Deciphering the Noise: The Welfare Costs of Noisy Behavior

Aleksandr Alekseev, Glenn W. Harrison, Morten Lau and Don Ross*

March 2019

Abstract

Theoretical work on stochastic choice mainly focuses on the sources of choice randomness, and less on its economic consequences. We attempt to close this gap by developing a method of extracting information about the monetary costs of noise from structural estimates of preferences and choice randomness. Our method is based on allowing a degree of noise in choices in order to rationalize them by a given structural model. To illustrate the approach, we consider risky binary choices made by a sample of the general Danish population in an artefactual field experiment. The estimated welfare costs are small in terms of everyday economic activity, but they are considerable in terms of the actual stakes of the choice environment. Higher welfare costs are associated with higher age, lower education, and certain employment status.

Keywords: stochastic choice, choice under risk, welfare costs, behavioral welfare economics

JEL codes: D61, D81, C93

*Economic Science Institute, Chapman University, USA (Alekseev); Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA (Harrison); Copenhagen Business School, Denmark (Lau); School of Sociology, Philosophy, Criminology, Politics and Government, University College Cork, Ireland; School of Economics, University of Cape Town, South Africa; and Center for Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA (Ross). Harrison is also affiliated with the School of Economics, University of Cape Town. E-mail contacts: alekseev@chapman.edu, gharrison@gsu.edu, mla.eco@cbs.dk and don.ross@uct.ac.za. We are grateful to the Danish Social Science Research Council (Project #12-130950) for financial support.

1 Introduction

Stochastic choice has become an active area of research in recent years, motivated primarily by two considerations. First, a large body of empirical evidence shows that stochastic choice is a robust empirical phenomenon,[1] and much work has been devoted to explaining this behavior.[2] Second, models of stochastic choice provide researchers with econometric tools to estimate structural models in a broad range of applications. The primary interest in applying a model of stochastic choice is to recover the structural parameters of the deterministic part of a model, such as risk or time preferences. Little attention has been given, however, to the systematic economic interpretation of the parameter estimates of the stochastic part, which determine the magnitude of choice randomness. The interpretation of these parameters is important for understanding the economic value of choice randomness, which has implications for the quality of decision making, and also for a better understanding of the underlying "source" models of stochastic choice. We study the economic consequences of stochastic choice by developing an intuitive method of translating the estimates of the stochastic part into economically tractable terms.

Consider a generic structural model of discrete choice that uses a standard multinomial logit model of stochastic choice,[3] which assigns each discrete alternative a choice likelihood $\mathbb{P}$ according to

$$\mathbb{P}(a \mid \beta, \mu) = \frac{\exp(U(a \mid \beta)/\mu)}{\sum_{a' \in A} \exp(U(a' \mid \beta)/\mu)}. \tag{1}$$

In this expression, a and a' are alternatives, such as lotteries or dated outcomes, from a set of all alternatives A. The deterministic part of this structural model is parametrized by a vector of behavioral parameters β, which could represent, for instance, an agent's risk or time preferences. For example, in the case of risk preferences, β could be a risk

[1] Nogee and Mosteller (1951) provide the earliest evidence of stochastic choice, followed by Tversky (1969), Starmer and Sugden (1989), Camerer (1989), and Ballinger and Wilcox (1997).

[2] Wilcox (2008) provides an excellent overview of many popular stochastic models of choice under risk. Recent examples include Swait and Marley (2013), Wallin, Swait, and Marley (2018), Matêjka and McKay (2015) and Agranov and Ortoleva (2017).

[3] Also known in the literature as the strong utility model or the Fechnerian model.

aversion parameter and U could be the expected utility of a risky alternative; in the case of time preferences, such as the quasi-hyperbolic discounting model, β would comprise the exponential and hyperbolic discounting parameters and U would be the discounted utility of an income stream. The behavioral parameters determine the aggregate utility function U assigned to each alternative.

The stochastic part of the model is parametrized by μ, often called the *noise* parameter.[4] The noise parameter determines how sensitive choice likelihoods are to the maximization of utility U according to a given structural model. As noise tends to zero, an agent will almost surely choose the alternative with the highest utility. When noise goes to infinity, the agent will assign equal likelihoods to choosing each alternative regardless of their utilities. Higher values of μ thus imply a higher magnitude of choice randomness in this popular specification.

Three issues arise with the interpretation of the estimates of the noise parameter. First, while the effect of μ on choice likelihoods is clear, one cannot readily interpret a particular estimate of noise in economic terms.[5] A monetary value assigned to a noise estimate, on the other hand, would provide clear information about the economic consequences of choice randomness. Second, since the noise parameter is unbounded from above, it is difficult to judge whether the randomness of an agent's choices is high or low. A value defined on the unit interval would solve this problem.[6] Third, the raw estimates of μ are not well suited for interpersonal comparisons, since behavioral parameters β also change across people. Having choice randomness expressed in common units, such as money, and taking into account the interpersonal differences in β would help to overcome this issue. Aspects of these three issues

[4] In the game theory literature on Quantal Response Equilibrium due to McKelvey and Palfrey (1995), which applies stochastic choice to strategic settings, it is common to use an alternative parametrization $\lambda \equiv 1/\mu$.

[5] In the existing literature (von Gaudecker *et al.*, 2011; Bland, 2018), an estimate of noise is sometimes interpreted as the likelihood of choosing the best alternative (among the two available) for a given difference in utilities (or certainty equivalents) between them. While this number is informative of the economic consequences of choice randomness, it does not provide a monetary measure of the welfare costs associated with stochastic choice.

[6] The parameter of the tremble model of stochastic choice (Harless and Camerer, 1994) has this property and thus allows one to evaluate the relative magnitude of choice randomness. However, an estimate of the tremble parameter would still require an economic interpretation. See Carbone and Hey (2000) for a comparison between the tremble model and the Fechnerian model.

arise not only in the standard multinomial logit model but also in its modifications, such as the contextual utility model of Wilcox (2011) or specifications that substitute the utilities of alternatives for their certainty equivalents, such as von Gaudecker *et al.* (2011).

We address these issues by converting an estimate of μ into two intuitive measures.[7] The first measure, absolute welfare cost (AWC), puts a dollar value on choice randomness. It shows how much money, in certainty equivalent terms, an agent would be allowed to "waste" compatibly with rationalization of her choices by an underlying structural model.[8] The second measure, relative welfare cost (RWC), scales the absolute welfare cost by the monetary value at stake in a choice context. The relative welfare cost is thus defined on the unit interval. It shows what proportion of the total monetary value at stake an agent would be allowed to waste compatibly with rationalization of her choice by the model.[9]

Our approach rests on a careful interpretation of the concepts of "noise" and "waste." We follow the descriptive, structural literature on risk preferences by assuming a specific model of the manner in which choice randomness is rationalized. In the language of Infante, Lecouteux, and Sugden (2016, p. 21), this is

> ...not an inference about the hypothetical choices of the client's inner rational agent, but rather a way of *regularising* the available data about the client's preferences so that it is compatible with the particular model of decision-making that the professional wants to use. Regularisation in this sense is almost always needed when a theoretical model comes into contact with real data.

In our case the subject being evaluated is the "client," and we are the "professional." Thus we consistently use the expression "noise," or some synonym, rather than "error." When it comes to us using this regularised model of the agent, we may then adapt the "inten-

[7] While the discussion below focuses on the multinomial logit model and its modifications, a similar logic can be applied to other models of stochastic choices, such as the trembles model (Harless and Camerer, 1994) or the random preferences model (Loomes and Sugden, 1995; Gul and Pesendorfer, 2006).

[8] While our discussion focuses on individual decision-making, our method can also be used to study stochastic choice in group decision-making (Bone, Hey, and Suckling, 1999).

[9] Other ways to measure the welfare costs of stochastic choice might exist, however we find that using monetary measures based on certainty equivalents to be intuitive and transparent. It might be the case that, depending on a particular research question, one might be more interested in an absolute measure than a relative measure, or *vice versa*. Our goal here is to provide the general tools, which can then be adapted to a particular research question.

tional stance" towards the evaluation of an agent's behavior, using a philosophical approach developed by Dennett (1987), theoretically interpreted for use in economics by Ross (2014, ch. 4), and explicitly applied to behavioral welfare economics by Harrison and Ross (2018, § 5). This perspective, which has become the dominant one in the philosophy of psychology, emphasizes that preferences and beliefs are not fixed internal states of people, but are rationalizations of choice behaviors that people rely on to interpret one another. This applies *mutatis mutandis* to self-interpretation. Preference and belief attributions pick out "real patterns" in choice behaviors (Dennett, 1991), and these patterns, which typically involve some noise, are the basis for assessing people's goals, and hence, for economics, their welfare. Only then can we use the expression "waste." Similarly, when we characterize behavior as being "imperfectly rational" below, that also reflects our intentional stance, rather than a claim that the agent has made an error in cognitive processing or problem representation.

Our measures of the welfare costs of noisy behavior are consistent with the model-based approach advocated by Manzini and Mariotti (2014). This means that in order to calculate our welfare cost measures, we assume specific deterministic and stochastic models of the decision-making process. These assumptions allow us to derive precise (in the sense of being point estimates) and efficient (in the sense of efficiently using available data, explained in Section 2.3) values of welfare costs. We recognize the potential sensitivity of our results to these assumptions, and address them in Appendix A.

Our absolute and relative welfare cost measures allow one to conveniently evaluate the economic significance of choice randomness, its relative magnitude, and to compare the magnitude of choice randomness across people. The implications of these measures for an agent's behavior, however, will ultimately depend on the underlying model of the source of choice randomness adopted by a researcher. This is an important point since different "source" models of stochastic choice often lead to the same choice likelihoods, such as the likelihoods generated by the multinomial logit model presented above. For instance, the Random Utility model due to Marschak (1960) assumes that when an agent makes an optimal choice, the choice randomness is due to the perturbations in her utility function that are unobservable to a researcher. The noise parameter in the Random Utility model is then proportional to the variance of the unobserved component of utility. High estimated welfare costs would imply that the stochastic part of the structural model dominates the deterministic part, i.e., the structural model cannot explain the agent's choices well. The welfare costs can then be viewed as measures of a model's fit.[10]

Some studies, such as von Gaudecker *et al.* (2011), interpret choice randomness induced by the Random Utility model as behavioral "mistakes." This interpretation usually arises in experiments on choice under risk. In these experiments, all the decision-relevant features, such as payoffs and probabilities, are observable to both researcher and subject, while the time frames are small enough to rule out meaningful changes in preferences. Provided that the assumed structural model is close to the true data-generating process, the estimates of welfare costs can then be interpreted as the magnitude of behavioral "mistakes." In practice it is likely that the model will be misspecified by a researcher (e.g. an Expected Utility model is estimated while the true model is Rank Dependent Utility) in addition to any potential behavioral "mistakes" a subject might make. The welfare costs will then capture both a model's fit and behavioral "mistakes." The correct interpretation of the estimates of welfare costs in this case is an upper bound on actual behavioral "mistakes."

Recent studies offer an alternative view on choice randomness as an optimal response to costly frictions in the decision-making process. For example, these frictions may be caused by the need to collect the relevant information to make a choice, as in Rational Inattention models of Caplin and Dean (2015) and Matêjka and McKay (2015). The noise parameter in a Rational Inattention model represents marginal information costs. The estimates of welfare costs in this type of models can then be interpreted as aggregate information costs, or losses that an agent incurs relative to an ideal case of no information costs. Another example of frictions is the pursuit of multiple goals that cannot be obtained simultaneously

[10] Recent work by Halevy *et al.* (2018) provides a promising example of how welfare costs can be used as a measure of fit.

(Swait and Marley, 2013; Wallin *et al.*, 2018). An agent is assumed to balance the goal of choosing the best available alternative with the goal of having diversity in choices. Noise parameters in this model represent the relative weight of the second goal. The estimates of welfare costs in this type of models can be interpreted as the economic value that an agent places on the goal of having diversity or, alternatively, as the loss an agent incurs relative to a case of having a single goal of choosing the best alternative.

We apply our method to the data from an artefactual field experiment in Denmark. The subjects came from a sample of the general Danish population and were asked to make a series of choices between two risky alternatives. Each subject answered a detailed demographic survey, which we use to characterize the effects of demographic characteristics on the observed heterogeneity in the AWC and RWC. We find that the average AWC are around 67 Danish kroner (\$10)[11] and thus negligible for the subjects' natural economic environment. However, the RWC are quite significant, at 0.87 on average. There is also considerable variation among the subjects in terms of their AWC and RWC. Regression analysis shows that certain demographic characteristics are associated with higher costs. In particular, subjects who are older, less educated, and have a particular employment status, have larger welfare costs. Females have higher AWC than males, but do not differ in RWC.

Section 2 describes the method of converting an estimate of noise into welfare costs measured in monetary terms and provides an explicit algorithm for computation in a binary choice case. Section 3 applies the method to data from an artefactual field experiment in Denmark involving choice under risk and studies the properties of the welfare costs, as well as their demographic correlates. Section 4 discusses connections with previous literature. Section 5 concludes.

[11] Throughout the text, we use an exchange rate of 1 Danish krone = \$0.15 that was prevalent at the time of the experiment.

2 Method

We first look at a general case when the set of alternatives is continuous. This case allows us to clearly demonstrate the logic behind our method of extracting the welfare cost information from a noise estimate. Then we turn to a more common discrete case with two alternatives and explicitly describe the algorithm to implement our method.

2.1 General Case

Consider an agent choosing from a set of alternatives indexed by real numbers on a compact interval $A = [\,a_l, a_h\,]$. Each alternative generates a lottery[12]

$$l(a) = \{x_1(a), \ldots, x_k(a);\ q_1(a), \ldots, q_k(a)\},$$
$$a \in A,\ x_i \in \mathbb{R},\ q_i \in \mathbb{R}_+,\ \forall i = 1, \ldots, k,\ \sum_{i=1}^{k} q_i = 1,$$

where x_i are monetary outcomes and q_i are respective probabilities of obtaining those outcomes.

This setting could represent allocating resources between two state-contingent accounts, as in Choi *et al.* (2007). Each allocation in this example is an alternative with $k = 2$ outcomes, $x_1(a)$ and $x_2(a)$, and equal probabilities of each outcome. The minimum ($a_l = 0$) and maximum ($a_h > 0$) amounts an agent can allocate to account 1 will define the interval of alternatives A. Then $x_1(a) = a$ and $x_2(a) = b(a_h - a)$, where $-b < 0$ is the slope of the budget line and $q_1(a) = q_2(a)$, $\forall a \in A$.[13]

The risk elicitation task of Gneezy and Potters (1997) is another example of such a setting. In this example, the minimum ($a_l = 0$) and maximum ($a_h > 0$) amounts a subjects can allocate to a risky asset define the set of alternatives A, where a_h is the initial endowment.

[12] The lottery itself does not need to be discrete. An alternative can generate a continuous probability density.

[13] In an actual experiment, the set of alternatives is, of course, discrete. This choice set, however, comes close to being continuous.

A subject's choice of how much of the endowment to allocate to a risky asset, a, generates lotteries with two outcomes given by $x_1(a) = a^h - a$ (the asset yields no return) and $x_2(a) = a^h + a(k-1)$ (the asset yields a positive return $k-1$). The probabilities of outcomes are given exogenously and do not depend on a.

Each alternative a has an aggregate utility $U(a) \equiv U(l(a))$ defined by an assumed structural model of choice under risk. Monetary outcomes are transformed using $u : \mathbb{R} \mapsto \mathbb{R}$, the von Neumann-Morgenstern utility function. Each value of $U(a)$ can be translated into a certainty equivalent $m(a)$, defined by $u(m(a)) = U(a)$. The ordering of alternatives is preserved for the certainty equivalent transformation: $U(a) \geqslant U(b) \Leftrightarrow m(a) \geqslant m(b)$, $\forall a, b \in A$.

Assume that U is concave and reaches its unique maximum (minimum) at a^* (a_*), as does the certainty equivalent function. Define the maximum certainty equivalent as $m^* \equiv m(a^*)$, and the minimum certainty equivalent as $m_* \equiv m(a_*)$. If the agent always chooses the best alternative a^*, we call this behavior *perfectly rationalizable* (by an assumed model of choice under risk). On the other extreme, if the likelihood of choosing a^* is the same as for any other alternative, we call such a behavior *non-rationalizable*. We are concerned with the behavior in between, which is neither perfectly rationalizable nor non-rationalizable, a behavior that we call *imperfectly rationalizable*.

The degree of this imperfection[14] is characterized by a number ε, $0 \leqslant \varepsilon \leqslant \Delta m$, with $\Delta m \equiv m^* - m_*$. Choices that lead to certainty equivalents within ε distance from the maximum certainty equivalent can be viewed, from the perspective of a model, as imperfectly rationalizable.[15] These choices form an optimal region A^* defined by

$$A^*(\varepsilon) = \left\{ a \in A \mid m(a) \geqslant m^* - \varepsilon \right\}. \tag{2}$$

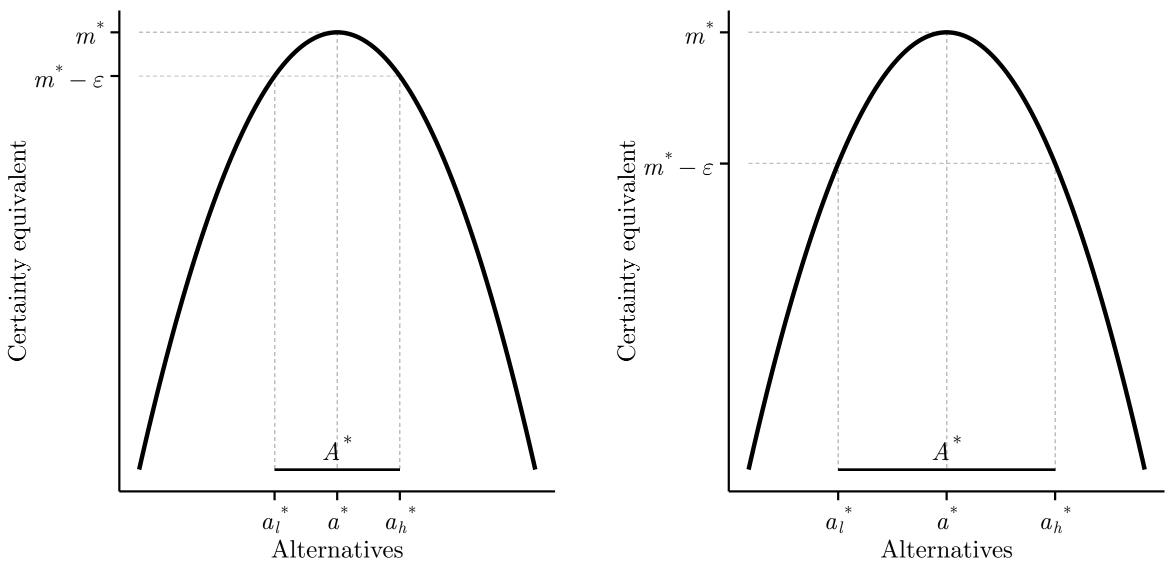
The degree of imperfection ε shows how much monetary welfare an agent would be allowed to

[14] This term should be understood as an imperfection of a given model to regularise data, rather than a statement about an agent making decision errors.

[15] The idea of allowing an agent some degree of imperfection in choices is not new. For example, Harrison (1994) introduces a similar quantity based on an agent's subjective cost of choosing one alternative versus the other to explain many EUT violations.

waste to make her choices rationalizable by the model, and effectively includes these choices in the optimal region. In other words, ε represents the welfare costs measured in monetary units. Our goal is to link these costs to noise.

The allowed degree of imperfection co-varies with the width of the optimal region. If ε is set to 0, the optimal region will consist only of the best alternative a^*. If ε is high enough, the optimal region will coincide with the whole set of alternatives A. Figure 1 illustrates how the optimal region varies with the degree of imperfection. Geometrically, the optimal region is the line segment $\left[a_l^*, a_h^*\right]$.



(a) Low Degree of Imperfection

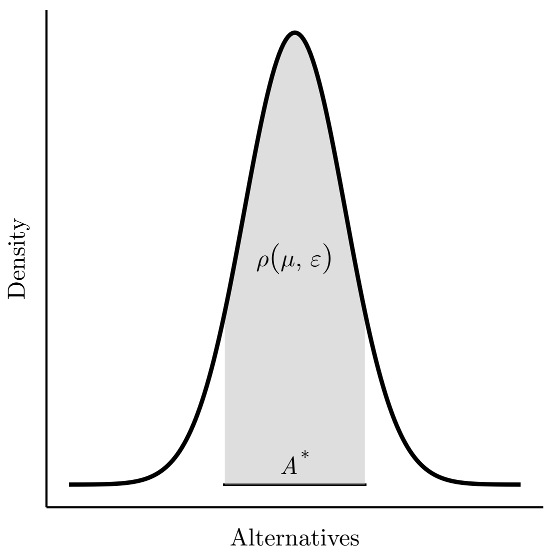
(b) High Degree of Imperfection

Figure 1: Optimal Region and Degree of Imperfection

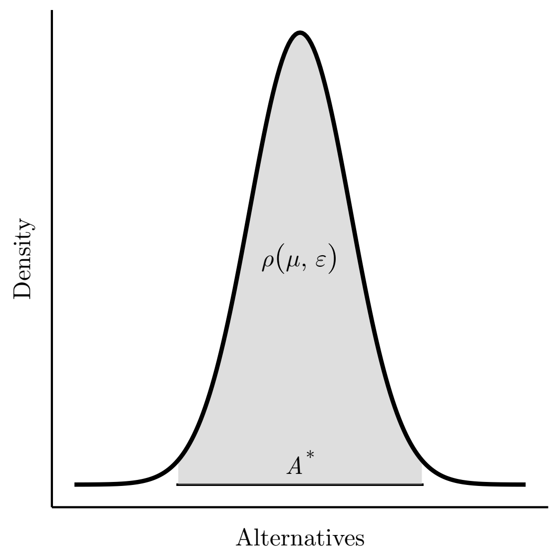
The optimal region and the degree of imperfection are the first two components that we need to interpret an estimate of noise. The third component comes from a stochastic model $p : A \mapsto \mathbb{R}_+$, which generates choice likelihoods over the set of alternatives. Some alternatives fall into the optimal region, by definition. By integrating the density $p(a)$ over this region we get the proportion of choices that are counted, from the perspective of a model, as imperfectly rationalizable for a given degree of imperfection. We call this measure a *degree of rationalizability* (DoR):

$$\rho(\mu, \varepsilon) = \int_{A^*(\varepsilon)} p(a)\, \mathrm{d}a. \tag{3}$$

The DoR has several intuitive properties, two of which turn out to be crucial for our analysis, and can be represented graphically. Figure 2 shows that as the degree of imperfection increases, the optimal region expands and the DoR, represented by the gray shaded area, increases. Figure 3 shows that as the noise goes up, the density flattens out and the probability mass shifts from the optimal region to the outside area, reducing the DoR.



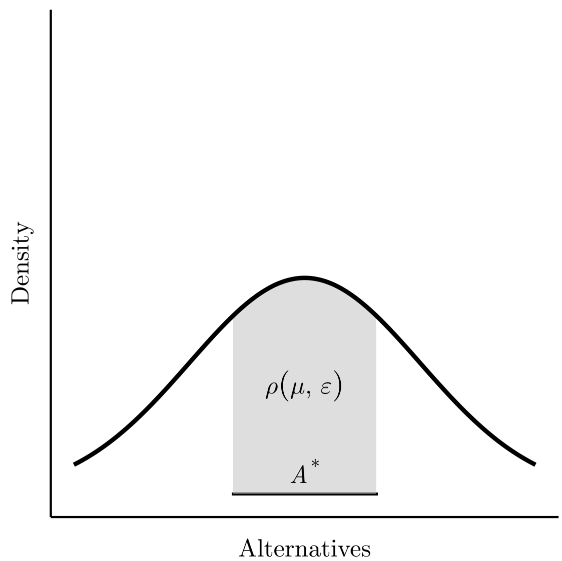
(a) Low Degree of Imperfection



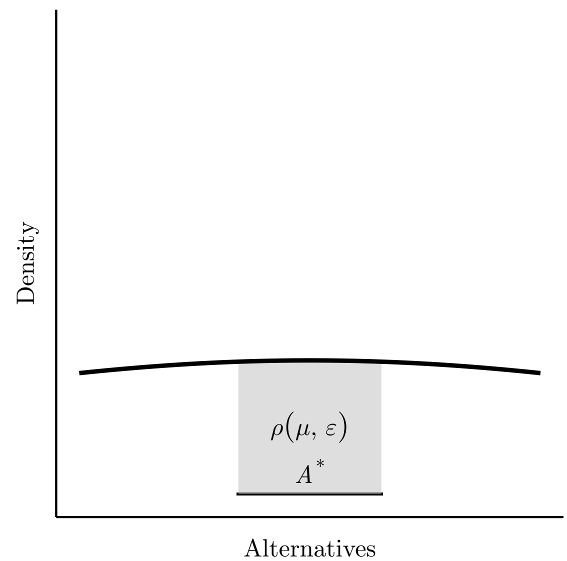
(b) High Degree of Imperfection

Figure 2: Degree of Rationalizability and Imperfection

The DoR for certain values of noise and imperfection has attractive interpretations. The quantity $\rho(\infty, \varepsilon)$ tells us what proportion of choices are counted as rationalizable for a given imperfection ε when they are, in fact, non-rationalizable. It represents a Type II error in a test to detect rationalizability, and the quantity $1 - \rho(\infty, \varepsilon)$ is the power of this test. This power will decrease as the allowed degree of imperfection increases or as the set of alternatives shrinks. The value of DoR at $\rho(\hat{\mu}, 0)$ measures the proportion of rationalizable choices for an estimated level of noise $\hat{\mu}$ and no imperfection. We refer to it as the *default* *degree of rationalizability* or DDoR.



(a) Low Noise



(b) High Noise

Figure 3: Degree of Rationalizability and Noise

We now have all the tools to decipher the noise. We do this by linking an estimate of μ, the value of which is hard to interpret, to the degree of imperfection, a monetary measure that has an intuitive economic interpretation as the welfare cost, or monetary welfare required to rationalize the agent's choices by a model. In order to link them, we need to reverse the steps we followed so far. Currently, we introduced a degree of imperfection ε that defines an optimal region A^*. The optimal region combined with a stochastic model, parameterized by μ, yields a value of DoR. Now suppose that instead we start with a DoR measure and fix it at some target level α. Let an estimated value of the noise be $\hat{\mu}$. The question is how much imperfection should be allowed for $100 \times \alpha\%$ of the choices to be rationalized for a given noise. In other words, we need to find ε that satisfies

$$\rho(\hat{\mu}, \varepsilon) = \alpha. \tag{4}$$

This equation establishes an implicit function, $\varepsilon(\hat{\mu}; \alpha)$. For the purpose of our analysis, the following property of this function is important.

Proposition 1. *For a given α, the degree of imperfection as a function of noise, $\varepsilon(\mu;\alpha)$, is monotonically increasing:*[16]

$$\frac{d\varepsilon}{d\mu} \geqslant 0.$$

Proof. See Appendix B. □

This property implies that noise and imperfection are in a direct and monotonic relation.[17] This property is important since more noise should imply higher welfare costs, which in our case are measured by imperfection. If imperfection and noise were not in a direct and monotonic relation, such an interpretation would be impossible. The relation between ε and μ comes from the fact that the DoR is decreasing in noise and increasing in imperfection. From these properties it also follows that higher values of α imply higher values of ε. The more choices we wish to rationalize, for a given value of a noise, the more imperfection we should allow. The choice of the target α is left to the discretion of a researcher. In our empirical analysis we use the values of $0.9, 0.95$, and 0.99, which appear to be reasonable targets.

So far we have focused on a single choice context, but in practice we observe agents make choices over a series of rounds of a choice task. Suppose we observe an agent's choices over n rounds indexed by $j = 1, \ldots, n$, and in each round the mapping of alternatives a into lotteries $l_j(a)$ is different. In the context of an allocation task, the variation is introduced by changing the slope of a budget line b^j and a maximum amount a_h^j that can be allocated to one of the accounts: $A_j = [0, a_h^j]$, $x_1^j(a) = a$, $x_2^j = b^j(a_h^j - a)$. We can repeat all the previous steps in deriving the DoR, but now it will differ by the round: $\rho_j(\mu, \varepsilon)$. What remains common across rounds, however, is the degree of allowed imperfection ε. We assume that μ and preferences

[16] We note that in the case when alternatives are discrete rather than continuous, as discussed below, the DoR as a function of imperfection will not be continuous and thus it will not be possible to match the target DoR α perfectly. We address this issue by using a discrete grid for ε and an interpolated version of $\rho(\hat{\mu}, \varepsilon)$. The interpolated DoR function on a discrete grid is continuous and thus Proposition 1 applies.

[17] This property holds for a given agent, or rather given risk preferences. This property will not hold perfectly across agents whose preferences are different.

remain fixed for the duration of a choice task. We can then aggregate the DoR from all the choices by averaging across the DoR for single choices:

$$\rho(\mu, \varepsilon) = \frac{1}{n} \sum_{j=1}^{n} \rho_j(\mu, \varepsilon). \tag{5}$$

Naturally, the average follows all the properties of the DoR for a single choice. In particular, it increases in ε and decreases in μ. We can then use the aggregate DoR in (5) to calculate the imperfection needed to reach a target α in equation (4).

After calculating the value of ε that satisfies equation (4), $\varepsilon(\hat{\mu}, \alpha)$, it makes sense to adjust this value to take into account the fact that the degree of imperfection should not exceed the difference between the maximum and minimum certainty equivalents for a given choice. Since a common ε is applied to all the rounds of a choice task, for some rounds it can actually exceed Δm. Increasing imperfection beyond this difference does not have any effect on the DoR and would imply that we allow an agent to waste more monetary welfare than there actually is. This issue can be addressed by bounding ε by Δm, and then averaging across all the rounds:

$$\bar{\varepsilon}(\hat{\mu}, \alpha) = \frac{1}{n} \sum_{j=1}^{n} \min \left\{ \varepsilon(\hat{\mu}, \alpha), \Delta m_j \right\}. \tag{6}$$

We call the resulting measure of imperfection *Absolute Welfare Costs* (generated by noise $\hat{\mu}$, with $100 \times \alpha\%$ of choices rationalized), or AWC. It represents the monetary welfare that the agent would be allowed to give up for exactly $100 \times \alpha\%$ of her choices to be rationalized by the model, given noise $\hat{\mu}$. For any estimated value of noise and any desired proportion of choices we would like to rationalize we can, therefore, always find a precise dollar value of the welfare costs.

We can go further and translate the welfare costs into relative terms, to compare these costs with the actual stakes of a choice context. For example, an AWC of \$1 may not look like much, but if Δm_j are close to \$1 in all the rounds, almost all the welfare would have to be sacrificed to rationalize an agent's choices. We divide the degree of imperfection by the difference between the maximum and minimum certainty equivalents for every round, and average across all the choices:[18]

$$\tilde{\varepsilon}(\hat{\mu},\alpha) = \frac{1}{n}\sum_{j=1}^{n} \min\left\{\frac{\varepsilon(\hat{\mu},\alpha)}{\Delta m_j}, 1\right\}. \tag{7}$$

The resulting degree of imperfection represents *Relative Welfare Costs* (generated by noise $\hat{\mu}$, with $100 \times \alpha\%$ of choices rationalized), or RWC. Another benefit of this measure is that it allows one to appreciate the relative magnitude of noise, since RWC are bound between 0 and 1, while a raw estimate of noise is unbounded from above. If rationalizing $100 \times \alpha\%$ of the choices requires on average almost all the difference between the maximum and the minimum certainty equivalents, in which case RWC are close to 1, that clearly indicates that the choices are close to being non-rationalizable, from the perspective of the model. On the other hand, if it requires only a small fraction of this difference, in which case RWC are near 0, then choices are close to being perfectly rationalizable, from the perspective of the model.

2.2 Binary Choice

An important special case arises when an agent has only two alternatives to choose from. This is one of the most common experimental designs in risk elicitation tasks.[19] In this case the set of alternatives in each round is $A = \{a_1, a_2\}$. Without loss of generality, assume that alternative a_2 always gives the highest utility, so that $U_j(a_2) > U_j(a_1)$, $j = 1, \ldots, n$, i.e., $a_j^* = a_2$, using the notational convention $U_j(a) \equiv U\left(l_j(a)\right)$. The maximum and the minimum certainty equivalents in each round j are $m_j^* \equiv m_j(a_2)$ and $m_{j*} \equiv m_j(a_1)$, respectively. The

[18] Since the resulting quantity has to be a fraction, we bound this ratio by 1.

[19] For example, the risk elicitation tasks developed and popularized by Hey and Orme (1994) and Holt and Laury (2002) apply to the binary choice case.

optimal region and the DoR can then take only two values:

$$A_j^*(\varepsilon) = \begin{cases} a_2, & \varepsilon < \Delta m_j, \\ A, & \varepsilon \geqslant \Delta m_j, \end{cases} \qquad \rho_j(\mu, \varepsilon) = \begin{cases} p_j(a_2), & \varepsilon < \Delta m_j, \\ 1, & \varepsilon \geqslant \Delta m_j, \end{cases} \tag{8}$$

where $p_j(a_2)$ is the likelihood of choosing alternative a_2 in round j.

Suppose we observe a series of binary choices made by a subject and estimate a structural model of risk preferences in which $\hat{\gamma}$ is a vector of estimated risk parameters and $\hat{\mu}$ is an estimate of noise. The $\hat{\gamma}$ vector in the Expected Utility Theory (EUT) case is typically just the relative risk aversion. In the case of Cumulative Prospect Theory (CPT) $\hat{\gamma}$ includes the risk aversion parameter(s), the probability weighting parameter(s), and the loss aversion parameter.[20] The computation of AWC and RWC (rationalizing $100 \times \alpha\%$ of the choices) from these data can be performed using the following algorithm.

1. For each round, compute the aggregate utilities of both alternatives, $U_j(a_1; \hat{\gamma}), U_j(a_2; \hat{\gamma})$, $j = 1, \ldots, n$.

2. Compute the certainty equivalents of both alternatives, $m_j(a_1), m_j(a_2)$, using the inverse transformation, $m_j(a) = u^{-1}(U_j(a; \hat{\gamma}); \hat{\gamma}), a \in A$, and the difference between them, Δm_j.

3. Compute the likelihoods of each alternative using the stochastic model, $p_j(a; \hat{\gamma}, \hat{\mu})$, $a \in A$.

4. Start with $\varepsilon = 0$. Compute the DoR in each round $\rho_j(\hat{\mu}, \varepsilon)$ using (8). Compute the aggregate DoR $\rho(\hat{\mu}, \varepsilon)$ using (5).

5. If $\rho(\hat{\mu}, \varepsilon) < \alpha$, increase ε by a small number $\Delta\varepsilon > 0$.

[20] The parametrization will also depend on the utility and probability weighting functions used. For example, if an expo-power utility function is used, it will have two parameters rather than one.

6. Repeat Step 5 until the aggregate DoR reaches the target level of α.[21]

7. Compute the AWC $\bar{\varepsilon}(\hat{\mu}, \alpha)$ using (6). Compute the RWC $\tilde{\varepsilon}(\hat{\mu}, \alpha)$ using (7).

2.3 Alternative Measures

Note that the proposed computation of welfare costs does not involve actual choices. After estimating risk parameters and noise, we ignore whether the actual choices corresponded to the maximum certainty equivalent or not. A question then arises: what choices do we rationalize, if not actual choices? This question also suggests an equivalent computation based on actual choices rather than on likelihoods.

Consider the following alternative algorithm. Start by computing the implied (by the model) decisions based on certainty equivalents. Compare actual and implied decisions by looking at the proportion of times when implied and actual decisions coincide. This proportion gives the actual default DoR. Next, calculate the vector of the differences in the certainty equivalents (CE differences) for the cases when implied and actual decisions disagree. These are the "mistakes," from the perspective of the model, we need to "correct," or regularise by adding a structural model of behavioral noise. Start with $\varepsilon = 0$ and increase it by a small positive amount. When ε is lower than the CE difference, the DoR in that round is zero; otherwise it equals one, meaning that implied and actual decisions become equivalent. After that, compute the relative proportion of times when rationalized decisions coincide with the actual ones. Increase ε until this proportion reaches the target level. Compute the average of the bounded (by CE difference) ε for the absolute *actual* welfare costs, and the average of their ratios to CE differences for the relative *actual* welfare costs.

Although this alternative algorithm is almost identical to the previous algorithm, there is a subtle difference. This difference makes us choose in favor of the method described in §2.2, which involves rationalizing *potential* choices, as opposed to *actual* ones. Consequently,

[21] In practice, due to the discreteness of $\rho(\hat{\mu}, \varepsilon)$ it will often be impossible to match the target α exactly. We use linear interpolation to handle this issue.

we obtain the estimates of the potential welfare costs, while the alternative method would give us the actual welfare costs. The key difference between the two methods lies in the fact that the likelihoods of choices represent what could have been chosen if the same options were presented many times. We view the approach of using potential choices as extracting more information from the same data points. The informational gain is obtained through the introduction of a particular structure that describes the choice likelihoods.

Of course, if the two methods gave completely different estimates, one would need a stronger argument in favor of one method against the other. Comparing the potential and actual welfare costs, however, shows that the measures are tightly associated in practice (not reported here). In principle, one could easily substitute one method for another.

Another alternative method of computing the absolute welfare costs would arise if we reconsidered equation (6), which involves bounding the value of imperfection by the difference in certainty equivalents. This is not required and we could, as well, have computed the *unbounded* absolute welfare costs.[22] One might expect that we would obtain higher estimates of the AWC in that case. Indeed, our calculations show (not reported here) that the unbounded AWC are on average twice as large as the bounded AWC, and both measures are tightly associated. We prefer to use the bounded measure, however, since it represents only the welfare costs that can be potentially incurred, while the unbounded measure allows wasting more monetary welfare than there actually is.

3 Empirical Analysis

3.1 Data

We present the results for 218 adult Danes, a subsample of a larger field study by Harrison, Jessen, Lau, and Ross (2018). The subjects for the original study were recruited from two

[22] The computation of the RWC must involve bounding, since they represent a fraction that must lie in the unit interval.

internet-based panels with 165,000 active members combined. The sample frame consisted of 65,592 adult Danes between ages 18 and 75. The sample was stratified by sex and age across three regions of Denmark: greater Copenhagen, Jutland, and Funen and Zealand.[23] The completed sample consisted of 8,405 respondents, or 12.8% of the sample frame. Invitations were sent out by email, and the subjects could participate in a survey using internet-browsers on their computers or mobile devices. The experiment was implemented as an artefactual field experiment (Harrison and List, 2004).

Table 1 provides a summary of the socio-demographic characteristics of our subsample who were invited to participate in an experiment after completing the online survey. Slightly less than half of the sample were females and the average age was just less than 50 years. The majority of the sample had college education, and the distribution of income across different income brackets was roughly equal. Most of the participants were either employed as public servants or retired. More than 75% of our sample comes from the Greater Copenhagen area.

The subjects made binary choices across 60 pairs of lotteries and answered a set of demographic questions. Once all the lottery choices were made, one of the choices was selected randomly for payoff. Table C.1 in the Appendix contains the battery of lotteries that were given to the subjects. This battery is based on designs by Loomes and Sugden (1998), Wakker, Erev, and Weber (1994) and Cox and Sadiraj (2008). Together they provide a powerful test of EUT and RDU.

3.2 Estimation Procedure

Computation of the welfare costs relies on structural estimates of risk preferences γ and noise μ. We implement the estimation in the standard fashion by maximizing the Bernoulli

[23] Greater Copenhagen area was assigned a weight of 50%, and the other two regions were assigned equal weights of 25%.

Table 1: Socio-Demographic Characteristics of the Sample

Characteristic	Mean
Female	0.46
Age	48.06
Education	
Vocational training	0.19
Low level of formal education	0.21
College, less than 3 years	0.09
College, 3 to 4 years	0.27
College, 5 or more years	0.24
Annual household income, before tax	
Less than 300k DKK	0.23
300k–500k DKK	0.23
500k–800k DKK	0.23
More than 800k DKK	0.17
Not reported	0.14
Occupation	
Public servant	0.42
Student	0.12
Unemployed	0.04
Retired	0.23
Skilled worker	0.03
Unskilled worker	0.06
Self-employed	0.06
Other	0.04
Family	
Has children	0.25
Lives with a partner	0.54
Geographic area	
Copenhagen	0.78
Central Denmark	0.07
Zealand	0.09
Southern Denmark	0.06

log-likelihood function at the level of a subject:[24]

$$(\hat{\gamma}, \hat{\mu}) = \arg\max_{\gamma,\mu} \sum_{j=1}^{n} \big(y_j \ln p_j(a_2;\gamma,\mu) + (1-y_j) \ln p_j(a_1;\gamma,\mu) \big) \, ,$$

where $y_j \equiv \mathbb{I}(a = a_2)_j$ is an indicator variable that takes a value of 1 whenever an alternative a_2 is chosen in round j. The alternative a_2 is taken to be the one on the right side of the screen without loss of generality, and we no longer assume that it gives the highest aggregate utility in all the rounds.

We assume that the choice probability $p_j(a_2;\gamma,\mu)$ is given by the strong utility model in the logit form[25]

$$p_j(a_2;\gamma,\mu) = \frac{\exp\big(U_j(a_2;\gamma)/\mu\big)}{\exp\big(U_j(a_2;\gamma)/\mu\big) + \exp\big(U_j(a_1;\gamma)/\mu\big)} = \Lambda\left(\frac{U_j(a_2;\gamma) - U_j(a_1;\gamma)}{\mu}\right),$$

where $\Lambda(\cdot)$ denotes the logistic cumulative density function, and $p_j(a_1;\gamma,\mu) = 1 - p_j(a_2;\gamma,\mu)$.

We also assume that the lotteries are compared according to their expected utilities (dropping an index for the round)

$$U(a;\gamma) = \sum_{i=1}^{k} q_i(a) u(x_i(a);\gamma),$$

and the u function takes the constant relative risk aversion form:

$$u(x;\gamma) = \frac{x^{1-\gamma}}{1-\gamma}.$$

[24] We find that the estimation procedure successfully converges for 183 subjects (84% of the sample). For the rest of our subjects, the estimation procedure terminates after a number of iterations and yields the best parameter values at the time of termination. The results in subsequent sections are reported for the full sample of subjects. Using only the subset of subjects with successful convergence yields quantitatively similar results, e.g., compare Figure 4 that uses the full sample and Figure D.7 (in Appendix) that uses the subset of subjects.

[25] One could alternatively use certainty equivalent functions m instead of the aggregate utility functions U in the specification of the stochastic model, as in Bruhin *et al.* (2010) or von Gaudecker *et al.* (2011). Using this alternative specification only changes the scale of the estimates of the noise parameter, and does not change the estimates of risk parameters or the estimated likelihoods of choosing each alternative. The algorithm for computing welfare costs and their magnitude would remain the same.

Nothing in our approach relies on assuming an EUT model or a strong utility model in the logit form. In fact, we could have proceeded in a way suggested by Harrison and Ng (2016) and estimated different models for different subjects, classifying our subjects as EUT or RDU, for example. Alternatively, as suggested by Monroe (2017), we could have assumed an RDU model for all the subjects, since correct classification has significant data requirements. Appendix A demonstrates this important generalization by regenerating all results assuming an RDU model of risk preferences, as well as assuming a different utility function, the expo-power utility function, and a different model of stochastic choice, the contextual utility model of Wilcox (2011).

3.3 Welfare Costs

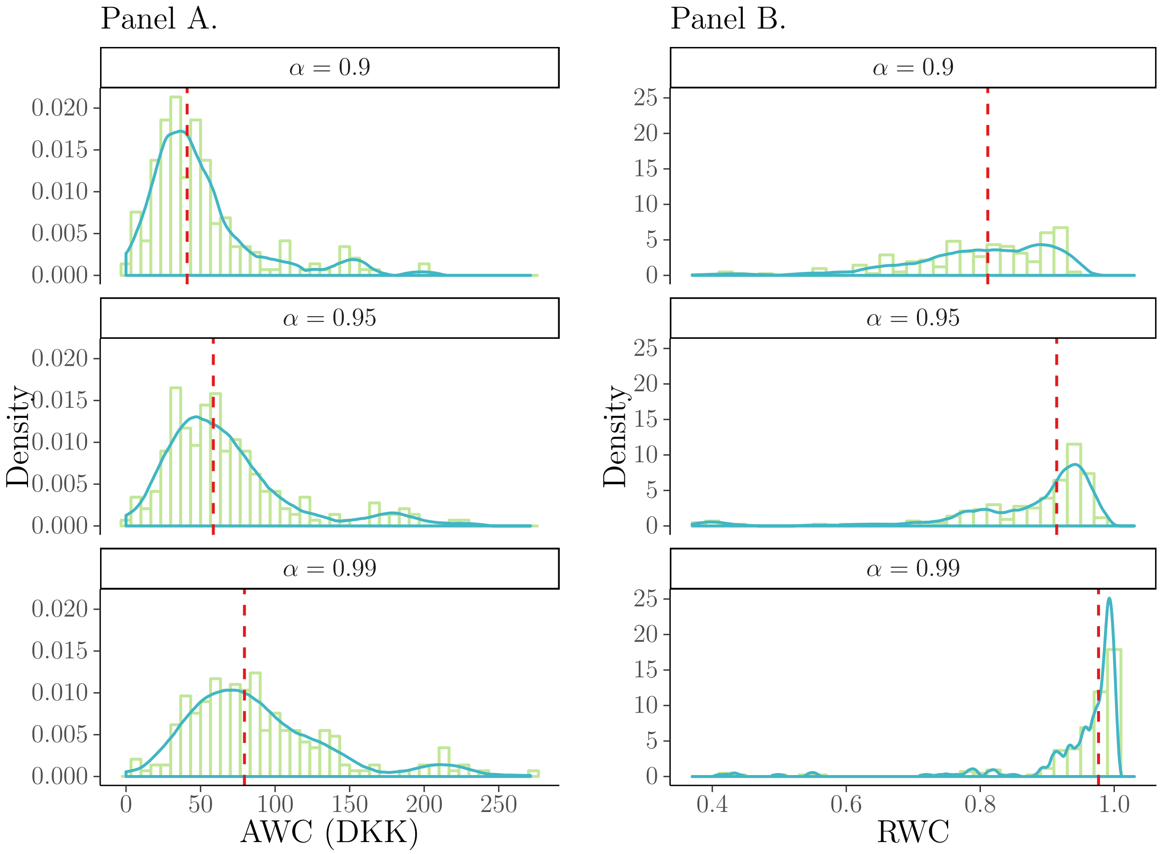
Figure 4 (Panel A) shows the distribution of the individual-level estimates of AWC (in Danish kroner, DKK) and Table 2 (Panel A) presents their summary statistics.[26] The distribution of the AWC is composed of two clusters: a major cluster on the left end and a minor cluster on the right end of the support, so that overall the distribution is right-skewed. The major cluster is bell-shaped and fairly symmetric. The minor cluster appears to be bell-shaped but the number of observation in it is small. As the target DoR increases, the distribution of the AWC flattens out and slides to the right end of the support.

The AWC are, on average, quite modest in size. For $\alpha = 0.95$, the mean AWC are only 66.96 DKK (10.04 USD) and the median AWC are even smaller, 58.56 DKK (8.78 USD). For 50% of the subjects, the AWC lie within 38.37 DKK (5.76 USD) and 80.76 DKK (12.11 USD) at this level of DoR. As the target level of DoR increases, the mean AWC also increase, as expected. For $\alpha = 0.99$, the mean AWC reach 88.66 DKK (13.3 USD), and the median AWC reach 79.44 DKK (11.92 USD).

There is substantial variation among subjects in their AWC. At $\alpha = 0.95$ the smallest

[26] As discussed in §2.3, we rationalize the *potential* choices, which allows us to use a fine grid for the target DoR. If we were rationalizing the *actual* choices instead, we would have to deal with the target DoR's that are fractions of 60 (the number of choice pairs in the experiment).

Figure 4: Distributions of AWC and RWC in the Sample



Note: The graph shows the distributions of the individual-level estimates of AWC (Panel A) and RWC (Panel B) for three target levels of α: 0.9, 0.95, and 0.99. The bars are the histograms and the smooth lines are the kernel density estimates. The dashed lines show the medians of the distributions. The AWC numbers are in DKK. For the RWC, we truncate the support at 0.4 to improve readability of the graph. This results in dropping 11 observations for which the RWC are below 0.4.

AWC are just 1.24 DKK (0.19 USD) while the maximum AWC are 224.23 DKK (33.63 USD), which is roughly 3 times as large as the mean AWC. The standard deviation of AWC at this level of DoR is 42.2 DKK (6.33 USD). The variation in AWC increases as the target DoR goes up, which is reflected in higher standard deviations and higher ranges. At $\alpha = 0.99$ the minimum AWC are still tiny, just 4.56 DKK (0.68 USD), while the maximum AWC become 271.3 DKK (40.69 USD), which is again roughly 3 times as large as the mean AWC at this level of DoR. The standard deviation reaches 48.05 DKK (7.21 USD) at this level of DoR.

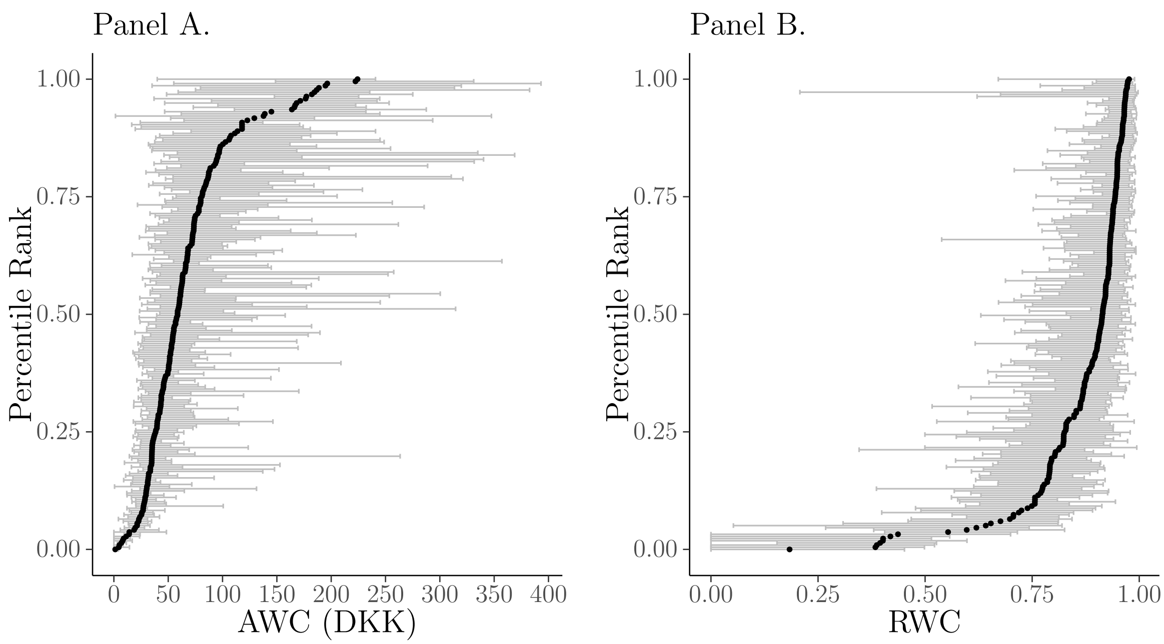
Table 2: Summary Statistics for AWC and RWC

α	Mean	SD	Min	Q1	Median	Q3	Max
Panel A. *AWC (DKK)*							
0.9	50.10	36.60	0	27.60	41.00	58.90	200.00
0.95	67.00	42.20	1.24	38.40	58.60	80.80	224.00
0.99	88.70	48.00	4.56	56.70	79.40	110.00	271.00
Panel B. *RWC*							
0.9	0.77	0.15	0	0.73	0.81	0.89	0.93
0.95	0.87	0.13	0.18	0.83	0.91	0.94	0.98
0.99	0.95	0.09	0.41	0.95	0.98	0.99	1.00

Notes: The table reports the summary statistics for the three samples of the individual-level estimates of AWC and RWC computed at different target levels of DoR: 0.9, 0.95, and 0.99. The AWC numbers are in DKK. RWC are measured as proportions.

While the variation in the AWC between subjects is substantial, there is also considerable uncertainty at the individual level. Figure 5 (Panel A) shows the point estimates of the AWC (for $\alpha = 0.95$) for each subject along with the 95% confidence interval around those estimates. The confidence intervals are computed using bootstrap methods. We rank subjects based on their point estimates of AWC. The vertical axis represents the percentile rank of each subject. The uncertainty in the individual-level estimates of AWC stems from the combined uncertainty in the estimates of risk aversion and noise and tends to increase with the percentile rank.

Figure 5: Uncertainty in the Individual-Level Estimates of AWC and RWC



Note: The graph shows the point estimates and the confidence intervals for AWC (Panel A) and RWC (Panel B) computed at $\alpha = 0.95$. Each point on the graph represents an individual-level estimate for a given subject. The estimates are ranked from lowest to highest, and the percentile rank of each subject is shown on the vertical axis. The horizontal error bars show the bootstrapped 95% confidence intervals around the point estimates.

Figure 4 (Panel B) shows the distribution of the individual-level estimates of RWC and Table 2 (Panel B) presents their summary statistics. At $\alpha = 0.9$ the distribution is very flat and has a long left tail resulting in a negative skew. As the target DoR increases, the distribution of the RWC shifts to right and becomes more concentrated while preserving a long left tail. The distribution features some observations on the left tail with unusually low RWC. For some of these outcomes, RWC are less than a half, even for the highest level of α.

In contrast to the AWC, the RWC are extremely high, which implies that while the AWC are modest in size, these costs represent a substantial portion of the monetary welfare available in the choice environment. For $\alpha = 0.95$, the mean RWC are 0.87 and the median RWC are 0.91, so that around 90% of the relative welfare has to be sacrificed in order to rationalize this proportion of choices. For 50% of the subjects, the RWC lie within 0.83 and 0.94 at this level of DoR. As the target level of DoR increases, the mean RWC increase even further. For $\alpha = 0.99$, the mean RWC reach 0.95, and the median RWC reach 0.98: almost all the welfare must be sacrificed in this case.[27]

The RWC numbers also show significant variation across subjects. At $\alpha = 0.95$ the smallest amount of RWC is 0.18, while the maximum amount is 0.98, which is roughly 1.13 times as large as the mean amount. The standard deviation at this level of DoR is 0.13. At $\alpha = 0.99$ the minimum RWC is slightly below a half, 0.41, while the maximum amount becomes 1, and the standard deviation is 0.09.

There is considerable uncertainty in the RWC at the individual level, just as we found for the AWC. Figure 5 (Panel B) shows the point estimates of RWC (at $\alpha = 0.95$) for each subject along with the 95% confidence interval around those estimates. Contrary to the AWC, however, the uncertainty in the RWC is highest at the lower percentile ranks. This uncertainty tends to decrease with the percentile rank.

The preceding analysis allow us to formulate the following result.

Result 1. *The welfare costs are low in terms of everyday economic activity, but are sub-*

[27] Rationalizing all the choices would definitionally require RWC of 1 for every subject with a non-zero noise, however small, which is the reason to use 0.99 as the highest level of α, and not 1.

stantial for the choice environment in which they occurred.

Comparing our results to Choi *et al.* (2014), we find that the subjects in our sample require a larger fraction of the total monetary welfare to rationalize their choices. This difference can be explained by the differences in methods. The GARP-based measure used by Choi *et al.* is well-known for its relatively mild requirements on choice consistency (Beatty and Crawford, 2011).[28] For example, their primary measure does not even require choices to satisfy first-order stochastic dominance. We find greater similarity between our results and the results reported by von Gaudecker *et al.* (2011), who also employ structural methods.

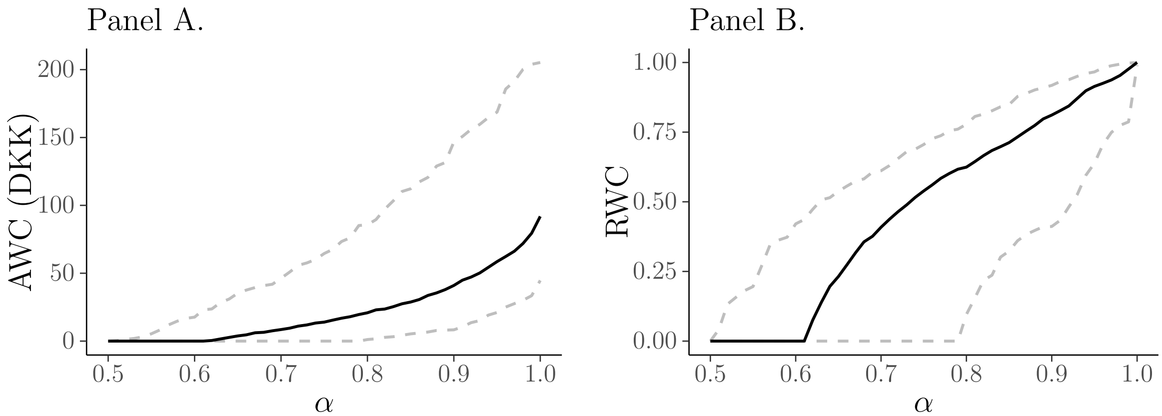
3.4 Marginal Welfare Costs

So far we have looked at the distributions of AWC and RWC for only three levels of the target DoR. This analysis does not tell us how quickly welfare costs grow as the rationalizability requirements become tighter, and in general what the shape of the costs as functions of α is. Figure 6 provides an answer to these questions by showing the median welfare costs as functions of α across all the subjects in the sample, with the dashed lines corresponding to the 5% and 95% empirical quantiles. The lowest possible target DoR in our context is 0.5, since the choice is binary. However, the welfare costs stay at 0 for the median subject until α crosses the 0.62 mark.

Panel A on Figure 6 shows the graphs of the AWC in relation to α. The median AWC tend to be a convex function of α: at first increasing the DoR requires relatively little AWC, but as the target becomes higher, each additional percentage point of DoR costs more and more in terms of AWC. The graph for the RWC on Panel B of Figure 6 is in a sense the mirror image of the AWC. The RWC tends to be a concave function of α. For small values of DoR extra percentage points of change require high welfare costs, but as the target increases these

[28] Another potential explanation could be that there are systematic differences in the samples used. The Choi *et al.* (2014) experiment was conducted in Netherlands, while our experiment was conducted in Denmark. We believe this explanation to be unlikely *a priori*. The results in Blow *et al.* (2008) support our claim, as it also employs a revealed preference approach and shows that Danes are generally consistent in other choice domains.

Figure 6: Welfare Costs as Functions of the Target DoR



Note: The graph shows the AWC (Panel A) and the RWC (Panel B) as functions of the target DoR (α). The black solid lines are the median welfare costs for each level of α. The dashed lines below (above) the solid line represent the 5% (95%) quantiles of the welfare costs.

extra points become less costly in relative terms. These observations allow us to formulate the next result.

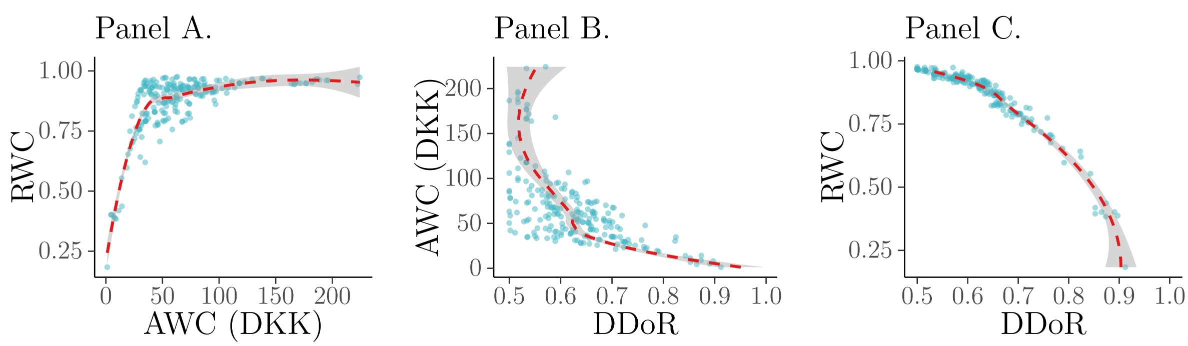
Result 2. *The marginal absolute (relative) welfare costs are increasing (decreasing) with the increase in the target DoR.*

This result can be explained by the way our measures of welfare costs are computed. Starting from a given default DoR, $\rho(\hat{\mu}, 0)$, we gradually increase ε until the DoR reaches the target, $\rho(\hat{\mu}, \varepsilon) = \alpha$. The low marginal AWC at low α targets imply there are many choices that can be easily rationalized by small ε, since the difference in the certainty equivalents between the alternatives must be low. At high target DoR more choices have to be rationalized, but no "easily rationalizable" choices are left. Increasing DoR requires tapping into choice pairs with higher differences in certainty equivalents, and hence higher marginal AWC. The implications for the RWC graphs are the converse. At low target DoR the marginal RWC are high, since rationalizing many choices with small differences in certainty equivalents requires the whole difference. At high targets fewer such choice pairs remain and the marginal RWC decrease.

3.5 Relationship Between the Measures

We now turn to the relationship between the two welfare costs measures and the default DoR (DDoR). We ask whether people with lower AWC also have lower RWC, and formally test the previous observation that people with lower DDoR tend to have higher costs. The motivation behind these questions is that it is intuitive to expect the positive relation between AWC and RWC. It does not follow, however, directly from the method of their computation. Only if preferences are held constant must higher AWC imply higher RWC, but there is no such prediction when preferences are not constant, as is typically the case when making comparisons across subjects. Likewise, even though it is natural to expect that people with higher DDoR have lower costs we cannot formally expect this observation to hold *a priori*.

Figure 7: Relationship Between AWC, RWC, and DDoR



Note: The graph shows the scatterplots between three pairs of measures: AWC, RWC, and DDoR. The welfare costs are evaluated at $\alpha = 0.95$. The dots represent individual subjects. The dashed lines are the smooth fitted lines estimated using local polynomial regressions, and the shaded regions are the estimated 95% confidence intervals.

Figure 7 (Panel A) shows a scatterplot of the RWC (y-axis) against the AWC (x-axis) computed at $\alpha = 0.95$.[29] A clear positive association between the two measures can be observed. The Kendall rank correlation between the two measures is 0.45 (the ranking of subjects by the two measures is the same 73% of the time) and is highly significant, with p-value < 0.001. The relation between RWC and AWC is non-linear, and has a concave shape.

[29] The results in this section remain quantitatively similar if we use $\alpha = 0.9$ or 0.99.

Figure 7 (Panel B) shows the scatterplot of the AWC (y-axis) against the DDoR (x-axis), and confirms our earlier observation from the analysis of marginal costs. There is a moderate negative association between the DDoR and the AWC. The Kendall rank correlation between the two measures is -0.5 (the ranking of subjects by the two measures is the opposite 75% of the time) and is highly significant, with p-value < 0.001. The relation between them is again non-linear, and has a convex shape.

Figure 7 (Panel C) shows the scatterplot of the RWC (y-axis) against the DDoR (x-axis). We can immediately see a very tight negative association between the two measures. The Kendall rank correlation between them is -0.85 (the ranking of subjects by the two measures is the opposite 93% of the time) and is highly significant, with p-value < 0.001. The relation is again slightly non-linear and has a concave shape.

These observations allow us to formulate the following result.

Result 3. *People with higher absolute welfare costs tend to have higher relative welfare costs. People with higher default degree of rationalizability tend to have lower absolute welfare costs and relative welfare costs.*

This result implies that there is a certain degree of consistency between the measures we introduce. Moreover, this consistency works in the way we expect. This is a nice property, but it could not have been deduced from the method by which these measures are constructed. If risk preferences were the same across subjects, higher AWC must have implied higher RWC, but we cannot say much about the case when preferences and noise are different across subjects. It is possible, that a subject with high AWC has preferences such that the differences in the certainty equivalents are even higher, and the RWC are actually low. We do, in fact, observe such cases. But the general tendency is for the subjects to have the same ordering, whether it is measured according to the absolute or relative measure of welfare costs.

There is also a negative relationship between the DDoR and the welfare two costs measures. This implies that people who make more consistent choices, measured by the DDoR, also require less welfare costs to rationalize their choices. This is an intuitive property, but it is hard to see *a priori* why it should hold even though the data indicate that it does, with the relation between the default DoR and the RWC being particularly strong. The relative strength of this relationship, compared to the relationship with the AWC can be partially attributed to the fact that both the default DoR and the RWC are relative measures defined on the unit interval. Nonetheless, such a strong relationship is remarkable, given that the two measures address two very different questions.

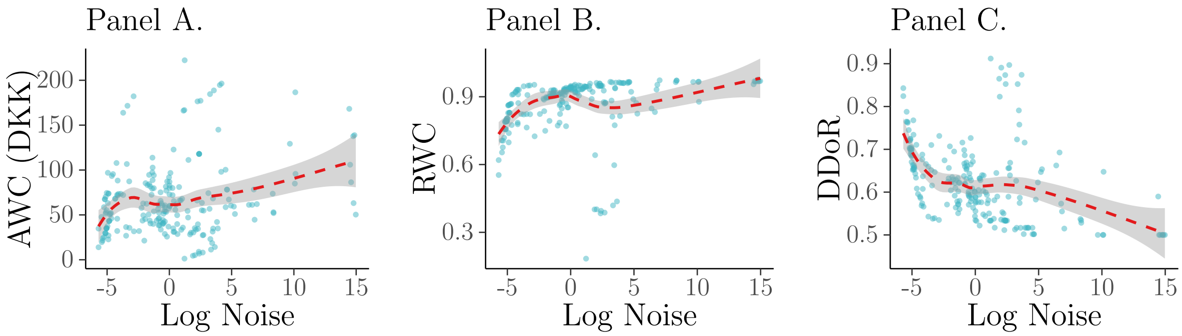
3.6 Welfare Costs and Noise

Our approach is in part motivated by the desire to attach an economic meaning to the noise parameter. It is, therefore, of interest to look at the relationship between the two welfare cost measures we introduced and noise, as well as the relationship between the DDoR and noise. Higher noise does translate into higher welfare costs if preferences are kept constant, but no prediction is available for comparisons between subjects, whose preferences are not kept constant. It is natural to expect, however, that this property should also hold between subjects. Given the negative association between the DDoR and the costs, it is also natural to expect that higher noise translates into lower default DoR, but whether it does is an empirical question.

Figure 8 shows the scatterplots of (from left to right) the AWC and RWC and the DDoR (on the y-axis) against the logarithm of noise (x-axis).[30] The three panels confirm our hypotheses. We do see that higher noise is associated with higher AWC and RWC and lower DDoR, although the strength of this association differs across the measures. It is small, though statistically significant, for the AWC. The Kendall rank correlation between the two measures is 0.13 with p-value $= 0.003$ (the ranking of subjects by the two measures is the same 57% of the time). The weakness of the association can be seen by the substantial variation in the AWC at the high values of noise, which means that there are many subjects

[30] We truncate the logarithm noise at 15, in order to make the graph more readable. This excludes one subject.

Figure 8: Relationship Between Welfare Costs and Noise



Note: The graph shows the scatterplots between three measures (AWC, RWC, DDoR) and (the logarithm of) noise. The welfare costs are evaluated at $\alpha = 0.95$. The dots represent individual subjects. The dashed lines are the smooth fitted lines estimated using local polynomial regressions, and the shaded regions are the estimated 95% confidence intervals. The graphs drop one subject with log noise higher than 15.

with high estimates of noise but low AWC. The association with the RWC is much stronger. The Kendall rank correlation between the two measures is 0.48 with p-value < 0.001 (the ranking of subjects by the two measures is the same 74% of the time). Finally, the association with the DDoR (in absolute terms) is slightly weaker than the association with the RWC, but much stronger than the association with the AWC. The Kendall rank correlation between the two measures is -0.41 with p-value < 0.001 (the ranking of subjects by the two measures is the opposite 70% of the time).

A notable feature in these results, most pronounced in the relationship between noise and the DDoR and the RWC, is that there is an outer boundary that constrains the values. On Panel B of Figure 8 this boundary constrains the values of the RWC from above, and on Panel C of Figure 8 this boundary constrains the values of the DDoR from below. This pattern suggests that for given noise the RWC (DDoR) cannot be higher (lower) than a certain value, defined by this boundary.

These findings lead us to the next result.

Result 4. *People with higher noise tend to have higher absolute and relative welfare costs and lower default degree of rationalizability. For any given value of noise there appears to exist a* *maximum (minimum) amount of absolute and welfare costs (degree of rationalizability) that one can have.*

The first part of this result confirms our intuitive guesses. We do see some association between noise and welfare costs, which implies that noise contains some information about welfare costs and choice consistency, but this information is imprecise. Despite big differences in noise estimates between some subjects, their RWC need not be that different. Similarly, some subjects might appear to have high welfare costs based on the noise measure, while in fact their AWC are not nearly as large.

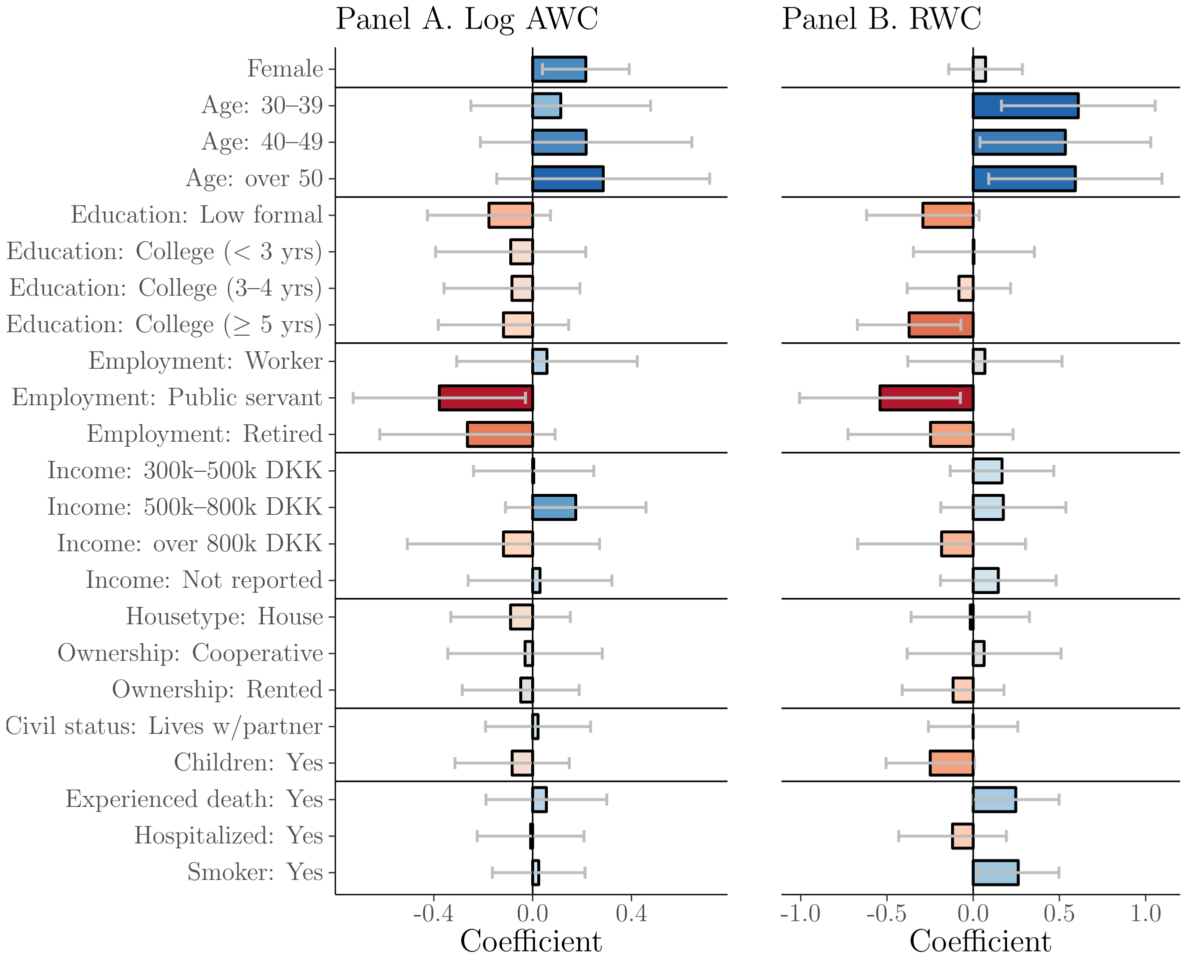
The second part of the result is unexpected and remarkable. It says that there is a regularity in the relation between noise, welfare costs and default DoR. This regularity is in the form of a boundary that constraints the possible values. The existence of such a boundary is likely to be related to the estimation and computation procedures, however it is not clear why it exists and what determines its shape. We leave this question for further research.

3.7 Socio-Demographic Covariates of Welfare Costs

We have seen that the estimates of welfare costs vary substantially between subjects in our sample. Here we attempt to attribute this variability to the observable socio-demographic characteristics of the subjects. We focus on sex, age, education, work, income, housing, family, and health characteristics. The demographic covariates are defined as indicator variables, relative to a base category. The base category is male, age 18–29, vocational training, employed as a student, household income less than 300,000 DKK, living in an apartment, owning apartment/house, living alone, no children, has not experienced death, has not been hospitalized, and not smoking.

Figure 9 provides descriptive regression results by plotting the estimates of regression coefficients along with 95% confidence intervals (using robust standard errors). The model

Figure 9: Regression Results



Note: The graph shows descriptive regression results for the AWC (Panel A, OLS) and RWC (Panel B, fractional regression). Bars correspond to coefficient estimates and error bars show 95% confidence intervals based on robust standard errors. Number of observations: 217.

on Panel A uses the logarithm of AWC, computed at $\alpha = 0.95$, as the dependent variable,[31]

$$\ln(AWC)_i = constant + \beta\, Demographic\ controls_i + \epsilon_i,$$

and is estimated using OLS. The model on Panel B uses the RWC, computed at $\alpha = 0.95$, as the dependent variable. Since the RWC are defined only on the unit interval, we use a fractional regression model due to Papke and Wooldridge (1996) to estimate the coefficients.

Several patterns emerge from Figure 9. Females tend to have higher AWC than males. The RWC, however, are not significantly different between males and females. Welfare costs tend to be higher for older subjects. The AWC tend to increase monotonically with the age group, but the effect is not precisely estimated. The RWC are higher for subjects older than 30 years than for younger subjects, but there is no statistically significant difference between the three age groups above 30 years. College education has a beneficial impact on the welfare costs relative to vocational training. The effect is most pronounced for the RWC and subjects with 5 or more years of college. Interestingly, subjects who are employed as public servants have significantly lower AWC and RWC than subjects employed as students or workers. Retired subjects tend to have lower AWC and RWC, on average, but the effect is not statistically different from zero. The effect of income is mixed and not precisely estimated. Subjects with medium and medium-high levels of income tend to have higher welfare costs, while subjects with very high levels of income tend to have lower welfare costs. The type of housing a subject occupies and the type of ownership does not appear to have a meaningful impact on welfare costs. Similarly, the effect of parenting status is small and not statistically significant, except for the effect having children on the RWC. Subjects show some systematic variation by their health status with the effects most pronounced for RWC. For instance, smokers tend to have higher RWC than non-smokers.

These observations lead us to the following result.

[31] Using alternative target DoR, 0.9 or 0.99, produces quantitatively similar results.

Result 5. *Having higher welfare costs is associated with higher age, lower education, and particular employment status. The RWC are not significantly different for males and females, although the AWC for females tend to be higher.*

Overall, even the rich set of socio-demographic characteristics that we use does little to explain the observed variance in welfare costs. The regression for the AWC, for instance, is able to explain only 12% of the observed variation. After correcting for the number of covariates included, the R^2 actually becomes negative. Such low explanatory power of socio-demographic characteristics for elicited economic variables is typical in the literature (l'Haridon *et al.*, 2018; Noussair *et al.*, 2014; Choi *et al.*, 2014; von Gaudecker *et al.*, 2011). One explanation for low predictive power of socio-demographic characteristics is the sampling error in the estimates on the left-hand side.[32] On the other hand, part of the heterogeneity in the estimates of welfare costs that we observe might be truly idiosyncratic, which in our view is not necessarily an undesirable property as suggested by some, such as l'Haridon *et al.* (2018). If an elicited economic quantity (such as welfare costs, in our case) could be perfectly decomposed into a linear combination of socio-demographic characteristics, this quantity would have nothing to contribute to explaining variation in other behavioral outcomes.

4 Related literature

Our approach connects to a large theoretical literature on stochastic choice, which we briefly summarize. The early work on stochastic choice dates back to Fechner (1860) and Thurstone (1927). It was subsequently developed into the Random Utility Model (RUM) by Marschak (1960) and summarized by McFadden (2001). Luce (1959) introduced and axiomatized the strong utility (or multinomial logit) model, as well as other models of stochastic choice. McFadden (1976) established necessary and sufficient conditions under which a RUM is equivalent to the multinomial logit model.

[32] Using weighted OLS in the AWC regression in which weights are proportional to the inverses of the squared standard errors substantially improves fit.

Wilcox (2011) extends the standard multinomial logit model by allowing for the noise heterogeneity that is caused by the range of monetary stakes in a choice context. This extension allows one to preserve the deterministic notion of being more risk averse in a stochastic setting. The stronger utility model developed by Blavatskyy (2014) also allows for noise heterogeneity, but focuses on preserving the first-order stochastic dominance relation in a stochastic choice setting. Gul, Natenzon, and Pesendorfer (2014) modify the multinomial logit model by considering the attributes of choice alternatives rather than alternatives themselves, to address some of the criticism of the original formulation. Apesteguia, Ballester, and Lu (2017) characterize the RUM that satisfies a single-crossing property.

Conceptually, our measures are similar to the Critical Cost Efficiency Index (CCEI) of Afriat (1972), which is used to evaluate the degree of consistency with the Generalized Axiom of Revealed Preference (GARP). Just like our relative cost measure, CCEI is defined on the unit interval, and its complement shows what proportion of monetary value an agent should be allowed to waste in order to rationalize her choices by some utility function. While GARP provides qualitative statements, we put more structure on it in a flexible manner to complement it and provide quantitative evidence.

Viewing our approach as a structural extension of GARP allows us to position our approach again in a broader methodological setting. Ross (2014, ch. 4) carefully lays out the full case for interpreting economic experimentation as an application of the intentional stance of Dennett (1987), noted earlier. This is the methodology that Ross (2014) calls "neo-Samuelsonian," a label that tries to nudge economists toward seeing that the intentional stance is what they have always been doing when they applied Revealed Preference Theory to actual, finite, choice data. In other words: our approach is not novel, exotic economic methodology. Instead we view it as just a sophisticated, structural interpretation of the good old-time religion for economists.

The intuition behind the computation of our measures also links it to a literature on payoff dominance in experiments (Harrison, 1994, 1992; Harrison and Morgan, 1990; Harrison, 1989). This literature shows that allowing for small deviations from optimal behavior, just as we do, allows one to rationalize supposedly anomalous effects observed in experimental studies.

Harrison and Ng (2016, 2018) use an approach similar to ours in order to evaluate the loss of consumer surplus resulting from suboptimal insurance choices. Harrison and Ross (2018) apply the same approach to evaluate suboptimal portfolio investments. Their measure of lost consumer surplus is similar to our AWC measure, with both being based on computing certainty equivalents. One key difference, however, is that these studies use two experimental tasks: one for preference estimation and the other for welfare evaluation, while we rely on a single task to estimate welfare costs resulting from stochastic choice. The approach that we take in this study does not rely on an independent risk metric, as is the case in Harrison and Ng (2016, 2018) and Harrison and Ross (2018), but rather relies on a specific noise structure to "bootstrap" a measure of welfare costs.

Our approach is closely related to studies that estimate structural models of choice under risk and over time. Holt and Laury (2002) study subjects' choices under risk in a laboratory experiment. Subjects make choices between a "safe" and a "risky" lottery across different pairs of lotteries, in which the probabilities of lottery outcomes vary from one pair to the next. HL estimate the Expected Utility model with a flexible Expo-Power utility function using the strict utility model of stochastic choice.[33] Andersen, Harrison, Lau, and Rutström (2008) also use the strict utility model to structurally estimate risk and time preferences of a representative sample of the Danish population. They note that noise estimates are higher in the risk task than in the discounting task. von Gaudecker *et al.* (2011) uses a representative sample of the Dutch population to estimate subjects' risk preferences using a model of stochastic choice that is a hybrid between the multinomial logit and tremble models, and thus features two measures of choice randomness: noise and trembles. While these studies typically focus on estimates of risk and time preferences, and do not interpret the estimates

[33]The strict utility model of Luce (1959) differs from the multinomial logit model in the way the noise parameter enters choice likelihoods.

of the stochastic part, such as noise or tremble parameters,[34] we explicitly focus on the estimates of the stochastic part and provide a systematic approach to economically interpret the estimates of choice randomness. Finally, Bland (2018) considers mixture specification over pooled choices, contrasting one "rational" model as one of the data generating processes (DGP) with a "behavioral" model as the other DGP. He then calculates CE of choices using the deterministic core of the "rational" model DGP, thereby evaluating potential welfare losses from using a "behavioral" DGP as well as the existence of noise for both DGP. We reject the simplistic identification of one model as "rational" and the explicit assumption that the "behavioral" model is therefore "irrational." But the general logic of allowing the estimated structural model of noise to provide a basis for welfare evaluation is consistent with our approach.

We provide economic measures of choice randomness (or consistency), which link to studies on the quality of decision-making. Choi *et al.* (2007) study decision-making under risk in a laboratory experiment in which they present subjects with convex budget sets for two Arrow securities. This design allows them to gauge the subjects' decision-making quality using a measure of GARP-consistency, a standard technique in the revealed preference approach to consumer demand. They find that subjects' behavior is highly consistent with GARP. Choi *et al.* (2014) expand the analysis by using a representative panel of the Dutch population. They also find a high degree of GARP-consistency in risky choices, which varies, however, with education, sex, and age. Beatty and Crawford (2011) show that while behavior in a wide range of situations is highly GARP-consistent, this might be a result of a misspecified measure of consistency. They propose an alternative to the traditional CCEI measure, which is based on predictive success, and show that the CCEI measures of GARP-consistency are overinflated, and hence that the actual consistency of choices is much lower. Hey (2001) studies decision-making quality in a laboratory experiment on choice under risk

[34] von Gaudecker *et al.* (2011) is an exception, which provides a brief discussion of the economic significance of the estimates of the tremble parameter. In particular, they give an example of what the estimated parameters of the stochastic part of the model imply for the relation between the difference in the certainty equivalents and the likelihood of choosing the higher-valued lottery.

and asks whether choice consistency improves with experience. He finds mixed evidence of a positive effect of experience on choice consistency. We rely on a parametric measure of choice consistency and find a lower degree of consistency than in the studies that use the non-parametric revealed preference approach.

Finally, our approach is also related to recent literature on rational inattention. Matêjka and McKay (2015) show that when an agent faces information costs, optimal behavior is stochastic choice, and that under certain conditions choice likelihoods are represented by the multinomial logit specification. Cheremukhin, Popova, and Tutino (2015) apply a model of rational inattention to risky choices and estimate the shape of the cost-of-information function in a laboratory experiment with student subjects. Caplin and Dean (2015) develop a revealed preference test of rational inattention theories with general cost-of-information functions. Since the noise parameter in the rational inattention models has the interpretation of marginal information costs, our method allows one to convert these costs into monetary or percentage terms.

5 Conclusion

Stochastic choice has become an active area of both theoretical and empirical research. While the existing literature mainly focuses on the sources of choice randomness, its economic consequences are less well understood. We develop tools to assess the economic significance of noise and apply them to a sample from the general Danish population in an artefactual field experiment.

We introduce three interconnected concepts: rationalizing imperfection, optimal region, and degree of rationalizability. Fixing the degree of rationalizability at a certain target level, we vary the amount of imperfection, which in turn affects the optimal region, to make the proportion of subjects' choices falling in the optimal region equal the target level. This amount of imperfection represents the welfare costs, or monetary welfare allowed to be wasted, that is required to rationalize by a model a given proportion of choices. The resulting welfare costs can be expressed both in absolute (dollar) and in relative (to the actual stakes of the choice environment) terms.

We compute the absolute welfare costs and relative welfare costs at the individual level in an experiment with binary-choice lotteries. Several patterns emerge from our analysis, some of which coincide with previous findings, and some of which are new. We find that the AWC are not economically significant in our sample, while the RWC are economically significant. In other words, the welfare costs are tiny if viewed from a broad perspective of economic activity, but they are substantial if viewed from the perspective of this particular choice experiment. As compared to Choi *et al.* (2014), who employ a relative measure based on the consistency with GARP, our estimates of RWC are much larger. However, our results for choice consistency are comparable with that of von Gaudecker *et al.* (2011), who also employ structural methods. We attribute the difference in results to the difference in the methods, with our method imposing stricter requirements.

Since our welfare costs measures depend on the target level of rationalizability α, we study the shape of the relation between α and these welfare costs. We find that the AWC increase in α at an increasing speed, while the RWC increase in α at a decreasing speed. The difference in these two relations is explained by the way our method of computation works. Subjects with higher AWC tend to have higher RWC. Also, a lower DDoR is associated with higher AWC and RWC: subjects who start out with low default degree of rationalizability require a higher cost to reach a given degree of rationalizability. Looking at the relationship between our cost measures and raw estimates of noise reveals that they are positively associated, though our measures do not have such a wide range, which allows for sensible comparisons across subjects and allows us to make judgments about the magnitudes of choice inconsistencies. The analysis of observable heterogeneity and its role in predicting welfare costs suggests patterns similar to those reported by von Gaudecker *et al.* (2011) and Choi *et al.* (2014). We find that welfare costs increase with age, decline with education, and are lower for certain occupations.

Finally, we take seriously the need for consistent methodological and philosophical positions when it comes to undertaking behavioral welfare economics. The reason is simple: one cannot question the consistency of observed choices by agents on the one hand and then turn around and effortlessly infer the preferences of those agents on the other hand. This isolates the deep normative challenge raised by the core descriptive insight of behavioral economics, as stressed by Ross (2014, ch. 4), Infante, Lecouteux, and Sugden (2016), and Harrison and Ross (2018, § 5). Dennett (1987)'s intentional stance, as applied to economics by Ross (2014)'s "neo-Samuelsonian" methodology, provides a general and consistent approach to address this challenge, and permits concrete applications illustrated by Harrison and Ng (2016, 2018), Harrison and Ross (2018) and the present study.

References

Afriat SN (1972). "Efficiency Estimation of Production Function." *International Economic Review*, **13**(3), 568–98.

Agranov M, Ortoleva P (2017). "Stochastic Choice and Preferences for Randomization." *Journal of Political Economy*, **125**(1), 40–68.

Andersen S, Harrison GW, Lau MI, Rutström EE (2008). "Eliciting Risk and Time Preferences." *Econometrica*, **76**(3), 583–618.

Apesteguia J, Ballester MA, Lu J (2017). "Single-Crossing Random Utility Models." *Econometrica*, **85**(2), 661–674.

Ballinger TP, Wilcox NT (1997). "Decisions, Error and Heterogeneity." *Economic Journal*, **107**(443), 1090–1105.

Beatty TKM, Crawford IA (2011). "How Demanding Is the Revealed Preference Approach to Demand?" *American Economic Review*, **101**(6), 2782–95.

Bland JR (2018). "The Cost of Being Behavioral in Risky Choice Experiments." *Working paper*, The University of Toledo, Department of Economics.

Blavatskyy PR (2014). "Stronger Utility." *Theory and Decision*, **76**(2), 265–286.

Blow L, Browning M, Crawford I (2008). "Revealed Preference Analysis of Characteristics Models." *Review of Economic Studies*, **75**(2), 371–389.

Bone J, Hey J, Suckling J (1999). "Are Groups More (or Less) Consistent Than Individuals?" *Journal of Risk and Uncertainty*, **18**(1), 63–81.

Bruhin A, Fehr-Duda H, Epper T (2010). "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion." *Econometrica*, **78**(4), 1375–1412.

Camerer CF (1989). "An Experimental Test of Several Generalized Utility Theories." *Journal of Risk and Uncertainty*, **2**(1), 61–104.

Caplin A, Dean M (2015). "Revealed Preference, Rational Inattention, and Costly Information Acquisition." *American Economic Review*, **105**(7), 2183–2203.

Carbone E, Hey JD (2000). "Which Error Story is Best?" *Journal of Risk and Uncertainty*, **20**(2), 161–176.

Cheremukhin A, Popova A, Tutino A (2015). "A Theory of Discrete Choice with Information Costs." *Journal of Economic Behavior & Organization*, **113**, 34–50.

Choi S, Fisman R, Gale D, Kariv S (2007). "Consistency and Heterogeneity of Individual Behavior under Uncertainty." *American Economic Review*, **97**(5), 1921–1938.

Choi S, Kariv S, Müller W, Silverman D (2014). "Who Is (More) Rational?" *American Economic Review*, **104**(6), 1518–1550.

Cox JC, Sadiraj V (2008). "Risky Decisions in the Large and in the Small: Theory and Experiment." In JC Cox, GW Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008).

Dennett DC (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett DC (1991). "Real Patterns." *The Journal of Philosophy*, **88**(1), 27–51.

Fechner GT (1860). *Elements of Psychophysics*. Amsterdam: Bonset.

Gneezy U, Potters J (1997). "An Experiment on Risk Taking and Evaluation Periods." *The Quarterly Journal of Economics*, **112**(2), 631–645.

Gul F, Natenzon P, Pesendorfer W (2014). "Random Choice as Behavioral Optimization." *Econometrica*, **82**(5), 1873–1912.

Gul F, Pesendorfer W (2006). "Random Expected Utility." *Econometrica*, **74**(1), 121–146.

Halevy Y, Persitz D, Zrill L (2018). "Parametric Recoverability of Preferences." *Journal of Political Economy*, **126**(4), 1558–1593.

Harless DW, Camerer CF (1994). "The Predictive Utility of Generalized Expected Utility Theories." *Econometrica*, **62**(6), 1251–1289.

Harrison GW (1989). "Theory and Misbehavior of First-Price Auctions." *American Economic Review*, **79**(4), 749–762.

Harrison GW (1992). "Theory and Misbehavior of First-Price Auctions: Reply." *American Economic Review*, **82**(5), 1426–1443.

Harrison GW (1994). "Expected Utility Theory and the Experimentalists." *Empirical Economics*, **19**, 223–253.

Harrison GW, Jessen LJ, Lau M, Ross D (2018). "Disordered Gambling Prevalence: Methodological Innovations in a General Danish Population Survey." *Journal of Gambling Studies*, **34**(1), 225–253.

Harrison GW, List JA (2004). "Field Experiments." *Journal of Economic Literature*, **42**(4), 1009–1055.

Harrison GW, Morgan P (1990). "Search Intensity in Experiments." *Economic Journal*, **100**(401), 478–486.

Harrison GW, Ng JM (2016). "Evaluating the Expected Welfare Gain from Insurance." *Journal of Risk and Insurance*, **83**(1), 91–120.

Harrison GW, Ng JM (2018). "Welfare effects of insurance contract non-performance." *The Geneva Risk and Insurance Review*, **43**(1), 39–76.

Harrison GW, Ross D (2018). “Varieties of Paternalism and the Heterogeneity of Utility Structures.” *Journal of Economic Methodology*, **25**(1), 42–67.

Hey JD (2001). “Does Repetition Improve Consistency?” *Experimental Economics*, **4**(1), 5–54.

Hey JD, Orme C (1994). “Investigating Generalizations of Expected Utility Theory Using Experimental Data.” *Econometrica*, pp. 1291–1326.

Holt CA, Laury SK (2002). “Risk Aversion and Incentive Effects.” *American Economic Review*, **92**(5), 1644–1655.

Infante G, Lecouteux G, Sugden R (2016). “Preference Purification and the Inner Rational Agent: A Critique of the Conventional Wisdom of Behavioural Welfare Economics.” *Journal of Economic Methodology*, **23**(1), 1–25.

l'Haridon O, Vieider FM, Aycinena D, Bandur A, Belianin A, Cingl L, Kothiyal A, Martinsson P (2018). “Off the Charts: Massive Unexplained Heterogeneity in a Global Study of Ambiguity Attitudes.” *The Review of Economics and Statistics*, **100**(4), 664–677.

Loomes G, Sugden R (1995). “Incorporating a Stochastic Element into Decision Theories.” *European Economic Review*, **39**(3), 641–648.

Loomes G, Sugden R (1998). “Testing Different Stochastic Specifications of Risky Choice.” *Economica*, **65**(260), 581–598.

Luce DR (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York.

Manzini P, Mariotti M (2014). “Welfare Economics and Bounded Rationality: The Case for Model-Based Approaches.” *Journal of Economic Methodology*, **21**(4), 343–360.

Marschak J (1960). “Binary-Choice Constraints and Random Utility Indicators.” In K Arrow (ed.), *Stanford Symposium on Mathematical Methods in the Social Sciences* (Stanford, CA: Stanford University Press).

Matêjka F, McKay A (2015). “Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model.” *American Economic Review*, **105**(1), 272–98.

McFadden DL (1976). “Quantal Choice Analaysis: A Survey.” In SV Berg (ed.), *Annals of Economic and Social Measurement* (pp. 363–390). NBER.

McFadden DL (2001). “Economic Choices.” *American Economic Review*, **91**(3), 351–378.

McKelvey RD, Palfrey TR (1995). “Quantal Response Equilibria for Normal Form Games.” *Games and Economic Behavior*, **10**(1), 6–38.

Monroe BA (2017). *Stochastic Models in Experimental Economics*. Ph.D. thesis, University of Cape Town.

Nogee P, Mosteller F (1951). "An Experimental Measure of Utility." *Journal of Political Economy*, **59**, 371–404.

Noussair CN, Trautmann ST, van de Kuilen G (2014). "Higher Order Risk Attitudes, Demographics, and Financial Decisions." *The Review of Economic Studies*, **81**(1), 325–355.

Papke LE, Wooldridge JM (1996). "Econometric Methods for Fractional Response Variables with an Application to 401 (K) Plan Participation Rates." *Journal of Applied Econometrics*, **11**(6), 619–32.

Prelec D (1998). "The Probability Weighting Function." *Econometrica*, **66**(3), 497–528.

Quiggin J (1982). "A Theory of Anticipated Utility." *Journal of Economic Behavior & Organization*, **3**(4), 323–343.

Ross D (2014). *Philosophy of Economics*. London: Palgrave Macmillan.

Starmer C, Sugden R (1989). "Probability and Juxtaposition Effects: An Experimental Investigation of the Common Ratio Effect." *Journal of Risk and Uncertainty*, **2**(2), 159–78.

Swait J, Marley AAJ (2013). "Probabilistic Choice (Models) as a Result of Balancing Multiple Goals." *Journal of Mathematical Psychology*, **57**(1–2), 1–14.

Thurstone LL (1927). "A Law Of Comparative Judgment." *Psychological Review*, **34**(4), 266–270.

Tversky A (1969). "Intransitivity of Preferences." *Psychological Review*, **76**(1), 31.

von Gaudecker HM, van Soest A, Wengstrom E (2011). "Heterogeneity in Risky Choice Behavior in a Broad Population." *American Economic Review*, **101**(2), 664–94.

Wakker P, Erev I, Weber EU (1994). "Comonotonic Independence: The Critical Test Between Classical and Rank-Dependent Utility Theories." *Journal of Risk and Uncertainty*, **9**(3), 195–230.

Wallin A, Swait J, Marley AAJ (2018). "Not Just Noise: A Goal Pursuit Interpretation of Stochastic Choice." *Decision*, **5**(4), 253–271.

Wilcox NT (2008). "Stochastic Models for Binary Discrete Choice Under Risk: A Critical Primer and Econometric Comparison." In JC Cox, GW Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008).

Wilcox NT (2011). "Stochastically More Risk Averse: A Contextual Theory of Stochastic Discrete Choice Under Risk." *Journal of Econometrics*, **162**(1), 89–104.

Wilcox NT (2015). "Error and Generalization in Discrete Choice Under Risk." *Working paper*, Chapman University.

Appendices

A Robustness Checks

Here we present additional results derived from alternative assumptions about risk preferences and stochastic choice.

First, we consider an alternative to the EUT, the Rank-Dependent Utility (RDU) model due to Quiggin (1982), which allows for probability weighting. The RDU model has been used extensively in applied and theoretical work. Under this alternative assumption the aggregate utilities of the lotteries are computed as

$$\begin{aligned} U(a;\gamma_u,\gamma_q) = \\ &= \sum_{i=1}^{k} \Big(\omega \left(q_{(1)}(a) + \ldots + q_{(i)}(a); \gamma_q \right) - \omega \left(q_{(1)}(a) + \ldots + q_{(i-1)}(a); \gamma_q \right) \Big) \times \\ &\times u \left(x_{(i)}(a); \gamma_u \right), \end{aligned}$$

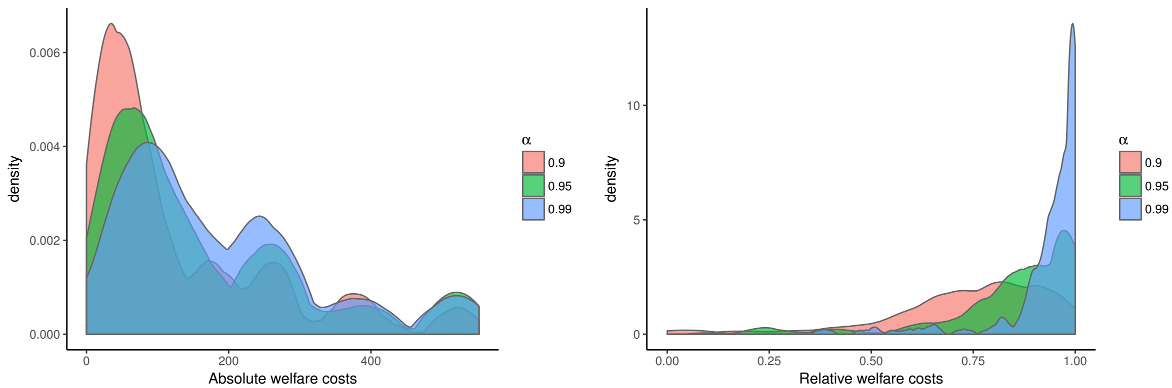
where $\omega : [0,1] \mapsto [0,1]$ is the probability-weighting function, and outcomes are ranked from highest $x_{(1)}$ to lowest $x_{(k)}$, with corresponding probabilities. We assume that ω is the two-parameter (Prelec, 1998) probability weighting function,[35]

$$\omega(q; \gamma_q^1, \gamma_q^2) = \exp(-\gamma_q^2 (-\ln q)^{\gamma_q^1}).$$

Figure A.1 shows the calculated absolute and relative welfare costs under the assumption of the RDU model for each individual. Figure A.1 shows that the distributions look very similar to those under EUT, Figure 4.

Taking a closer look at the differences between the EUT and RDU-based calculations, we can see from Figure A.2a that the AWC calculated using the EUT model are lower. For

[35] We do not restrict the shape parameter γ_q^1 to the unit interval, and thus do not impose an inverse-S shape on the probability weighting function.

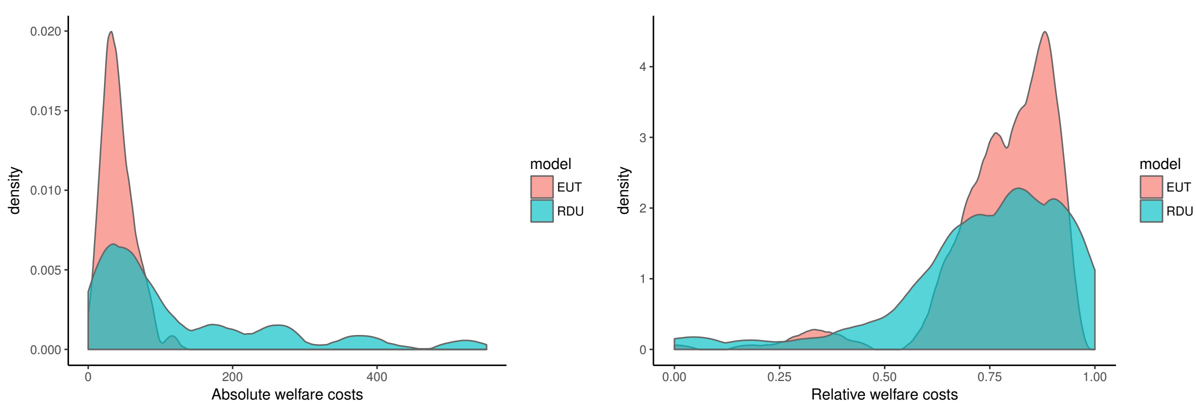


(a) (b)

Figure A.1: Absolute and Relative Welfare Costs for Three Levels of α, RDU.

$\alpha = 0.9$ the difference in the medians between the AWC calculated using EUT vs. RDU is -59.04 (Wilcoxon signed rank test, p-value < 0.001). The mean of the differences is -84.89 DKK (approximately -13 USD): RDU-based AWC are almost 3 times higher on average.

The RWC, however, are slightly higher under EUT, as shown in Figure A.2b. The difference in the medians between the RWC calculated using EUT vs. RDU is 0.02 (Wilcoxon signed rank test, p-value $= 0.02$). The mean of the differences is 0.03. The difference in the RWC for RDU and EUT disappears at higher values α, while the difference in the AWC persists. All the other qualitative results on marginal welfare costs, relations between the measures, and observable heterogeneity hold under the RDU assumption.



(a) (b)

Figure A.2: Absolute and Relative Welfare Costs for EUT vs. RDU, $\alpha = 0.9$.

Second, we consider a different specification for the utility function under EUT, an expo-power (EP) utility, which generalizes the CRRA and CARA utility functions

$$u(x; \gamma_a, \gamma_r) = \frac{1 - \exp(-\gamma_a x^{1-\gamma_r})}{\gamma_a},$$

where γ_a and γ_r are the two parameters to be estimated. This specification does not do so well in modeling subjects' risk preferences in our data. For a large (40%) fraction of subjects the estimation procedure yields unreasonably high parameter values, which impedes the calculation of certainty equivalents and welfare costs. We use CRRA specification for these subjects when presenting the results on Figure A.3. They look very similar to the baseline specification with the CRRA utility function.

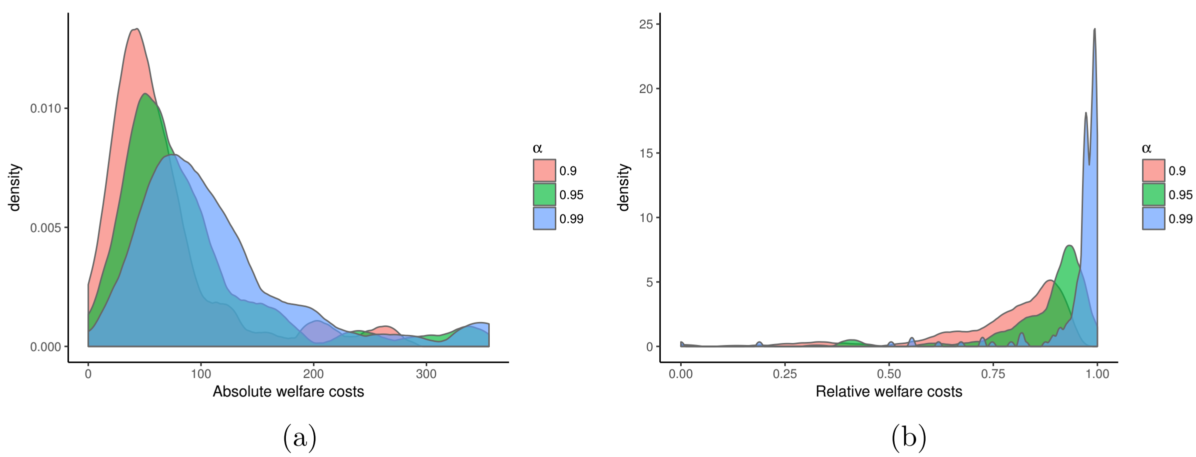
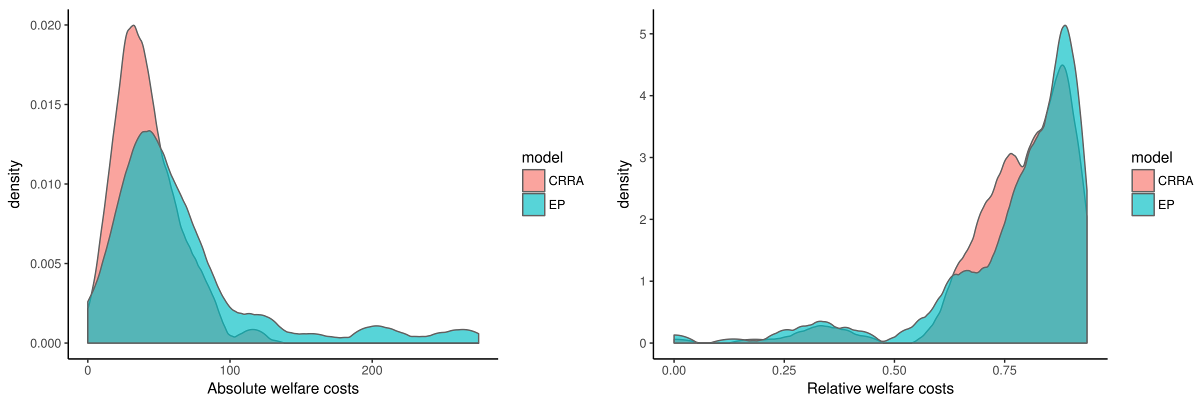


Figure A.3: Absolute and Relative Welfare Costs for 3 Levels of α, EP.

Looking at the differences between the AWC calculated under the two utility specifications, our baseline specification again provides lower values (see Figure A.4a). The difference in the medians between the AWC calculated using CRRA vs. EP is -16.17 (Wilcoxon signed rank test, p-value < 0.001), for $\alpha = 0.9$. The mean of the differences is -24.7 DKK (approximately -4 USD). The AWC under the EP utility function are roughly 60% higher than in the baseline, which is even higher than in the case of the RDU model as an alternative.

At the same time there are no significant differences in the RWC between the two utility specifications (Wilcoxon signed rank test, p-value $= 0.52$). The same pattern of results hold for other values of α. Under the EP-utility assumption the marginal welfare costs have a similar shape, but the association between the measures becomes weaker, as do the effects of observable heterogeneity.



(a) (b)

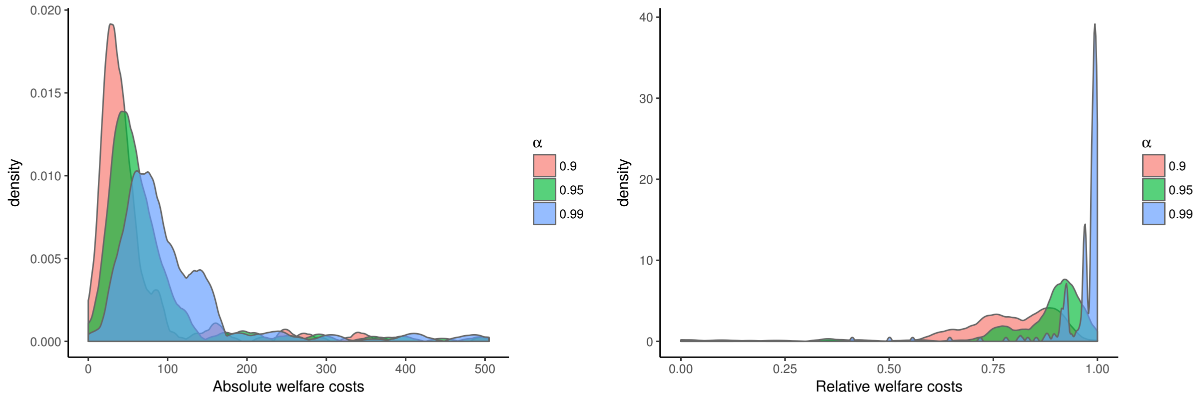
Figure A.4: Absolute and Relative Welfare Costs for CRRA vs. EP, $\alpha = 0.9$.

Finally, we look at an alternative stochastic choice specification, the contextual utility model due to Wilcox (2011), which allows for a heterogeneous noise term and preserves the "more risk averse" relation in the stochastic domain. This specification of noise has been shown by (Wilcox, 2015) to have good out-of-sample predictive power. Under the assumption of contextual utility the choice probabilities become

$$p(a_2; \gamma, \mu) = \Lambda \left(\frac{U(a_2; \gamma) - U(a_1; \gamma)}{\mu \left(u(x_{(1)}; \gamma) - u(x_{(k)}; \gamma) \right)} \right),$$

where we drop the index for the decision round, and $p(a_1; \gamma, \mu) = 1 - p(a_2; \gamma, \mu)$. As before, $x_{(1)}$ and $x_{(k)}$ denote the highest and lowest outcomes, but this time they are defined only among the outcomes that occur with positive probabilities, and outcomes are ranked *across both lotteries* in the choice.

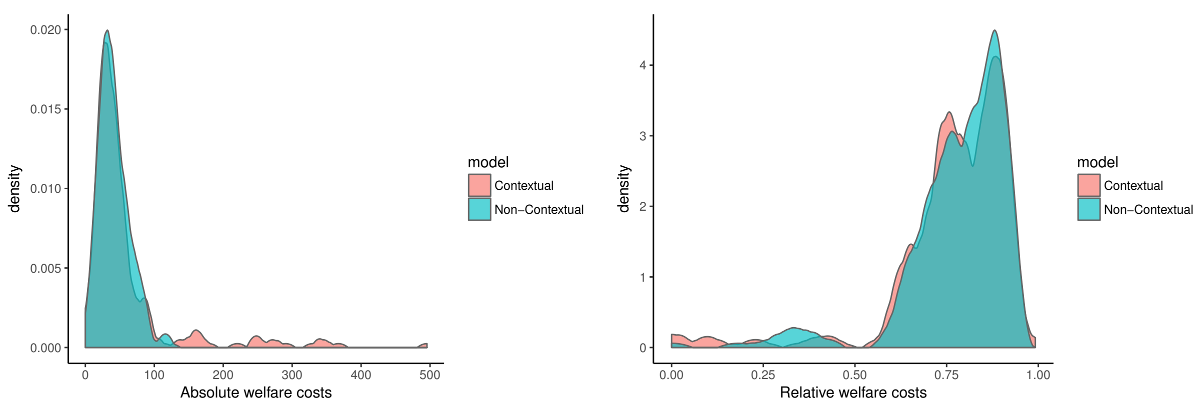
Figure A.5 shows the calculated AWC and RWC under the assumption of contextual utility. These graphs, again, look very similar to those under EUT and no contextual utility (Figure 4), except that the right tails in the distributions of the AWC become thicker.



(a) (b)

Figure A.5: Absolute and Relative Welfare Costs for Three Levels of α, Contextual Utility.

Figure A.6 contrasts the AWC and RWC for the baseline and alternative specifications of noise. The densities of the AWC are very much alike, except for a thicker right tail in the case of contextual utility, which leads to higher welfare costs. The difference in the medians between the AWC calculated using non-contextual vs. contextual models is -1.87 (Wilcoxon signed rank test, p-value < 0.001). The mean of the differences is -16.33 DKK (approximately -2 USD). This result is comparable to the non-contextual noise specification with RDU as an alternative. Again, there is no significant difference between the RWC for these two models (Wilcoxon signed rank test, p-value ≈ 0.47). All the results reported for the baseline model hold in the case of contextual utility model as well.



(a) (b)

Figure A.6: Absolute and Relative Welfare Costs for Non-contextual vs. Contextual Utility, $\alpha = 0.9$.

B Proofs

Consider an implicit function $\rho(\mu, \varepsilon) = \alpha$. From the implicit function theorem, it follows that

$$\frac{d\varepsilon}{d\mu} = -\frac{\partial\rho/\partial\mu}{\partial\rho/\partial\varepsilon}.$$

The denominator of this expression is

$$\frac{\partial\rho}{\partial\varepsilon} = \frac{\partial}{\partial\varepsilon}\int_{a_l^*(\varepsilon)}^{a_h^*(\varepsilon)} p(a)da = p(a_h^*(\varepsilon))a_h^{*'}(\varepsilon) - p(a_l^*(\varepsilon))a_l^{*'}(\varepsilon) \geqslant 0,$$

since $a_h^{*'}(\varepsilon) \geqslant 0$, and $a_l^{*'}(\varepsilon) \leqslant 0$.

In order to show the sign of the numerator, we restrict our attention to the binary choice case, since it is the setting of our primary interest. Recall that

$$p(a_2; \gamma, \mu) = \Lambda\left(\frac{U(a_2; \gamma) - U(a_1; \gamma)}{\mu}\right).$$

Then

$$\frac{\partial p(a_2; \gamma, \mu)}{\partial\mu} = \Lambda'\left(\frac{U(a_2; \gamma) - U(a_1; \gamma)}{\mu}\right)(U(a_2; \gamma) - U(a_1; \gamma))(-\mu^2) < 0,$$

since alternative a_2 gives the highest certainty equivalent by our assumption. Therefore,

$$\frac{\partial\rho}{\partial\mu} = \begin{cases} \frac{\partial p(a_2;\gamma,\mu)}{\partial\mu}, & \varepsilon < \Delta m, \\ 0, & \varepsilon \geqslant \Delta m, \end{cases}$$

so that $\partial\rho/\partial\mu \leqslant 0$. Together the two results imply that $d\varepsilon/d\mu \geqslant 0$.

C Additional Tables

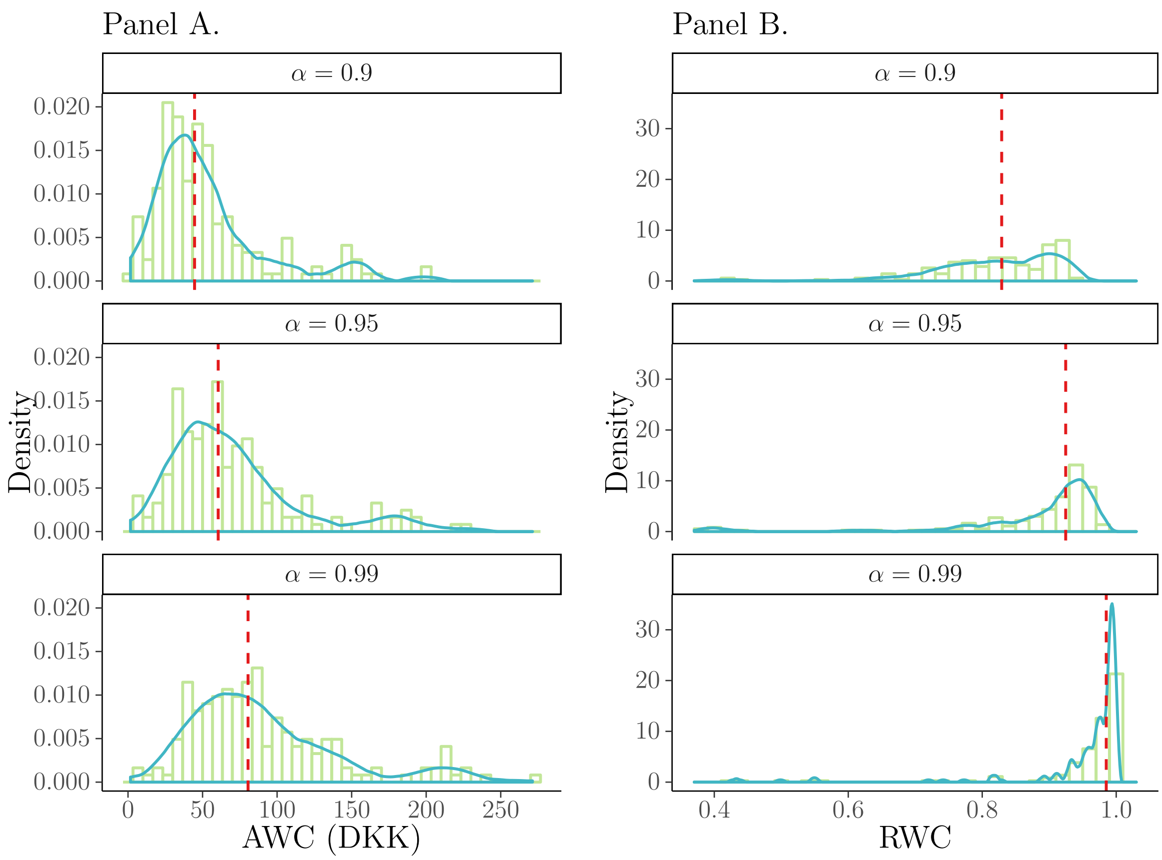
Table C.1: The Battery of Lotteries

ID	La1	Lp1	La2	Lp2	La3	Lp3	La4	Lp4	Ra1	Rp1	Ra2	Rp2	Ra3	Rp3	Ra4	Rp4
1	450	0.50	1,350	0	2,250	0.50	0	0	450	0.10	1,350	0.80	2,250	0.10	0	0
2	450	0.50	1,350	0	2,250	0.50	0	0	450	0	1,350	1	2,250	0	0	0
3	450	0.10	1,350	0.80	2,250	0.10	0	0	450	0	1,350	1	2,250	0	0	0
4	450	0.70	1,350	0	2,250	0.30	0	0	450	0.50	1,350	0.40	2,250	0.10	0	0
5	450	0.70	1,350	0	2,250	0.30	0	0	450	0.40	1,350	0.60	2,250	0	0	0
6	450	0.50	1,350	0.40	2,250	0.10	0	0	450	0.40	1,350	0.60	2,250	0	0	0
7	450	0.40	1,350	0	2,250	0.60	0	0	450	0.10	1,350	0.75	2,250	0.15	0	0
8	450	0.40	1,350	0	2,250	0.60	0	0	450	0	1,350	1	2,250	0	0	0
9	450	0.30	1,350	0	2,250	0.70	0	0	450	0.15	1,350	0.25	2,250	0.60	0	0
10	450	0.10	1,350	0.75	2,250	0.15	0	0	450	0	1,350	1	2,250	0	0	0
11	450	0.70	1,350	0	2,250	0.30	0	0	450	0.60	1,350	0.25	2,250	0.15	0	0
12	450	0.70	1,350	0	2,250	0.30	0	0	450	0.50	1,350	0.50	2,250	0	0	0
13	450	0.60	1,350	0.25	2,250	0.15	0	0	450	0.50	1,350	0.50	2,250	0	0	0
14	450	0.40	1,350	0	2,250	0.60	0	0	450	0.20	1,350	0.60	2,250	0.20	0	0
15	450	0.40	1,350	0	2,250	0.60	0	0	450	0.10	1,350	0.90	2,250	0	0	0
16	450	0.20	1,350	0.60	2,250	0.20	0	0	450	0.10	1,350	0.90	2,250	0	0	0
17	450	0.60	1,350	0	2,250	0.40	0	0	450	0.50	1,350	0.30	2,250	0.20	0	0
18	450	0.30	1,350	0	2,250	0.70	0	0	450	0	1,350	0.50	2,250	0.50	0	0
19	450	0.60	1,350	0	2,250	0.40	0	0	450	0.40	1,350	0.60	2,250	0	0	0
20	450	0.50	1,350	0.30	2,250	0.20	0	0	450	0.40	1,350	0.60	2,250	0	0	0
21	450	0.25	1,350	0	2,250	0.75	0	0	450	0.10	1,350	0.60	2,250	0.30	0	0
22	450	0.25	1,350	0	2,250	0.75	0	0	450	0	1,350	1	2,250	0	0	0
23	450	0.10	1,350	0.60	2,250	0.30	0	0	450	0	1,350	1	2,250	0	0	0
24	450	0.50	1,350	0.20	2,250	0.30	0	0	450	0.40	1,350	0.60	2,250	0	0	0
25	450	0.55	1,350	0	2,250	0.45	0	0	450	0.40	1,350	0.60	2,250	0	0	0
26	450	0.55	1,350	0	2,250	0.45	0	0	450	0.50	1,350	0.20	2,250	0.30	0	0
27	450	0.15	1,350	0.25	2,250	0.60	0	0	450	0	1,350	0.50	2,250	0.50	0	0
28	450	0.15	1,350	0.75	2,250	0.10	0	0	450	0	1,350	1	2,250	0	0	0
29	450	0.60	1,350	0	2,250	0.40	0	0	450	0	1,350	1	2,250	0	0	0
30	450	0.60	1,350	0	2,250	0.40	0	0	450	0.15	1,350	0.75	2,250	0.10	0	0
31	135	0.55	1,620	0.25	1,890	0.20	0	0	135	0.55	1,215	0.25	2,430	0.20	0	0
32	810	0.40	675	0.40	1,620	0.20	0	0	810	0.40	405	0.40	2,025	0.20	0	0
33	1,485	0.40	675	0.40	1,620	0.20	0	0	1,485	0.40	405	0.40	2,025	0.20	0	0
34	2,160	0.40	675	0.40	1,620	0.20	0	0	2,160	0.40	405	0.40	2,025	0.20	0	0
35	675	0.70	1,485	0.10	2,835	0.20	0	0	675	0.70	945	0.10	3,375	0.20	0	0
36	1,620	0.70	1,485	0.10	2,835	0.20	0	0	1,620	0.70	945	0.10	3,375	0.20	0	0
37	2,565	0.70	1,485	0.10	2,835	0.20	0	0	2,565	0.70	945	0.10	3,375	0.20	0	0
38	3,510	0.70	1,485	0.10	2,835	0.20	0	0	3,510	0.70	945	0.10	3,375	0.20	0	0
39	0	0.50	540	0.10	540	0.40	0	0	0	0.50	0	0.10	810	0.40	0	0
40	540	0.50	540	0.10	540	0.40	0	0	540	0.50	0	0.10	810	0.40	0	0
41	1,080	0.50	540	0.10	540	0.40	0	0	1,080	0.50	0	0.10	810	0.40	0	0
42	945	0.55	1,620	0.25	1,890	0.20	0	0	945	0.55	1,215	0.25	2,430	0.20	0	0
43	1,620	0.50	540	0.10	540	0.40	0	0	1,620	0.50	0	0.10	810	0.40	0	0
44	540	0.50	1,080	0.10	1,080	0.40	0	0	540	0.50	540	0.10	1,350	0.40	0	0
45	1,080	0.50	1,080	0.10	1,080	0.40	0	0	1,080	0.50	540	0.10	1,350	0.40	0	0
46	1,620	0.50	1,080	0.10	1,080	0.40	0	0	1,620	0.50	540	0.10	1,350	0.40	0	0
47	2,160	0.50	1,080	0.10	1,080	0.40	0	0	2,160	0.50	540	0.10	1,350	0.40	0	0
48	1,755	0.55	1,620	0.25	1,890	0.20	0	0	1,755	0.55	1,215	0.25	2,430	0.20	0	0
49	2,565	0.55	1,620	0.25	1,890	0.20	0	0	2,565	0.55	1,215	0.25	2,430	0.20	0	0
50	135	0.65	945	0.20	1,485	0.15	0	0	135	0.65	810	0.20	1,620	0.15	0	0
51	675	0.65	945	0.20	1,485	0.15	0	0	675	0.65	810	0.20	1,620	0.15	0	0
52	1,215	0.65	945	0.20	1,485	0.15	0	0	1,215	0.65	810	0.20	1,620	0.15	0	0
53	1,755	0.65	945	0.20	1,485	0.15	0	0	1,755	0.65	810	0.20	1,620	0.15	0	0
54	135	0.40	675	0.40	1,620	0.20	0	0	135	0.40	405	0.40	2,025	0.20	0	0
55	0	0	0	0	0	0	1,200	1	0	0	0	0	975	0.50	1,440	0.50
56	0	0	0	0	0	0	1,275	1	0	0	0	0	1,155	0.50	1,410	0.50
57	0	0	0	0	0	0	450	1	0	0	0	0	225	0.50	690	0.50
58	0	0	0	0	0	0	1,950	1	0	0	0	0	1,725	0.50	2,190	0.50
59	0	0	0	0	0	0	2,025	1	0	0	0	0	1,905	0.50	2,160	0.50
60	0	0	0	0	0	0	225	1	0	0	0	0	105	0.50	360	0.50

Notes. The columns are coded as follows: "L" and "R" denote left and right lottery, "a" denotes amounts (in DKK) and "p" denotes probabilities. The amounts in the table are baseline amounts. In addition to these amounts, 1.5x and 2x amounts were used. The subjects were randomized across the baseline, 1.5x and 2x amounts.

D Additional Graphs

Figure D.7: Distributions of AWC and RWC in the Subset of Subjects



Note: The graph shows the distributions of the individual-level estimates of AWC (Panel A) and RWC (Panel B) for three target levels of α: 0.9, 0.95, and 0.99. The sample is restricted to include only the subjects for whom the estimation procedure successfully converged. The bars are the histograms and the smooth lines are the kernel density estimates. The dashed lines show the medians of the distributions. The AWC numbers are in DKK. For the RWC, we truncate the support at 0.4 to improve readability of the graph. This results in dropping 8 observations for which the RWC are below 0.4.