chi Square Test: p-value = .1175

3. ME of Similar Objects (19:2)

		Scaled-down		
		S'	R'	TOTAL
Scaled-up	S	25	**19**	44
	R	**2**	14	16
	TOTAL	27	33	60

McNemar's Chi Square Test: p-value < .01

4. Dissimilar Objects (9:4)

		Scaled-down		
		S'	R'	TOTAL
Scaled-up	S	23	**9**	32
	R	**4**	24	28
	TOTAL	27	33	60

McNemar's Chi Square Test: p-value = .2673

5. ME of Dissimilar Objects (30:4)

		Scaled-down		
		S'	R'	TOTAL
Scaled-up	S	15	**30**	45
	R	**4**	11	15
	TOTAL	19	41	60

McNemar's Chi Square Test: p-value < .01

The left half of the figure displays the objects sets, and the right half shows their counterparts, where the consequences were the objects' monetary equivalents. According to EUT and alternative models such as CPT, the right and the left sections of the figure should match, given that the probabilities in the prospects, and the subjective values of their consequences (as stated by each participant) were the same. However, that was not what we found. The classic CRE, with £30 and £40 payoffs, was replicated. This effect (i.e., the asymmetry between the

[19] The p-values reported in Table 8 were computed using a continuity correction. The significance of the tests does not change without the continuity correction or with an exact version of the test (Fay, 2011) ran using the "exact2x2" R package. It is worth noting that in the case of the similar objects set, the p-value without continuity correction drops to .0736. However, we consider this irrelevant as our significance threshold is $\alpha = .05$. The difference in p-values between the similar and dissimilar objects sets (.1175 vs. .2673) cannot be explained by value differences between the outcomes in the pair, differences in the deviation from the target ratio, or differences in preference confidence – see summary statistics of these metrics and statistical tests in Table 31, Table 32 and Table 33 in Appendix B.

[20] The use of McNemar's Chi-squared test to determine whether the CRE is systematic or the deviations from consistency are random is well-established (see Moffatt, 2015). It is worth noting that this test underlies the assumption that if the CRE was caused by noise, deviations from consistency would be equally likely in both directions (S-R' and R-S').

number of preference switches, with more of the SR' kind) persisted for the other money sets, but its statistical significance disappeared for the objects sets. In isolation, the choice patterns observed when the consequences were objects could be taken as evidence of the validity of EUT. However, because we did observe preference shifts with money, we cannot claim our results support the validity of this theory. Most importantly, our results indicate that choice complexity, induced by whether the consequences were money amounts or objects, did have an effect on choice patterns.

3.5. Discussion

EUT and alternative models such as CPT predict that as long as the subjective value of the consequences offered by the prospects, and their (weighted) probabilities are the same, then respondents' choice patterns should also be the same regardless of how complex the choice is. In this study, we set out to explore, under full incentive compatibility, whether the complexity of the choice, brought by the comparability of the consequences, would also influence decisions. We elicited participants' money equivalents for 10 objects as carefully as possible, refining them through several different tasks. We then asked participants to make choices under risk with the objects as outcomes, and the same questions but with the monetary equivalents participants had stated instead. We found that choice patterns were different with monetary and non-monetary payoffs (i.e., consumer goods) that they had stated were equivalent. This evidence suggests that it is not only the subjective value of the consequence what enters people's decisions.

We replicated the classic CRE and also found this reversal with monetary payoffs, even when the amounts did not have an exact ¾ value ratio. However, when the CRE questions were made incentive compatible and involved payoffs that are difficult to compare to each other, we found that the established CRE was greatly attenuated. The choice patterns that we found with objects are closer to those predicted by EUT axioms than the ones we found with money, yet this cannot be taken as evidence of the descriptive validity of EUT, as participants did not behave as expected utility maximisers with monetary outcomes. In fact, these results might be pointing out a void in the decision making models literature, which fails to explicitly account for the effect of the (un)comparability of the consequences dimensions. The need to include features that increase the complexity of choices in models of decision making has been highlighted in previous work (e.g., DeShazo & Fermo, 2002; Swait & Adamowicz, 2001b).

A possible explanation for these choice patterns is that respondents resorted to heuristics (such as "choose the prospect with the preferred outcome, regardless of its probability") when comparing the consequences was too hard. This rationale is consistent with literature that suggests that when decisions are too difficult to make, people use heuristics - see, for example, Gigerenzer and Goldstein (1996), or Camerer (2003). One way of simplifying decisions is "by ignoring some potentially relevant information and by avoiding trade-offs among values" (Payne, Bettman & Johnson, 1988, p. 3). In line with the previous statement and closely related to our results, Payne (1976) found evidence that increased decision complexity resulted in respondents using heuristics to reduce cognitive strain.

Comparing multi-dimensional goods such as everyday objects may be harder than comparing money amounts, which are unidimensional and have cardinal properties. When confronted with a choice involving two objects, respondents may have been able to tell which object was better, but not *how much* better it was. This would have made trading off the subjective value of the consequences with the probability of receiving them very difficult, leading respondents to make their decisions on the basis of heuristics instead. In the scaled-up question, which is arguably the hardest out of the two (as the prospects differ in both the outcome and the probability dimensions), the choice patterns we observed are consistent with the use of the heuristic mentioned above. Respondents were more likely to choose the risky prospect in this question when the payoffs were objects than when they were money. The CRE involves choosing the safe prospect in the scaled-up question, and because that happened much less with objects, there was less room for the CRE to take place with objects compared to money. This might explain the difference in choice patterns between the objects and money sets, and why the CRE was found for the latter and not for the former.

With regards to the complexity manipulation within the objects, we found no evidence that choices between the similar objects and their money equivalents were more consistent than those between the dissimilar objects and their money counterparts. However, we did find more consistent cases with the dissimilar objects than with the similar objects (40 vs. 47 out of 60). The consistency we observed, once again, may have been a result of complexity, which made making respondents resort to heuristics.

To conclude, the findings from this study support the claim that preferences and attitudes are domain-specific: our participants made different choices when the consequences were money than when they were objects, even though they had indicated the two were equivalent. This has important implications regarding the generalisability of the behaviour that we test in

laboratory experiments, and the validity of money questions to represent real life decisions. These findings also provide empirical support to the notion that outcome domain may influence preferences, which has been shown to occur in the case of attitudes to risk by Weber, Blais, and Betz (2002).

Chapter 4.
The Efficiency-Equity Trade-off, Self-Interest, and Moral Principles in Health and Safety Valuation

Abstract

Governments attempt to respect public preferences when making trade-offs between policy options. Yet most estimates of the value of health and safety reflect only individuals' self-interested preferences. This approach may neglect or misrepresent individuals' preferences over the *distribution* of public resources (for a review, see Guth and Kocher, 2014). We conduct an experiment in which participants rate and choose between policy options that differ in terms of their efficiency (expected number of lives a policy would save or the cases of ill health it would prevent) and their equity (balance of risk reductions for different parts of the population). Participants considered several interventions to clean up a hypothetical city's water supply. Different options would reduce the risk of death or ill health for people in different areas of the city, and the size of the risk reduction varied across options. In order to see whether personally benefitting from the risk reduction would affect this trade-off, we told half of the sample to imagine that they lived in a specified area of the city. Our results suggest that efficiency is the most important factor in determining which policy option is preferred, but participants were almost as influenced by equity as by efficiency. When participants were included in the scenarios, they were influenced by efficiency in the area where they (hypothetically) lived, favouring policies that particularly benefitted them. However, the effect of self-interest was smaller than that of the general concern for efficiency. Our findings suggest that preferences for efficiency and equity are not substantially affected by self-interest, and contribute to the growing evidence that the number of people helped is not all that matters to decision makers evaluating health interventions.

Keywords: efficiency, equity, self-interest, moral principles

4.1. Introduction

When deciding how to allocate resources to maximise societal welfare, difficult trade-offs are inevitable. Should resources be spent on education or on healthcare? Are some health improvements more valuable than others? Such illustrative trade-offs involve weighing benefits and costs of different magnitudes, at different times, and to different recipients.

Governments attempt to respect public preferences when making trade-offs between different policy options. One way they do this is by using estimates of the monetary value of policy outcomes, which can be obtained using methods such as contingent valuation (e.g., Beattie et al., 1998). However, most of these estimates (e.g., fatalities prevented, improvements to health outcomes, or environmental damage prevented or rectified) reflect only the preferences that individuals express for their personal benefits and costs. The measures reflect individuals' willingness to pay for reductions in their own risk, improvements in their own health, or benefits to their own enjoyment of environmental amenities, or their willingness to accept compensation for harms in these contexts. To some extent, this self-interested approach is legitimate, since it avoids double counting of the benefits of policies (for a useful discussion in relation to the Value of Statistical Life, see Jones-Lee, 1991[21]). However, there is good reason to suppose people have preferences that extend beyond concern for themselves. These preferences include concern for efficiency (maximising the expected number of fatalities or cases of illness prevented) and equity (balancing risk reductions for different parts of the population[22]). Focusing exclusively on individuals' values for their own gains or losses may neglect their preferences over the distribution of societal resources and thereby fail to adequately represent important sources of societal welfare.

Accounting for preferences over efficiency and equity *per se* is not enough. To ensure that resource allocations maximise society's welfare, policies must reflect the trade-offs that individuals would make between self-interest and these distributional concerns. After all,

[21] Jones-Lee concludes that societal desire (or willingness to pay) for safety could be double counted when we account for the level of safety that we want for others in excess of what they want for themselves. He shows that this would be the case if altruism took a *pure* form (i.e., when people care about all determinants of others' welfare) rather than being *safety-focused* (i.e., when they only care about others' safety). The provision of 'double-counted' safety would be done at the expense of other determinants of welfare, and hence be suboptimal.

[22] Throughout this chapter, we use 'equity' to refer to 'gains egalitarianism' rather than 'outcome egalitarianism' (read more about this distinction on Tsuchiya and Dolan (2009)). This approach allows us to measure efficiency and equity on the same space: gains. We ask respondents to trade off more gains overall with a more equal distribution of gains.

policies that influence the level of health and safety in a population usually do not apply only to some group of "others" to which the typical member of the public does not belong. Instead, the important preferences are those of members of the affected population. As such, attempting to respect equity preferences by relying on Person Trade-Off surveys[23] where the respondent takes the role of the social planner cannot directly shed light on their preferences for the equity and efficiency of the policies that affect them. Nor can we use the experimental evidence from the laboratory on the preference for self and others in a monetary context, since the policy contexts are far removed from the distribution of a monetary endowment in the laboratory. Despite the critical importance of this question (i.e., how preferences for welfare interact with each other), the existing literature does not provide a satisfactory answer. We present evidence that begins to address this issue: we elicit preferences in the context of health and safety policy choices; we use structured experimental designs that can isolate the relevant effects; and we explicitly test the effect of self-interest in determining preferences for the different policy options.

Section 2 discusses the relevant literature. Section 3 presents our experimental design. In section 4 we outline our results, and in Section 5 we discuss their implications.

4.2. Theoretical Background

Standard economic theory, which provides the theoretical foundation for classic preference elicitation methods (Braga & Starmer, 2005), assumes that people are *self-interested, utility-maximising* agents. However, people's behaviour is often not well-described by this assumption. This has been shown through experiments and surveys in a range of policy-relevant contexts, which we outline in this section. Yet, little is understood about how self-interest modifies the equity-efficiency trade-off.

At the most abstract level, experimental economists have provided evidence that people are sometimes willing to reduce their own payoff in order to improve the payoff of others, which is contrary to the assumption of self-interested agents. This is demonstrated in Dictator Games, in which a participant receives some endowment of experimental tokens or cash and chooses how to distribute it between themselves and a randomly assigned, anonymous partner, the

[23] In the Person Trade-Off method, respondents are asked to consider two different groups of people, each affected by a different ailment, and consisting of a different number of patients. They need to make choices about which group to help as the number of people in each group varies, up to the point when they consider both groups to be equally deserving to be helped. This method was first used by Patrick, Bush and Chen (1973), but it was Nord (1992) who gave it its current name.

"receiver". Although the self-interested Nash Equilibrium predicts that the endowed player will keep the whole endowment for themselves, non-zero offers are the norm (see a 2011 meta-analysis by Engel). The Ultimatum Game has one more stage than the Dictator game: the receiver can reject the offer of the other player, in which case both receive nothing. Again, non-zero offers are pervasive, and receivers often reject low offers, reducing the total payoff to zero rather than accepting an unfair or inequitable distribution. Evidence along these lines is outlined in Guth et al. (1982); Kahneman et al. (1986); Roth et al. (1991); Camerer and Thaler (1995), and a useful review is provided by Guth and Kocher (2014). This laboratory evidence is accompanied by a wealth of theoretical models that account for preferences for others' welfare (often modelled as interdependent preferences, such as in Bergstrom, 1999) and for distributional concerns (e.g., fairness models including Rabin, 1993; Fehr and Schmidt, 2000; and Bolton and Ockenfels, 2000). Human aversion to inequality has even been explained at the neural level by Tricomi et al. (2010).

This laboratory evidence illuminates how individuals are willing to sacrifice self-interest to improve overall efficiency, and also to achieve more equitable distributions. However, preferences have been found to depend on the context in which they are elicited (see an early formalisation of this by Tversky and Simonson, 1993). Also, there is mixed evidence on whether behaviour in context-free experiments is the same as in experiments with framing, and in the field. Laury and Taylor (2008), Voors, Turley, Kontoleon, Bulte and List (2012), and Galizzi and Navarro-Martínez (2018) investigated the correlation between behaviour in context-free experiments (they used popular games such as the dictator and public good games) and behaviour in the field. The latter found no correlation, whereas the former found correlations in some instances only. In a context of unethical behaviour involving tax evasion and bribery respectively, Alm, McClelland and Schulze (1992) and Abbink and Hennig-Schmidt (2006) found that the experiment framing did not make a difference. As such, we cannot make confident inferences from the laboratory evidence when considering the equity-efficiency trade-off in policy relevant contexts such as safety and health risk mitigation.

There is a large body of work on the equity-efficiency trade-off in more realistic policy contexts. In the context of safety, there is a long and rich literature on the notion of fairness in public policymaking. In an early discussion, Charles Fried (1969, pp. 1426-1427) argues that:

> *"The most obvious objection to a pure efficiency argument is an argument from fairness. […] it is not sufficient to justify the choice of a particular life-saving strategy that it leads to the least net loss of lives in the long run. It must also be shown that the risk of death*

is fairly apportioned among the relevant population […] Fairness, thus is an important […] constraint on efficiency".

Writing in *Science,* Arrow et al. (1996, p. 222) state:

> *"Although benefit-cost analysis should focus primarily on the overall relation between benefits and costs, a good analysis will also identify important distributional consequences."*

Despite the recognition that distributional consequences matter, it is not straightforward to define what they mean. For example, Viscusi (2008) sets out difficulties of being "fair" – for example if every expected fatality prevented is considered to be equally valuable, this implies that not all life years are equally valued, as some people are expected to live longer than others (although see Jones-Lee et al. (2015) for a discussion). Viscusi (p. 316) summarises with the insight that *"what is viewed as equitable and what is not hinges quite critically on how one frames the fairness issue."*

More progress has been made in the health economics literature. Bobinac et al. (2012) provide a useful review of the literature about the trade-offs between efficiency and equity when cost-effectiveness is measured in quality adjusted life years (QALYs). They advocate consideration of both efficiency and equity concerns, where equity can be determined by patient characteristics such as age and whether or not they have dependents. Light (1992) discusses the efficiency-equity trade-offs embedded in the healthcare system, beginning with the premise that equity and efficiency are both desirable. Wagstaff (1991) proposes a Social Welfare Function approach to incorporate these efficiency and equity concerns into health resource allocation decisions. However, neither Light nor Wagstaff provide empirical evidence about public preferences over the trade-off between efficiency and equity. Lindholm, Rosen and Emmelin (1996) examine the requirements for meaningful empirical estimates of the trade-off, and in a pilot study, they found that at least two thirds of the 68 Swedish politicians responsible for healthcare that took part in their study stated that they would be willing to give up efficiency to achieve more equity.

Patrick and colleagues (1973) and Nord (1992) present the Person Trade-Off approach, a framework for eliciting trade-offs. These trade-offs are embedded in choices between helping different people to achieve different levels of health. Dolan (1998) extends this method and provides a small experimental test demonstrating a general preference for fairness. In his study, it was impossible to distinguish fairness in terms of treating the person who was initially worse off, versus fairness in terms of reaching an equitable final distribution of health. More recently,

Bleichrodt et al. (2005) present a method for non-parametric estimation of preferences for equality. These trade-offs were elicited in terms of cohorts of newborns with different potential QALYs over the lifetime. The results suggest that people are averse to inequalities in the domain of health, and the authors propose that policy values for health effects should be adjusted using equity weights.

Ubel et al. (1996a) conduct a survey to elicit equity and efficiency trade-offs from a sample including the general public and from specialists including medical ethicists and experts in Medical Decision Making. They found that:

> "Fifty-six percent of the prospective jurors, 53 percent of the medical ethicists, and 41 percent of the experts in medical decision making recommended offering [a] less effective screening test to everyone, even though 100 more lives would have been saved by offering [a] more expensive test to only a portion of the population." (Ubel et al., 1996a, p. 1174)

In follow up studies, Ubel et al. (2000) provide evidence that suggests the trade-off between efficiency and equity is not continuous, but rather the preference for equity is 'all or none'. When respondents were offered a choice between a more efficient and a more equitable policy, if the equitable option did not help 100% of the population, then respondents chose the efficient one. On the other hand, Johannesson and Gerdtham (1996) had collected empirical evidence of the magnitude of the efficiency-equity trade-off, and they had found that respondents were "willing to give up 1 QALY in the group with more QALYs to gain 0.45 QALYs in the group with fewer QALYs" (pp. 365-366). They used a veil of ignorance approach, and a sample of 80 students.

All of the health studies mentioned elicit social preferences from impartial observers through Person Trade-Off or similar methodologies, setting aside self-interest. However, Ubel et al. (1996b) present an experiment that directly compares self-interested utility scores with impartial person trade-off responses to ask whether "people place the same values on health care conditions when thinking of their own health as when thinking about health care policy" (p. 109). They find strong evidence that the social and the self-interested perspectives generate different implied distributions of health. The utilities of the worst conditions implied from their impartial choices were much lower than those elicited when participants took a self-interested perspective. Similarly, Nord et al. (1999) consider different distributional preferences including aversion to inequalities in health and discuss how these preferences can be combined with a self-interested measure of health state utilities, proposing the "Cost-Value Approach".

However, in both studies the social perspective and self-interested perspective are elicited, or proposed to be elicited, using completely different methods. In Ubel et al. (1996b), it is not clear to what extent any differences derive from differences in perspective, or from differences in the type of task. Specifically, when in the self-interested tasks, there was no scope to express preferences for equity.

We present a study in which participants trade off different distributional concerns within the same task. In some cases, the participants had to trade off their own self-interest against these distributional concerns, while in other cases, they acted impartially. With a single type of task, we are able to explore the influence that self-interest has on the efficiency-equity trade-offs that people made.

In addition to the main trade-offs, we explore whether the stake size can influence individuals' willingness to trade efficiency for equity. We do this by manipulating whether the policies under consideration would prevent illnesses (low stakes) or fatalities (high stakes). It is plausible that pro-social preferences including the preference for equity are more evident when the stakes are lower than when the stakes are high. Evidence that changing monetary stakes can influence social preferences is presented by Forsythe et al. (1994) and by List and Cherry (2008). Camerer and Hogarth (1999) provide a useful review. However, to the best of our knowledge, no comparable research has been conducted in the context of health and physical risk.

The factors explored in our study – individuals' private interest and concerns for efficiency and equity - are integral features of the trade-offs that society must make between policy outcomes. These trade-offs imply a set of underlying moral principles. For example, if individuals chose the policy that gives everyone in the population a risk reduction, despite offering a lower expected risk reduction overall, that implies that they are favouring equity over efficiency. A related but underexplored question is whether these principles can be elicited directly from members of the public; for instance, by asking them whether they agree or disagree with statements involving the principles. If so, would the policy decisions implied by people's directly elicited moral principles be equivalent to the decisions implied by their explicit choices between policy outcomes? If that is the case, using moral principles to guide difficult trade-offs in policy making could be a fruitful avenue for future research.

4.3. Method

The methods presented here were developed on the basis of the findings of a pilot study (n=107) conducted with student participants at the University of Durham. Results of that pilot study are available on request.

4.3.1. Scenario

Participants were asked to imagine a city made up of an East and a West zone. The city's inhabitants were at risk of potential adverse health effects resulting from a contaminated water supply, and participants expressed their preferences concerning different policy options to improve the situation. Participants were told that 100,000 people lived in each zone of the city, and that the city's water supply came from two reservoirs. The reservoirs were contaminated by three types of bacteria (E, W and B) which differed in two respects: the zone they affected, and the anticipated number of people they would harm. The initial risks were not set to be equal across the zones: the West zone had a higher baseline risk than the East zone. The baseline case before policy intervention is presented in Table 9.

Table 9. Baseline Risks

Bacteria Name	Baseline Risk in East Zone	Baseline Risk in West Zone
Bacteria E	10/100,000	0/100,000
Bacteria W	0/100,000	20/100,000
Bacteria B	8/100,000	8/100,000
	Total = 18/100,000	Total = 28/100,000

We manipulated self-interest between subjects by telling half of the participants that they lived in the East zone. The other half were not told they lived in the city at all, so can be assumed to adopt an impartial perspective[24], analogous to that of a social planner. We can therefore use the choices of the impartial participants (who did not live in the city) to investigate trade-offs between efficiency and equity, and compare these with the choices of those who might be influenced by self-interest, to investigate how self-interest affects these trade-offs.

Between subjects, we manipulated the harm that would be caused by the bacteria. For half of subjects, the bacteria would cause fatality. For the other half, the bacteria would cause gastroenteritis, a health complaint described as follows:

[24] Even if participants did assume they lived in the city, there is no reason for them to suppose they lived in the East zone rather than the West zone or vice-versa.

Gastroenteritis caused by these bacteria is not contagious between people and it cannot kill you, but people infected will have the following symptoms for approximately two weeks: Watery diarrhoea; Abdominal cramps and pain; Nausea, vomiting, or both; Occasional muscle aches or headache; Low-grade fever.

We defined four policy options, which are outlined in Table 10. The policies differed in which bacteria they targeted (and hence which zone they helped), as well as in the magnitude of the risk reductions.

Table 10. Policy Options

Policy Option	EAST ZONE (Baseline Risk = 18/100,000)		WEST ZONE (Baseline Risk = 28/100,000)	
	Risk Reduction	**Final Risk**	**Risk Reduction**	**Final Risk**
Policy E	10/100,000	8/100,000	0	28/100,000
Policy W	0	18/100,000	20/100,000	8/100,000
Policy B	8/100,000	10/100,000	8/100,000	20/100,000
Policy X	3/100,000	15/100,000	4/100,000	24/100,000

Policy E eliminates bacteria E, which is expected to affect 10 people all in the East zone but does not help the West zone. Policy W eliminates bacteria W, which is expected to affect 20 people all in the West zone but does not help the East zone. Therefore, option W is the most efficient (because it offers the greatest expected reduction in harm) but option E is the self-interested option for those subjects who were asked to assume they lived in Zone E. Options B and X improve safety in both zones. Policy option B eliminates bacteria B and is equitable in the *absolute* sense, since it reduces the risk in both zones by the same amount. Option X reduces the risk in both zones by a smaller amount that is close to proportional to the initial risk in each zone.

See the instructions, bacteria and policy descriptions as participants saw them in Appendix C.1.

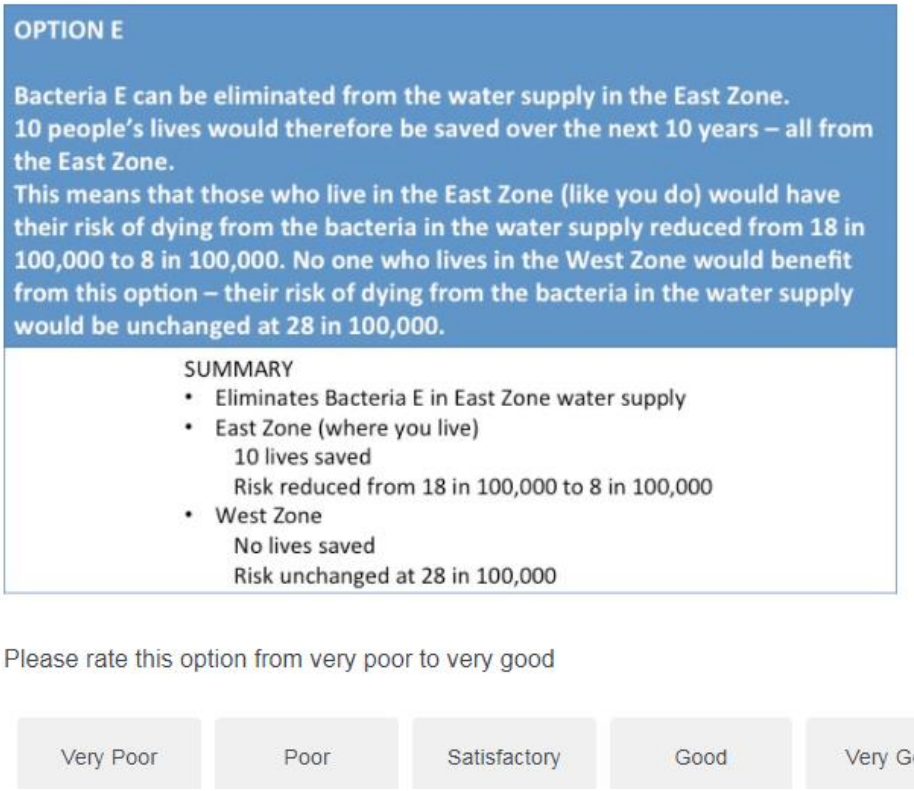
4.3.2. Tasks

The study consisted of three sections: the policy rating and comprehension task, where participants became familiarised with the policies; the stated principles task, in which participants rated their agreement with a series of relevant statements of moral principles; and the preferences task, in which participants made choices between policy options. First, we presented the rating and comprehension tasks. Then, we randomised the order in which the stated principles and the preference tasks were presented between subjects.

Policy rating and comprehension task

Each of the water cleaning options was presented on the screen, and the participant rated each policy as *very poor*, *poor*, *satisfactory*, *good*, or *very good* (see example in Figure 13).

Figure 13. Example Rating Question



This gives us an initial assessment of how good the policy options were regarded to be. Also, it ensured that participants carefully considered each of the options before they were asked to complete the stated principles and preferences tasks. To check whether the important properties of the policies were understood, participants answered either four or five comprehension questions (depending on the condition). These are outlined in Table 11. In all cases, the information about the policies was displayed on screen, and the respondent had to identify the correct policy in each comparison. After the decision was made, participants were shown the correct answer with an explanation. The questions were presented in random order, and the order of the options on the screen was also randomised.

Table 11. Comprehension Questions

	Comparison	Question	Correct answer
1	**Policy E vs. Policy W**	Which option would make the water safer for the people who are most at risk from the bacteria?	Policy W
2	**Policy W vs. Policy X**	Which option would make the water a lot safer for some of the people who live in the city rather than a little safer for everyone who lives in the city?	Policy W
3	**Policy B vs. Policy E**	Which option saves lives (avoids gastroenteritis cases) in both zones of the city rather than in just one zone of the city?	Policy B
4	**Policy W vs. Policy B**	Which option would save the most lives (avoid the most gastroenteritis cases)?	Policy W
5	**Policy B vs. Policy X***	Which option would make the water safest for you and your household?	Policy E

Note. The comparison between B and X (marked with an asterisk) was only asked of those people in the self-Interest condition.

Stated Principles task

Participants were presented with six pairs of statements (see Table 12). Each statement is a trade-off between two relevant principles, which were written to represent opposite ways of thinking about how options should be prioritised. Each statement favoured one principle over the other, forcing participants to explicitly express a relative preference between competing moral principles (or indicate indifference between them). This design also prevented participants from agreeing with all principles due to a desire to conform to social norms. Participants indicated which of the two statements in the pair they favoured, and how strongly. Figure 14 shows an example question.

Figure 14. Example Stated Principles Question

The following two statements represent two opposite ways of thinking about which option should be prioritised. Pick the option that best describes your opinion.

Statement A	Statement B
The chosen option should save the most lives even if that means that the water is made safer for only some of the people who live in the city.	The chosen option should make the water safer for everyone who lives in the city even if that means that fewer lives are saved.

Strongly favour A over B	Moderately favour A over B	Slightly favour A over B	Equally favour A and B	Slightly favour B over A	Moderately favour B over A	Strongly favour B over A

We asked participants to trade off the following principles: (i) efficiency (to maximise the expected number of cases prevented), (ii) equity (to evenly spread the risk reduction across everyone in the city), (iii) self-interest (to prioritise own risk reduction), (iv) helping those most at risk (i.e., to reduce the risk of those with the highest baseline), and (v) offering a higher risk reduction to only a few citizens.

On the basis of this task, we constructed an efficiency preference score that indicates the preference for efficiency against other concerns. The score is the sum of the times efficiency is favoured over the concern it is traded-off against. The trade-offs that did not involve efficiency were included to induce participants to consider principles in the wider sense, and not only as a trade-off against efficiency. The preferences captured by the efficiency preference score will be compared with those elicited in the preference task (see below). However, we wanted to rule out the possibility that consistency between these two tasks was merely a result of participants' desire to appear coherent rather than a reflection of the consistency of true underlying preferences. By randomising whether the stated principles task was presented before or after the preference task, we are able to observe systematic differences in responses by order of presentation, which would indicate this effect.

Table 12. Stated Principles Task

	Trade-Off	Principle A	Principle B
1	**Equity vs. Efficiency**	The chosen option should make the water safer for everyone who lives in the city even if that means that fewer lives are saved.	The chosen option should save the most lives even if that means that the water is made safer for only some of the people who live in the city.
2	**Helping those most at risk vs. Efficiency**	The chosen option should make the water safer for the people who are most at risk from the bacteria even if that means that fewer lives are saved.	The chosen option should save the most lives even if that means that the water is not made safer for the people who are most at risk from the bacteria.
3	**Equity vs. Inequity**	The chosen option should make the water a little safer for everyone rather than a lot safer for only some of the people in the city.	The chosen option should make the water a lot safer for only some of the people in the city rather than a little safer for everyone.
4	**Equity vs. Helping those most at risk**	The chosen option should make the water a little safer for everyone rather than a lot safer for the people who are most at risk from the bacteria.	The chosen option should make the water a lot safer for the people who are most at risk from the bacteria rather than a little safer for everyone.
5	**Self-interest vs. Helping those most at risk***	The chosen option should make the water safer for me even if that means that the water is not made safer for the people who are most at risk from the bacteria.	The chosen option should make the water safer for the people who are most at risk from the bacteria even if that means that the water is not made safer for me.
6	**Self-interest vs. Efficiency***	The chosen option should make the water safer for me even if that means fewer lives are saved.	The chosen option should save the most lives even if that means the water is not made safer for me.

Note. Questions marked with an asterisk were shown to participants in the self-interest condition only.

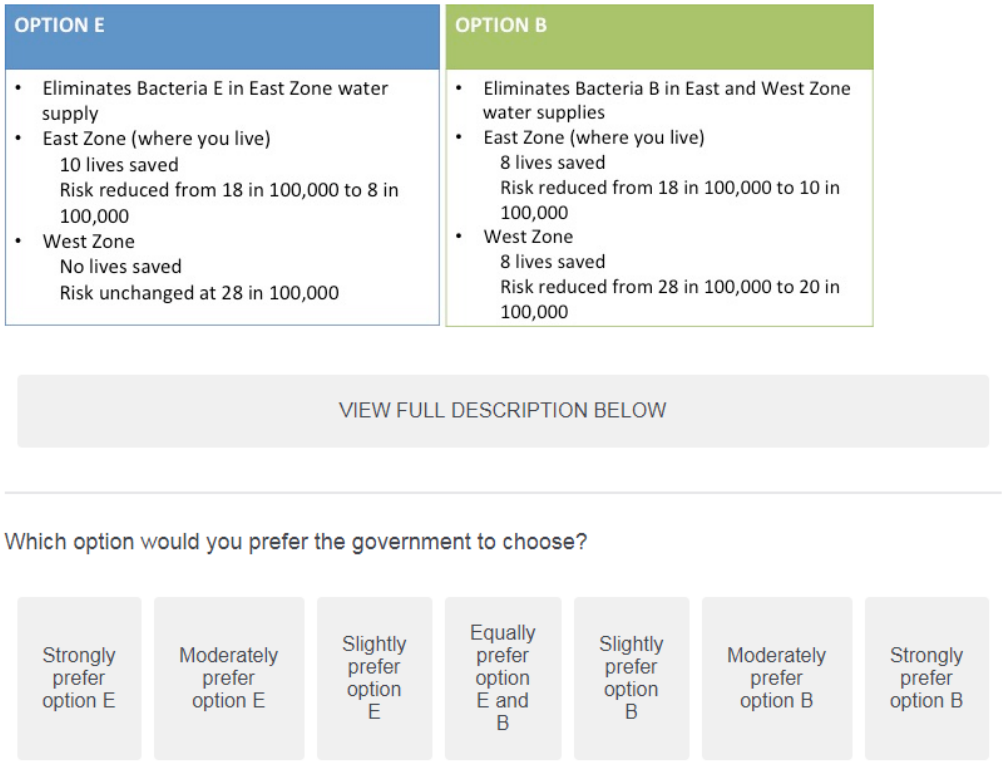
Preference Choice task

The questions in the preference task asked respondents to choose between policies. We informed participants that due to the cost of treating the bacteria in the reservoirs, the government would not be able to implement all the policies. We told them: "The following questions will ask you to tell us your opinion about which of the options you would prefer the government to choose[25]. There are no right or wrong answers. We are simply interested in your honest opinion." The questions presented the options in pairs. With four policies, this generates six comparisons (B vs W; E vs W; X vs W; E vs B; X vs B; X vs E). Every respondent saw every

[25] This framing allows us to use the same wording in both the self-interested and impartial conditions. It also reflects the goal of the study, which is to elicit preferences to inform policymaking.

comparison, and the order in which the comparisons were presented was randomised for each participant.

Figure 15. Preference Elicitation Task Example



Note. In this example, the participant was in the self-interest condition. For these participants, we highlighted that they lived in the East zone to ensure they understood which of the policies would directly affect them, and by how much. For other participants, the parentheses and their contents were omitted.

Figure 15 shows a typical preference elicitation screen. Participants saw summaries of the policies, and more details were available if they clicked "view full description below". We randomised which options were presented on the left and right of the display. We asked participants to select which policy they would prefer the government to choose, and to indicate the strength of their preference.

To summarise, the study employed a 2x2x2 between-subjects design. The first distinction was the manipulation of self-interest. Specifically, half of subjects were told they lived in the East zone, whilst the other half were not told they lived in the city at all, so can be assumed to adopt an impartial perspective. The second distinction was the type of consequence the risk had: for half of the participants, the risk was contracting gastroenteritis, whilst for the other half it was the risk of death. The final distinction randomised whether participants undertook the preference choice task before the stated principles task or vice versa.

4.3.3. Participants

The study was completed online by a non-student sample of UK residents. We obtained 322 completed responses, assigned to treatments as outlined in Table 13. Our sample is 48% female, ranging in age from 20 to 71 (M = 38.6, SD = 10.83), and with an average household size of 4 (M = 3.9, SD = 1.57). Household members included any close relatives living in the household (e.g., children, spouse or partner, parents). The randomisation to treatments was successful, with no significant differences in age, gender, number of dependants, or the percentage of comprehension questions correctly answered (87% overall) between conditions. The study was implemented in Qualtrics and distributed through Prolific Academic, an online labour market. The median time taken to complete this study was 13 minutes. Participants received a fixed payment of £2.50 for their participation.

Table 13. Assignment to Treatments

	Self-Interested (n=160)		**Impartial** (n=162)	
	Principles Before (n=82)	**Principles After** (n=78)	**Principles Before** (n=81)	**Principles After** (n=81)
Gastroenteritis (n=160)	40	40	40	40
Death (n=162)	42	38	41	41

4.4. Results

4.4.1. Descriptive Results

Ratings

Participants rated each of the four policy options on a five-point scale from “very poor” to “very good”. See the distribution of the ratings for each of the policy options in Figure 16.

The highest rated policy option was option B, which eliminates bacteria β reducing the risk in each of the two zones by 8/100,000. This option is the most equitable, and also relatively efficient since it reduces risk by 16/200,000 overall. The ratings of option B were significantly higher than those of the second highest rated option, W, as indicated by a Wilcoxon signed-rank test (p-value < .01). Option W is the most efficient, since it saves 20 lives, and these lives are of those people most at risk; but it is not equitable, since all of the benefit occurs in the West Zone. Option W was ranked significantly higher than option X (p-value < .01) and E (p-value < .01). Option X reduces the risk by 7/200,000 in total, spread across both zones, and option E reduces

the risk in the East zone alone by 10/10,000. There were no significant differences in the ratings of options X and E (p-value = .931). Overall, this generates the preference ordering B>W>X=E.

Figure 16. Ratings of the Policy Options

Figure 16 shows the overall distribution of ratings, but does not distinguish between those by people in the different treatments. Table 14 provides this breakdown, showing that there is a significant difference in the rating of option E and B depending on whether respondents were self-interested or impartial. In both cases, the ratings by participants in the impartial condition were slightly lower. There are no other significant differences between the ratings of participants in the self-interested and impartial conditions, nor between the ratings of participants in the death and the gastroenteritis treatments.

Table 14. Mean (SD) Ratings, by Condition

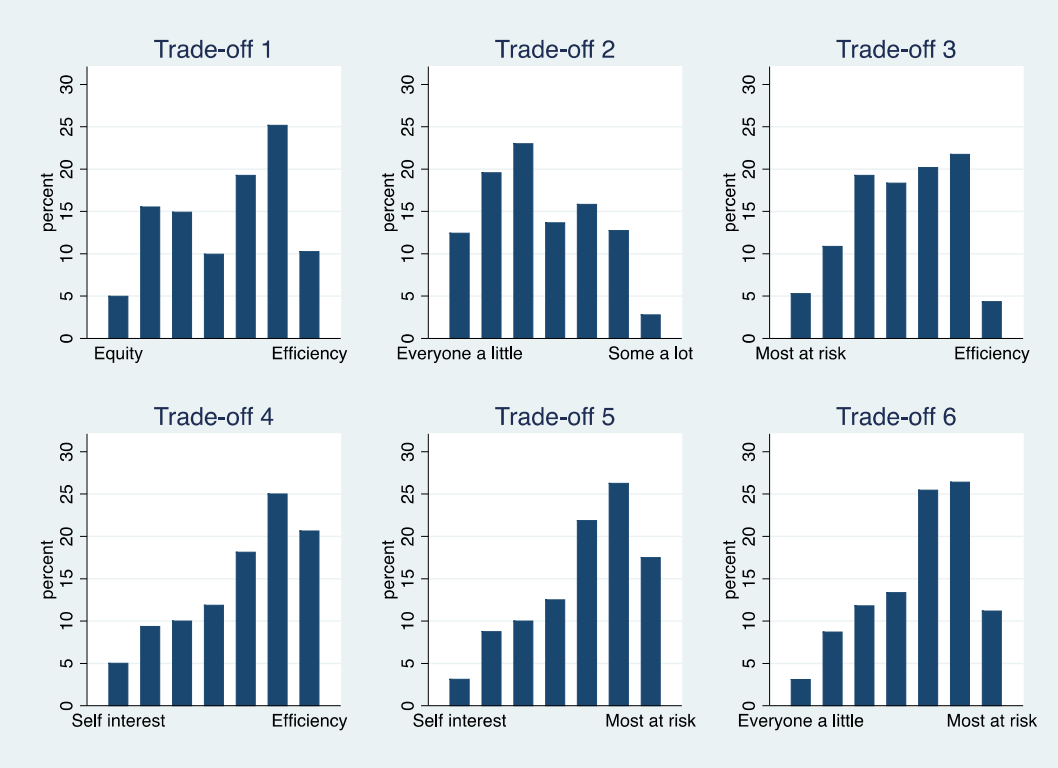
Policy Option	Mean (SD); (1 = very poor; 5 = very good)	WILCOXON RANK-SUM TEST P-VALUE	
		H_a: Self-Interested ≠ Impartial	H_a: Death ≠ Gastroenteritis
Policy W	2.91 (0.01)	n.s.	n.s.
Policy E	2.53 (0.99)	2.64 (0.97) vs. 2.41 (0.99)*	n.s.
Policy B	3.29 (0.92)	3.41 (0.90) vs. 3.17 (0.94)*	n.s.
Policy X	2.54 (0.97)	n.s.	n.s.

Note. *** < .001, ** < .01, * < .05, n.s.: not significant

Stated Principles

Participants stated the strength of preference (strong, moderate, or slight) with which they favoured one principle over the other in each trade-off. They could use the midpoint of the scale to indicate indifference. The distribution of the answers to the principle trade-offs are shown in Figure 17. Each plot presents a trade-off between the two principles on the x-axis. For example, on Trade-off 1 the first bar presents the percentage of participants who indicated that they strongly favoured equity over efficiency.

Figure 17. (Dis)agreement with Statements of Principle



The more polarised preferences were those in the pairs of statements where respondents had to trade off self-interest against another concern. In the case of self-interest vs. "helping those most at risk", 43.8% of respondents moderately or strongly favoured helping those most at risk over self-interest; in the case of efficiency, this percentage increased to 45.6%. These trade-offs might have been the easiest to make, given that there is a clear virtuous response not to prioritise oneself. That is not the case when trading off the other principles. The most finely balanced trade-off was between efficiency and helping those most at risk: almost 20% of participants equally favoured the statements, just less than 20% slightly favoured efficiency and just over 20% slightly favoured helping those most at risk. This indicates that this may be the most difficult trade-off.

A one-sample Wilcoxon signed-rank test rejected the hypothesis of indifference between principles in all six comparisons. The answers, split by condition, are shown in Table 15. We found no significant difference between participants in the self-interest and the impartial conditions, and no difference depending on whether the principle statements were completed before or after the preference task. However, the preference for efficiency over equity, and of efficiency over helping those most at risk was stronger for participants in the condition with the risk of death, compared to gastroenteritis.

These principles appear to be transitive, as an order of preference can be established from the answers to the principle trade-off questions: efficiency ≻ helping those most at risk ≻ equity ≻ self-interest. Moreover, this task supports the results of the rating task in suggesting that equity and efficiency are both important, and that self-interest is not sufficient to override these preferences.

Table 15. (Dis)agreement with Statements of Principle, by Condition

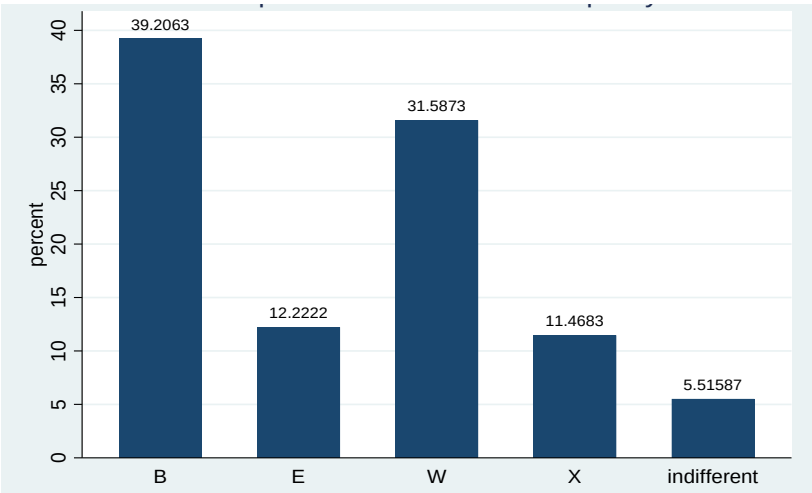
	Trade-off between Principle Statements	Mean[a] (SD); Wilcoxon signed-rank test p-value H_a: Mean ≠ 0	WILCOXON RANK-SUM TEST P-VALUE		
			H_a: Self-Interested ≠ Impartial	H_a: Death ≠ Gastroenteritis	H_a: Before ≠ After
1	Equity vs Efficiency	0.39*** (1.79)	n.s.	0.64 (1.64) vs 0.15 (1.89)*	n.s.
2	Helping those most at risk vs Efficiency	0.20* (1.59)	n.s.	0.48 (1.55) vs. -0.09 (1.58)**	n.s.
3	Equity vs Inequity	-0.50*** (1.68)	n.s.	n.s.	n.s.
4	Equity vs Helping those most at risk	0.73*** (1.60)	n.s.	n.s.	n.s.
5	Self-interest vs Helping those most at risk	0.90*** (1.67)	n/a	n.s.	n.s.
6	Self-interest vs Efficiency	0.86*** (1.80)	n/a	n.s.	n.s.

Notes. [a] -3 = strongly favour the 1st concern over the 2nd; 3 = strongly favour the 2nd concern over the 1st); *** < .001, ** < .01, * < .05, n.s.: not significant.

Preference Choices

Initially, we present the proportion of all choices in which a given policy option was chosen (i.e., when it was strongly, moderately, or slightly preferred to its comparator). These are presented in Figure 18.

Figure 18. Policy Option Choice Proportions



In line with the ratings data, option B was the most preferred, followed by W, with options E and X being chosen in a similar proportion of choices. To look more deeply into the choice data, we present the breakdown of the results for each comparison. In each one, participants selected a point from "strongly prefer Y" to "strongly prefer Z" where Y and Z are policy options under consideration. Option Y is always the more *efficient* option of the pair, where efficiency is the greater magnitude of risk reduction. For example, in the W vs. B comparison, a strong preference for Y would mean that the more efficient option W was strongly preferred to the less efficient option B. Table 16 presents the mean responses for each of the policy choices.

Table 16 also shows the means for the different groups in a treatment, in the cases when they are significantly different. This was the case in four of the choices for the self-interested and the impartial participants. When the most efficient option and the one that benefited the self-interested participants the most were not the same, then self-interested participants still chose the most efficient one, but their strength of preference was significantly lower than that of the impartial participants. The exception to this is the trade-off between policy B and W: on average, self-interested participants had a slight preference for B (which is the one that benefitted them the most), while impartial participants had a slight preference for W, which reduced risks by 4 more cases overall.

There were no significant differences in the answers of respondents who considered fatality risks and those who considered the risk of gastroenteritis. Recall that in the principle statements, we did find a difference between participants in these two conditions. It may be that preferences elicited with a choice task are less context-dependent than principles.

We also found a significant difference in the choices between policies X and B for those participants who answered the principle statements task before the choices, and those who answered after. On average, both groups chose the most efficient alternative, but their strength of preference varied. Even though this difference was statistically significant, it is very small in magnitude. It appears that participants who had thought about trade-offs between moral principles before this task were slightly more sensitive to the differences in risk baselines. This might explain why their strength of preference for option B, which offered an equal risk reduction in both zones, was lower than that for X, which offered a higher risk reduction to those most at risk.

Table 16. Policy Options Preferences, split by condition

	Policy Choice		Mean (SD); Wilcoxon signed-rank test p-value H_a: Mean ≠ 0	WILCOXON RANK-SUM TEST P-VALUE		
	Z (-3)	Y (3)		H_a: Self-Int'd ≠ Impartial	H_a: Death ≠ Gastro	H_a: Before ≠ After
1	**B^E**	**W**	-0.08 (1.87) n.s.	-0.51 (1.93) vs. 0.36 (1.71)***	n.s.	n.s.
2	**E^E**	**W**	1.40 (1.53)***	0.94 (1.69) vs. 1.85 (1.19)***	n.s.	n.s.
3	**X^E**	**W**	1.14 (1.85)***	0.78 (1.99) vs. 1.51 (1.64)**	n.s.	n.s.
4	**E^E**	**B**	1.96 (1.33)***	1.76 (1.53) vs. 2.17 (1.05)*	n.s.	n.s.
5	**X**	**B^E**	2.21 (1.15)***	n.s.	n.s.	2.09 (1.22) vs. 2.34 (1.06)*
6	**X**	**E^E**	0.44 (1.73)***	n.s.	n.s.	n.s.

Notes. Superscript E denotes the option that was better for the East zone population, i.e., the self-interested participants; *** < .001, ** < .01, * < .05, n.s.: not significant.

4.4.2. Summary regressions

So far, our analysis has hinted that efficiency, equity and self-interest all play a role in determining the ratings of, and preferences between, different policy options. However, until now we have been unable to quantify the effects of different motivations, particularly that of self-interest. Nor have we investigated in depth the magnitude of the differences between the gastroenteritis and fatality subsamples, or that of answering the principle statements task before or after the main preference questions. To address all of these issues, we next present a series of logistic regression models.

In the five logistic regression models presented in Table 17, the dependent variable is choosing the most efficient option in the pair of policy options. We present odds ratios, which can be interpreted as the relative likelihood that the efficient option is chosen, given that independent variable. Because every participant answered 6 questions, standard errors were clustered at the participant level. Model (1) includes only demographic variables, and none significantly alters the odds of choosing the most efficient option. Model (2) adds the main effects. Model (3) adds the two-way interactions and model (4) adds a three-way interaction. Last, model (5) accounts for the effect of participants' preference for efficiency, as stated in the principles task, and its interaction with their preference for efficiency and equity inferred from their choices in the preference task.

We included a dummy variable capturing the effect of being in the fatality scenario compared to the gastroenteritis scenario. Perhaps surprisingly, we found no significance, either for the main effect or for any interactions. This implies that individuals make the same trade-offs between equity, efficiency and self-interest regardless of whether they are answering questions about gastroenteritis or about fatality. This may indicate that, within the domain of health and physical risk, preferences are not stake-sensitive. Alternatively, it might indicate that the gastroenteritis condition was severe enough that both treatments were perceived to have high stakes.

We also included a dummy variable controlling for the order in which participants answered the preference questions compared to the statements questions. Again, there were no significant main effects or interactions. This is reassuring, implying that preferences and statements of principle are reasonably well aligned, and that response to choice questions in this context is not altered as a result of stating moral principles. Last, we included a dummy to account for whether the participant was in the self-interest or the impartial condition (see discussion of this effect below).

Turning to the characteristics of the policies (see a summary of these values for each policy in Appendix C.2), *efficiency* is defined as the number of cases prevented by Policy Y (the most efficient option in any given comparison) minus the number of cases prevented by Policy Z (the other option in the comparison). In all cases, *efficiency* is a positive and significant predictor. For each additional incidence of harm prevented by Policy Y compared to Policy Z, respondents are between 11.5% and 18.5% more likely to choose the efficient option. This effect is robust across model specifications.

Equity is defined as the difference between the numbers of cases prevented in the West and in the East by policy Y, less the difference between the numbers of cases prevented in the West and in the East by policy Z. Thus, higher e*quity* indicates that the efficient option is also the most equitable. *Equity* is another robust, positive indicator for choice, so that when the efficient policy is also more equitable than the other policy, it is more likely to be chosen. Specifically, for a one-person difference in *equity* between the policies, the efficient option is between 8% and 11% more likely to be chosen. A Wald test indicated that the coefficients of the variables capturing efficiency and equity concerns were significantly different (p-value < .05).

Half the participants were told they lived in the East zone, and the main effect of this is captured by the dummy *self-interested*. We refer to these participants as 'self-interested' because they had personal reasons to prefer options with the largest reductions in risk in that zone[26]. We found that they were significantly less likely to choose the most efficient option than impartial participants; this implies that they took into account considerations other than efficiency. This interpretation is supported by the coefficients relating to *benefit in the East zone*. *Benefit in the East zone* is defined as the number of lives saved in the East zone by policy Y, minus the number of lives saved in the East zone by policy Z. There is no significant main effect of this predictor at the 5% level. However, as expected, the odds ratio of the interaction between the *self-interested* dummy variable and the *benefit to the East zone* is greater than 1 and significant: participants prefer policies that benefit the East zone, but only if they are in the self-interested condition. Specifically, whilst impartial participants are marginally less likely to select the efficient option if it benefits the East zone[27], self-interested participants were around 9% *more* likely to opt for it. A Wald test indicated that the effect of the interaction of self-interest and efficiency in the East was significantly smaller than that of overall efficiency; this suggests that self-interest does not override affected participants' concern for efficiency.

The variable *favour efficiency* corresponds to the efficiency preference score, defined based on the stated principles task. The score captures the extent to which participants favoured efficiency over the concerns it was traded-off against (i.e., equity, self-interest and helping those most at risk). Model (5) suggests that for every additional principle efficiency was favoured over,

[26] If self-interest was the only motivation, the preference ordering would be E>B>X>W, since these reduce the East zone risk by 10, 8, 3 and 0 in 100,000 respectively. In contrast if Efficiency is the only motivator the ranking would be W>B>E>X.

[27] At the 10% level, people are around 4% less likely to choose the efficient option if it saves an additional life in the East zone. This is likely due to the West zone being the one where people are most at risk to begin with.

the odds of choosing policy Y over Z increased by 58%. This corroborates the fact that the preferences stated in the principles trade-offs tasks are strongly aligned with those that can be inferred from choices in the preference task.

The odds ratios of the interactions between the *favour efficiency* variable and the main effects *efficiency* and *equity* can be regarded as multiplier effects. At a 5% significance level, every additional concern efficiency was favoured over in the statements task did not have multiplier effect on the preference for efficiency[28]. In the case of *equity*, the higher the efficiency preference score, the more the odds of choosing the most equitable option were lowered. One of the principles that efficiency was traded-off against was equity, hence the higher the number of principles efficiency was favoured over, the more likely it was that they preferred it against equity. This made choosing the most efficient policy 1.3% less likely, plausibly in favour of the more equitable option.

[28] As intuition would suggest, this effect was positive, but it was only significant at a 10% significance level.

Table 17. Odds Ratios (robust standard errors, in 322 clusters at the participant level)

Choice of the efficient option	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
Main effects of choice scenario					
Efficiency *(Difference in total efficiency between Y and Z)*		1.151***	1.185***	1.185***	1.115***
		(0.013)	(0.029)	(0.029)	(0.030)
Equity *(Difference in equity between Y and Z)*		1.084***	1.109***	1.109***	1.114***
		(0.008)	(0.019)	(0.019)	(0.014)
Benefit East *(Difference in benefit to the East zone between Y and Z)*		1.004	0.962	0.961	
		(0.0102)	(0.021)	(0.023)	
Main effects of treatment					
Self-interested *(Dummy: assigned to living in the East zone treatment = 1)*		0.502***	0.569*	0.569*	
		(0.085)	(0.146)	(0.146)	
Fatality scenario *(Dummy: assigned to fatality scenario treatment = 1)*		1.113	1.077	1.076	
		(0.186)	(0.252)	(0.252)	
Statements first *(Dummy: assigned to answer Stated Principles task first treatment = 1)*		1.254	1.259	1.259	
		(0.211)	(0.297)	(0.297)	
Interactions					
Efficiency*Self-interested			0.957	0.957	
			(0.023)	(0.023)	
Efficiency*Fatality scenario			1.001	1.001	
			(0.024)	(0.024)	
Efficiency*Statements first			0.990	0.990	
			(0.024)	(0.024)	
Equity*Self-interested			0.969	0.969	
			(0.016)	(0.016)	
Equity*Fatality scenario			0.996	0.996	
			(0.015)	(0.015)	
Equity*Statements first			0.999	0.999	
			(0.015)	(0.015)	
Benefit East*Self-interested			1.093***	1.094**	
			(0.022)	(0.030)	
Benefit East*Fatality scenario			1.008	1.009	
			(0.020)	(0.026)	
Benefit East*Statements first			0.973	0.973	
			(0.019)	(0.019)	
Benefit East *Self-interested* Fatality scenario				0.998	
				(0.033)	
Principles					
Favour Efficiency					1.580***
					(0.192)
Efficiency*Favour Efficiency					1.029
					(0.016)
Equity*Favour Efficiency					0.987*
					(0.007)
Demographics					
Age (years)	1.010	1.014	1.014	1.014	
	(0.007)	(0.009)	(0.009)	(0.009)	
Female (dummy, 1=yes)	0.889	0.859	0.857	0.857	
	(0.129)	(0.144)	(0.146)	(0.146)	
Household size	0.964	0.966	0.965	0.965	
	(0.046)	(0.054)	(0.055)	(0.055)	
Constant	2.182*	1.807	1.799	1.799	
	(0.728)	(0.771)	(0.855)	(0.856)	
Observations	1,920	1,920	1,920	1,920	

Notes. *** < .001, ** < .01, * < .05, n.s.: not significant.

4.5. Discussion

When considering different policy options, decision makers must trade different principles against one another. These principles include efficiency and equity. To respect the preferences of the affected population, it is important to understand the extent to which individuals would be willing to trade efficiency for equity. In addition to that, it is also important to understand whether these trade-offs are robust to the influence of self-interest. We presented a between-subjects experiment designed to test how individuals trade off these competing principles and to understand how self-interest modifies these trade-offs.

On balance, we found evidence that efficiency is the most important factor. Policies should be implemented if they are the most effective, saving the most lives or preventing the most harm. However, equity was also an important concern and participants in our sample were almost as highly influenced by equity as by efficiency. In direct comparisons of principles, efficiency was slightly preferred to equity, but in the direct choice between policy W (the most efficient) and policy B (the most equitable) there was a very even split in preference between the two options. Overall, this suggests that whilst efficiency matters, equity is also sufficiently important to warrant attention when choosing between different policy options.

This result holds even when participants are included in the scenarios and stand to benefit from different policies. We observed that participants in the self-interest condition favoured those policies that would benefit them particularly. However, the effect of self-interest was significantly smaller than that of the efficiency and not differentiable from that of the equity concerns. These results, quantifying the difference between preferences of 'affected' and 'impartial' respondents, contribute to the debate on whose preferences should be taken into account to inform policy (see, for instance, Dolan (1999)). They provide empirical evidence of the extent to which these differ, and hence what the trade-offs will be when we try to satisfy either of these groups. More generally, these results provide further evidence that metrics that capture private interest only do not provide an accurate representation of people's preferences.

Aside from self-interest, we also manipulated whether the risks that respondents were trading off were fatal or not (i.e., risk of gastroenteritis). We did not find significant differences in the preferences reported by participants in the two conditions. In the health context, differently to laboratory settings (Camerer & Hogarth, 1999), stakes do not seem to matter.

We also explored the relationship between individuals' directly elicited moral principles and their choices in a preference task that involved implicitly trading off these same concerns.

Respondents' answers to the principle statement trade-offs revealed a transitive ordering of the principles we focused on (efficiency $\succ$ equity $\succ$ self-interest). This ordering was consistent with the one underlying choices in the preference task, and also aligned with the one suggested by the ratings. These findings are encouraging – they are a first indication that moral principles could be explored as a substitute for, or complement to, stated preferences elicited through choice. This unlocks the possibility of answering policy questions that are difficult to approach empirically with choice-based methods, such as whether and how to incorporate altruistic concerns into policy valuations. A rich theoretical discussion about these issues can be found in Jones-Lee (1991; 1992) and Johansson (1992; 1994).

As discussed, our study answers policy-relevant questions about the role of self-interest in influencing the trade-off between efficiency and equity in policy applications; and about the use of principles as a policy tool. Our results also raise thought-provoking questions about the mechanisms at play when individuals make complex moral trade-offs. Take respondents in the impartial treatment: these respondents' choices must be driven by their preferences for the *distribution* of risk reductions across the population. That is, they cannot be influenced by self-interest. In contrast, respondents in the self-interest condition cannot be revealing purely distributional concerns – they also have to take account of their own outcomes. This changes the interpretation of expressed preferences for equitable or efficient outcomes that do not favour the area where they (hypothetically) lived. The preferences of the participants in the self-interest condition must also embody altruistic concern for others, since the respondent is showing a willingness to forgo benefit to themselves in order to benefit others in the scenario. Our results suggest that altruism does characterise individuals' preferences for policy outcomes, since they are indeed willing to forgo risk reduction to achieve greater efficiency and equity overall. However, there is opportunity for future investigations to extend this empirical work, using the experimental paradigm we have developed to further our understanding of this rich and complex empirical question.

Finally, the answers in this study, as those in most studies involving health outcomes, are hypothetical. This might raise doubts about their validity, as aside from respondents' preferences over the distribution of the risk reductions, respondent's wish to act in a socially desirable manner may have influenced answers too. In this study, participants were asked to indicate which policy options they would prefer the government to choose. Hence, it was up to respondents to decide whether this corresponded to their *true preferences* or what they considered is desirable for society (in cases when they differed). Based on this, one could

therefore argue that socially desirable answers may be as valid a basis for public policy as preferences that have been elicited under incentive-compatible conditions are.

Chapter 5. Conclusion

In this thesis, we set out to explore some of the issues that emerge when we estimate the societal value of life and health. In particular, those that may be observed due to the lack of descriptive validity of economic theory, which underlies the elicitation methods and the value metrics rely on. First, expected utility theory may not provide a valid description of people's behaviour under risk. Second, the account that people exclusively care about maximising their own utility, disregarding that of others, may not be accurate. Furthering our understanding of these issues is crucial to ensure that public budgets can be allocated to maximise the population's welfare. In what follows, the most important findings from each chapter are highlighted.

In Chapter 2, we studied the disparities between the utility estimates elicited directly and indirectly using the standard gamble method. They may be caused by the failure of the reduction of compound lotteries axiom, but they also may be (at least partially) caused by the context in which the estimates are elicited. We moved away from the affective nature of health and incentivised answers, and found that when choice complexity was the lowest the disparities persisted. This suggests that context may play a role, but it is not sufficient to explain the disparities. As other studies before ours (e.g., Oliver, 2003; Abdellaoui, Klibanoff, & Placido, 2015; Harrison, Martin-Correa, & Swarthout, 2015), we provide evidence of the failure of the reduction of compound lotteries. We are the first to do this eliciting probability equivalents in an incentivised study outside of the monetary domain. These findings do not provide direct evidence against the use of the standard gamble method, as they were not obtained in the context of health. However, they do call for caution when using this method, and for future research in this area. A natural extension of this study would be to use the same experimental paradigm to test our hypotheses in a setup that progressively converged to that of health and life valuation studies.

The second set of results from this study concerns choice complexity, which we conceptualised as the difficulty involved in comparing outcomes to each other when a lot of their attributes are not shared. According to expected utility theory, the probability and subjective value of the outcomes are the only two features of the choice that should enter decisions. We kept these two features constant across questions and found that the complexity of the choice (which is lower when the outcomes are money, and higher when they are consumer goods) did influence choice behaviour. In particular, participants displayed higher risk aversion for monetary payoffs

than for consumer goods, despite each participant had indicated that the two were equivalent in value. This evidence speaks against the descriptive validity of expected utility theory.

In Chapter 3, in order to find out more about the role of choice complexity, we tested whether the common ratio effect would persist with consumer goods. This effect is a violation of expected utility theory independence axiom, widely replicated in the laboratory with monetary outcomes. This experimental paradigm may fall short when looking for a broader descriptive account of behaviour. People's everyday decisions involve complex, multi-faceted consequences (rather than simpler, unidimensional ones as money). Therefore, models built on evidence solely from choices involving monetary amounts may fail to capture the nuances of actual decision making.

As many before us had done, we replicated the common ratio effect with money. Contrary to what the only two existing studies testing this effect in a non-monetary domain had found (Kahneman & Tversky, 1979; Keller, 1985b), we did not find the effect with objects. The choice patterns we observed with consumer goods could be regarded as evidence of the validity of expected utility theory. However, this theory predicts that as long as the subjective value of the outcomes (and the probabilities associated to them, possibly weighted) are the same, choice patterns in decisions involving money and equivalent objects should be the same, as that is not what we found. Instead, the greater consistency found with objects may be caused by participants resorting to heuristics (such as 'always pick the prospect that offers the best payoff') when decisions are too difficult to make because they are too complex. These findings are in line with those in the previous chapter. These results support the idea that preferences and attitudes may be domain-specific, which is already acknowledged in the case of risk attitudes (Weber, Blais, & Betz, 2002). Also, our results question the use of money to proxy other goods in the lab (as having the same subjective value will not warrant equivalent decisions); especially in instances where the goal is to develop and/or to test models of decision making that we aim to generalise to goods beyond money.

In the remainder of the thesis, we focused on the study of people's preferences. In Chapter 4, we aimed to provide an empirical account of the preferences that ought to be captured by value metrics such as the value of statistical life. This is the only study to date to look at how self-interest affects the trade-off between efficiency and equity in the context of physical and fatality risk reductions. We found that *efficiency* (maximising the expected number of fatalities or cases of illness prevented) matters the most, with (gains) *equity* (evenly distributing risk reductions for different parts of the population) following closely. For every case of illness or fatality a

(hypothetical) policy prevented overall, participants were around 18% more likely to choose it; whereas when the additional case achieved a more equitable distribution, this percentage dropped to 11%. With regards to *self-interest*, having one's own risk reduction at stake did not alter one's concern for overall efficiency or equity, but participants whose risk reduction was at stake did give more weight to efficiency in their area. The magnitude of this effect was smaller than that of the preference for overall efficiency, and so the general preference of efficiency over equity prevailed for all participants. Policy makers should take these preferences into consideration when making allocative decisions or designing health interventions. These results also contribute to the debate of whose preferences should be considered in order to inform policy: those of the affected population, or the general population (e.g., Dolan, 1999). Our findings suggest that in terms of the trade-offs between efficiency and equity concerns, the welfare difference that could be achieved by eliciting either group's preference would not differ substantially.

An interesting finding was the null effect of risk severity in the trade-offs participants made. One of our experimental conditions involved the risk of gastroenteritis, whereas the other condition involved the risk of death. We did not find significant differences in the preferences reported by participants in the two conditions. This evidence indicates that in the health context, differently to context-free laboratory settings (Camerer & Hogarth, 1999), stakes do not appear to matter.

In the same study, we also explored the relationship between individuals' elicited moral principles and their choices. We did this in a preference task that involved implicitly trading off these same concerns. We elicited trade-offs between different principles and found that, in the aggregate, these trade-offs gave rise to a transitive ordering of the principles that was consistent to that revealed by respondents' choices. This might be the first step taken in a promising avenue of valuation research, where statements of principle could also be used to elicit societal preferences.

This research contributes to the literature on health and life valuation, decision making under risk, and social preferences. It answered some questions, and it raised some new queries. With regards to the disparities within the standard gamble method, there are still some underexplored factors that could have an effect on them. With our experimental paradigm, we could study preference imprecision. Using participants' answers about how sure they were about their answers, aided by a behavioural, rather than not self-reported, measure such as response time, we could gain valuable insights about the variability of people's answers even when context has been as simplified as possible. With regards to choice complexity, the current

studies could be considered exploratory initial steps. The current literature does not acknowledge the full extent of its influence on decision making, but our results suggest that this factor does have an effect. Future work should formalise the concept so that its influence can be tested using appropriate economic techniques. This strand of work could benefit from an improved item pool to test the hypotheses (for example, outcomes that differ in a very specific number of attributes but are exactly the same with regards to their other features).

Finally, when studying social preferences, it may be worth presenting a broader range of scenarios to participants. It is plausible that some respondents may have favoured a principle over the other, but not on the terms we presented them (i.e., with the trade-offs involving the exact number of people in our scenarios). The important policy implications of this sort of studies call for thorough robustness checks of the results before they can be implemented. A finding that should be probed further in future research is the main effect of efficiency gains in the self-interested zone. Although this effect was only significant at a 10% level, its direction makes it interesting. The fact that respondents were less likely to choose a policy when it offered a higher efficiency gain in the East zone may mean that they were concerned about reducing the risk for those with the higher baseline risk instead. In our study, the concern for those most at risk could not be disentangled from a general preference for efficiency, as the largest gain would always be accrued by those that were initially at a higher risk. However, future work involving a formal test of this difference would be a valuable contribution to the literature about the nature of people's concern for others.

Bibliography

Abbink, K., & Hennig-Schmidt, H. (2006). Neutral versus loaded instructions in a bribery experiment. *Experimental Economics*, 9(2), 103-121.

Abdellaoui, M., Klibanoff, P., & Placido, L. (2015). Experiments on compound risk in relation to simple risk and to ambiguity. *Management Science*, *61*(6), 1306-1322.

Abellán-Perpiñán, J. M., Martínez-Pérez, J. E., Méndez-Martínez, I., Pinto-Prades, J. L., & Sánchez- Martínez, F. I. (2011). El valor monetario de una vida estadística en España. Retrieved from: http://195.64.186.10/profesionales/saludPublica/prevPromocion/Lesiones/JornadaDecenioAcc ionSeguridadVial/docs/InformeVVEJorgeMartinez.pdf

Allais, M. (1953). 'Le comportement de l'homme rationnel devant le risque, critique des postulats et axiomes de l'école américaine.' Econometrica, vol. 21, pp. 503-46.

Alm, J., McClelland, G. H., & Schulze, W. D. (1992). Why do people pay taxes? *Journal of Public Economics*, 48(1), 21-38.

Arentze, T., Borgers, A., Timmermans, H., & DelMistro, R. (2003). Transport stated choice responses: effects of task complexity, presentation format and literacy. *Transportation Research Part E: Logistics and Transportation Review*, 39(3), 229-244.

Arrow, K. J., Cropper, M. L., Eads, G. C., Hahn, R. W., Lave, L. B., Noll, R. G., Portney, P. R., Russell, M., Schmalensee, R., Smith, V. K., & Stavins, R. N. (1996). Is there a role for benefit-cost analysis in environmental, health, and safety regulation? *Science*, 272(5259), 221-222.

Ball, C. (1997). A comparison of single-step and multiple-step transition analyses of multiattribute decision strategies. *Organizational Behavior and Human Decision Processes*, 69(3), 195-204.

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance*, *9*(3), 396-406.

Beattie, J., Covey, J., Dolan, P., Hopkins, L., Jones-Lee, M. W., Loomes, G., Pidgeon, N., Robinson, A., & Spencer, A. (1998). On the contingent valuation of safety and the safety of contingent valuation. Part 1: Caveat Investigator. *Journal of Risk and Uncertainty*, 17, 5-26.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Systems Research and Behavioral Science*, 9(3), 226-232.

Bennett, K.J., & Torrance, G.W. (1996). Measuring health state preferences and utilities: rating scale, time trade-off and standard gamble techniques, In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials*, 2nd edition (pp. 253-265). Philadelphia, PA: Lippincott-Raven.

Bergstrom, T. C. (1999). Systems of benevolent utility functions. *Journal of Public Economic Theory*, 1(1), 71-100.

Bernasconi, M., & Bernhofer, J., (2017). Catch Me If You Can: Testing the reduction of compound lotteries axiom in a tax compliance experiment. Working Paper, Department of Economics, Ca'Foscari University of Venice.

Bernasconi, M., & Loomes, G. (1992). Failures of the reduction principle in an Ellsberg-type problem. *Theory and Decision*, 32(1), 77-100.

Bleichrodt, H. (2001). Probability weighting in choice under risk: an empirical test. *Journal of Risk and Uncertainty*, 23(2), 185-198.

Bleichrodt, H., Doctor, J., & Stolk, E. (2005). A nonparametric elicitation of the equity-efficiency trade-off in cost-utility analysis. *Journal of Health Economics*, 24(4), 655-678.

Bleichrodt, H., Pinto-Prades, J. L. & Abellán-Perpiñán, J. M., (2003). A consistency test of the time trade-off. *Journal of Health Economics,* 22**,** 1037-1052.

Blomquist, G. C. (2004). Self-protection and averting behavior, values of statistical lives, and benefit cost analysis of environmental policy. *Review of Economics of the Household*, 2(1), 89-110.

Bobinac, A., van Exel, N. J. A., Rutten, F. F., & Brouwer, W. B. (2012). Inquiry into the relationship between equity weights and the value of the QALY. *Value in Health*, 15(8), 1119-1126.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 166-193.

Braga, J., & Starmer, C. (2005). Preference anomalies, preference elicitation and the discovered preference hypothesis. Environmental and Resource Economics, 32(1), 55-89.

Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*, 21(2), 271-292.

Brooks, R. (1996) EuroQol: The current state of play. *Health Policy*. 37:53–72.

Budescu, D. V., & Fischer, I. (2001). The same but different: an empirical investigation of the reducibility principle. *Journal of Behavioral Decision Making*, 14(3), 187-206.

Butler, D. J., & Loomes, G. C. (2007). Imprecision as an account of the preference reversal phenomenon. *American Economic Review*, 97(1), 277-297.

Camerer, C. (1995). Individual decision-making, In Kagel, J., & Roth, A. (Eds.) *Handbook of Experimental Economics* (pp. 587-703). Princeton, NJ: Princeton University Press.

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences*, 7(5), 225-231.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19(1-3), 7-42.

Camerer, C., & Thaler, R. H. (1995). Anomalies: Ultimatums, dictators and manners. *The Journal of Economic Perspectives*, 9(2), 209-219.

Carthy, T., Chilton, S., Covey, J., Hopkins, L., Jones-Lee, M. W., Loomes, G., Pidgeon, N., Robinson, A., & Spencer, A. (1999). On the contingent valuation of safety and the safety of contingent valuation: Part 2-The CV/SG" chained" approach. *Journal of Risk and Uncertainty*, 17(3), 187-214.

Caussade, S., de Dios Ortúzar, J., Rizzi, L. I., & Hensher, D. A. (2005). Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation Research Part B: Methodological*, 39(7), 621-640.

Chilton, S., & Spencer, A. (2001). Empirical evidence of inconsistency in standard gamble choices under direct and indirect elicitation methods. *Revue Suisse d'Economie Politique et de Statistique*, 137(1), 65-86.

Cropper, M., Hammitt, J. K., & Robinson, L. A. (2011). Valuing mortality risk reductions: progress and challenges. *The Annual Review of Resource Economics*, 3, 313-336.

Danthurebandara, V. M., Yu, J., & Vandebroek, M. (2011). Effect of choice complexity on design efficiency in conjoint choice experiments. *Journal of Statistical Planning and Inference*, 141(7), 2276-2286.

Dave, C., Eckel, C. C., Johnson, C. A., & Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, 41(3), 219-243.

Davis, R. K. (1963). The value of outdoor recreation: an economic study of Maine woods. *Unpublished Ph. D. dissertation, Harvard University*.

De Blaeij, A., Florax, R. J., Rietveld, P., & Verhoef, E. (2003). The value of statistical life in road safety: a meta-analysis. *Accident Analysis & Prevention*, 35(6), 973-986.

Deck, C., & Schlesinger, H. (2016). On the robustness of higher order risk preferences. Chapman University, Economic Science Institute working papers repository. Retrieved from: https://www.chapman.edu/research/institutes-and-centers/economic-science-institute/_files/WorkingPapers/On-the-Robustness-of-Higher-Order-Risk-Attitudes.pdf

Dellaert, B. G., Brazell, J. D., & Louviere, J. J. (1999). The effect of attribute variation on consumer choice consistency. *Marketing Letters*, 10(2), 139-147.

Dellaert, B. G., Donkers, B., & Soest, A. V. (2012). Complexity effects in choice experiment–based models. *Journal of Marketing Research*, 49(3), 424-434.

Department for Transport (2018). Transport Analysis Guidance (TAG) Unit A1.1. London: Transport Appraisal and Strategic Modelling (TASM) Division, Department for Transport. Retrieved from:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/712699/tag-unit-a1.1-cost-benefit-analysis-may-18.pdf

Desaigues, B., & Rabl, A. (1995). Reference values for human life: an econometric analysis of a contingent valuation in France. In *Contingent valuation, transport safety and the value of life* (pp. 85-112). Springer, Dordrecht.

DeShazo, J. R., & Fermo, G. (2002). Designing choice sets for stated preference methods: the effects of complexity on choice consistency. *Journal of Environmental Economics and Management*, 44(1), 123-143.

Dolan, P. (1998). The measurement of individual utility and social welfare. *Journal of health economics*, 17(1), 39-52.

Dolan, P. (1999). Whose preferences count? *Medical Decision Making*, 19(4), 482-486.

Dolan, P. (2011). *Using happiness to value health*. London, UK: Office of Health Economics. Retrieved from: https://www.ohe.org/system/files/private/publications/362%20-%20Using%20Happiness%20Dolan%20Nov%202011.pdf

Dolan, P., Gudex, C., Kind, P., & Williams, A. (1996). Valuing health states: a comparison of methods. *Journal of Health Economics*, *15*(2), 209-231.

Dolan, P., Layard, R., & Metcalfe, R. (2011). *Measuring subjective well-being for public policy*. Newport, UK: Office for National Statistics. Retrieved from: http://eprints.lse.ac.uk/35420/1/measuring-subjective-wellbeing-for-public-policy.pdf

Dolan, P., & White, M. P. (2007). How can measures of subjective well-being be used to inform public policy? *Perspectives on Psychological Science*, 2(1), 71-85.

Donaldson, C., Mason, H., & Shackley, P. (2012). Contingent valuation in health care, In Jones, A. (Ed.), *The Elgar companion to health economics*, 2nd edition (pp. 425-434). Cheltenham, UK: Edward Elgar.

Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., & Torrance, G. W. (2015). *Methods for the economic evaluation of health care programmes*. New York, NY: Oxford university press.

Ebert, S., & van de Kuilen, G. (2016). Measuring multivariate risk preferences. Working paper, retrieved from SRNN: https://ssrn.com/abstract=2637964

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 643-669.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, 14(4), 583-610.

Fasolo, B., Carmeci, F. A., & Misuraca, R. (2009). The effect of choice complexity on perception of time spent choosing: When choice takes longer but feels shorter. *Psychology & Marketing*, 26(3), 213-228.

Fay, M. P. (2011). Exact McNemar's Test and Matching Confidence Intervals. Retrieved from: http://www2.uaem.mx/r-mirror/web/packages/exact2x2/vignettes/exactMcNemar.pdf

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817-868.

Fehr, E., & Schmidt, K. M. (2001). *Theories of fairness and reciprocity – evidence and economic applications*. Working Paper. Institute for Empirical Research in Economics, University of Zurich. Retrieved from: https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/146616/eth-25578-01.pdf?sequence=1

Fischer, G. W., Carmon, Z., Ariely, D., & Zauberman, G. (1999). Goal-based construction of preferences: Task goals and the prominence effect. *Management Science*, 45(8), 1057-1075.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3), 347-369.

Fox, C., Erner, C. & Walters, D. (2015). Decision under risk: from the field to the laboratory and back. In Keren, G. and Wu, G. (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (pp. 43-88), John Wiley & Sons, Chichester, UK.

Fried, C. (1969). The value of life. *Harvard Law Review*, 82(7), 1415-1437.

Froberg, D. G., & Kane, R. L. (1989). Methodology for measuring health-state preferences—II: Scaling methods. *Journal of Clinical Epidemiology*, 42(5), 459-471.

Gafni, A. (1994). The standard gamble method: what is being measured and how it is interpreted. *Health Services Research*, *29*(2), 207.

Galizzi, M. M., & Navarro-Martínez, D. (2018). On the external validity of social preference games: a systematic lab-field study. Management Science. Available from https://doi.org/10.1287/mnsc.2017.2908.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, 103(4), 650.

Greifeneder, R., Scheibehenne, B., & Kleber, N. (2010). Less may be more when choosing is difficult: Choice complexity and too much choice. *Acta Psychologica*, 133(1), 45-50.

Güth, W., & Kocher, M. G. (2014). More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior & Organization*, 108, 396-409.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4), 367-388.

Gyrd-Hansen, D., Kjær, T., & Nielsen, J. S. (2012). Scope insensitivity in contingent valuation studies of health care services: should we ask twice?. *Health Economics*, *21*(2), 101-112.

Harrison, G. W., Martínez-Correa, J., & Swarthout, J. T. (2015). Reduction of compound lotteries with objective probabilities: Theory and evidence. *Journal of Economic Behavior & Organization*, *119*, 32-55.

Hajimoladarvish, N. (2018). How do people reduce compound lotteries? *Journal of Behavioral and Experimental Economics*, 75, 126-133.

Hausman, J. (2012). Contingent valuation: from dubious to hopeless. Journal of Economic Perspectives, 26(4), 43-56.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65-70.

Horowitz, J. K. (2006). The Becker-DeGroot-Marschak mechanism is not necessarily incentive compatible, even for non-random goods. *Economics Letters*, *93*(1), 6-11.

Jacoby, J., Speller, D. E., & Berning, C. K. (1974). Brand choice behavior as a function of information load: Replication and extension. *Journal of Consumer Research*, 1(1), 33-42.

Jacoby, J., Speller, D. E., & Kohn, C. A. (1974). Brand choice behavior as a function of information load. *Journal of Marketing Research*, 63-69.

Jamison, D. T., Mosley, W. H., Measham, A. R. & Bobadilla, J. L. (1993). *Disease control priorities in developing countries*. New York, NY: Oxford University Press.

Johannesson, M., & Gerdtham, U. G. (1996). A note on the estimation of the equity-efficiency trade-off for QALYs. *Journal of Health Economics*, 15(3), 359-368.

Johannesson, M., & Johansson, P. O. (1996). To be, or not to be, that is the question: an empirical study of the WTP for an increased life expectancy at an advanced age. *Journal of Risk and Uncertainty*, 13(2), 163-174.

Johannesson, M., & Johansson, P. O. (1997). Quality of life and the WTP for an increased life expectancy at an advanced age. *Journal of Public Economics*, 65(2), 219-228.

Johannesson, M., & Jönsson, B. (1991). Economic evaluation in health care: is there a role for cost-benefit analysis? *Health Policy*, 17(1), 1-23.

Johansson, P. O. (1992). Altruism in cost-benefit analysis. *Environmental and Resource Economics*, 2(6), 605-613.

Johansson, P. O. (1994). Altruism and the value of statistical life: empirical implications. *Journal of Health Economics*, 13(1), 111-118.

Jones-Lee, M. W. (1976). *The value of life: an economic analysis*. Chicago, IL: University of Chicago Press.

Jones-Lee, M. W. (1989). *The value of avoidance of non-fatal road injuries: an exploratory study*. Report to the Department for Transport. University of Newcastle Upon Tyne.

Jones-Lee, M.W. (1991). Altruism and the value of other people's safety. *Journal of Risk and Uncertainty,* 4(2), 213-219.

Jones-Lee, M. W. (1992). Paternalistic altruism and the value of statistical life. *The Economic Journal*, 102(410), 80-90.

Jones-Lee, M. W., Chilton, S., Metcalf, H., & Nielsen, J.S. (2015). Valuing gains in life expectancy: clarifying some ambiguities. *Journal of Risk and Uncertainty,* 51(1), 1-21.

Jones-Lee, M. W., Hammerton, M., & Philips, P. R. (1985). The value of safety: results of a national sample survey. *The Economic Journal*, 95(377), 49-72.

Jones-Lee, M. W., Loomes, G., & Philips, P. R. (1995). Valuing the prevention of non-fatal road injuries: Contingent valuation vs. standard gambles. *Oxford Economic Papers*, 676-695.

Kahneman, D., & Knetsch, J. L. (1992). Valuing public goods: the purchase of moral satisfaction. Journal of environmental economics and management, 22(1), 57-70.

Kahneman, D., Knetsch, J. L., & Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *The American Economic Review*, 728-741.

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

Kartman, B., Stålhammar, N. O., & Johannesson, M. (1996). Valuation of health changes with the contingent valuation method: a test of scope and question order effects. *Health Economics*, 5(6), 531-541.

Keller, L. R. (1985a). Testing of the 'reduction of compound alternatives' principle. *Omega*, 13(4), 349-358.

Keller, L. R. (1985b). The effects of problem representation on the sure-thing and substitution principles. *Management Science*, 31(6), 738-751.

Keller, K. L., & Staelin, R. (1987). Effects of quality and quantity of information on decision effectiveness. *Journal of Consumer Research*, 14(2), 200-213.

Knox, F. (1960). The doctrine of consumers' sovereignty. *Review of Social Economy*, 18(2), 138-149.

Laury, S. K., & Taylor, L. O. (2008). Altruism spillovers: Are behaviors in context-free experiments predictive of altruism toward a naturally occurring public good? *Journal of Economic Behavior & Organization*, 65(1), 9-29.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, 89(1), 46.

Light, D. W. (1992). Equity and efficiency in health care. *Social Science & Medicine*, 35(4), 465-469.

Lindholm, L., Rosen, M., & Emmelin, M. (1996). An epidemiological approach towards measuring the trade-off between equity and efficiency in health policy. *Health Policy*, 35(3), 205-216.

Lindman, H. R. (1971). Inconsistent preferences among gambles. *Journal of Experimental Psychology*, 89(2), 390.

List, J. A., & Cherry, T. L. (2008). Examining the role of fairness in high stakes allocation decisions. *Journal of Economic Behavior & Organization*, 65(1), 1-8.

Llewellyn-Thomas, H., Sutherland, H. J., Tibshirani, R., Ciampi, A., Till, J. E., & Boyd, N. F. (1982). The measurement of patients' values in medicine. *Medical Decision Making*, 2(4), 449-462.

Llewellyn-Thomas, H., Sutherland, H. J., Tibshirani, R., Ciampi, A., Till, J. E., & Boyd, N. F. (1984). Describing health states: methodologic issues in obtaining values for health states. *Medical Care*, 543-552.

Loomes, G., & Pogrebna, G. (2016). Do preference reversals disappear when we allow for probabilistic choice?. *Management Science*, 63(1), 166-184.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*(368), 805-824.

Loomes, G., & Sugden, R. (1998). Testing different stochastic specifications of risky choice. *Economica*, 65(260), 581-598.

Machina, M. J. (1982). "Expected Utility" Analysis without the Independence Axiom. *Econometrica*, 277-323.

Machina, M. J. (1987). Choice under uncertainty: Problems solved and unsolved. *Journal of Economic Perspectives*, 1(1), 121-154.

Mazzotta, M. J., & Opaluch, J. J. (1995). Decision making when choices are complex: a test of Heiner's hypothesis. *Land Economics*, 500-515.

McIntosh, E., Clarke, P. M., Frew, E., & Louviere, J. J. (2010). *Applied methods of cost-benefit analysis in health care*. Oxford, UK: Oxford University Press.

Miao, B., & Zhong, S., (2012). An experimental study of attitude towards compound lottery. Working paper, Department of Economics, National University of Singapore.

Mishan, E. J. (1971). Evaluation of life and limb: a theoretical approach. *Journal of Political Economy*, 79(4), 687-705.

Moffatt, P. G. (2016). *Experimetrics: Econometrics for experimental economics*. Croydon: Macmillan International Higher Education.

Murray, C. J. L. (1996). Rethinking DALYs, In Murray, C. J. L. & Lopez, A. D. (Eds.), *The Global Burden of Disease* (pp. 1-98). Cambridge, MA: Harvard University Press.

Mushkin, S. (1978). Cost of disease and illness in the United States in the year 2000. *Public Health Reports*, 93, 493.

National Institute for Health and Care Excellence (NICE) (2013). Guide to the methods of technology appraisal. London: UK National Health Service, National Institute for Health and Care Excellence.

Neumann, P. J., Sanders, G. D., Russell, L. B., Siegel, J. E., & Ganiats, T. G. (2016). *Cost-effectiveness in health and medicine*. New York, NY: Oxford University Press.

Nord, E. (1992). Methods for quality adjustment of life years. *Social Science & Medicine*, 34(5), 559-569.

Nord, E., Pinto, J. L., Richardson, J., Menzel, P., & Ubel, P. (1999). Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Economics*, 8(1), 25-39.

Oliver, A. (2003). The internal consistency of the standard gamble: tests after adjusting for prospect theory. *Journal of Health Economics*, 22(4), 659-674.

Olshavsky, R. W. (1979). Task complexity and contingent processing in decision making: A replication and extension. *Organizational Behavior and Human Performance*, 24(3), 300-316.

Patrick, D. L., Bush, J. W., & Chen, M. M. (1973). Methods for measuring levels of well-being for a health status index. *Health Services Research*, 8(3), 228.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance*, *16*(2), 366-387.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3), 534.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York, NY: Cambridge University Press.

Perkins, D. W., & Carey, M. P., (2017). *Cost-benefit analysis and financial regulator rulemaking*. Washington DC: Congressional Research Service. Retrieved from: https://fas.org/sgp/crs/misc/R44813.pdf

Pinto-Prades, J. L., Sánchez-Martínez, F. I., Abellán-Perpiñán, J. M., & Martínez-Pérez, J. E. (2018). Reducing preference reversals: The role of preference imprecision and nontransparent methods. *Health Economics*, doi:10.1002/hec.3772

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 1281-1302.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34-55.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review*, 1068-1095.

Rubinstein, A. (1988). Similarity and decision-making under risk (Is there a utility theory resolution to the Allais paradox?). *Journal of Economic Theory*, *46*(1), 145-153.

Rutten-van Mölken, M. P., Bakker, C. H., van Doorslaer, E. K., & Van Der Linden, S. (1995). Methodological issues of patient utility measurement: experience from two clinical trials. *Medical Care*, 922-937.

Savage, L. J. (1954). *The Foundations of Statistics*. New York, Wiley.

Schelling, T. C. (1968). The life you save may be your own, In Bhase, S. B. (Ed.), *Problems in Public Expenditure Analysis* (pp. 127-176). Washington, DC: Brookings Inst.

Schoemaker, P. J. (1982). The expected utility model: Its variants, purposes, evidence and limitations. *Journal of Economic Literature*, 20(2), 529-563.

Shiloh, S., Koren, S., & Zakay, D. (2001). Individual differences in compensatory decision-making style and need for closure as correlates of subjective decision complexity and difficulty. *Personality and Individual Differences*, 30(4), 699-710.

Slovic, P., & Lichtenstein, S. (1968). Relative importance of probabilities and payoffs in risk taking. *Journal of Experimental Psychology*, 78 (3 part 2), 1.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *The American Economic Review*, *72*(5), 923-955.

Smith, R. D. (2003). Construction of the contingent valuation market in health care: a critical assessment. *Health Economics*, 12(8), 609-628.

Spencer, A. (1998), Value elicitation techniques: methodological issues involved in two-stage procedures, D.Phil. thesis, University of York.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, 38(2), 332-382.

Stiggelbout, A. M., Kiebert, G. M., Kievit, J., Leer, J. W. H., Stoter, G., & De Haes, J. C. J. M. (1994). Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Medical Decision Making*, 14(1), 82-90.

Sugden, R. (2004). Alternatives to expected utility: Foundations. In *Handbook of Utility Theory* (pp. 685-755). Springer, Boston, MA.

Swait, J., & Adamowicz, W. (2001a). The influence of task complexity on consumer choice: a latent class model of decision strategy switching. *Journal of Consumer Research*, 28(1), 135-148.

Swait, J., & Adamowicz, W. (2001b). Choice environment, market complexity, and consumer behavior: a theoretical and empirical approach for incorporating decision complexity into models of consumer choice. *Organizational Behavior and Human Decision Processes*, 86(2), 141-167.

Thaler, R., & Rosen, S. (1976). The value of saving a life: evidence from the labor market In Terleckyj, N. E., (Ed.), *Household Production and Consumption* (pp. 265-300). New York, NY: Columbia University Press. Retrieved from: http://www.nber.org/chapters/c3964.pdf

Timmermans, D., & Vlek, C. (1992). Multi-attribute decision support and complexity: An evaluation and process analysis of aided versus unaided decision making. *Acta Psychologica*, 80(1-3), 49-65.

Torrance, G. W. (1986). Measurement of health state utilities for economic appraisal: a review. *Journal of Health Economics*, *5*(1), 1-30.

Torrance, G. W., Feeny, D. H., & Furlong, W. J. (2001). Visual Analog Scales: do they have a role in the measurement of preferences for health states? *Medical Decision Making*, 21, 329-334.

Torrance, G. W., Feeny, D. H., Furlong, W. J., Barr, R. D., Zhang, Y., & Wang, Q. (1996). Multiattribute utility function for a comprehensive health status classification system: Health Utilities Index Mark 2. *Medical Care*, *34*(7), 702-722.

Torrance, G. W., Thomas, W. H., & Sackett, D. L. (1972). A utility maximization model for evaluation of health care programs. *Health Services Research*, *7*(2), 118.

Treasury (2018) *The Green Book: Appraisal and Evaluation in Central Government*. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/685903/The_Green_Book.pdf

Tricomi, E., Rangel, A., Camerer, C. F., & O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature*, 463(7284), 1089-1091.

Tsuchiya, A., & Dolan, P. (2009). Equality of what in health? Distinguishing between outcome egalitarianism and gain egalitarianism. *Health Economics*, 18(2), 147-159.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297-323.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, 95(3), 371.

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management Science*, 39(10), 1179-1189.

Ubel, P. A., DeKay, M. L., Baron, J., & Asch, D. A. (1996a). Cost-effectiveness analysis in a setting of budget constraints—is it equitable?. *New England Journal of Medicine*, 334(18), 1174-1177.

Ubel, P. A., Loewenstein, G., Scanlon, D., & Kamlet, M. (1996b). Individual utilities are inconsistent with rationing choices: A partial explanation of why Oregon's cost-effectiveness list failed. *Medical Decision Making*, 16(2), 108-116.

Ubel, P. A., Baron, J., Nash, B., & Asch, D. A. (2000). Are preferences for equity over efficiency in health care allocation" all or nothing"?. *Medical Care*, 38(4), 366-373.

US Office of Management and Budget (OMB) (2003). *Circular A-4, Regulatory analysis*.

Viscusi, W. K. (2008). How to value a life. *Journal of Economics and Finance*, 32(4), 311-323.

Viscusi, W. K., & Aldy, J. E. (2003). The value of a statistical life: a critical review of market estimates throughout the world. *Journal of Risk and Uncertainty*, 27(1), 5-76.

Viscusi, W. K., Magat, W. A., & Huber, J. (1991). Communication of ambiguous risk information. *Theory and Decision*, *31*(2-3), 159-173.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behaviour*. Princeton, NJ: Princeton University Press.

Voors, M., Turley, T., Kontoleon, A., Bulte, E., & List, J. A. (2012). Exploring whether behavior in context-free experiments is predictive of behavior in the field: Evidence from lab and field experiments in rural Sierra Leone. *Economics Letters*, 114(3), 308-311.

Wagstaff, A. (1991). QALYs and the equity-efficiency trade-off. *Journal of Health Economics*, 10(1), 21-41.

Weber, E. U., Blais, A. R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15(4), 263-290.

Williams, A. (1974). The cost-benefit approach. *British Medical Bulletin*, 30(3), 252-256.

Appendix A – Chapter 2

A.1. Pilot Studies

The final design of the study was developed after pre-tests done in 4 pilot studies.

The first one was an online study (n=105) that consisted of the ranking and valuation tasks only. The goals of this study were two-fold. First, it was used to find out about the value people placed on 30 objects we had pre-selected, in order to select 10 out of them, which would be used in the actual study. We wanted the valuations of the final selection of objects to be spread out rather than clustered together, but the values of the six objects of interest (the three that had some comparable dimensions, and the three that did not) to be approximately at par. We also wanted to check the relationship between the monetary valuations and market price (see Table 18). We found a broad correlation between the two: cheaper items were usually in the bottom of respondents' rankings, and more expensive ones were at the top. However, their ranking within these two groups and their monetary valuations did not match market price. This indicated that valuations did not exclusively depend on it and suggested that participants were introspecting to come up with those values. The second goal of this study was to check that the instructions were clear, and the interface was easy to use, for the ranking and the valuation tasks. We found no problems with regards to this.

Table 18. Objects Price in Amazon (July 2015)

Alarm Clock	£4.60	Toaster	£18
Radio Alarm Clock	£19	Bottle	£10.60
Radio-speaker Alarm Clock	£20	Headphones	£33
Air Mattress	£33	Mouse	£2.59
Clicker	£8	Suitcase	£37

Pilot 2 was run in the lab (n=10) and featured the final 10 objects selection, and the ranking, valuation, and lotteries tasks. It was run to test the clarity of the instructions and the ease of use of the interface for the whole study. After participants had finished with these tasks on the computer, we held two focus groups and asked participants about the instructions and interface, and to walk us through the way they completed each of the tasks. Participants agreed that the tasks were clear, and the interface was easy to use. Some participants stated that in the ranking and valuation tasks, they could not avoid thinking about market price, whereas in the gambles, they only considered their preferences. This pointed out the need to modify the first part of the design to minimise the influence of market prices on people's monetary equivalents. In order to do that, we decided to introduce a pairwise choice task before the

ranking, where respondents would be making choices between every possible pair of objects, and the objects and some money amounts. This task should also make it easier for participants to introspect about their preferences, as instead of building the objects ranking from scratch, the computer would infer one from their answers to the binary choices, and they would just need to adjust it.

Pilot 3 was an online study (n=159) aimed to test whether having participants consider money amounts in the binary choice questions would induce them to pay additional attention to market price, and whether that would override their preferences as a criterion for choice. Participants went through the pairwise choice and ranking tasks. Half of them encountered objects only, and the other half made decisions involving objects and money amounts. We found no significant differences between the objects rankings of participants in the condition where they saw questions with objects only, compared to the one where they saw questions that also included money amounts. Hence, we concluded that seeing money amounts did not change people's stated preferences about their monetary values for the objects, compared to not seeing money amounts. This confirmed that further changes to the study design were not necessary.

Pilot 4 (n=21) was ran to ensure that there were not technical issues with the study before we recruited the whole sample of participants. Participants received a show-up fee as compensation for their time, but their answers were not incentivised. For that reason, the data collected in this pilot was not pooled with the data from the actual study.

A.2. Instructions and Interface

This appendix includes a sample of screenshots of the interface, and instructions for each task of the study.

Figure 19. Welcome

Welcome, and thank you for taking part in this study.

We hope to learn more about how people make decisions involving everyday objects. For instance, we will ask you to consider which of two items you would rather receive if you were given the choice.

Understanding how people make decisions like these may have many useful implications, so please ensure your answers are as accurate as possible.

The study has 9 blocks of questions and should take about 1 hour.

Before the study starts, you are going to see the objects that you will encounter during the session, together with two questions about each.

Answer the questions and take your time to familiarize yourself with each object and its features.

Press "Next" to proceed.

Figure 20. Familiarisation Question Example

Object Overview **Object 1**

Radio Alarm Clock

- FM/AM radio tuner
- Standard, dual alarm or projector model options and extendable snooze setting
- Daylight saving: Light sensor enables clock to give soft lights (can have this function off and press snooze/light button when needed)

Do you already own this object, or something similar?

○ Yes
◉ No

How much would you like to receive this object?

Not at all | ○ ○ ○ ○ ○ ◉ ○ | Very Much

Figure 21. Riskless Pairwise Choice Test Question Example

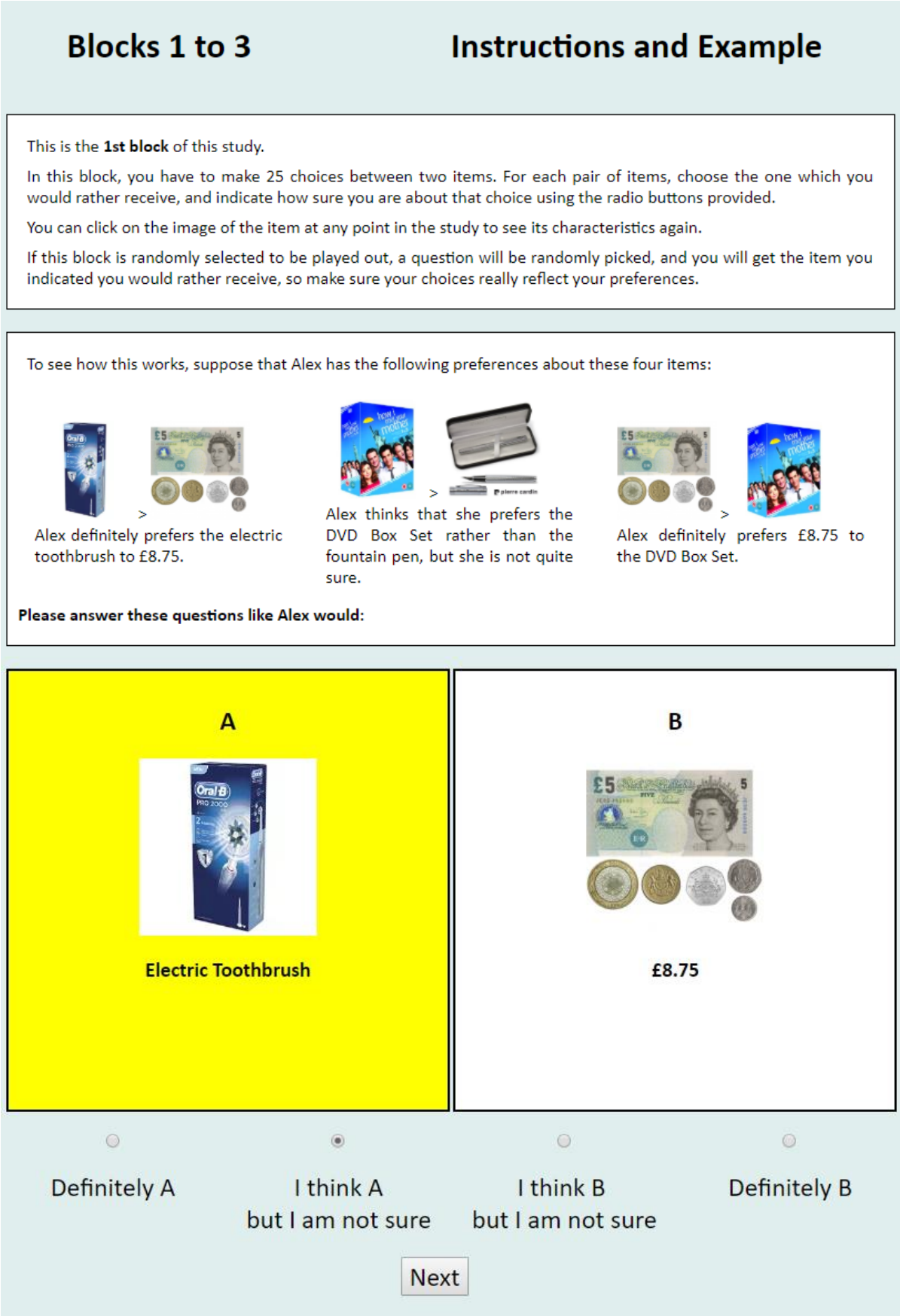
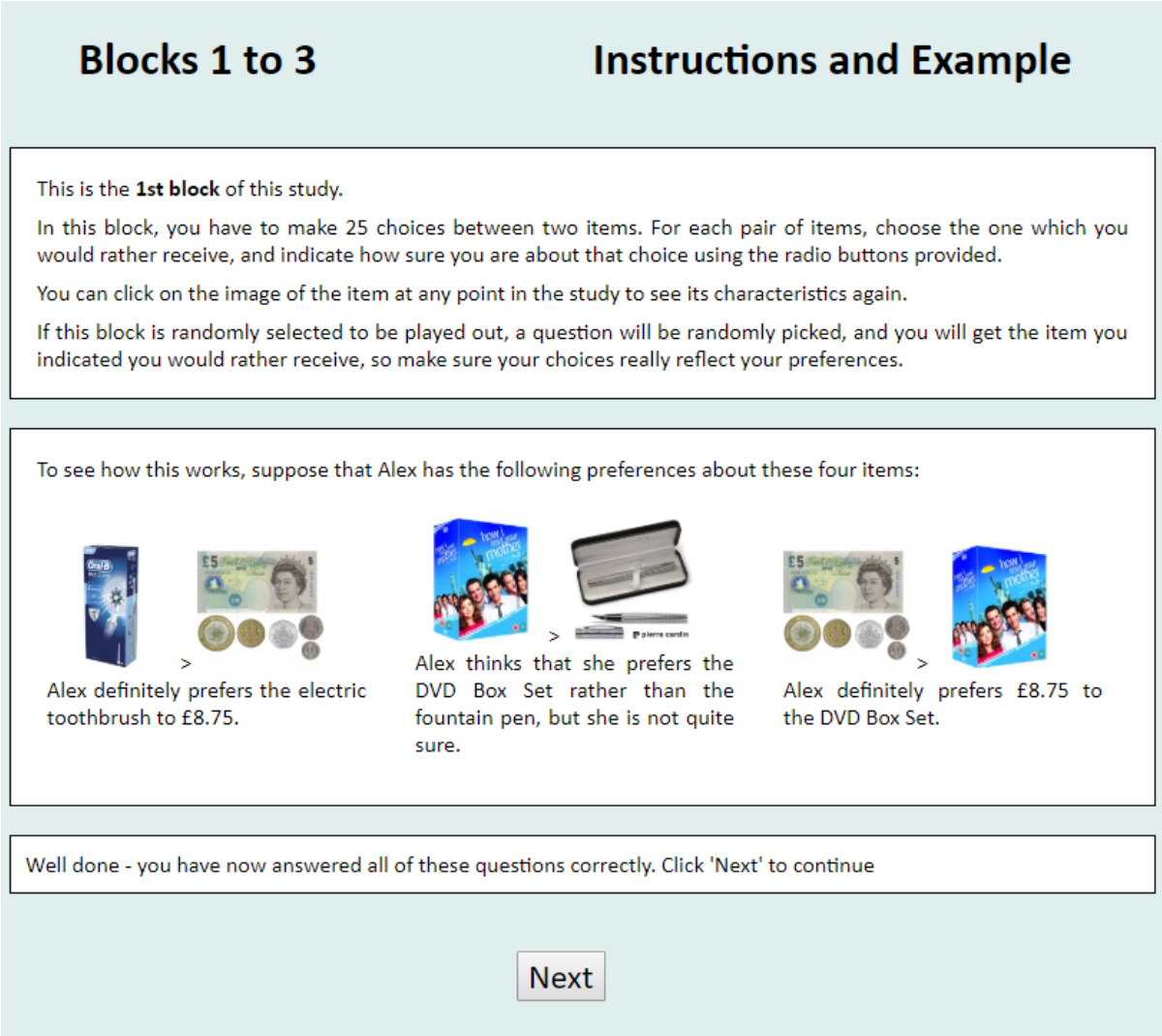


Figure 22. Riskless Pairwise Choice Instructions



Blocks 1 to 3 Instructions and Example

This is the **1st block** of this study.

In this block, you have to make 25 choices between two items. For each pair of items, choose the one which you would rather receive, and indicate how sure you are about that choice using the radio buttons provided.

You can click on the image of the item at any point in the study to see its characteristics again.

If this block is randomly selected to be played out, a question will be randomly picked, and you will get the item you indicated you would rather receive, so make sure your choices really reflect your preferences.

To see how this works, suppose that Alex has the following preferences about these four items:

Alex definitely prefers the electric toothbrush to £8.75.

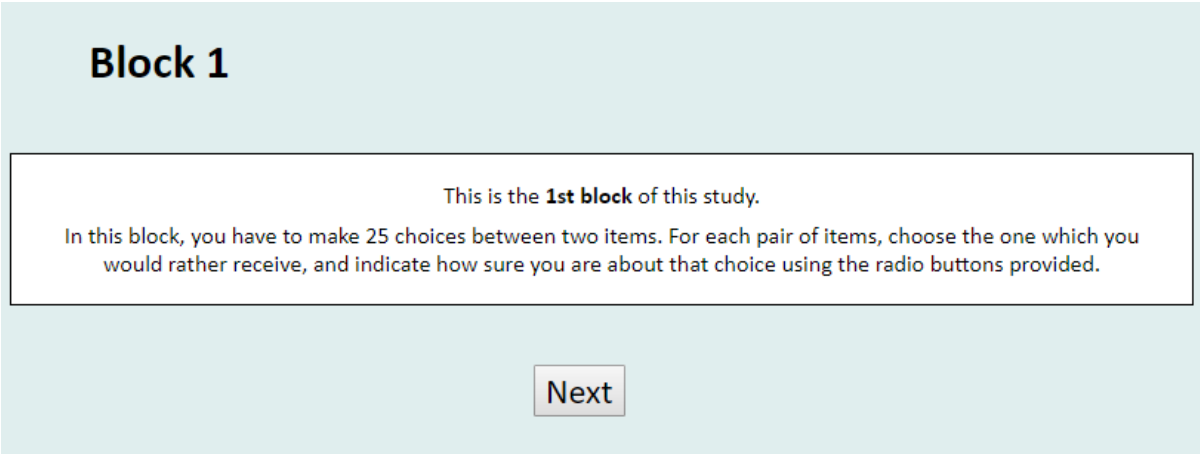
Alex thinks that she prefers the DVD Box Set rather than the fountain pen, but she is not quite sure.

Alex definitely prefers £8.75 to the DVD Box Set.

Well done - you have now answered all of these questions correctly. Click 'Next' to continue

Next

Figure 23. Introduction to Riskless Pairwise Choice Block



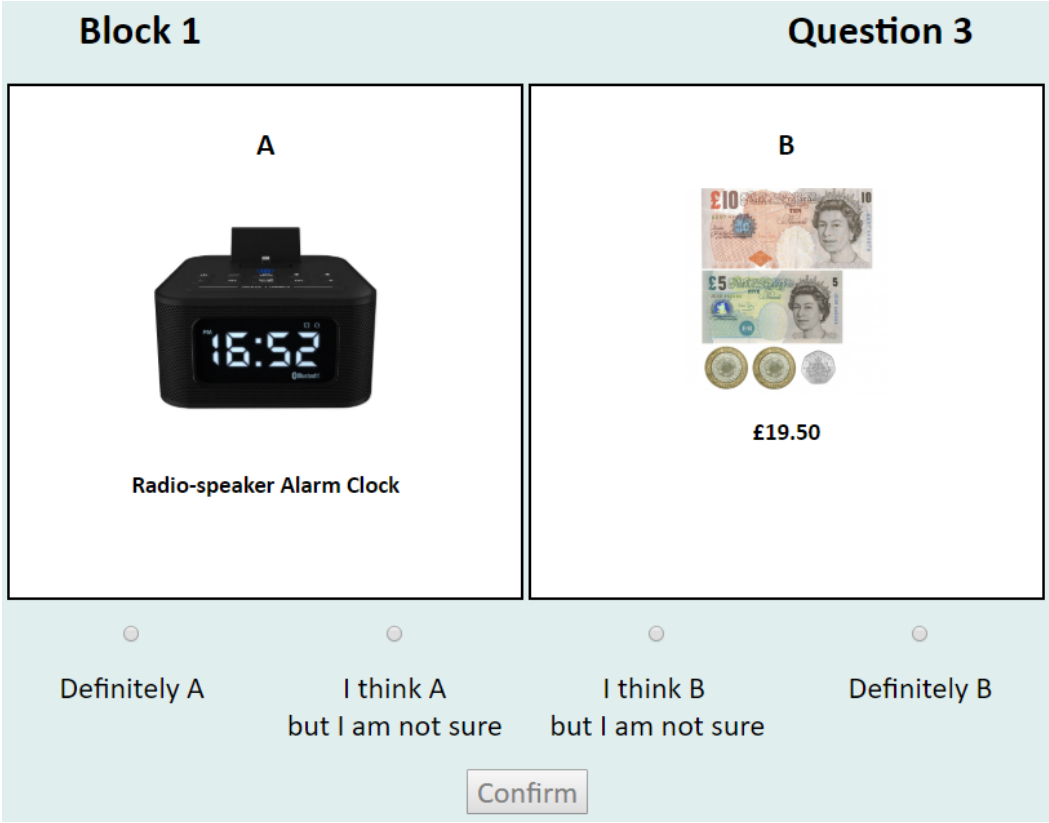
Block 1

This is the **1st block** of this study.

In this block, you have to make 25 choices between two items. For each pair of items, choose the one which you would rather receive, and indicate how sure you are about that choice using the radio buttons provided.

Next

Figure 24. Riskless Pairwise Choice Examples



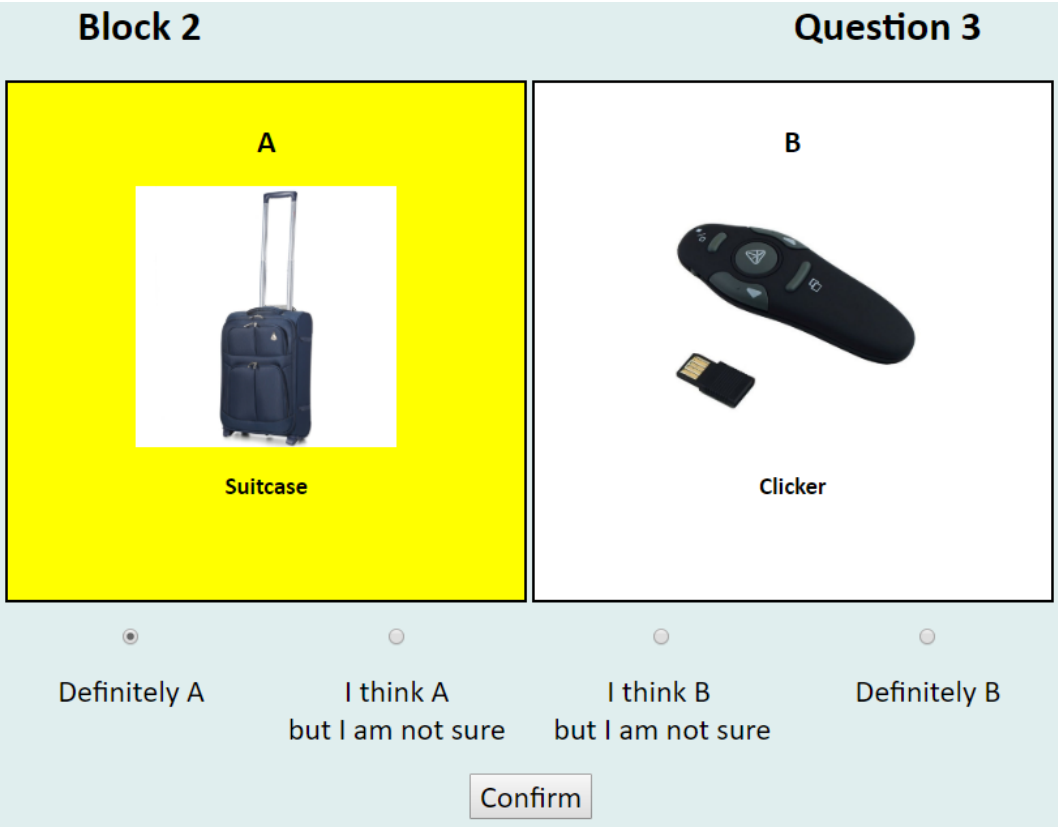
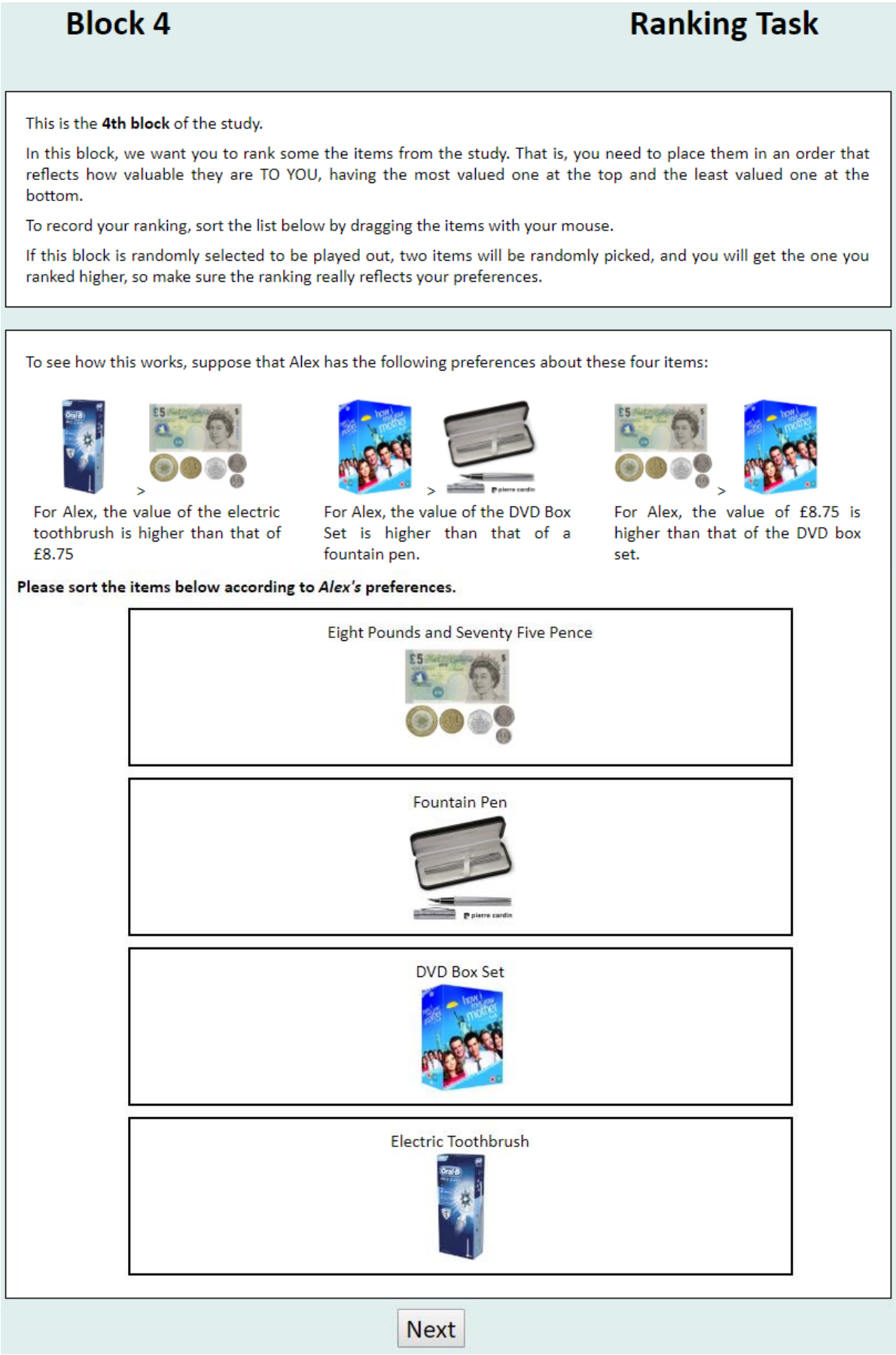


Figure 25. Ranking Task Practice Question Example



Block 4 Ranking Task

This is the **4th block** of the study.

In this block, we want you to rank some the items from the study. That is, you need to place them in an order that reflects how valuable they are TO YOU, having the most valued one at the top and the least valued one at the bottom.

To record your ranking, sort the list below by dragging the items with your mouse.

If this block is randomly selected to be played out, two items will be randomly picked, and you will get the one you ranked higher, so make sure the ranking really reflects your preferences.

To see how this works, suppose that Alex has the following preferences about these four items:

For Alex, the value of the electric toothbrush is higher than that of £8.75

For Alex, the value of the DVD Box Set is higher than that of a fountain pen.

For Alex, the value of £8.75 is higher than that of the DVD box set.

Please sort the items below according to *Alex's* preferences.

Eight Pounds and Seventy Five Pence

Fountain Pen

DVD Box Set

Electric Toothbrush

Next

Figure 26. Ranking Test Questions

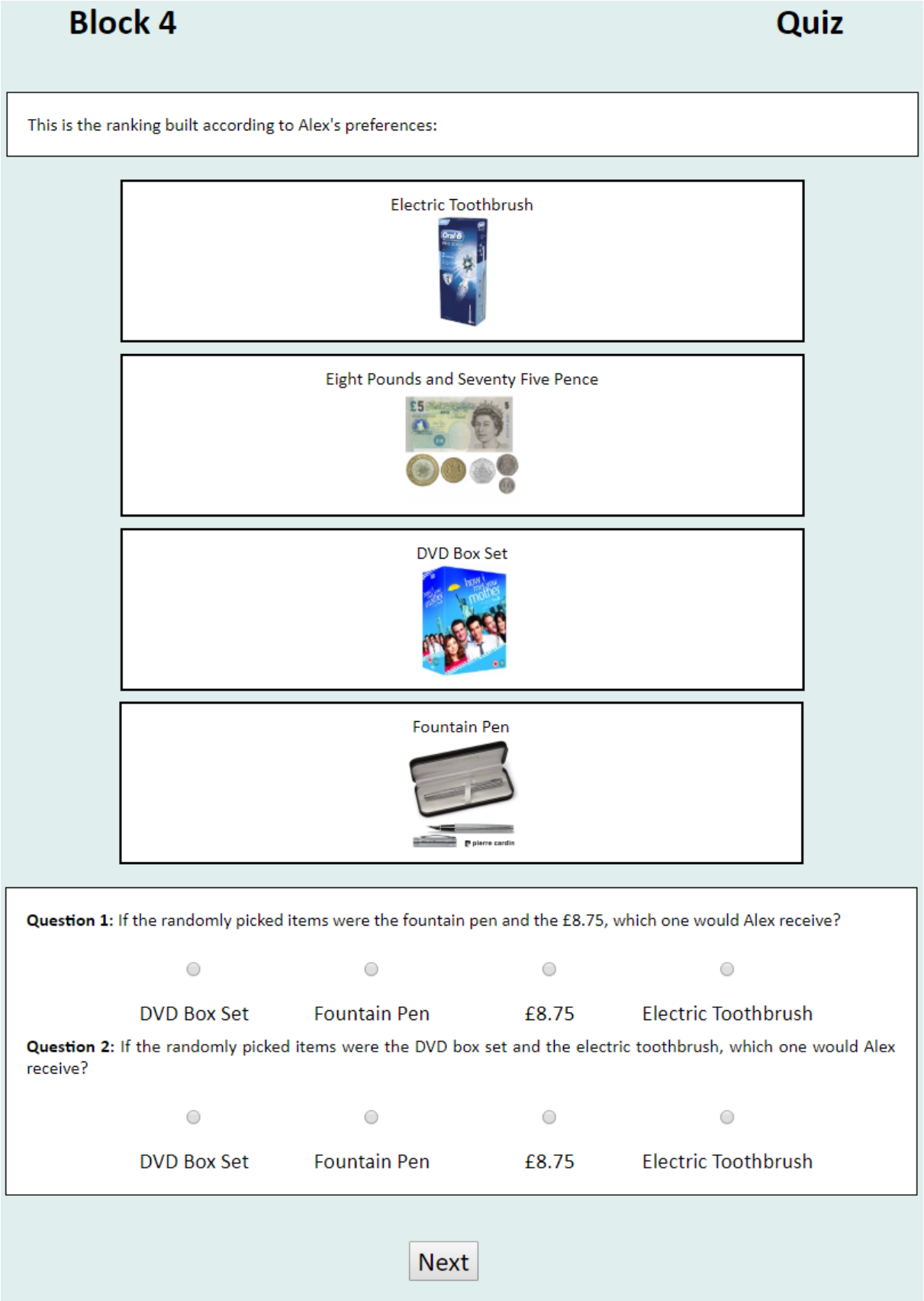


Figure 27. Ranking (reduced to fit the page)

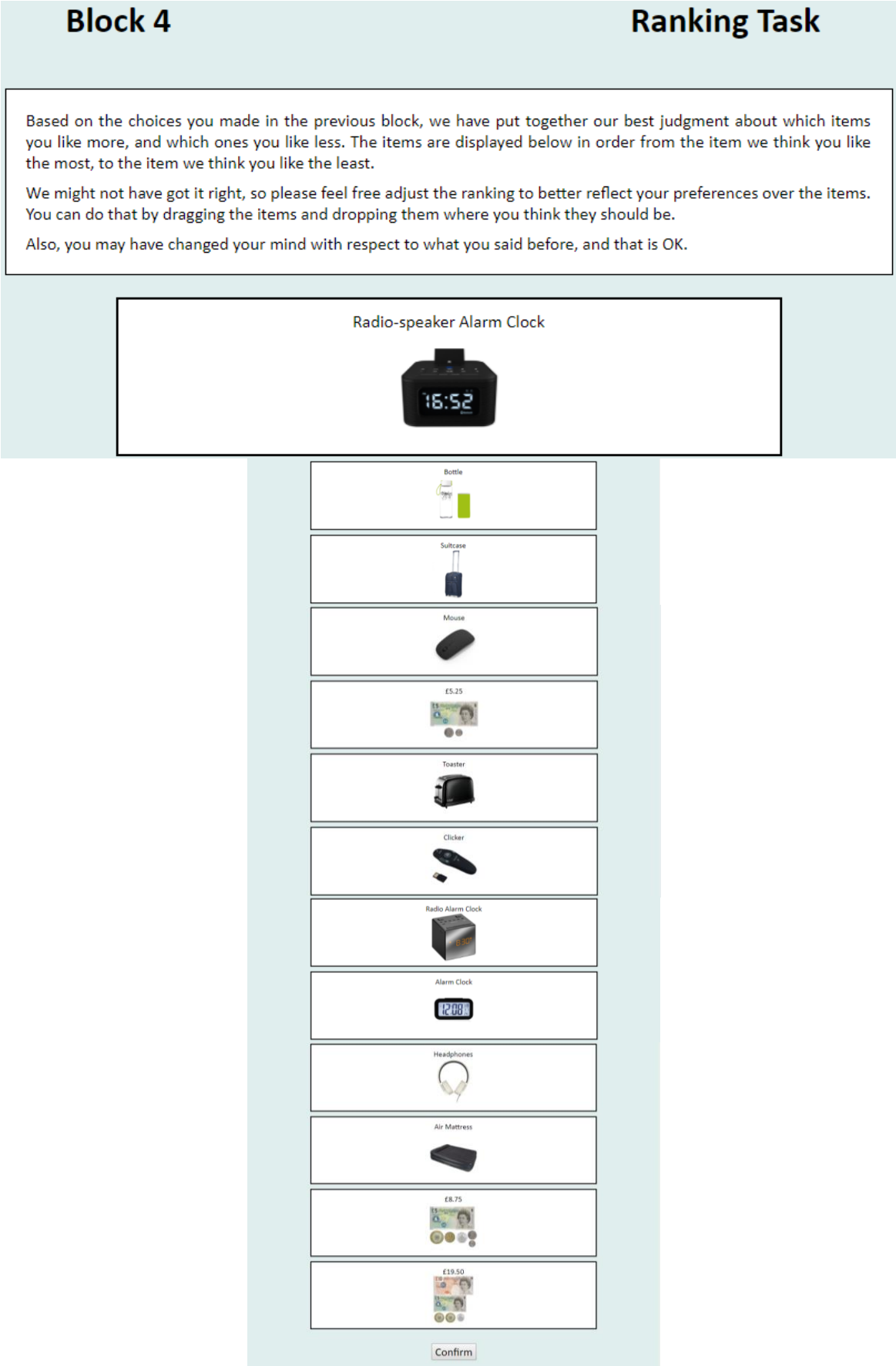


Figure 28. Detail of Object Information Available throughout the Study

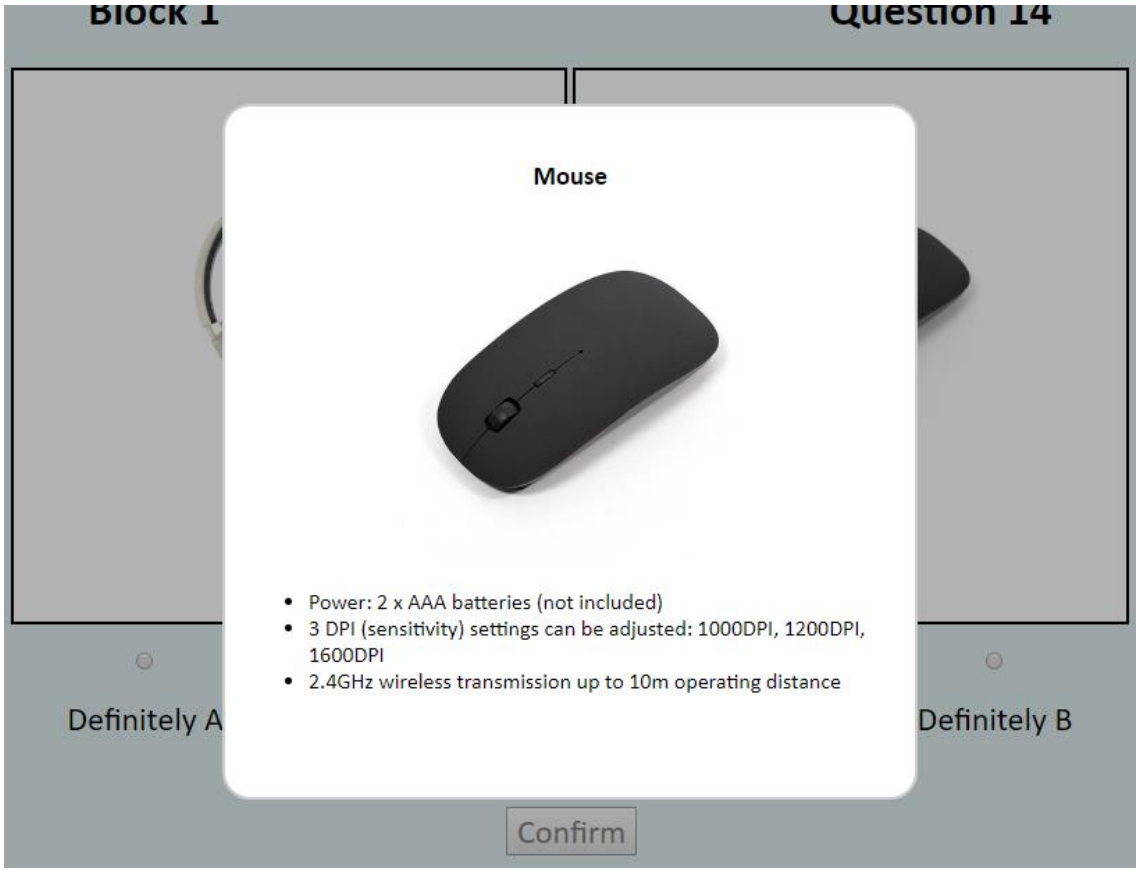
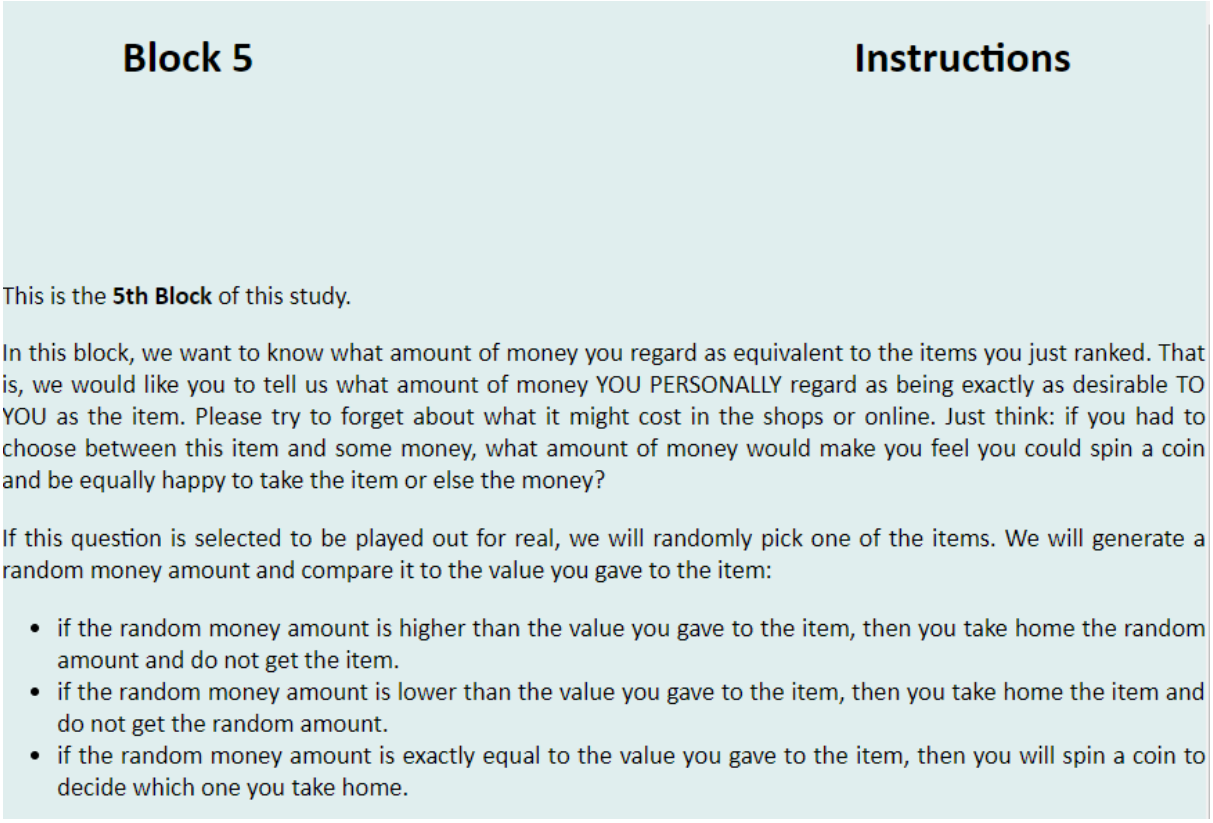


Figure 29. Valuation Instructions



Block 5 **Instructions**

This is the **5th Block** of this study.

In this block, we want to know what amount of money you regard as equivalent to the items you just ranked. That is, we would like you to tell us what amount of money YOU PERSONALLY regard as being exactly as desirable TO YOU as the item. Please try to forget about what it might cost in the shops or online. Just think: if you had to choose between this item and some money, what amount of money would make you feel you could spin a coin and be equally happy to take the item or else the money?

If this question is selected to be played out for real, we will randomly pick one of the items. We will generate a random money amount and compare it to the value you gave to the item:

- if the random money amount is higher than the value you gave to the item, then you take home the random amount and do not get the item.
- if the random money amount is lower than the value you gave to the item, then you take home the item and do not get the random amount.
- if the random money amount is exactly equal to the value you gave to the item, then you will spin a coin to decide which one you take home.

Figure 30. Valuation Test Questions

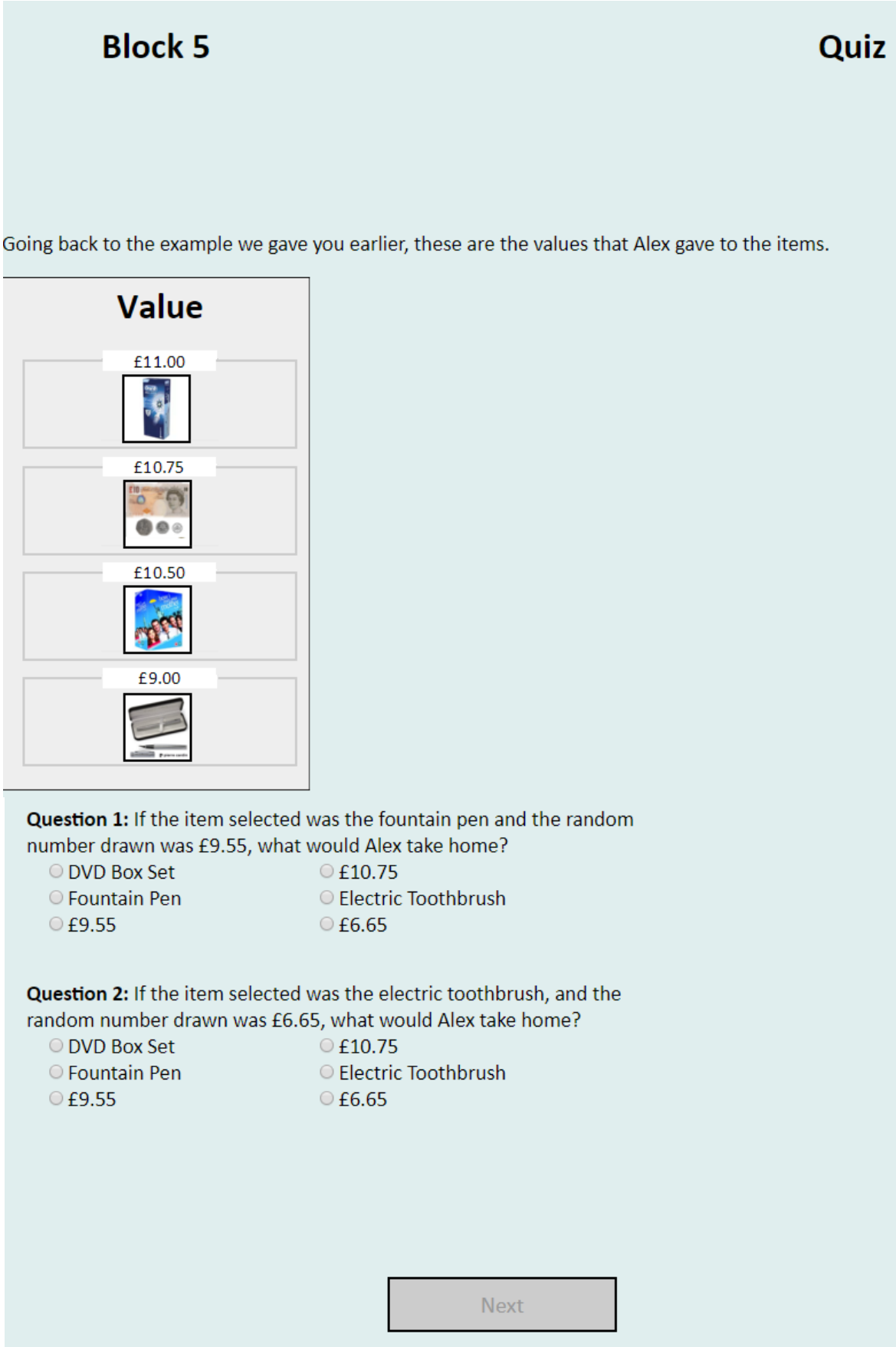


Figure 31. Valuation Task I

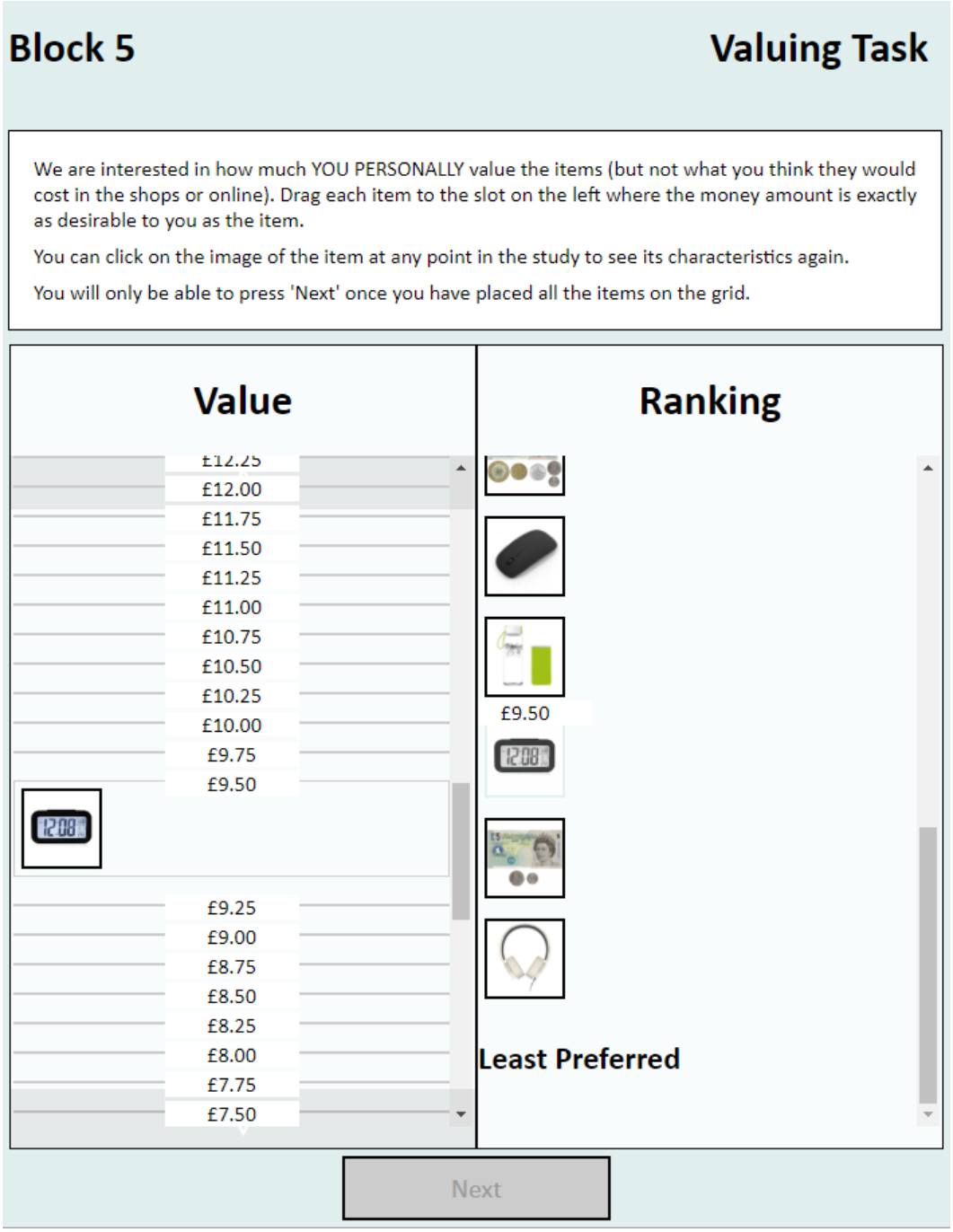


Figure 32. Valuation Task II (reduced to fit the page)

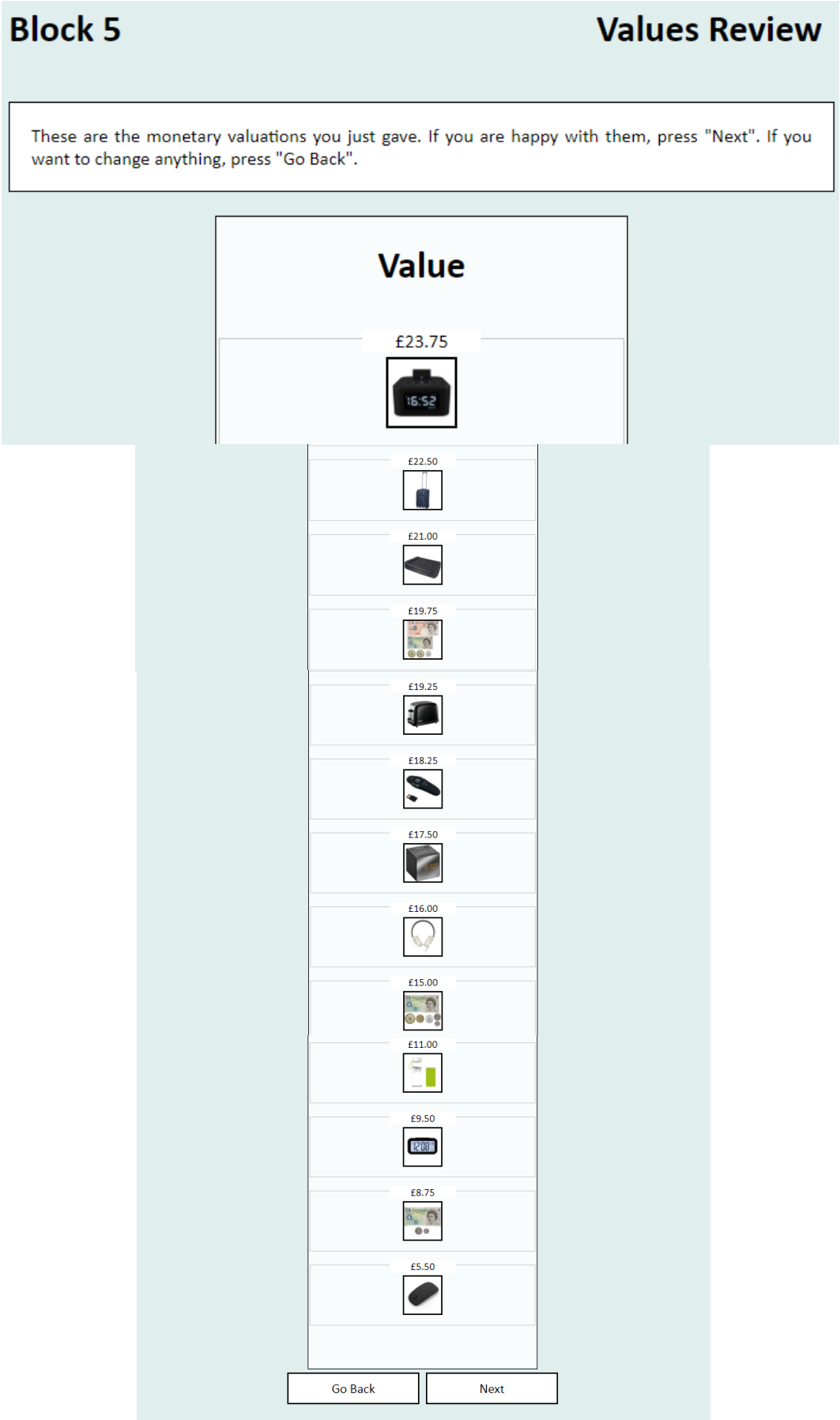
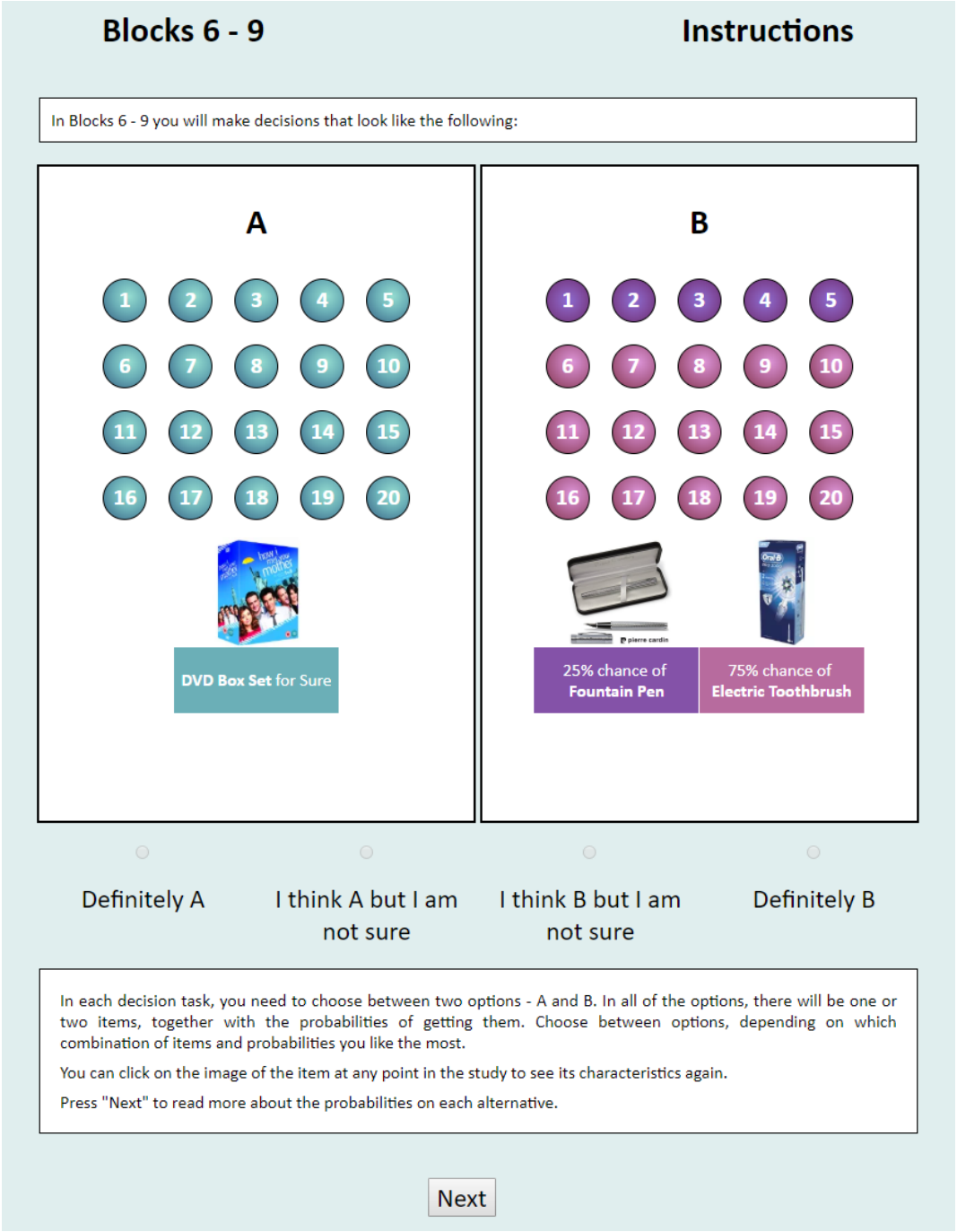


Figure 33. Risky Pairwise Choice Instructions



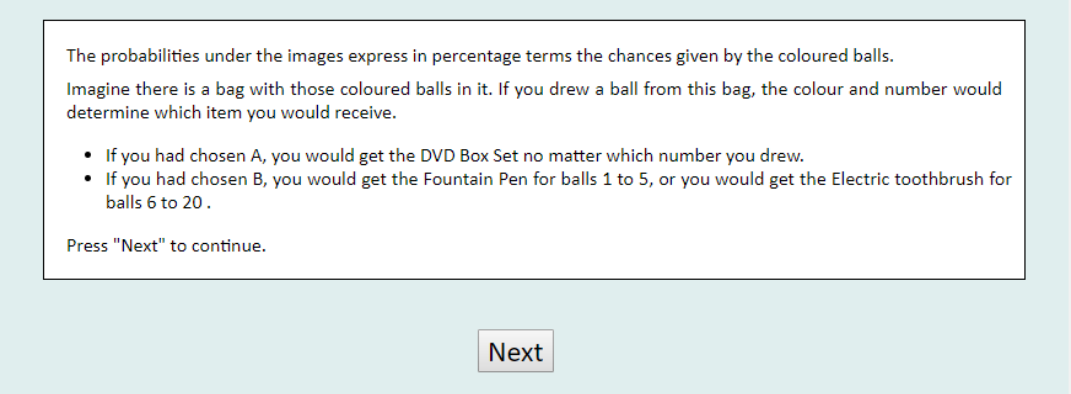


Figure 34. Risky Pairwise Choice Test Questions

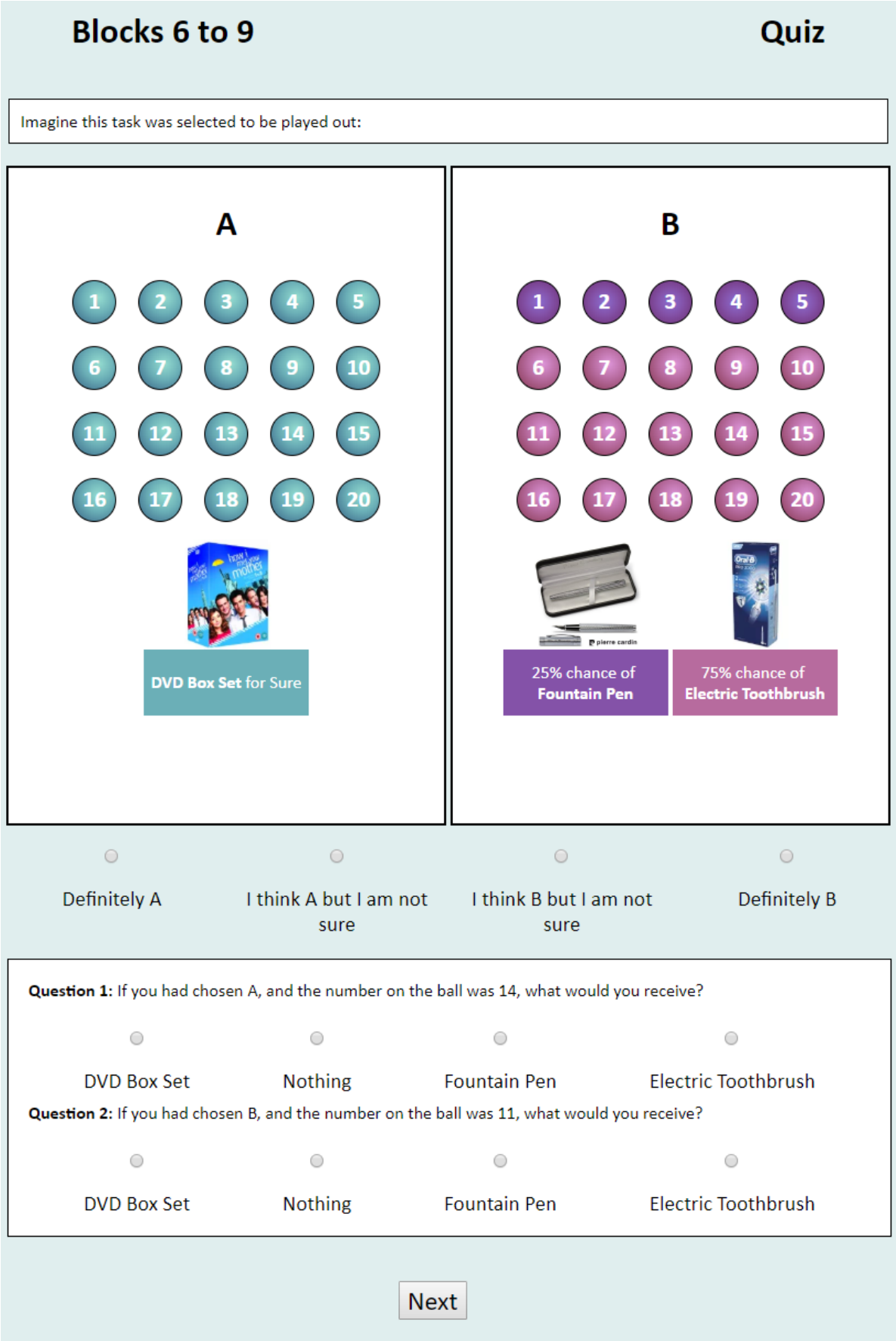


Figure 35. Introduction to Risky Pairwise Choice Block

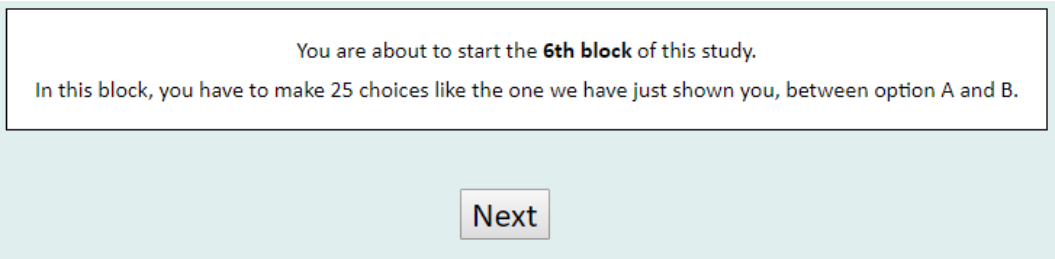
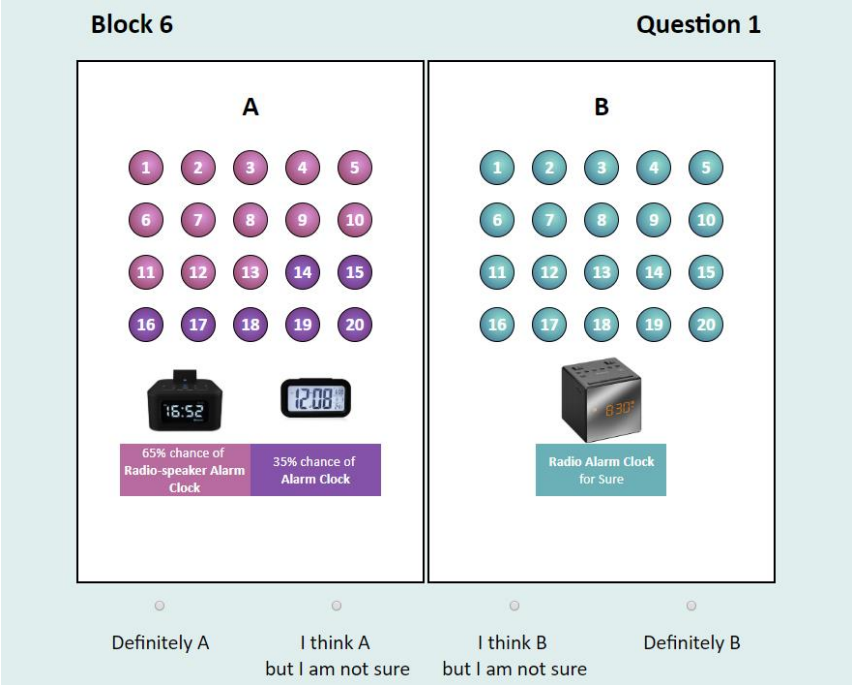
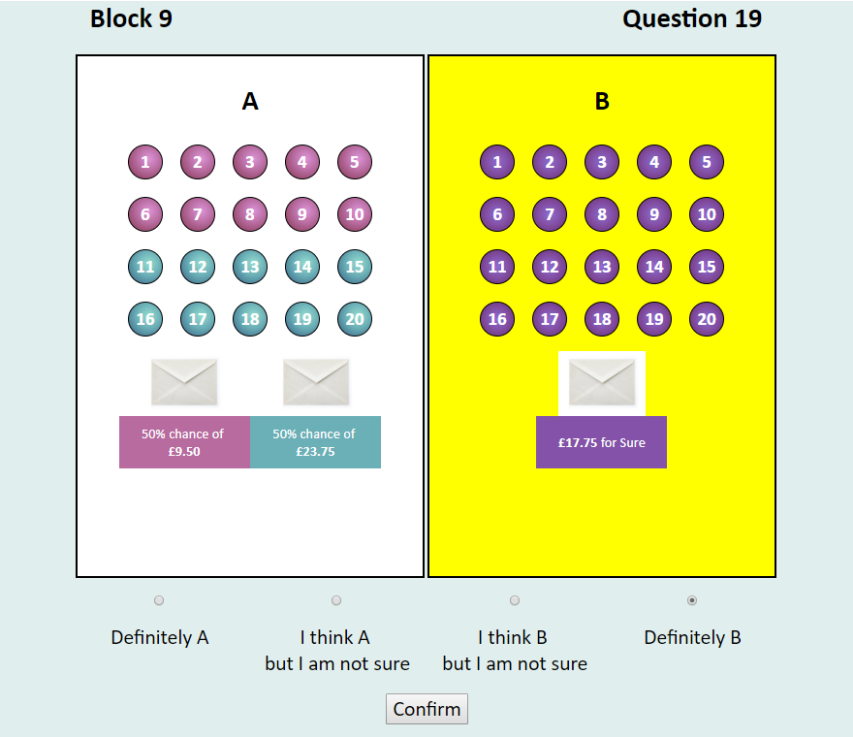


Figure 36. Risky Pairwise Choice Examples





Demographics

"Thank you for your answers to far!

Here are some questions about you, which will help us understand the different ways that people make decisions like these. Once you have filled in this information, the session will be over.

Gender: Male / Female

Age: [open box]

Highest education level attained:

- GCSE/O Levels
- A Levels or equivalent
- Bachelor's Degree
- Master's Degree
- PhD

Currently studying? Yes/No; Subject: [open box]

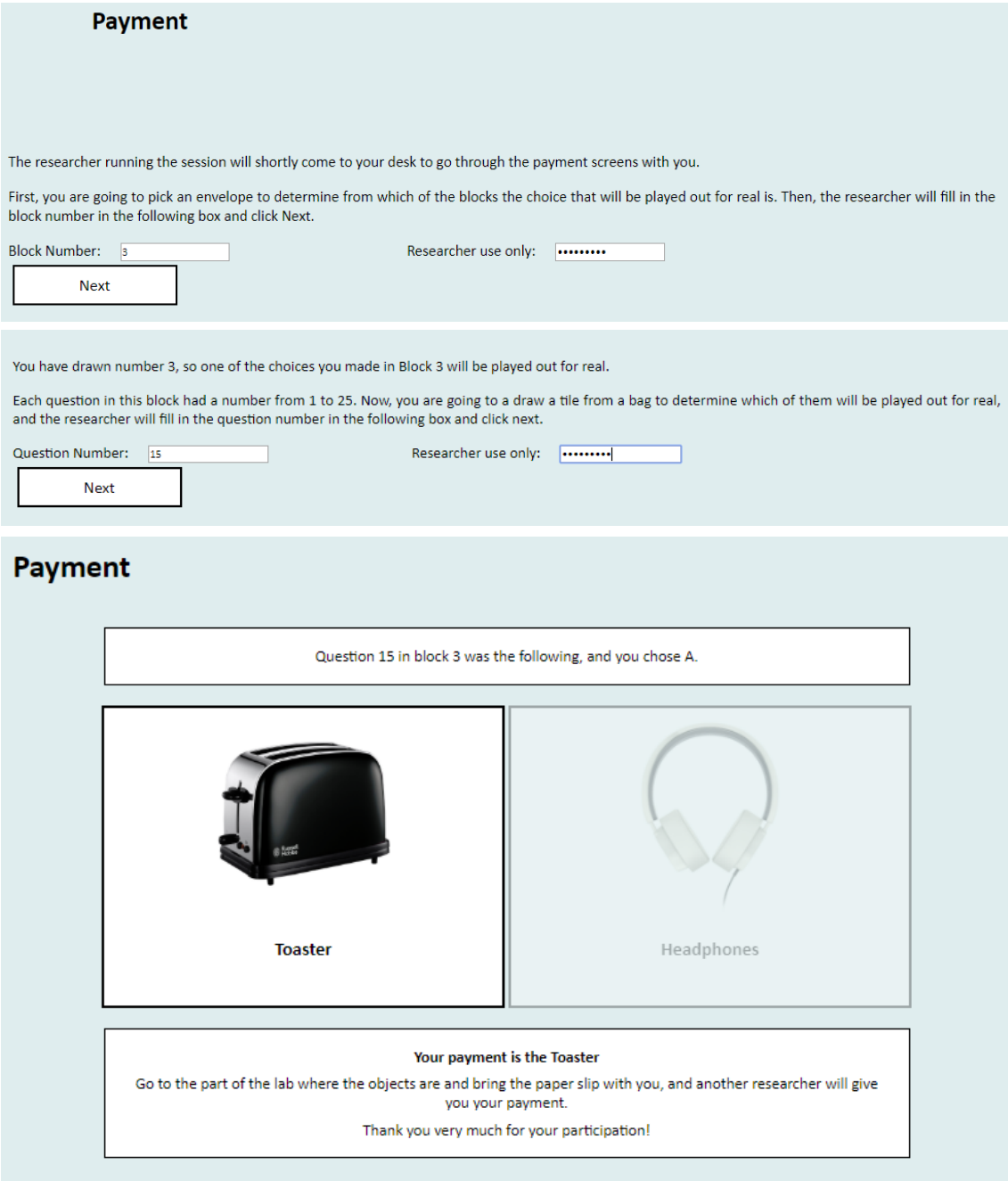
Employed: Yes/No; Full Time/Part Time

Personal monthly income before tax:

- Up to £1,000
- £1,001 to £1,500
- £1,501 to £2,000
- £2,001 to £3,000
- £3,001 to £4,000
- More than £4,000

Help us improve our survey. Was any aspect of what we asked you to do unclear or overly complicated? If there was anything you didn't like, how could we improve it? If you wrote everything clear, please write 'all fine'. [open box]"

Figure 37. Payment Screens



"The researcher running the session will shortly come to your desk to go through the payment screens with you. First, you are going to pick an envelope to determine from which of the blocks the choice that will be played out is. Then, the researcher will fill in the block number in the following box and click Next.

[Example] You have drawn number 3, so one of the choices you made in Block 3 will be played out for real. Each question in this block had a number from 1 to 25. Now, you are going to draw a tile from a bag to determine which of them will be played out for real, and the researcher will fill in the question number in the following box and click Next.

Question 15 in block 3 was the following and you chose A. Your payment is the Toaster. Go to the part of the lab where the objects are and bring the paper slip with you, and another researcher will give you your payment. Thank you very much for your participation!"

A.3. Probability Trees

Figure 38. Probability Tree – Starting Probability = 0.50

(The numbers on the left correspond to the probabilities of receiving the best payoff)

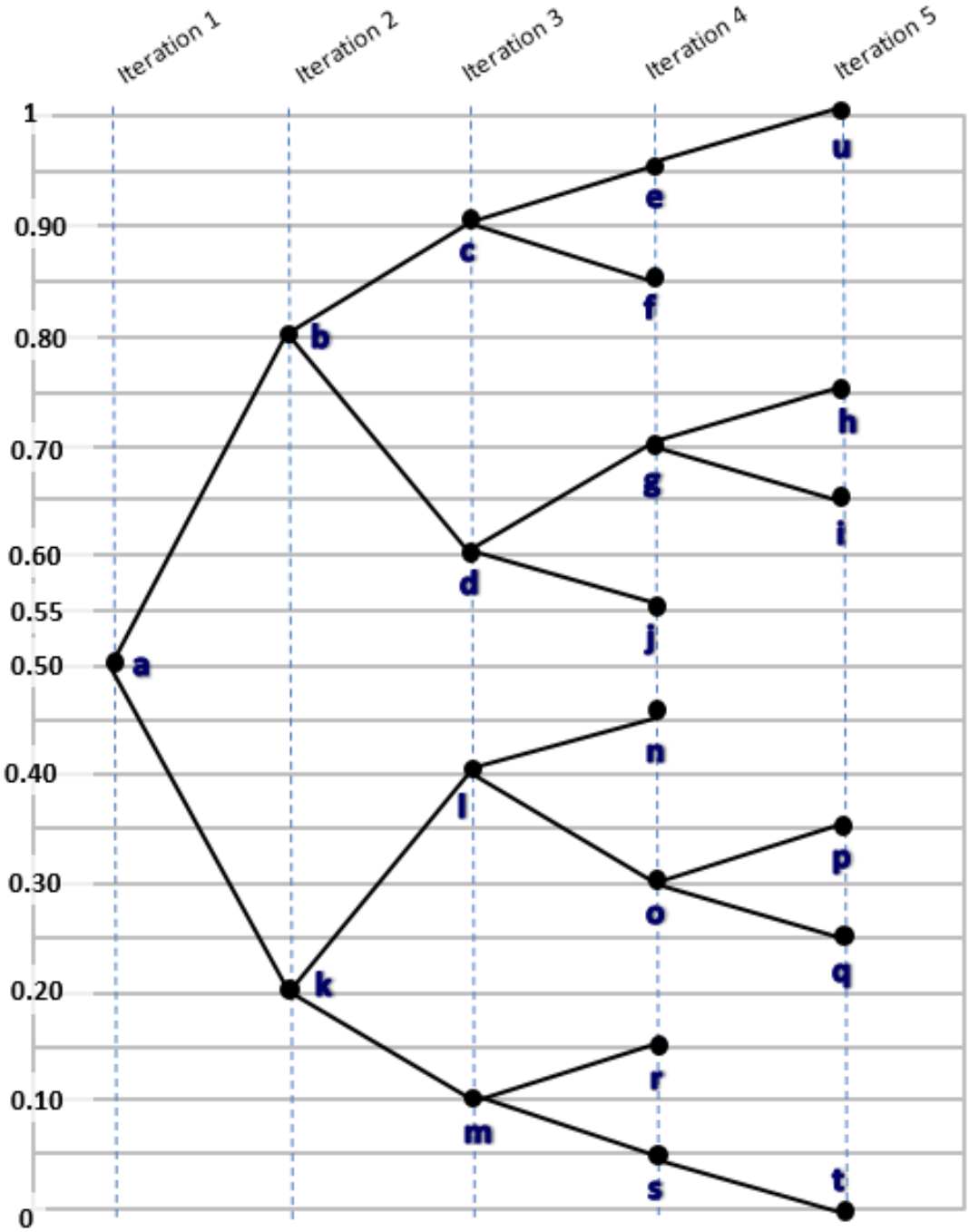


Figure 39. Probability Tree – Starting Probability = 0.65

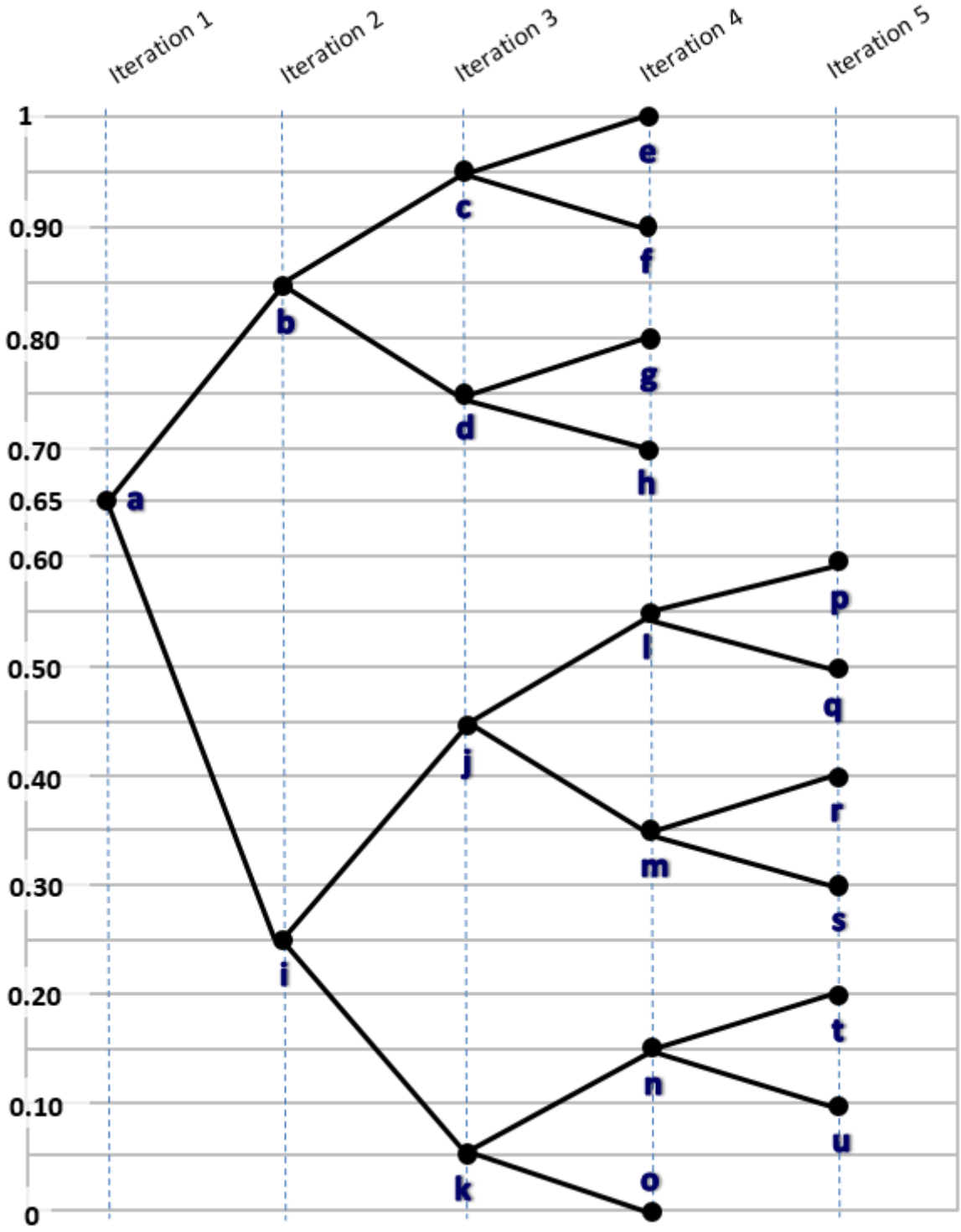
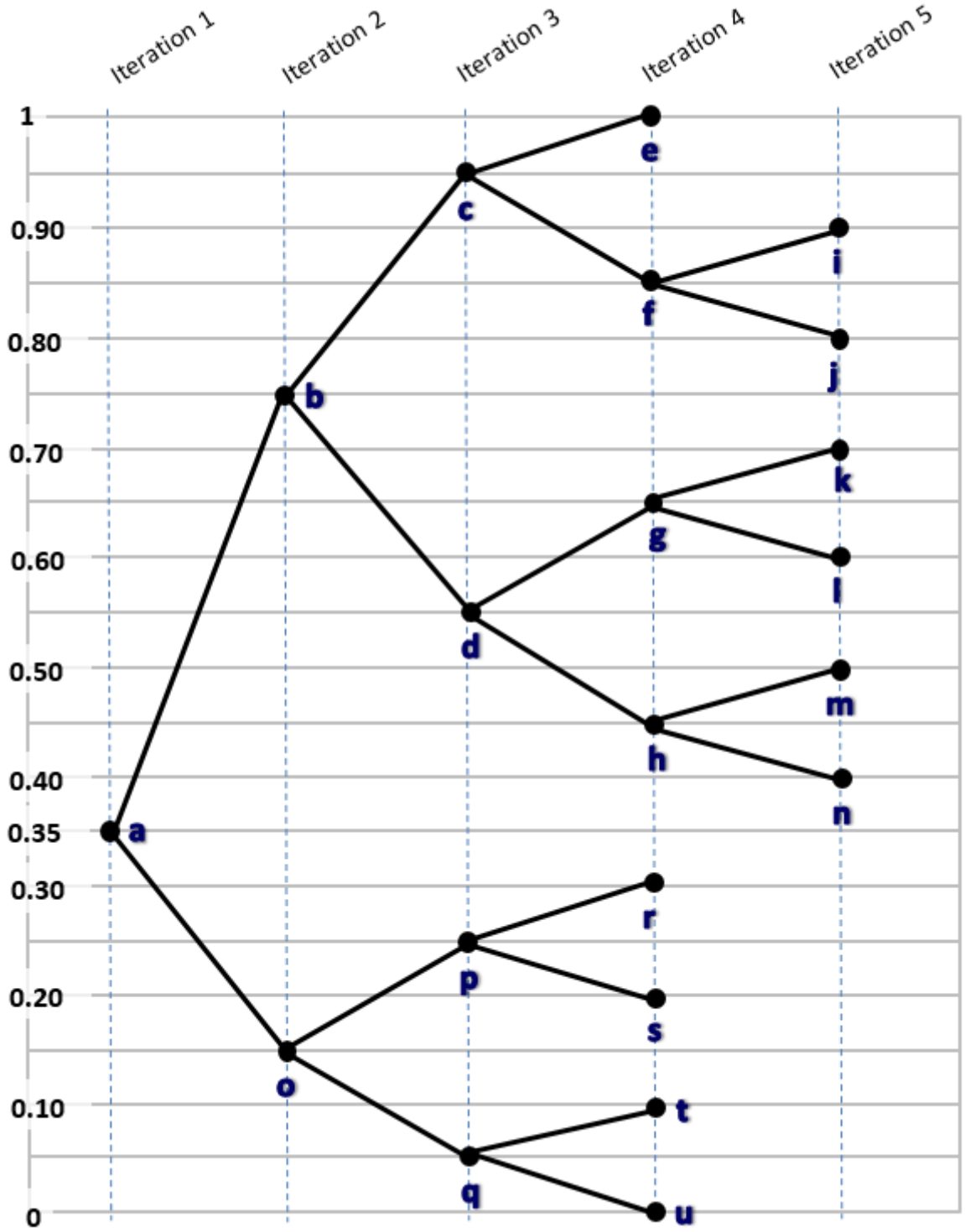


Figure 40. Probability Tree – Starting Probability = 0.35



A.4. Risky Pairwise Choice Randomisation

Types of Questions

The risky pairwise choice questions in this study can be classified in 6 categories:

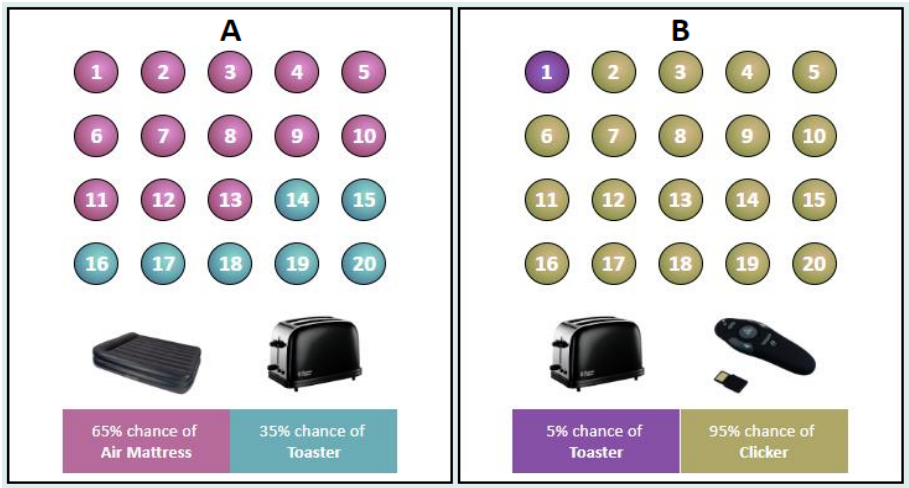
- X: Questions of interest for Chapter 2
- CR: Questions of interest for Chapter 3
- A: Attention checks
- M: Modified standard gamble questions
- B: Bound distance manipulation questions
- R: Repeated questions

More information about questions X and CR can be found in their respective chapters, and about the attention check questions, in Appendix A.5.

<u>Modified Standard Gamble Questions (M)</u>

These questions differ from the gamble scenarios in Chapter 2 in that they do not have a riskless prospect. See example in Table 19. The formerly 'sure' prospect has been modified by adding a 5% risk of the worst outcome. This modification was used by Carty et al. (1999) when valuing health in order to avoid participants choosing the 'sure' prospect driven by their desire to avoid risk rather than by their preferences over the health states. These questions were included in this study to test the extent of this effect with non-health goods; that is to see how much the percentage of 'sure' prospect choices decreased when a small possibility of receiving the worst outcome was added to the prospect. These results are not relevant for this thesis but are available upon request.

Table 19. Modified Standard Gamble Question



The gamble scenarios that were presented with and without the modification were scenarios P and Q with dissimilar objects. Here, these are labelled using the shortening Mp and Mq respectively, followed by a number (1 to 5) for the iteration. Some of these questions (Mp1, Mp2, Mp3, Mq1 and Mq2) always appeared in the same order for all participants, but the rest (Mp4, Mq3, Mq4, and Mp5 and Mq5 if the iterations had not finished after 4 questions) would not always be presented to participants.

Bound distance manipulation questions (B)

These questions were included in the study to test the hypothesis that the distance between the new bound and the bound they substituted (here, the worst object and receiving nothing) may have an effect on the direction of the disparity between the direct and indirect SG estimate. This was tested by using the same original bounds, but manipulating the distance between the new bound and the old bound.

The questions used to test this had different payoffs A, B and C than the questions in Chapter 2. Based on each participant's valuations, a ranking of preference for the objects was established. The best (A) and second best (B) payoffs were the 1st and 4th in the ranking. Payoff C, the worst object and new bound, was varied across 'conditions' or questions. In one question it was the 10th object (so that the distance between the new bound and the old bound, receiving nothing, was the lowest possible); in the other question, it was the 7th object (so that the distance was higher).

Table 20 includes the questions required to answer our research question. The first gamble scenario, which corresponds to scenario P, is the same for both questions: we call it Bpi, where i corresponds to the question number. Scenarios Q and R vary in the worst and the best payoffs respectively, so in order to differentiate them, we include the question number in their name (for example, B1qi and B2qi). The following questions appeared in a fixed order: Bp1-3, B1q1-3, B2q1-2, B1r1-3. After these, the order in which they would appear is: B2q3, Bp4, B1q4, B1r4, B2q4, and if the iterations had not finished, Bp5, B1q5, B1r5, B2q5. Due to an error in data collection, the gamble scenario B2Ri was not presented to participants. Hence, these questions could not be used to answer the intended question.

Table 20. Bound Distance Manipulation Gamble Scenarios

n	Bpi		Bnqi		Bnri	
1	Object #4	Object #1	Object #4	Object #1	Object #10	Object #1
				Object #10		Nothing
2		Nothing	Object #4	Object #1	Object #7	Object #1
				Object #7		Nothing

Pre-set questions

Some participants valued one or more of the similar or dissimilar objects above £50. In the money equivalent gamble scenarios, instead of having their own object valuations, they were shown questions with the following pre-set amounts:

- Money Equivalent of Similar Objects: £10.75 for certain vs. £2 or £17.50
- Money Equivalent of Dissimilar Objects: £8.75 for certain vs. £5.25 or £19.50
- Common Ratio Effect: £9 for certain vs. £0 or £12 (note that the ¾ ratio is maintained)

Question Randomisation

The order of the questions is as shown in Table 21, where every row corresponds to a round. Rounds were not made explicit to participants, who completed four blocks of 25 questions each. In every round, one iteration of the questions of interest for Chapter 2 was completed – remember that these iterations could take 4 or 5 questions.

The first round includes 24 questions; the questions of interest for Chapter 2 and the rest were alternated. In subsequent rounds, there was only one of the *other* questions for every three of interest for Chapter 2. In the last two, there are several questions per cell, indicating that there were several options which would be presented in the order that they appear in the cell. For example, consider question 10 of round V. The first option was the fifth iteration of one of the Chapter 2 gamble scenarios (X5). If the iteration had finished in 4 rounds, then the question presented would be a modified standard gamble question (M). If the iteration of the modified scenario had finished too, a random Chapter 2 question would be repeated (R).

In every round, the order of Chapter 2 questions was randomised. These questions are identified using an X and a number that corresponds to their iteration. There were 10 common ratio effect (CR) questions, which were all presented in random order, and hence labelled just using CR. There were 4 attention check questions, identified as A1 (the first one) and then simply A (to indicate that it could be any of the remaining 3). There were up to 10 modified standard gamble questions (M), if both iterations took 5 questions. With regards to the bound distance manipulation questions (B), up to 20 of them were presented.

Starting Probability Randomisation

Each participant's Chapter 2 gamble scenarios were randomly allocated to one of the white round 1 cells (see Table 21). Each of these cells had a probability associated to them, which had been selected to add variety to the first round. These were the probabilities at which each gamble scenario would start the iteration.

Table 21. Question Randomisation

Round / Question																								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
I	X1	Bp1	X1	B1q1	X1	CR	X1	B1r1	X1	Mp1	X1	CR	X1	B2q1	X1	Bp2	X1	CR	X1	Mq1	X1	B1q2	X1	CR
	1		**2**		**3**	**4**	**5**		**6**		**7**	**8**	**9**		**10**		**11**	**12**	**13**		**14**		**15**	**16**
II	X2		X2		X2	B1q2	X2		X2		X2	CR	X2		X2		X2	Mp2	X2		X2		X2	CR
III	X3		X3		X3	B2q2	X3		X3		X3	Mq2	X3		X3		X3	A1	X3		X3		X3	Rp3
IV	X4		X4		X4	CR	X4		X4		X4	B1q3	X4		X4		X4	Mp3	X4		X4		X4	R1r3
V	X5 / B		X5 / M		X5 / B	M	X5 / B		X5 / M		X5 / B	B	X5 / A		X5 /M /R		X5 / B/ R	CR	X5 / B / R		X5 / M / R		X5 / A	CR
	1		**2**		**3**		**4**		**5**		**6**		**7**		**8**		**9**		**10**		**11**		**12**	
VI	CR		A		B/A/M/R		A/B/M/R		B/A/M/R		A/B/M/R		B/A/M/R		A/B/M R		B/A/M/R		A/B/M/R		B/A/M/R		A/B/M/R	

Table 22. Starting Probability Randomisation

Round I Qn.	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***	***18***	***19***	***20***	***21***	***22***	***23***	***24***
Probability	0.65	-	0.35	-	0.50	-	0.50	-	0.35	-	0.35	-	0.65	-	0.65	-	0.50	-	0.35	-	0.65	-	0.50	-

A.5. Attention Checks

Two sets of two attention checks questions each were built into our design (see Table 23).

Table 23. Attention Check Questions

1	Alarm Clock, p=0.50 Nothing, p=0.50	Radio-speaker Alarm Clock, p=0.55 Nothing, p=0.45
2	Radio-speaker Alarm Clock, p=1	Alarm Clock, p=0.95 Nothing, p=0.05
3	Bottle, p=1	Suitcase, p=0 Nothing, p=1
4	Bottle, p=1	Radio-speaker alarm clock, p=0 Nothing, p=1

The first one featured the similar objects and also has "nothing" as a possible outcome. For participants that ranked the alarm clocks in the expected order (radio-speaker alarm clock $\succ$ radio alarm clock $\succ$ alarm clock), the rational, (expected) utility maximising answer would be to choose the right side on gamble 1a, and left side on 1b. These sides correspond to the gambles as depicted on the table above, but in the study, they were assigned to the right or left side randomly. For those that did not rank the clocks in that order, they should choose the left side on 1a as long as the following is true

$$WTP(Radio\ Speaker\ Alarm\ Clock) < \left(\frac{0.50}{0.55}\right) \cdot WTP(Alarm\ Clock) \tag{8}$$

and right side on 1b as long as

$$WTP(Radio\ Speaker\ Alarm\ Clock) < 0.95 \cdot WTP(Alarm\ Clock). \tag{9}$$

One participant did not answer question 1a as a rational participant would have, but the rest of participants did. This participant only failed 1 out of 4 attention checks so we do not consider necessary to exclude her answers from the analysis.

Gambles on the second set offered an object that was generally viewed as relatively bad for certain, and a very good one with probability 0, and nothing with probability 1 on the other side. If participants chose the left side, then we know that they were not disregarding probabilities and paying attention to the items only, even in cases when the difference in value of the objects was large. There was only one participant that failed this attention check, and because this did only happen in one out of their four questions, we did not consider necessary to exclude their answers.

Table 24. Self-Reported Measures about the Objects: Desire to Receive Them

Object	DESIRE (0=not at all, 6=very much)		
	Median	**Mean**	**SD**
Headphones	4	4.02	1.96
Radio-speaker Alarm Clock	4	3.53	2.10
Suitcase	3	3.12	2.27
Radio Alarm Clock	3	2.88	2.15
Air Mattress	3	2.83	2.17
Mouse	2	2.10	2.07
Alarm Clock	1	1.81	1.89
Bottle	1	1.64	1.65
Toaster	1	1.62	1.89
Clicker	1	1.50	1.84

Note: Objects ordered in decreasing order according to the median desire.

Table 25. Self-Reported Measures about the Objects: Already Own Them

Object	ALREADY OWN THEM	
	No	**Yes**
Clicker	52	6
Air Mattress	46	12
Radio-speaker Alarm Clock	43	15
Radio Alarm Clock	40	18
Alarm Clock	26	32
Headphones	26	32
Toaster	26	32
Mouse	19	39
Bottle	18	40
Suitcase	10	48

Note: Objects ordered in decreasing order according to the number of participants that did not own them.

A.6. Additional Results

Table 26. Self-Reported Measures about the Objects: Desire to Receive Them

Object	DESIRE (0=not at all, 6=very much)		
	Median	**Mean**	**SD**
Headphones	4	4.02	1.96
Radio-speaker Alarm Clock	4	3.53	2.10
Suitcase	3	3.12	2.27
Radio Alarm Clock	3	2.88	2.15
Air Mattress	3	2.83	2.17
Mouse	2	2.10	2.07
Alarm Clock	1	1.81	1.89
Bottle	1	1.64	1.65
Toaster	1	1.62	1.89
Clicker	1	1.50	1.84

Note: Objects ordered in decreasing order according to the median desire.

Table 27. Self-Reported Measures about the Objects: Already Own Them

Object	ALREADY OWN THEM	
	No	**Yes**
Clicker	52	6
Air Mattress	46	12
Radio-speaker Alarm Clock	43	15
Radio Alarm Clock	40	18
Alarm Clock	26	32
Headphones	26	32
Toaster	26	32
Mouse	19	39
Bottle	18	40
Suitcase	10	48

Note: Objects ordered in decreasing order according to the number of participants that did not own them.

Table 28. OLS Regression Output: Randomised Design Features

Dependent variable: Indifference Probability **Independent variables:**	**Coefficients**
Starting probability 50	-1.029
	(2.486)
Starting probability 65	-0.948
	(2.507)
Certain colour pink	0.269
	(1.809)
Certain colour purple	1.691
	(1.930)
Risky better colour pink	-1.188
	(1.818)
Risky better colour purple	-0.551
	(1.742)
Risky worse colour pink	-
Risky worse colour purple	-
Risky option displayed on right side	-0.758
	(1.280)
Constant	57.97***
	(3.031)
Observations	3,113[29]
Clusters	58
R-squared	0.001

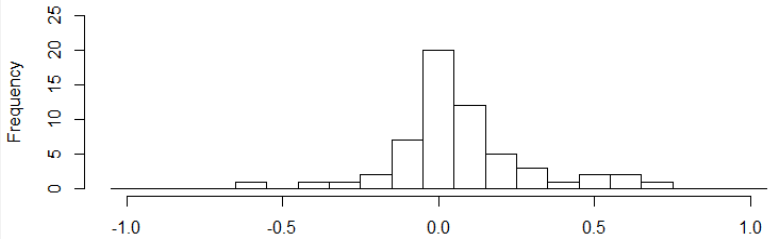
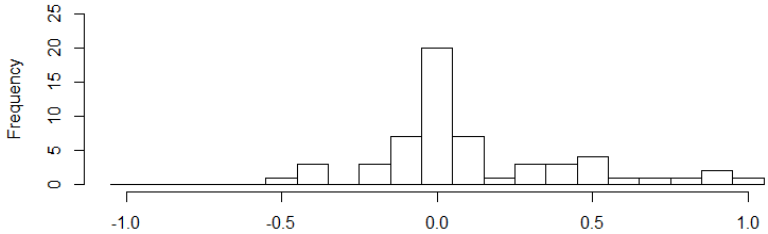
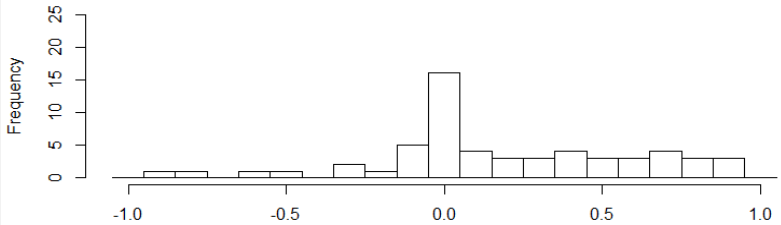
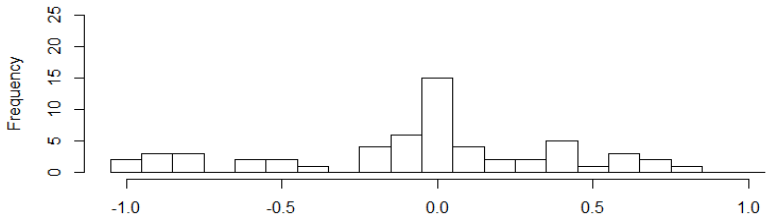
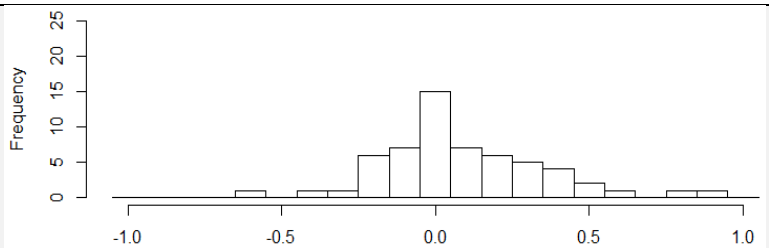
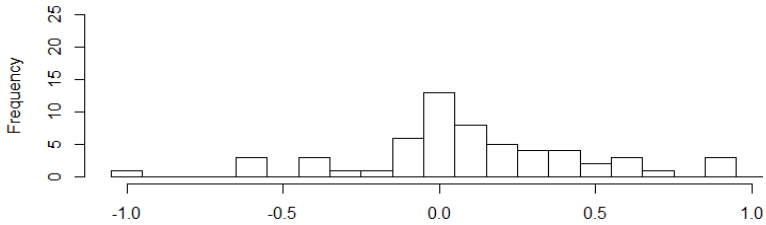
Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Notes:

- The reference categories were starting probability of 35, turquoise colour, and risky option displayed on the left-hand side.
- "Certain colour" refers to the balls in the alternative where the participant would get the payoff for sure, and "risky" refers to the ones used in the one where they could either get the best or the worst payoffs.
- The "risky worse colour" variables were omitted. They capture the colour of the balls that was used to represent the worse outcome in the risky alternative of the gamble. The colours of the best and worst options were automatically generated by an algorithm that followed a systematic sequence, and so the colours for the worst payoff do not add variance on top of that added by the colours of the best payoff.

[29] These are only the risky pairwise choices of interest, which are the ones we analysed in Chapter 2.

Table 29. Objects vs. Money Equivalent Estimates

Estimate	Type of Good	u(£) - u(Object)	Wilcoxon signed-rank p-value
$u(B)_P$	**Similar**		.0012
	Dissimilar		.0045
$u(B)_Q$	**Similar**		.0008
	Dissimilar		.6390
$u(C)_R$	**Similar**		.0030
	Dissimilar		.0114

Effect of Differences in Outcome Value on Disparities

In order to explore this, we classified estimates in two groups only, 'consistent' and 'inconsistent'. There are two metrics that capture the value difference between the three payoffs in each set. First, the value range: the difference between the value of the best and worst objects (or their money equivalent). And second, the value of the best payoff, which is the only outcome that is common to the three gambles in each set. Table 30 includes summary statistics of these values for the consistent and the inconsistent cases. We used two-tailed Wilcoxon rank-sum tests to check whether there was a significant difference between the outcomes value differences of the consistent and the inconsistent cases. The p-values from these tests can also be found on Table 30.

Table 30. Mean (SD) of Value Difference metrics by Type of Object (£)

	VALUE RANGE (BEST – WORST) (n=58)			**BEST VALUE (n=58)**		
	Money	**Similar Objects**	**Dissimilar Objects**	**Money**	**Similar Objects**	**Dissimilar Objects**
Consistent	9.53 (9.18)	7.79 (8.54)	11.01 (11.33)	17.26 (10.73)	13.63 (10.01)	17.22 (12.39)
Inconsistent	11.42 (11.34)	13.46 (11.81)	8.97 (9.19)	16.30 (13.11)	19.92 (12.95)	14.20 (11.89)
Wilcoxon rank-sum test p-value	.3724	.0334	.5912	.3863	.0563	.3650

Because we ran 3 tests for each value metric, we applied the Holm-Bonferroni multiple testing correction (Holm, 1979). This correction adjusts the significance level alpha to account for the number of tests performed, using less stringent criteria for each test as the p-values increase. The more conservative test is hence the one for the lowest p-value (.0334 in this case): the former threshold α = .05 was divided by 3 (which is the number of tests performed), and became .017. We could not reject the null hypothesis of equality of the similar objects value range between the consistent and inconsistent cases. In the remaining tests, the p-values were already above .05, so reducing the significance level below .05 would not have changed the result that the difference was not statistically significant.

Appendix B – Chapter 3

Table 31. Object Valuations and Comparison of the Value Difference (£) (n=60)

		Median	Mean	SD	Wilcoxon Signed-Rank Test p-value H_a: Distance similar objects ≠ Distance dissimilar objects
Similar Objects	**Better**	15.00	15.92	12.42	.4092
	Worse	8.38	11.81	11.03	
	Difference	2.50[a]	4.12	4.55	
Dissimilar Objects	**Better**	9.50	13.56	11.89	
	Worse	7.25	10.30	9.08	
	Difference	2.25	3.26	2.87	

[a] Note that this is the median of the value difference, not the difference of the medians

Table 32. Comparison of the Deviation from Target Ratio (n=60)

	Median	Mean	SD	Wilcoxon Signed-Rank Test p-value H_a: Deviation similar objects ≠ Deviation dissimilar objects
Similar Objects	-0.0114	-0.0210	0.1640	.4512
Dissimilar Objects	0	0.0077	0.0317	

Table 33. Confidence ('sureness') in Preference by Similarity (n = 60)

		Dissimilar Goods		
		Unsure	**Sure**	**TOTAL**
Similar Goods	**Unsure**	7	3	10
	Sure	17	33	50
	TOTAL	24	36	60
McNemar's Chi-squared Test p-value < .01				

Appendix C – Chapter 4

C.1. Instructions

Figure 41 and Table 34 show how the bacteria and the policy options were presented to participants across conditions. The text in blue was only shown to those participants in the self-interest condition. The text in red corresponds to the alternative versions for the death and the gastroenteritis conditions.

Figure 41. Instructions: Bacteria Introduction & Policy Options Rating

Imagine a city with a population of 200,000 people. Half the population live in the East Zone and the other half in the West Zone. You live in the East Zone.

These zones are serviced by water supplies from two different reservoirs. These reservoirs are currently contaminated with three different types of bacteria. These bacteria are harmful and if nothing is done about them they are expected to kill / give gastroenteritis to 46 people over the next 10 years.

Gastroenteritis caused by these bacteria is not contagious between people and it cannot kill you, but people infected will have the following symptoms for approximately two weeks:
- Watery diarrhea
- Abdominal cramps and pain
- Nausea, vomiting, or both
- Occasional muscle aches or headache
- Low-grade fever.

- Bacteria E is affecting the reservoir which supplies the people who live in the East Zone (like you do). It is expected to kill / infect 10 people who live in the East Zone.
- Bacteria W is affecting the reservoir which supplies the people who live in the West Zone. It is expected to kill / infect 20 people who live in the West Zone.
- Bacteria B is affecting both reservoirs. It is expected to kill / infect 16 people - 8 who live in the East Zone and 8 who live in the West Zone.

The risk of dying /getting gastroenteritis from the three bacteria in the water supply is therefore 18 in 100,000 for those who live in the East Zone (like you do) and 28 in 100,000 for those who live in the West Zone.

However, treating the bacteria in the reservoirs is very expensive and due to cost constraints the local government must decide between four water treatment options.

Please, rate each option on a scale from very poor to very good.
[Answers: Very poor, poor, satisfactory, good, very good]

Table 34. Policy Options Descriptions for Participants

Policy B	Bacteria B can be eliminated from the water supply in the East Zone and the West Zone. 16 people's lives would therefore be saved / 16 fewer people will have gastroenteritis over the next 10 years – 8 lives / people from each zone. This means that those who live in the East Zone (like you do) would have their risk of dying / getting gastroenteritis from the bacteria in the water supply reduced from 18 in 100,000 to 10 in 100,000. Those who live in the West Zone would have their risk of dying / getting gastroenteritis from the bacteria in the water supply reduced from 28 in 100,000 to 20 in 100,000.
Policy E	Bacteria E can be eliminated from the water supply in the East Zone. 10 people's lives would therefore be saved / 10 fewer people will have gastroenteritis over the next 10 years – all from the East Zone. This means that those who live in the East Zone (like you do) would have their risk of dying / getting gastroenteritis from the bacteria in the water supply reduced from 18 in 100,000 to 8 in 100,000. No one who lives in the West Zone would benefit from this option – their risk of dying / getting gastroenteritis from the bacteria in the water supply would be unchanged at 28 in 100,000.
Policy W	Bacteria W can be eliminated from the water supply in the West Zone. 20 people's lives would therefore be saved / 20 fewer people will have gastroenteritis over the next 10 years – all from the West Zone. This means that those who live in the West Zone would have their risk of dying from the bacteria in the water supply reduced from 28 in 100,000 to 8 in 100,000. No one who lives in the East Zone (like you do) would benefit from this option – your/ their risk of dying from the bacteria in the water supply would be unchanged at 18 in 100,000.
Policy X	All three bacteria which affect the water supplies in the East Zone and West Zone can be treated. Although the bacteria would not be completely eliminated by this treatment, 7 people's lives will be saved / fewer people will have gastroenteritis over the next 10 years – 3 lives / people from the East Zone and 4 lives / people from the West Zone. This means that those who live in the East Zone (like you do) would have their risk of dying / getting gastroenteritis from the bacteria in the water supply reduced from 18 in 100,000 to 15 in 100,000. Those who live in the West Zone would have their risk of dying / getting gastroenteritis from the bacteria in the water supply reduced from 28 in 100,000 to 24 in 100,000.

C.2. Variable Specification

Table 35. Main Effects of Choice Scenario: Values

Choice	MAIN EFFECTS OF CHOICE SCENARIO		
	'Efficiency'	**'Equity'**	**'Benefit East'**
B vs. W	4	-8	-20
E vs. W	10	-10	-10
X vs. W	13	-3	-19
E vs. B	6	-2	10
X vs. B	9	5	1
X vs. E	3	7	-9