the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2021)

www.wileyonlinelibrary.com

Empirical underidentification in estimating random utility models: The role of choice sets and standardizations

Sebastian Olschewski[1,2*], Pavel Sirotkin[1] and Jörg Rieskamp[1]

[1]Center for Economic Psychology, University of Basel, Switzerland
[2]Warwick Business School, University of Warwick, Coventry, UK

A standard approach to distinguishing people's risk preferences is to estimate a random utility model using a power utility function to characterize the preferences and a logit function to capture choice consistency. We demonstrate that with often-used choice situations, this model suffers from empirical underidentification, meaning that parameters cannot be estimated precisely. With simulations of estimation accuracy and Kullback–Leibler divergence measures we examined factors that potentially mitigate this problem. First, using a choice set that guarantees a switch in the utility order between two risky gambles in the range of plausible values leads to higher estimation accuracy than randomly created choice sets or the purpose-built choice sets common in the literature. Second, parameter estimates are regularly correlated, which contributes to empirical underidentification. Examining standardizations of the utility scale, we show that they mitigate this correlation and additionally improve the estimation accuracy for choice consistency. Yet, they can have detrimental effects on the estimation accuracy of risk preference. Finally, we also show how repeated versus distinct choice sets and an increase in observations affect estimation accuracy. Together, these results should help researchers make informed design choices to estimate parameters in the random utility model more precisely.

1. Introduction

Measuring people's risk preferences is one of the main research interests in economics and psychology as well as in many everyday-life domains. For example, financial advisers need to assess the level of risk a client is willing to take to give sensible investment advice. Similarly, a physician has to know the patient's willingness to take risks when discussing surgery and comparing it to a conservative therapy. Other domains where personal risk preferences play a role are traffic psychology, the insurance market, career choices, and vacation destinations. In these domains people might not be able to fully understand an option's implied risk, for instance, the risk of a financial product or of a medical treatment, so they cannot identify the option that corresponds to their risk preferences. Therefore, expert advice in these areas is crucial and experts need to take people's risk preferences into account.

*Correspondence should be addressed to Sebastian Olschewski, Center for Economic Psychology, University of Basel, Missionsstrasse 62A, 4056 Basel, Switzerland (email: sebastian.olschewski@unibas.ch).

DOI:10.1111/bmsp.12256

People's risk preferences can be conceptualized as a (relatively) stable personal propensity across different life domains (cf. Frey, Pedroni, Mata, Rieskamp, & Hertwig, 2017; Weber, Blais, & Betz, 2002). Risk preferences can be measured with different risk-taking paradigms, such as repeated choices between risky gambles. When one assumes that people make consistent choices between risky gambles, their behaviour can be understood as the maximization of expected utility (von Neumann & Morgenstern, 1944). The parameter of the utility function that best captures observed behaviour then provides a quantitative measurement of risk preferences that can be generalized to other domains.

However, stochastic behaviour complicates the elicitation of risk preference, and a common way to deal with this is the implementation of random utility models (RUMs). In this paper we present challenges faced by this model and use simulation and recovery analysis as well as Kullback–Leibler divergence measures to examine methods that help meet these challenges.

1.1. Utility functions and stochastic behaviour

Since Bernoulli (1954 [1738]), most researchers have used a nonlinear utility function that maps objective outcomes to subjective utility. With a concave utility function, higher objective outcomes are discounted and the marginal utility of an additional unit of outcome decreases. The utility of the expected value of a risky lottery is thus higher than its expected utility. This implies that people prefer the expected value of a lottery as a certain outcome over playing that lottery, representing risk aversion. In contrast, with a convex utility function, the expected utility of a lottery is higher than the utility of the expected value, representing risk seeking. Finally, with a linear utility function, the expected utility of a lottery is equal to the utility of the expected value, representing risk neutrality.

The extent of the curvature of the utility function reflects the degree of risk aversion or risk seeking. Thus, estimating the utility function provides a quantitative measure of people's risk preferences. A utility function can be estimated from the elicited certainty equivalent for one lottery. However, as risky choices are stochastic (Mosteller & Nogee, 1951; Rieskamp, 2008), people do not always choose the same certainty equivalent (Schmidt & Hey, 2004), or when choosing repeatedly between lotteries they do not always make the same choices (Hey, 2001). When participants choose twice between the same lotteries, the percentage of lottery pairs where people choose the same lottery is a model-free measure of choice consistency. Measured this way, consistency in risky choice is estimated to be as low as around 75–85% on average (Glöckner & Pachur, 2012; Hey, 2001; Hey & Orme, 1994; Starmer & Sugden, 1989). This ignores that choice consistency can be a function of the choice situation, but gives a first approximation of the magnitude of the problem.

This lack of consistency translates into measurement errors of the utility function and makes it necessary to estimate this function based on many choices. However, there is no consensus in the applied literature on how many choices are required for reliable estimates of people's risk preferences. In some work, individual utility functions have been estimated based on 10–20 pairwise choices (Anderson, Harrison, Lau, & Rutström, 2007; Dohmen, Falk, Huffman, & Sunde, 2010; Holt & Laury, 2002), whereas others have gone up to almost 200 choices (e.g., 84 in Frey et al., 2017; 90 in Stott, 2006; 100 in Hey & Orme, 1994, and Hey, 2001; 180 in Rieskamp, 2008).

Similarly, whereas there is theoretical work about optimal experimental designs (Myung & Pitt, 2009; Navarro, Pitt, & Myung, 2004), there is no consensus in the applied literature on what choices to use. This might be because optimal design algorithms often cannot be straightforwardly applied to experiments without further assumptions. Often

in the applied literature, risk preferences have been measured with the choice set proposed by Holt and Laury (2002), but other sets have been used as well. In an attempt to compare different specifications of cumulative prospect theory, a different set of choices was created by Stott (2006), and to explore the stochastic nature of risky choice, yet another choice set was created by Rieskamp (2008). Finally, to examine the stability of parameter estimates in risky choice, a mixture of gambles from both Holt and Laury (2002) and Rieskamp (2008) (among others) were used by Glöckner and Pachur (2012). These choice sets are thought to be superior to choice sets that are created just randomly and usually invoke heuristic concepts of being informative, for example, by excluding situations with dominating gambles.

1.2. Random utility models

Given the stochasticity of choice behaviour, fitting an expected utility theory to data requires a mapping of utility differences to choice probabilities. RUMs characterize people's risk preferences as well as the consistency of their behaviour (for an overview, see Loomes & Pogrebna, 2014; Rieskamp, Busemeyer, & Mellers, 2006; Train, 2009).

Estimating people's choice consistency and exploring how choice consistency differs as a function of the choice environment are important research questions in their own right. Recently, economists and psychologists alike have shown interest in understanding how choice inconsistency can be derived from more basic principles of information perception, representation and processing (Bhui & Gershman, 2018; Polania, Woodford, & Ruff, 2019; Woodford, 2020). It has also been shown that consistency in risky choices is correlated with a person's cognitive abilities (Andersson, Holm, Tyran, & Wengström, 2016). Another approach is to see how differences in momentary cognitive resources can shape consistency. For example, dual-task or time-pressure manipulations affected choice consistency in risk taking (Olschewski & Rieskamp, 2021; Olschewski, Rieskamp, & Scheibehenne, 2018).

For our simulation analyses, we employed the class of power utility functions that has often been used in the decision-making literature (Stott, 2006; Tversky & Kahneman, 1992). The function U maps an outcome x_1 to its (average) subjective utility and has a free parameter, α, that captures risk preferences:

$$U(x_1) = x_1^{\alpha} + \varepsilon, \tag{1}$$

where $\alpha < 1$ signifies concave, $\alpha > 1$ convex, and $\alpha = 1$ linear utility, which correspond to risk-averse, risk-seeking, and risk-neutral preferences. In this model, utility is conceptualized as a random variable with the error term ε with mean 0 and constant variance. To illustrate the model's prediction, we consider two lotteries x and y with two outcomes each, x_1, x_2 and y_1, y_2, that occur with probability p_{x1}, $1 - p_{x1}$ and p_{y1}, $1–p_{y1}$, respectively. The corresponding expected utilities are

$$E[U(x)] = p_{x1} \cdot U(x_1) + (1 - p_{x1}) \cdot U(x_2),$$

$$E[U(y)] = p_{y1} \cdot U(y_1) + (1 - p_{y1}) \cdot U(y_2). \tag{2}$$

Assuming that the error term in equation (1) is extreme-value distributed implies a logit function to determine choice probabilities as a function of expected utility differences:

$$p(y) = \frac{1}{1 + \exp(-\theta \cdot (E[U(y)] - E[U(x)]))}, \tag{3}$$

where θ governs the amount of choice consistency, with higher θs meaning more consistent choices than lower θs for a given expected utility difference. As an illustration, consider the choice between 50 for certain or a 50% chance of winning 100 or else nothing. The expected value of both choice options is the same and thus a risk-averse person (e.g., one with $\alpha = 0.8$) would prefer the sure outcome, meaning that person would choose it more than 50% of the time. Conversely, a moderately risk-seeking person with $\alpha = 1.2$ would choose the lottery more than 50% of the time for all θs. The logit function is often applied in estimating risk preference from empirical choice data (Rieskamp, 2008; Scheibehenne & Pachur, 2015). It can also be derived from a fixed utility framework, assuming that utilities are deterministic and the error term only enters at the choice stage (Luce, 1957; also called softmax, Sutton & Barto, 2018).

1.3. Empirical underidentification

An important condition for every model is that its parameters, here α and θ, are identifiable. A model is identifiable if, for any observable behaviour, there is at most one set of parameters that predicts this behaviour (Bamber & van Santen, 2000). However, even when this technical condition is fulfilled for a model so that it is identifiable, a model can still have problems with empirical underidentification (Schmittmann, Dolan, Raijmakers, & Batchelder, 2010). This means that, given a particular set of stimuli, different sets of parameter estimates can lead to very similar predictions, thus making it difficult to estimate parameters precisely. To illustrate this problem, we take two choice situations from a choice set often used to estimate parameters of RUMs in the literature (Rieskamp, 2008). In situation 1, the choice is *A*: win 18 with a 37% chance, or else 41, or *B*: win 8 with a 2% chance, or else 56. In situation 2, the choice is *C*: win 27 with a 76% chance, or else 47, or *D*: win 29 with a 12% chance, or else 45. Note that in both situations there is no stochastic dominance and since we have a model with two free parameters and two observations, we should in principle be able to identify a unique parameter combination that explains a given choice pattern. However, taking again our prototypical risk-averse ($\alpha = 0.8$) and risk-seeking ($\alpha = 1.2$) agents, choice proportions can be empirically undistinguishable for both agents if we adjust θ accordingly. Take, for example, the parameter combinations $\alpha = 0.8$ and $\theta = 0.12$, and $\alpha = 1.2$ and $\theta = 0.0188$, which both lead to choice probabilities of 74% for *B* over *A* and 63% for *D* over *C*. Thus, although the choice proportions are not identical (probabilities were rounded above) the predictions are very similar and we cannot hope to distinguish them with a realistic amount of data (see also Alempaki et al., 2019; Stewart, Canic, & Mullet, 2019). As a consequence, this can lead to completely different interpretations of a person's behaviour when estimating the model's parameters. A risk-averse person could, for instance, be classified as a risk seeker.

One reason why a model can suffer from empirical underidentification is if the model's parameter estimates can (partly) trade off each other, given a particular set of observations (see also Krefeld-Schwalb, Pachur, & Scheibehenne, 2021; Spektor & Kellen, 2018). As the scale of the expected utility difference depends on α and is multiplied by θ, this leads to a correlation between α and θ, meaning that higher α estimates go along with lower θ estimates for the same level of consistency. Mechanistically, the higher is the estimate of α, the higher is the expected utility difference of a given pair of lotteries, and thus to have a

similar level of choice consistency, the estimate of θ must be smaller for higher than for lower α estimates (see Stewart, Scheibehenne, & Pachur, 2018). To illustrate this effect, consider lottery *E*: a 50% chance of winning 80, or else 20. For a risk-averse agent with $\alpha = 0.8$, this lottery has a certainty equivalent of approximately 48. If this agent now has to choose between lottery *E* and a sure amount of 46, which is two units below the agent's certainty equivalent, for a $\theta = 1$ this would result in a predicted choice proportion of the lottery of 68%. In contrast, for a risk-seeking agent with $\alpha = 1.2$, the certainty equivalent of lottery *E* is approximately 52. If this agent decides between lottery *E* and a certain amount of 50, that is, again two units lower than the agent's certainty equivalent, for the same θ this results in a choice proportion for the lottery of 99%. To get a choice proportion of 68% for the risk-seeking agent to choose the lottery, θ has to be reduced from 1 to 0.15. Consequently, it is not possible to compare θ estimates as an indicator of choice consistency across different levels of risk preference.

1.4. Research design

Given the problem of empirical underidentifiability, the question is how to estimate risk preference and choice consistency as accurately as possible. In this paper we identify three factors that affect estimation accuracy, namely, the *stimulus design*, the *estimation method*, and the *repetition of choice sets*. As shown in the examples above, choices between two gambles can be non-informative with respect to the mapping of choices to a utility function. For that reason, we would expect a randomly created choice set as stimuli for an experiment to perform poorly in estimating our model. This is in line with the idea of an optimal experimental design. An optimal experimental design is a choice set that best enables the measurement of the parameters of a given mathematical model or best distinguishes between two competing models (Myung & Pitt, 2009; Navarro et al., 2004). However, it is difficult to find the optimal choice set when multiple factors can affect estimation. In the case of risky choice, not only does the expected value differ between two lotteries, but also stochastic dominance, or more generally, the relation of the cumulative distribution functions of two lotteries, can affect the estimation accuracy of RUMs. Therefore, researchers so far have either relied on randomly created choice sets or used heuristics to create choice sets. In a recent comparison, these heuristically created choice sets did not outperform randomly created choice sets in estimating cumulative prospect theory parameters (Broomell & Bhatia, 2014). Here, we examine a new algorithm to create a choice set that improves estimation accuracy over randomly created choice sets and choice sets used in the literature so far, as we describe below.

We also illustrated that the scale on which choice consistency is measured differs depending on the risk preference. Therefore, standardizing the expected utility difference to be on a similar scale for risk-averse and risk-seeking agents in the estimation should help mitigate the correlation between estimates. This can be done because utility is usually measured on an interval scale, and thus the absolute values have no meaning (see Wakker, 2008). In the following we examine the four most prominent standardization approaches, which we call *utility*, *outcome*, *monetary equivalence* and *variance* standardization and define mathematically below. With parameter recoveries we probe these standardizations on their ability to improve estimation accuracy. Note that two of these standardizations, utility and outcome standardization, lead to context dependencies, meaning that the parameter estimates depend on the choice set. Therefore, these models do not satisfy an assumption of strict utility models, as the logit function in equation (3) does (Luce & Suppes, 1965; Wilcox, 2011). Regardless of whether strict utility models are descriptively

plausible (cf. Lieder, Griffiths, & Hsu, 2018; Wilcox, 2015), such an approach is viable if a researcher is interested predominantly in the measurement of preferences and/or consistency in similar choice contexts.

Finally, estimation accuracy should increase when we give the same choice set repeatedly, since that increases the number of observations. Yet, it is less clear whether it is more informative to give the same choice repeatedly or use distinct choices to estimate parameters. Also as illustrated, researchers vary greatly in the number of choices they deem sufficient to estimate RUMs. Therefore, we systematically examined how the number of choices affects parameter estimation accuracy.

2. Method

2.1. Estimation accuracy

We conducted parameter recoveries to examine the effects of the three factors identified above on estimation accuracy and correlations. To measure estimation accuracy, we defined the bias as the difference between parameter estimates and the data-generating values. In this measure, negative and positive deviations can cancel each other out. Hence, we additionally use the absolute deviation between parameter estimates and data-generating values to measure estimation accuracy. We interpret the average absolute deviation for one parameter as the expected measurement error when estimating the model. For this analysis, we picked two levels of risk preference, risk aversion ($\alpha = 0.8$) and risk seeking ($\alpha = 1.2$), and calibrated choice consistency to a value that led to approximately 80% of choices being consistent with the (average) utility order, a value in accordance with empirical findings, as mentioned in the Introduction. This choice consistency θ depends on the choice set as well as on the standardizations implemented and thus varies across different specifications. To meaningfully compare bias and estimation accuracy across different data-generating α- and θ-values, we divided both measures by the respective data-generating value to calculate the relative bias and estimation accuracy, respectively.

We simulated risky choices from one agent in 60 choice situations, which is a number of choices in the range of the amounts of data used in the applied literature to estimate risk preference. The choice simulation depends on the standardization: it corresponds to the model specified in equations (1) and (2) in the baseline case without standardization but varies for the respective standardizations as outlined below. After the choices were simulated, we tried to recover the parameters by means of maximum likelihood estimation with a standard algorithm in R using the data-generating values as starting points[1] (Nelder-Mead in *optim*; R Core Team, 2016; RStudio Team, 2015). Thus, we tried to recover the parameters of each simulated data set with exactly the same model specifications that were used for simulation. That way we focus on the recoverability and abstract away from the question of which model best describes empirical choice data. We repeated the simulation and estimation 10,000 times and report summary statistics. Overall, we implemented recoveries for three different choice sets, four different standardization methods in addition to the baseline case without standardization, and with various repetitions of the choice (sub)set.

[1] In experiments, the data-generating parameter values are unknown. However, in experimental applications, this method can be approximated by estimating the model repeatedly with randomly selected starting values and then selecting the parameter estimates with the highest likelihood. In our analyses the two methods led to similar results.

2.2. *Kullback–Leibler divergence*

Estimation accuracy as an outcome measure has the advantage that the absolute magnitude can be intuitively interpreted. As a disadvantage, the accuracy measures used above are not formally rooted in information theory and can depend on the concrete parameter values we chose for simulation. Therefore, we additionally estimated the Kullback–Leibler (KL) divergence between different parameter values for a given choice set (Chang & Ying, 1996). This measures the extent to which two parameter combinations mimic each other in the predicted choice proportions. In the following we use the notation and estimation method proposed in Broomell and Bhatia (2014) and consequently define the multivariate parameter discrimination (MPD) measure for a given choice set and standardization method as follows:

$$\mathrm{MPD}_{[\alpha,\theta]} = p(\alpha_0,\ \theta_0) \cdot p(\alpha_1,\ \theta_1) \cdot D_{\mathrm{KL}}[p(c|\alpha_0,\ \theta_0) \| \ p(c|\alpha_1,\ \theta_1)], \tag{4}$$

where D_{KL} is the KL divergence between the conditional probability of simulated data c given two different sets of free parameters (α_0, θ_0) and (α_1, θ_1). The first two factors in equation (4) are the prior probabilities of the respective parameter sets. For the power utility parameter α we specified a uniform probability distribution with a range from 0.01 to 1.99. Similarly, we specified a uniform probability distribution for choice consistency θ with a range calibrated to lead to consistencies between 51% and 99% for each choice set and each standardization. As a result of this specification, the prior probabilities are the same for every parameter combination and thus the first two factors in this equation reduce to a scaling variable.

One can also calculate divergence measures for individual parameters by fixing one of the two parameters in equation (4). For example, when fixing $\theta_0 = \theta_1$, we denote the resulting divergence measure by $\mathrm{MPD}_{[\alpha]}$. However, this measure does not take into account that the divergence could be affected by imprecise estimates of θ in the case where estimates of α and θ are correlated. Therefore, Broomell and Bhatia (2014) introduced the univariate parameter discrimination (UPD) measure that takes the effect of θ into account and can be calculated as follows:

$$\mathrm{UPD}_{[\alpha]} = 0.5 \cdot (\mathrm{MPD}_{[\alpha,\theta]} - \mathrm{MPD}_{[\theta]}) + 0.5 \cdot (\mathrm{MPD}_{[\alpha]} - 0). \tag{5}$$

Thus, UPD is a measure of the discriminability of the individual parameter (here α) under conditions where parameter estimates are correlated. This measure can similarly be calculated for θ.

Finally, from the individual and overall divergence measures one can calculate a measure of the percentage reduction in discrimination through the estimation inaccuracy of the other parameter, called percent reduced discrimination (PRD). This measure is between 0 and 1 and is higher the more one parameter estimate is affected by the estimation of the other parameter:

$$\mathrm{PRD}_{[\alpha]} = 1 - (\mathrm{UPD}_{[\alpha]}/\mathrm{MPD}_{[\alpha]}). \tag{6}$$

We simulated the KL divergence 100,000 times by randomly choosing two parameter sets from the specified prior distributions. As a disadvantage, this method depends critically on the outcome scale, so we cannot use this measure to examine the standardization methods.

2.3. Stimulus design: composition of choice sets

As suggested in the Introduction, the problem of empirical underidentification of our model depends on the choice set implemented. To demonstrate how to improve estimation accuracy, we constructed three choice sets, each with the same number of choice situations.

2.3.1. Random choice set

The first set was created by randomly selecting four outcomes between 1 and 99 drawn without replacement as the two outcomes for both lotteries. Then two numbers between .01 and .99 (rounded to two digits) were drawn without replacement, where the first draw was the probability of outcome 1 in the first lottery and the second draw the probability of outcome 1 in the second lottery. The probability for the second outcome in both lotteries just followed from the remaining probability value to add up to 1. No further criteria were invoked, and for each simulation a new random set was created. This choice set should function as a baseline of estimation accuracy.

2.3.2. No-dominance choice set

The second choice set consisted of 60 pairs of two-outcome lotteries in the gain domain, a set that has been frequently used in other publications (Glöckner & Pachur, 2012; Rieskamp, 2008; Scheibehenne & Pachur, 2015). It was created by choosing outcomes randomly between 0 and 100 and choosing outcome probabilities randomly between 0 and 1 (rounded to two digits). So far this resembles the creation method used for the first choice set, but two extra criteria were imposed: first, the set included no stochastically dominant lottery pairs; and second, the set included only lottery pairs where the ratio between the absolute expected value difference between the two lotteries and the smaller of the two expected values was less than 1. Both criteria should make choices more informative and thus increase estimation accuracy.

2.3.3. Switching choice set

Finally, as shown in the example in the Introduction, even non-dominant lottery pairs can fail to reliably distinguish between risk-averse and risk-seeking preferences. Therefore, we propose a new method to create lottery pairs: again, outcomes and probabilities were chosen randomly, but different bins of expected value differences as well as variance differences were created. This was done to obtain a spectrum of lottery pairs that was informative for very risk-averse as well as very risk-seeking people. Most importantly, in this choice set we added the constraint that the power utility function would switch the (average) ordinal utility order of the two lotteries (the lottery that is chosen with more than 50% probability) between $\alpha = 0.2$ and $\alpha = 2.8$. This means there were only lottery pairs in this set for which a choice was informative for the range of α estimates between 0.2 and 2.8. All conditions were implemented with accept–reject sampling, meaning that random combinations of numbers for outcomes and probabilities for a lottery pair were sampled repeatedly and combinations were accepted only when all the above conditions were fulfilled. The resulting choice set should allow for the highest estimation accuracy of the three proposed sets.

2.4. Estimation methods: standardizations of the utility scale

In the Introduction we also illustrated the problem of correlated estimates, which stems from the dependency of the expected utility scale on the estimated preference parameter.

Thus, a remedy for this correlation could be to standardize the expected utility scale and thus deviate from the standard logit framework. This means that for each α value the utility difference is on a similar scale, and consequently the absolute value of choice consistency θ leads to a similar proportion of choices that are consistent with the best-fitting utility curvature independent of α. In past work many researchers reported some form of standardization. However, different researchers proposed different *ad hoc* standardizations and they are usually not tested competitively against one another.

2.4.1. Utility standardization

A straightforward way of reducing the dependency of the expected utility scale on the risk preference parameter is to rescale expected utility to be always located between 0 and 1. Therefore, for a given α parameter the expected utilities of all lotteries in a choice set are calculated. Then a new expected utility $E[U_S]$ is calculated for each lottery:

$$E[U_S(x)] = \frac{E[U(x)] - \min_{z \in X} E[U(z)]}{\max_{z \in X} E[U(z)] - \min_{z \in X} E[U(z)]}, \tag{7}$$

where X comprises all lotteries present in the choice set under consideration. In this way the utility order and the relative distances between different lotteries in the choice set in terms of expected utility are preserved. At the same time the scale differences across different values of α are minimized.

2.4.2. Outcome standardization

A related but mathematically different approach is to rescale all outcomes to be between 0 and 1 (Olschewski et al., 2018). This has the effect that numbers between 0 and 1 stay between 0 and 1 even when taken to the power of a number larger than 1. This way the relation between a less concave (or convex) power function and the magnitude of utility is mitigated. We denote by O the set of all outcomes present in a particular choice set. The minimum and maximum of all outcomes in the set O are taken and each outcome x_i is transformed as follows:

$$x_{i,S} = \frac{x_i - \min_{z \in O} z}{\max_{z \in O} z - \min_{z \in O} z}. \tag{8}$$

2.4.3. Monetary equivalence standardization

Another approach to reduce the parameter correlation recently proposed by Stewart et al. (2018) is to retransform the expected utility scale back to the monetary scale and calculate the choice probabilities based on the monetary scale difference. This retransformation is calculated as follows:

$$E[U_M(x)] = [p_{x1} \cdot U(x_1) + (1 - p_{x1}) \cdot U(x_2)]^{1/\alpha}. \tag{9}$$

This way, within a lottery, large outcomes are transformed differently from smaller outcomes depending on α, but the expected utility is transformed back to the monetary scale and thus prevents higher values of α leading to higher inputs into the logit function.

2.4.4. Variance standardization

A fourth approach to decrease the correlation between α and θ is to divide the expected utility difference by the pooled variance of utilities. This idea was introduced by Busemeyer and Townsend (1993) to be part of the calculation of the drift rate in decision field theory, a sequential sampling model of decision-making, but it can also be used within RUMs, where the expected utility difference between two lotteries x and y reads

$$\Delta E[U_V(y,\ x)] = \frac{E[U(y)] - E[U(x)]}{\sqrt{\mathrm{Var}[U(y)]\ +\ \mathrm{Var}[U(x)]}},$$

$$p(y) = \frac{1}{1 + \exp(-\theta \cdot \Delta E[U_V(y,\ x)])}, \tag{10}$$

assuming independent lotteries. The reasoning behind this standardization is that higher values of the exponent lead to higher utility differences, but also to a higher pooled utility variance. Thus, dividing the difference by the pooled variance should in turn weaken the correlation between α and θ.

2.5. Number of repetitions

Giving participants the same choices repeatedly should increase estimation accuracy for the parameters of the RUM. This simply follows from the law of large numbers, namely, with 600 compared to 60 (independent) measurement points, that is, choices of a person, researchers should estimate any statistic with a smaller standard error. However, for pragmatic reasons (e.g., time, money, participant's attention), the number of choices that can be acquired is limited. Thus, trade-offs between amount of data and feasibility have to be made. Under these circumstances, we explore whether it is better to elicit data from 60 distinct choice situations or, for example, from 15 choice situations each four times. Intuitively, whereas it might be beneficial to have distinct choice situations to examine risk preferences, it could help estimates of choice consistency to have multiple choices for one choice situation. This intuition is built on the logic of observing behaviour in the same situations repeatedly as a model-free measure of consistency. In addition, we simulate how repeating the same set of 60 choices affects estimation accuracy when these estimates can be biased as well as how this interacts with the choice sets and the standardization of estimates.

3. Results

3.1. The effect of choice sets

The results of the parameter estimation accuracy analysis without any standardizations are presented for the risk preference parameter α in Table 1 and for the choice consistency parameter θ in Table 2. Both tables show substantial deviations from the data-generating parameters. In the random choice set, α was overestimated by approximately 10%. Biased α estimates were less of a problem in the no-dominance and the switching choice set, with on average 2% and less than 1% bias, respectively. Parameter θ was overestimated by over 400% in the random, by over 100% in the no-dominance, and by about 40% in the switching choice set.

Table 1. Results of basic parameter recovery for risk preference α

	Random choice set		No-dominance choice set		Switching choice set	
Variable	α = 0.8	α = 1.2	α = 0.8	α = 1.2	α = 0.8	α = 1.2
Mean estimate α	0.89 [0.24, 2.33]	1.31 [0.55, 3.00]	0.80 [0.40, 1.28]	1.24 [0.70, 2.07]	0.80 [0.56, 1.07]	1.21 [0.90, 1.55]
Relative bias	11% [−69%, 191%]	9% [−54%, 150%]	1% [−50%, 60%]	3% [−41%, 72%]	0.3% [−29%, 34%]	1% [−25%, 29%]
Bias |± 0.2|	70%	74%	33%	51%	12%	22%
Absolute deviation	0.42 [0.02, 1.53]	0.49 [0.02, 1.80]	0.17 [0.01, 0.52]	0.26 [0.01, 0.87]	0.09 [0.00, 0.29]	0.13 [0.01, 0.37]
Relative absolute deviation	52% [2%, 191%]	41% [1%, 150%]	21% [1%, 65%]	22% [1%, 72%]	13% [0%, 37%]	11% [0%, 31%]

Note. Mean values and empirical 95% quantile ranges in brackets for the estimation based on 10,000 simulations with 60 choices each. Choice sets are described in detail in the text. The corresponding θs and their estimates can be found in Table 2.

Table 2. Results of basic parameter recovery for choice consistency θ

	Random choice set		No-dominance choice set		Switching choice set	
Variable	θ = 0.18	θ = 0.03	θ = 0.41	θ = 0.05	θ = 0.40	θ = 0.06
Mean estimate θ	0.93 [0.00, 3.60]	0.14 [0.00, 0.54]	0.83 [0.04, 4.29]	0.12 [0.01, 0.63]	0.54 [0.10, 1.64]	0.09 [0.01, 0.29]
Relative bias	414% [−100%, 1,900%]	419% [−100%, 1,900%]	101% [−91%, 945%]	131% [−98%, 1,168%]	35% [−75%, 309%]	44% [−81%, 378%]
Absolute deviation	0.87 [0.01, 3.42]	0.13 [0.00, 0.51]	0.6 [0.01, 3.88]	0.09 [0.00, 0.58]	0.28 [0.01, 1.24]	0.05 [0.00, 0.23]
Relative absolute deviation	484% [8%, 1,900%]	410% [8%, 1,900%]	146% [3%, 945%]	189% [5%, 1,168%]	70% [2%, 309%]	82% [2%, 378%]

Note. Mean values and empirical 95% quantile ranges (in brackets) for the estimation based on 10,000 simulations. The corresponding αs were 0.8 for the first and 1.2 for the second column in each choice set.

The average absolute deviation of α was 46% for the random, 21% for the no-dominance, and 12% for the switching choice set. We additionally looked at absolute deviations above 0.2 that we assumed were psychologically important measurement errors, as they imply that a risk-averse person ($\alpha = 0.8$) is potentially classified as being risk seeking ($\alpha > 1$) and vice versa for a risk-seeking person. This was the case in 72% of all recoveries for the random and in about 42% and 17% of all simulations for the no-dominance and switching choice set, respectively. This means that roughly 35% of all recoveries resulted in a misclassification of a person as risk neutral or risk seeking when they were actually risk averse or the other way round using the random choice set. Similarly, for θ the absolute deviation was largest for the random (447%), lower for the no-dominance (168%), and least for the switching (76%) choice set.

Measuring MPD from equation (4), we calculated a KL divergence of 40.45 for the random, 86.00 for the no-dominance, and 191.01 for the switching choice set, corroborating that the switching set was most informative for estimating the parameter values across the whole range of plausible values. For comparison, we also estimated the MPD for the Holt and Laury (2002) choice set, transformed to the outcome scale of the other choice sets and repeated six times to achieve 60 trials. The MPD measure was 172.24 for the modified Holt and Laury (2002) set and thus lower than for the switching set, meaning that the switching set was better able to discriminate between parameter values in our model. We calculated the UPD separately for both parameters to be 188.28 for α and 2.72 for θ in the case of the switching set. Similar results were estimated for the other choice sets (see Table A1) and show that α is estimated more accurately than θ.

Finally, we examined the correlation between the estimates of the two parameters for all baseline simulations with the same choice set and the same data-generating parameters. For the random set, there was a substantial linear Pearson correlation of −.68 (Spearman rank correlation −.90). Correlations were similar for the no-dominance and slightly higher for the switching set. The shape of this correlation is illustrated for the switching set in Figure 1 and resembles an exponential relation. For the switching set, the PRD measures (based on KL divergence) were .01 for α and .35 for θ. This demonstrates that the accuracy of the θ estimate was more affected by the correlation than α. To sum up, as expected, the switching choice set showed the best estimation accuracy when we looked at bias, expected measurement error, and the KL divergence for both parameters in our model.

3.2. The effects of standardizations

The effects of the standardizations on estimation accuracy are summarized in four plots in Figure 2 using the switching choice set, as this set led to the most accurate estimation results when using no standardization method.

3.2.1. Utility standardization

The correlation between estimates of α and θ was strongly reduced to approximately −.07 (rank correlation −.09). This reduction had a different effect on the measurement accuracy of estimates of α and θ. The estimation bias for α was slightly smaller using the random and the no-dominance set but slightly higher using the switching set. In contrast, the bias for θ estimates strongly decreased for all choice sets to around 16% on average. This effect was especially strong for the random set that produced a bias of over 400% before standardization. Although the switching set remained the most accurate estimation

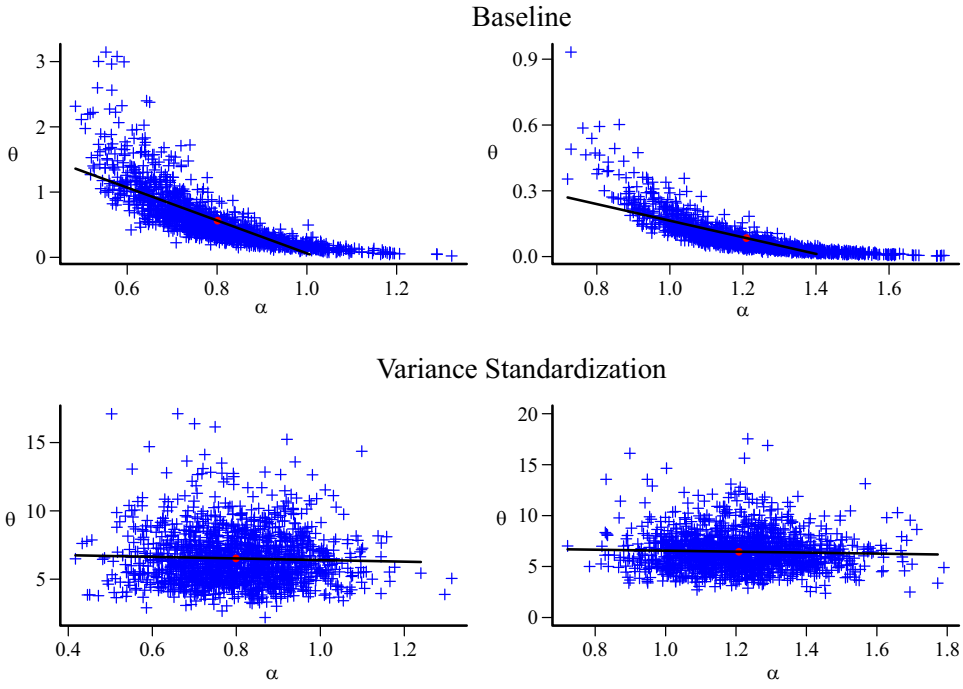

Figure 1. Recovered parameters α and θ with linear correlation line after 1,500 simulations. *Note*. Data-generating α = 0.8 (left) and α = 1.2 (right), indicated by a red dot. First row shows the parameter correlation in the baseline case, and the second row with variance standardization.

set, the difference in accuracy of θ estimates between the three sets almost vanished with the standardization.

A similar pattern emerged for the absolute deviation. The absolute deviation of α estimates improved slightly for the random set but got worse for the no-dominance and the switching set compared to the baseline case. In contrast, the absolute deviation for θ estimates strongly decreased to less than 30% for all choice sets.

From the KL divergence we computed that the influence of the estimate of one parameter on the estimation accuracy of the other parameter (PRD) decreased on average in line with the decrease in the correlation. However, this influence decreased only for θ (.12), and not for α (.03).

3.2.2. Outcome standardization

The effects of this standardization were similar to the effects reported for the utility standardization. There was only a weak correlation between the estimates, and the improvement of α estimates in terms of bias and absolute deviations was very small. Again, for the switching set there was a deterioration of the α accuracy for the absolute deviation, but this was not as strong as with the utility standardization. For θ, estimation was drastically improved for bias and absolute deviations for all choice sets.

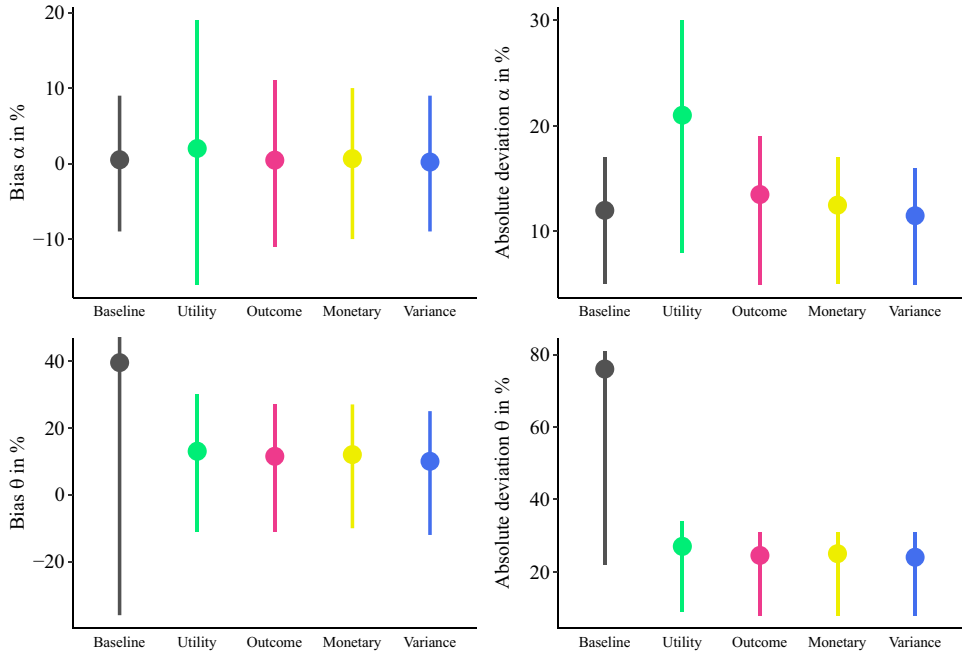

Figure 2. Percentage of bias (left) and absolute deviation (right) of the estimates from the true data-generating parameter values for risk preference α (top) and choice consistency θ (bottom). *Note*. Estimations were conducted with 60 choices from the switching set. The x-axis shows different standardization approaches in comparison to baseline. Lines show interquartile ranges from 10,000 samples. The exact data for the standardization results can be found in Tables A2–A9.

3.2.3. Monetary equivalence standardization

Similarly to the previous results, there was a weak correlation between the parameters, but the effect on α estimation accuracy was subtle. In the random choice set the α estimation bias decreased, but the absolute deviation of α estimates increased. For the other two choice sets, the α estimation accuracy stayed similar to the baseline case. As before, θ estimations improved drastically in terms of bias and absolute deviation for all choice sets.

3.2.4. Variance standardization

Again, there was a very weak correlation between the parameters. Interestingly, bias and absolute deviations of the α estimates were slightly smaller than for the other standardizations. Similarly, θ estimates drastically improved in bias and absolute deviations for all choice sets, and deviation was lowest across all standardizations for the switching choice set.

3.2.5. Summary of standardization results

A qualitatively similar pattern emerged for all four proposed standardizations. Standardizations strongly decreased the correlation of the parameter estimates, as can be seen in

the lower plots of Figure 1 for the example of variance standardization and the switching set. Estimation accuracy for θ was drastically improved by all standardizations. This was corroborated by a strong decrease in the effect of α estimation imprecision on θ estimates according to the KL divergence measures. Yet, standardizations failed to improve the estimation accuracy for α and in some cases even decreased it. With standardizations, estimation accuracy was highest for the switching set compared to the other two choice sets for both parameters. Yet, whereas estimation accuracy for α was still substantially higher for the switching compared to the other choice sets, this difference was not strong for θ, as can be seen in Figure A1 for the random and Figure A2 for the no-dominance set.

In the switching set, the variance standardization performed best with respect to estimation accuracy for θ (average 10% bias and 24% absolute deviation). Yet, there were no meaningful differences in the estimation accuracy of the utility, outcome and monetary equivalence standardizations for θ. Moreover, the variance standardization performed best with respect to the estimation accuracy for α compared to all other standardizations and at the same level as the baseline case (average absolute deviation 12%). However, for the other standardizations, there was a trade-off between estimation accuracy of α and θ in the methods examined, as estimation accuracy for α decreased compared to the baseline case. The utility standardization decreased estimation accuracy for α most among all standardizations for the switching set (26%).

The ranking of best-performing standardizations with respect to measurement accuracy changes when looking at the random set (see Figure A1). Here, the utility and the outcome standardization produced the same level of estimation accuracy for α as the baseline case without standardization, whereas the other two standardizations led to a lower estimation accuracy. The outcome standardization had a lower estimation accuracy for θ compared to all other standardizations. Consequently, in the case of the random set, the utility standardizations performed best for estimation accuracy for both α and θ. For the no-dominance choice set (see Figure A2), utility standardization showed the best estimation accuracy for α and the monetary equivalence standardization showed the best estimation accuracy for θ.

In summary, standardizations should be used carefully when the choice set is already informative and can be used more readily with random sets. Further, the effect of the standardization on measurement accuracy interacts with the choice set in a non-trivial way.

3.3. The effect of choice repetitions

It is important to know whether for a given number of observations it is better to present distinct choices or to present a subset of choices repeatedly. We examined this by directly comparing the estimation accuracy of the three full choice sets with 60 distinct choices with the estimation accuracy when randomly drawing 15 of the 60 choice situations and repeating these four times. In the latter case the 15 choices were different in every simulation round and the total number of choices was the same for both approaches (60). As a result, for all measures and gamble sets, the approach with distinct choice situations led to better or equivalent estimation accuracy compared to the approach with four times the same choice situation. The discrepancy in estimation accuracy between the approaches was higher for θ than for α estimates and for random and no-dominance sets compared to switching sets (see Tables A10 and A11 for full results). This shows that in particular when estimates were imprecise (as was the case for θ) and the information value of choice sets was low (as was the case for random sets), using distinct choice

situations had an advantage over repeating the same choices. Importantly and perhaps counterintuitively, choice consistency within our model was better measured with distinct than with repeated identical choices. Moreover, in practical applications with human participants, using distinct choices will help to retain interest in the task and to potentially avoid violations of independence between choice repetitions.

To check by how much estimation accuracy increased with higher sample size, we increased the number of times we simulated choices for our choice sets from 1 (with 60 simulated choices) up to 15 choice repetitions (with 900 simulated choices). Deviation and bias trended toward zero with 15 choice set repetitions for both α and θ estimates (see Figures A3–A5).[2] The linear part of the correlation between the estimates increased with the number of choice repetitions for the baseline (up to .91) and remained relatively low for the standardization procedures (between .01 and .17).

The dichotomy of the effect of standardization on α and θ estimation accuracy persisted for all numbers of repetitions. For estimates of risk preference α, bias and deviation of the baseline model were always of the magnitude of the best standardization results (and often even slightly better), whereas for choice consistency θ, bias and deviation were always lower for the standardizations compared to the baseline model. Finally, the rank order of estimation accuracy between the different standardizations remained the same by and large across all numbers of repetitions for both parameter estimates.

4. Discussion

We examined the estimation accuracy of an RUM with a power utility and a logit function. This model requires the estimation of two free parameters, risk preference α and choice consistency θ. Via simulations and parameter recovery, we demonstrated that RUMs suffer from empirical underidentification, meaning that choice sets often do not precisely differentiate between risk-averse and risk-seeking agents.

4.1. Stimulus design

The stimulus design strongly affected the estimation accuracy of the model parameters. Using random choice sets as stimuli led to an overestimation of α of approximately 10%, whereas when dominant gamble pairs were excluded and the expected value difference between gambles was kept low, α was estimated nearly without bias. The expected deviation from the true value in a single estimation for α again demonstrated the importance of the choice set. Here in particular the newly developed set that incorporated a switching point in expected utility order from a risk-averse to a very risk-seeking α-value for every lottery pair showed the best results. The expected absolute deviation in this choice set was 12% and this was a little more than half the deviation of a choice set that had only non-dominant gambles and a quarter of the deviation in a randomly created choice set.

The other parameter, choice consistency θ, was severely overestimated by on average 200% across all choice sets, and the expected absolute deviation was even higher. Yet again, there were huge differences between the choice sets, and the switching choice set

[2] In line with the previous paragraph, estimation accuracy was slightly higher for distinct than for repeated choices. However, this difference decreased and became trivial with higher numbers of observations.

fared better with a magnitude of one-tenth of the overestimation compared to the random choice set and one-third compared to the no-dominance choice set. These results were corroborated by numerically estimating the KL divergence, a measure of the discriminability of different parameter values of a model across all plausible parameter values (see Broomell & Bhatia, 2014). In addition, the KL divergence demonstrated that the switching choice set was also more informative than an adjusted version of the Holt and Laury (2002) set. Hence, a choice set created according to some measurement-theory guidelines can improve estimation accuracy compared to randomly created choice sets and the purpose-built choice sets often used in the literature (cf. Broomell & Bhatia, 2014). As a limitation, the KL divergence results depend on the prior distribution of parameter values. Here, we assumed that all parameter combinations within a plausible range were equally likely. It could be interesting for specific applications in future research to construct priors based on the actually observed distribution of parameter values in a target population.

Together, this shows the importance of a well-constructed choice set to estimate RUMs precisely. The good performance of the switching choice set can be explained by the fact that every choice is informative for the measurement of the utility function in the range of α-values between 0.2 and 2.8.[3] In contrast, other choice sets include choices where, although one gamble might not stochastically dominate another, every reasonable power parameter results in the same utility order. This means that either choosing the gamble with the higher utility does not distinguish between α-values in this range or choosing the gamble with the lower utility would lead to assuming either an extreme α-value or an increase in noise captured by the choice consistency parameter.

As a limitation to this, we cannot rule out that the estimation accuracy could be further improved by tailoring the choice set to the estimation task. Theoretically, an optimal experimental design to estimate parameters of a given model can be created (Myung & Pitt, 2009). However, this requires knowing exactly which design variables are connected to measurement accuracy, and in the case of multiple design variables it requires an extension of the Myung and Pitt framework. Another approach is to use an adaptive design that chooses the most informative lottery couple after observing choices from a given participant (Cavagnaro, Gonzalez, Myung, & Pitt, 2013; Toubia, Johnson, Evgeniou, & Delquié, 2013). We see our approach as complementary to an adaptive design in cases where the researcher does not want to make the theoretical assumptions necessary to determine the next most informative lottery or if such an approach is not feasible for practical reasons (see also Chang & Ying, 1996).

4.2. Estimation methods

In all baseline recoveries, there was a substantial linear correlation of the two parameter estimates of about −.70 and an even higher rank correlation of about −.90. This trade-off contributes to the empirical underidentification (Spektor & Kellen, 2018) and prevents meaningful comparisons between choice consistency values for different levels of risk preference (Stewart et al., 2018). We tested standardization techniques to check whether they could mitigate this correlation: these were the standardization of expected utility between 0 and 1, the standardization of outcomes between 0 and 1, the retransformation of expected utility differences back to the monetary scale, and the dividing of the

[3] This criterion should also help in the estimation of random parameter models (see Loomes & Sugden, 1995), as they also depend on the condition of a switch in utility rank order between choice options.

expected utility difference by the pooled utility variance. All four standardization techniques substantially reduced the correlation of the estimates and decreased the estimation bias for choice consistency θ by up to 70% for the switching choice set and even more strongly for the other two choice sets. Furthermore, all four approaches improved estimation accuracy for θ with the same order of magnitude. This demonstrates the importance of using a standardization to estimate choice consistency more precisely and to allow for meaningful comparisons of choice consistency parameters across different levels of risk preference. The standardization leading to the most accurate estimation results was the variance standardization for the switching choice set. For the other two choice sets, depending on the emphasis on estimation accuracy for risk preference or choice consistency, the utility or the monetary equivalence standardization performed best.

From a decision-theory point of view, utility and outcome standardizations lead to context dependencies. This means that parameter estimates depend on the specific choice set that is used for eliciting choices. If a researcher is only interested in measuring individual levels of risk preference and choice consistency within similar contexts, this is a viable approach. However, context dependency can be a measurement-theory problem when researchers want to aggregate parameter estimates over participants who have seen different choice sets or when researchers want to predict behaviour in new choice situations (see Stewart et al., 2019).

As a caveat, none of the standardization approaches improved estimation accuracy for the risk preference parameter α compared to the baseline case, and occasionally even deteriorated it. This holds true although in all cases we used the same model for data creation and data fitting, and it was particularly bad for the utility standardization in the switching choice set. Thus, although in an experimental context the researcher does not know how the observed data were created, we can say that from a measurement-theory point of view, the utility standardization can lead to low estimation accuracy for the risk preference parameter. Intuitively, this might be the case because bringing down the correlation helps stabilize θ estimates (which occasionally have high outliers) much more than α estimates (with fewer outliers). Thus, the choice of a standardization depends on the research question, and if a researcher is interested only in the estimation accuracy of risk preference and treats choice consistency as a nuisance parameter, one can defend not using a standardization in the estimation process as long as one has an informative choice set.

4.3. Choice repetitions

Finally, we showed that although parameter estimates traded off and choice consistency was estimated with a bias, the estimation accuracy of both parameters increased continuously with higher numbers of trials. As a limitation, the exact numbers depend on the implemented choice sets. However, even if a researcher is not using our recommended switching choice set, we want to draw attention to the range of measurement error to be expected for different estimation and stimulus design choices when estimating RUMs.[4]

[4] Full tables of results for all numbers of trials, choice sets and standardization methods can be found in the Supporting Information.

To give an idea of the impact of number of trials on estimation accuracy, we provide some examples. If one requires to estimate risk preference α with an expected deviation from the true parameter of about 5%, one needs 180 choices from the switching set. At the same time, such an estimation plan results in an overestimation of choice consistency θ of 14% on average and an expected measurement error of approximately 40%. In contrast, with the random and the no-dominance sets a similar accuracy of α estimation cannot be achieved with a realistic number of choices.

If one is interested in a precise estimation of the choice consistency parameter θ, a standardization is indispensable. Taking the same example as above, using 180 switching gambles for the estimation, a standardization decreases the bias of θ from 14% to 4% and the expected measurement error from 40% to 14%. An expected measurement error of 10% for θ can be achieved only with 300 choices and of 5% only with approximately 900 choices using any of the four standardizations. Unlike for risk preference parameter estimations, the precision of these θ estimates is also quite similar for the random and the no-dominance choice sets using any of the standardizations.

4.4. Beyond RUMs

Our model recovery analyses are based on a power utility function. The power utility function belongs to the class of constant relative risk aversion (CRRA, e.g., Holt & Laury, 2002; see Wakker, 2008) and is also implemented in cumulative prospect theory (CPT; Tversky & Kahneman, 1992). For gambles in the gain domain, CPT adds a weighting function for probabilities. We expect the problems we found with the simpler power utility model to become worse in the more complex CPT model that requires the estimation of (at least) one additional parameter (see Broomell & Bhatia, 2014; Krefeld-Schwalb et al., 2021; Scheibehenne & Pachur, 2015; Spektor & Kellen, 2018). The results presented here thus serve as an upper boundary on expected accuracy when estimating CPT.

There are also different classes of models, such as the mean–variance model, which estimate a linear combination of mean and variance of a lottery with a free parameter for the influence of variance on choices (for a comparison of the two utility models, see Olschewski et al., 2018, Experiment 1; Spiliopoulos & Hertwig, 2019). This framework has a lower parameter correlation between risk preference and consistency. As a disadvantage, estimating the parameters of a mean–variance model does not compare easily to estimating the parameters of a power utility approach, which is predominant in the literature.

We used a logit or Fechner choice function (Carbone, 1997). Other specifications have been discussed and our results hold as well for the probit choice function. The probit choice function differs from the logit in the assumption of a normal instead of an extreme-value distribution of the error term. Another possibility is to use a trembling hand error, which estimates the probability of choosing the on average inferior lottery. However, such a formulation is theoretically not very plausible, since it ignores the fact that some choices are easier than others, and it has also been rejected empirically (Blavatskyy & Pogrebna, 2010; Stott, 2006). Finally, random parameter models have been proposed for a long time (Becker, DeGroot, & Marschak, 1963; Loomes & Sugden, 1995) but have two problems: they cannot cope with choices of dominated lotteries and they can be difficult to estimate reliably. To circumvent these problems, two recent papers have proposed random parameter models with additional error sources (Apesteguia & Ballester, 2018; Bhatia & Loomes, 2017; see also Loomes, Moffatt, & Sugden, 2002). Yet, in these

specifications there are in total three parameters that have to be estimated: the mean power utility parameter, the variability of the power utility parameter, and a logit or trembling parameter. It is so far an open question whether additional parameters might reduce or increase the estimation accuracy of risk preference.

In summary, the present work illustrates that the estimation of a person's risk preference or choice consistency can be a demanding enterprise and requires careful experimental designs and estimation methods. The right choice set is most important for the estimation of risk preferences, whereas a standardization is most important for the estimation of choice consistency. These results should be taken into account when testing choice theories or when relying on characterizing people's preferences for practical interventions or treatments.

Acknowledgments

All simulation and recovery code as well as the code for the creation of choice sets and an example choice set ready to use are available at https://osf.io/p8dqs. We thank Stephen Broomell for sharing code for the Kullback–Leibler divergence calculations. This research was supported by Swiss National Science Fund grant P2BSP1_188188 to the first author.

Conflicts of interest

All authors declare no conflict of interest.

Author contribution

Sebastian Olschewski: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Methodology (equal); Project administration (equal); Software (equal); Writing – original draft (equal). **Pavel Sirotkin:** Data curation (equal); Formal analysis (equal); Software (equal); Validation (equal); Visualization (equal). **Jörg Rieskamp:** Conceptualization (equal); Methodology (equal); Project administration (equal); Writing – review & editing (equal).

Open research badges

This article has been awarded Open Data, Open Materials Badges. All materials and data are publicly accessible via the Open Science Framework at https://osf.io/p8dqs/

References

Alempaki, D., Canic, E., Mullett, T. L., Skylark, W. J., Starmer, C., Stewart, N., & Tufano, F. (2019). Reexamining how utility and weighting functions get their shapes: A quasi-adversarial collaboration providing a new interpretation. *Management Science*, *65*, 4841–4862. https://doi.org/10.1287/mnsc.2018.3170

Anderson, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2007). Valuation using multiple price list formats. *Applied Economics*, *39*, 675–682. http://doi.org/10.1080/00036840500462046

Andersson, O., Holm, H. J., Tyran, J. R., & Wengström, E. (2016). Risk aversion relates to cognitive ability: Preferences or noise? *Journal of the European Economic Association*, *14*, 1129–1154. http://doi.org/10.1111/jeea.12179

Apesteguia, J., & Ballester, M. A. (2018). Monotone stochastic choice models: The case of risk and time preferences. *Journal of Political Economy*, *126*(1), 74–106. http://doi.org/10.1086/695504

Bamber, D., & van Santen, J. P. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, *44*(1), 20–40. http://doi.org/10.1006/jmps.1999.1275

Becker, G. M., DeGroot, M. H., & Marschak, J. (1963). Stochastic models of choice behavior. *Behavioral Science*, *8*(1), 41–55. http://doi.org/10.1002/bs.3830080106

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, *22*(1), 23–36. Original work published 1738.http://doi.org/10.2307/1909829

Bhatia, S., & Loomes, G. (2017). Noisy preferences in risky choice: A cautionary note. *Psychological Review*, *124*, 678–687. http://doi.org/10.1037/rev0000073

Bhui, R., & Gershman, S. J. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review*, *125*, 985. http://doi.org/10.1037/rev0000123

Blavatskyy, P. R., & Pogrebna, G. (2010). Models of stochastic choice and decision theories: Why both are important for analyzing decisions. *Journal of Applied Econometrics*, *25*, 963–986. http://doi.org/10.1002/jae.1116

Broomell, S. B., & Bhatia, S. (2014). Parameter recovery for decision modeling using choice data. *Decision*, *1*(4), 252–274. http://doi.org/10.1037/dec0000020

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459. http://doi.org/10.1037/0033-295X.100.3.432

Carbone, E. (1997). Investigation of stochastic preference theory using experimental data. *Economics Letters*, *57*, 305–311. http://doi.org/10.1016/S0165-1765(97)00244-9

Cavagnaro, D. R., Gonzalez, R., Myung, J. I., & Pitt, M. A. (2013). Optimal decision stimuli for risky choice experiments: An adaptive approach. *Management Science*, *59*, 358–375. http://doi.org/10.1287/mnsc.1120.1558

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213–229. http://doi.org/10.1177/014662169602000303

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review*, *100*, 1238–1260. http://doi.org/10.1257/aer.100.3.1238

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, *3*(10), e1701381. http://doi.org/10.1126/sciadv.1701381

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition*, *123*(1), 21–32. http://doi.org/10.1016/j.cognition.2011.12.002

Hey, J. D. (2001). Does repetition improve consistency? *Experimental Economics*, *4*(1), 5–54. http://doi.org/10.1142/9789813235816_0002

Hey, J. D., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, *62*, 1291–1326. http://doi.org/10.2307/2951750

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*, 1644–1655. http://doi.org/10.1257/000282802762024700

Krefeld-Schwalb, A., Pachur, T., & Scheibehenne, B. (2021). Structural parameter interdependencies in computational models of cognition. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000285

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1–32. http://doi.org/10.1037/rev0000074

Loomes, G., Moffatt, P. G., & Sugden, R. (2002). A microeconometric test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty*, *24*(2), 103–130. http://dx.doi.org/10.1023/A:1014094209265

Loomes, G., & Pogrebna, G. (2014). Measuring individual risk attitudes when preferences are imprecise. *The Economic Journal*, *124*, 569–593. http://doi.org/10.1111/ecoj.12143

Loomes, G., & Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, *39*, 641–648. http://doi.org/10.1016/0014-2921(94)00071-7

Luce, R. D. (1957). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Luce, R. D., & Suppes, P. (1965). Preference, utility and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. *III*, pp. 249–410). New York: Wiley.

Mosteller, F., & Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, *59*, 371–404. http://doi.org/10.1086/257106

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*, 499–518. http://doi.org/10.1037/a0016104

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*(1), 47–84. http://doi.org/10.1016/j.cogpsych.2003.11.001

Olschewski, S., & Rieskamp, J. (2021). Distinguishing three effects of time pressure on risk taking: Choice consistency, risk preference, and strategy selection. *Journal of Behavioral Decision Making*, *34*, 541–554. https://doi.org/10.1002/bdm.2228

Olschewski, S., Rieskamp, J., & Scheibehenne, B. (2018). Taxing cognitive capacities reduces choice consistency rather than preference: A model-based test. *Journal of Experimental Psychology: General*, *147*, 462–484. http://doi.org/10.1037/xge0000403

Polania, R., Woodford, M., & Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience*, *22*(1), 134–142. http://doi.org/10.1038/s41593-018-0292-0

R Core Team (2016). R: A language and environment for statistical computing [Computer software manual]. Retrieved from https://www.R-project.org/

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1446–1465. http://doi.org/10.1037/a0013646

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, *44*, 631–661. http://doi.org/10.1257/jel.44.3.631

RStudio Team. (2015). *R* [Computer software manual]. Retrieved from http://www.rstudio.com/

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review*, *22*, 391–407. http://doi.org/10.3758/s13423-014-0684-4

Schmidt, U., & Hey, J. D. (2004). Are preference reversals errors? An experimental investigation. *Journal of Risk and Uncertainty*, *29*(3), 207–218. http://doi.org/10.1023/B:RISK.0000046143.10752.0a

Schmittmann, V. D., Dolan, C. V., Raijmakers, M. E., & Batchelder, W. H. (2010). Parameter identification in multinomial processing tree models. *Behavior Research Methods*, *42*, 836–846. http://doi.org/10.3758/BRM.42.3.836

Spektor, M. S., & Kellen, D. (2018). The relative merit of empirical priors in non-identifiable and sloppy models: Applications to models of learning and decision-making. *Psychonomic Bulletin & Review*, *25*, 2047–2068. http://doi.org/10.3758/s13423-018-1446-5

Spiliopoulos, L., & Hertwig, R. (2019). Nonlinear decision weights or moment-based preferences? A model competition involving described and experienced skewness. *Cognition*, *183*, 99–123. http://doi.org/10.1016/j.cognition.2018.10.023

Starmer, C., & Sugden, R. (1989). Violations of the independence axiom in common ratio problems: An experimental test of some competing hypotheses. *Annals of Operations Research*, *19*(1), 79–102.

Stewart, N., Canic, E., & Mullett, T. L. (2019). *On the futility of estimating utility functions: Why the parameters we measure are wrong, and why they do not generalize*. PsyArXiv preprint. https://doi.org/10.31234/osf.io/qt69m

Stewart, N., Scheibehenne, B., & Pachur, T. (2018). *Psychological parameters have units: A bug fix for stochastic prospect theory and other decision models*. OSF preprint. https://doi.org/10.31234/osf.io/qvgcd

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, *32*(2), 101–130. http://doi.org/10.1007/s11166-006-8289-6

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge: MIT Press.

Toubia, O., Johnson, E., Evgeniou, T., & Delquié, P. (2013). Dynamic experiments for estimating preferences: An adaptive method of eliciting time and risk parameters. *Management Science*, *59*, 613–640. http://doi.org/10.1287/mnsc.1120.1570

Train, K. E. (2009). *Discrete choice methods with simulation*. New York: Cambridge University Press. http://doi.org/10.1017/CBO9780511753930

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. http://doi.org/10.1007/BF00122574

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Wakker, P. P. (2008). Explaining the characteristics of the power (CRRA) utility family. *Health Economics*, *17*, 1329–1344. http://doi.org/10.1002/hec.1331

Weber, E. U., Blais, A. R., & Betz, N. E. (2002). A domain specific risk attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263–290. http://doi.org/10.1002/bdm.414

Wilcox, N. T. (2011). Stochastically more risk averse: A contextual theory of stochastic discrete choice under risk. *Journal of Econometrics*, *162*(1), 89–104. http://doi.org/10.1016/j.jeconom.2009.10.012

Wilcox, N. T. (2015). Error and generalization in discrete choice under risk [Working paper]. Chapman University, Economic Science Institute.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, *12*, 579–601. http://doi.org/10.1146/annurev-economics-102819-040518

Received 22 August 2020

Appendix

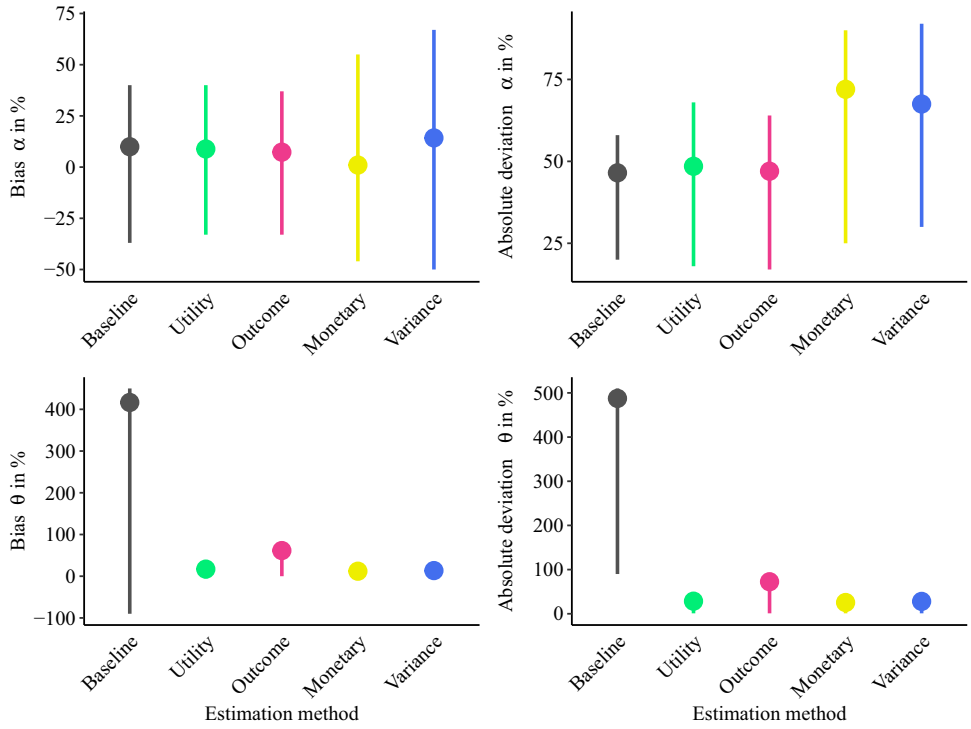

Figure A1. Random choice set: percentage of bias (left) and absolute deviation (right) of the estimates from the true data-generating parameter values for risk preference α (top) and choice consistency θ (bottom). *Note*. Estimation was conducted with 60 randomly created choice situations. The x-axis shows different standardization approaches in comparison to baseline. Lines show interquartile ranges from 10,000 rounds of estimation.

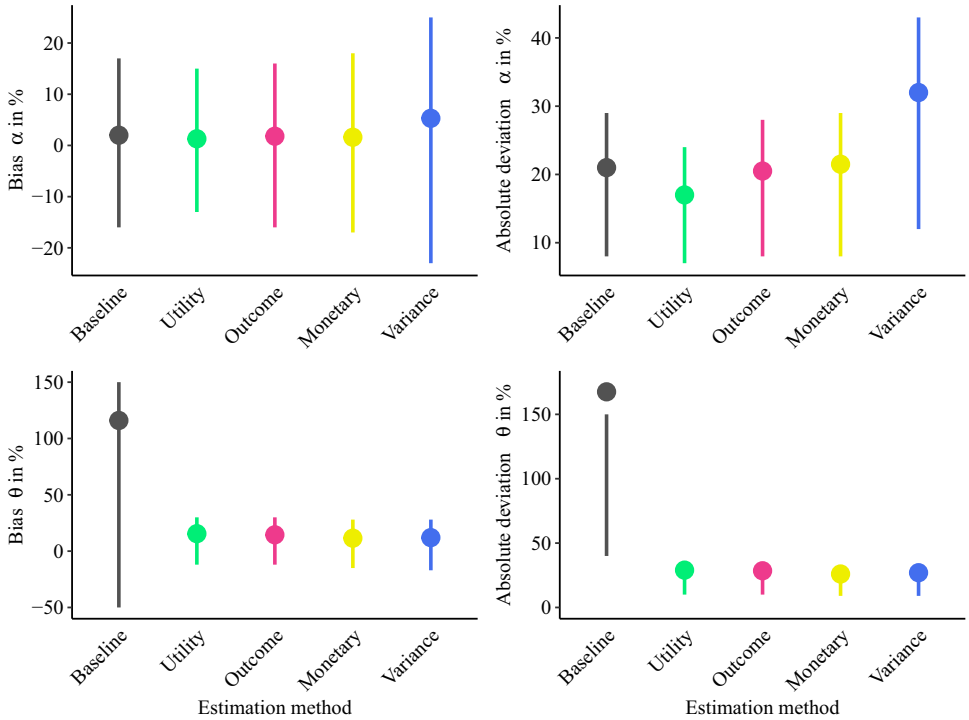

Figure A2. No-dominance choice set: percentage of bias (left) and absolute deviation (right) of the estimates from the true data-generating parameter values for risk preference α (top) and choice consistency θ (bottom). *Note*. Estimation was conducted with 60 choices from the no-dominance choice set. The x-axis shows different standardization approaches in comparison to baseline. Lines show interquartile ranges from 10,000 rounds of estimation.

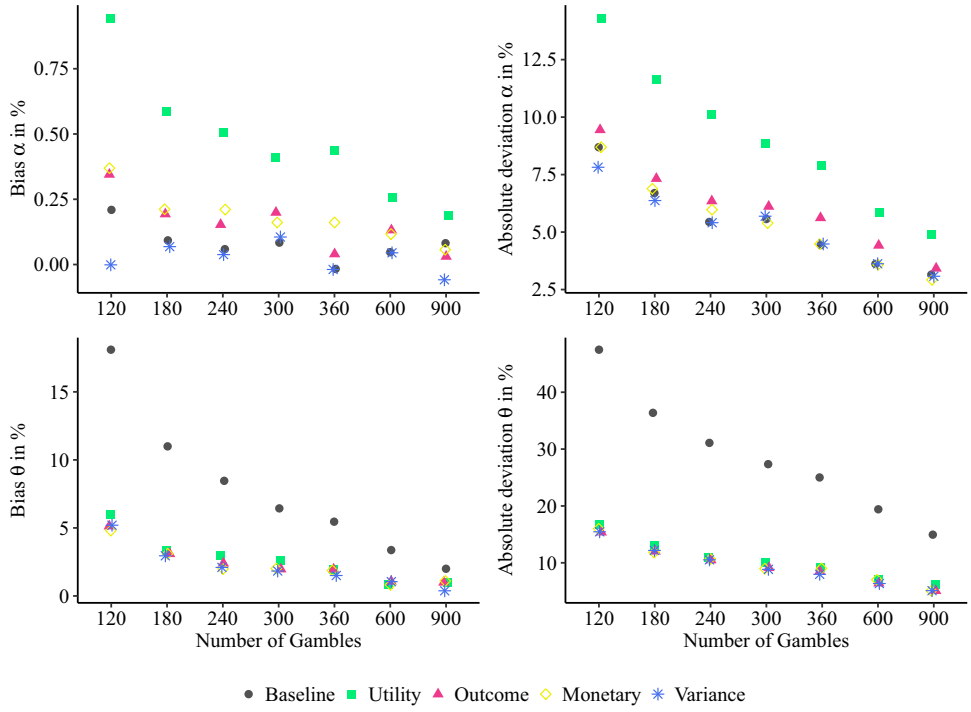

Figure A3. Switching choice set: estimation bias on the left and absolute deviation on the right for risk preference α (top) and choice consistency θ (bottom) for different numbers of repetitions. *Note*. The 60 choices from the switching set were used and repeated for higher numbers of gambles. Colours show different standardization approaches in comparison to baseline. The exact data can be found in Tables S1–S30 of the Supporting Information.

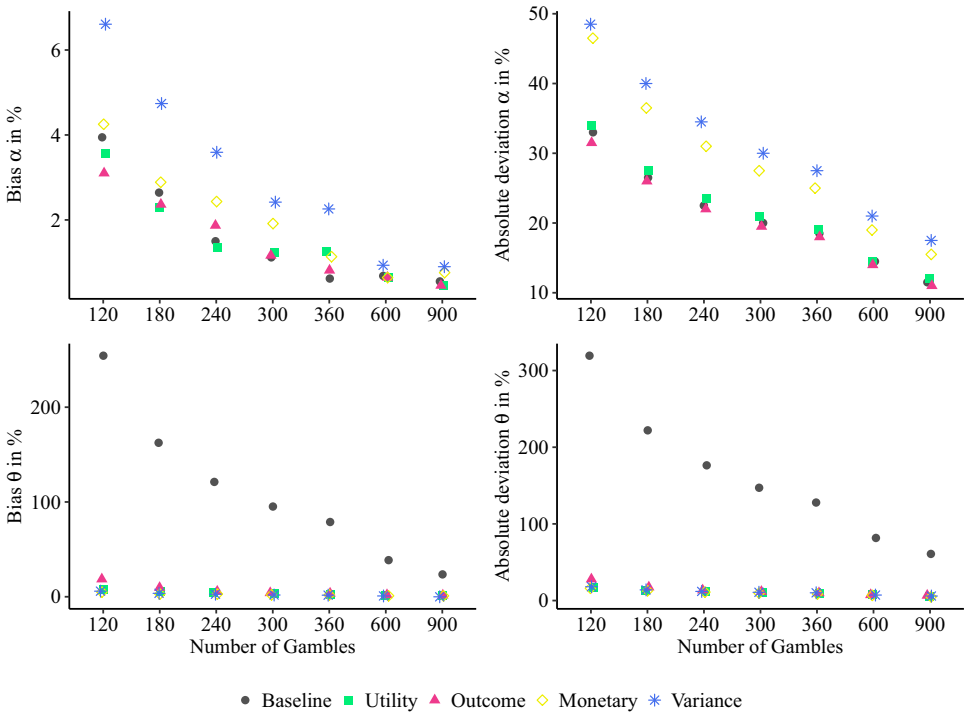

Figure A4. Random choice set: estimation bias on the left and absolute deviation on the right for risk preference α (top) and choice consistency θ (bottom) for different numbers of repetitions. *Note*. The 60 choices were randomly created and were repeated for higher numbers of gambles. Colours show different standardization approaches in comparison to baseline.

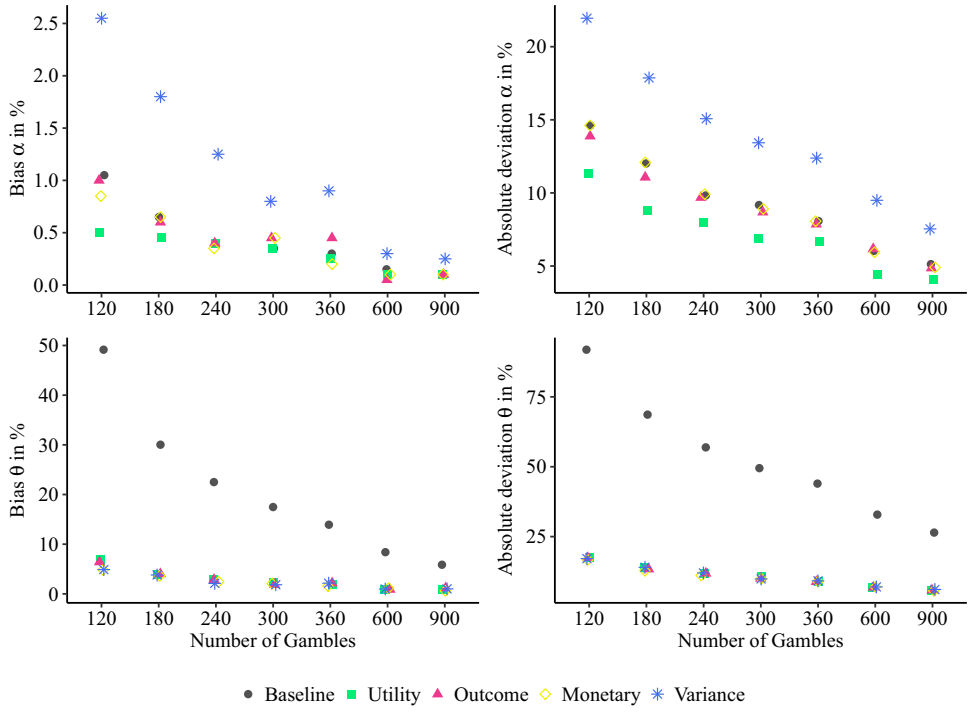

Figure A5. No-dominance choice set: estimation bias on the left and absolute deviation on the right for risk preference α (top) and choice consistency θ (bottom) for different numbers of repetitions. *Note*. The 60 choices from the no-dominance set were used and repeated for higher numbers of gambles. Colours show different standardization approaches in comparison to baseline.

Table A1. Kullback–Leibler divergence for all choice sets and baseline as well as all standardization methods

		Choice set		
Standardization method	Measure	Random	No-dominance	Switching
Baseline	MPD	40.45	86.00	191.01
	UPD α	37.86	83.56	188.28
	UPD θ	2.59	2.44	2.72
	PRD α	0.03	0.01	0.01
	PRD θ	0.30	0.32	0.35
Utility	MPD	23.70	83.87	54.66
	UPD α	11.02	73.83	43.62
	UPD θ	12.68	10.04	11.03
	PRD α	0.03	0.01	0.03
	PRD θ	0.03	0.12	0.12
Monetary	MPD	21.58	56.84	166.48
	UPD α	8.88	43.57	156.82
	UPD θ	12.70	13.27	9.66
	PRD α	0.03	0.03	0.02
	PRD θ	0.02	0.08	0.29
Variance	MPD	19.25	49.63	165.02
	UPD α	7.27	37.24	155.58
	UPD θ	12.01	12.38	9.43
	PRD α	0.04	0.03	0.02
	PRD θ	0.02	0.09	0.02
Outcome	MPD	27.51	65.01	142.15
	UPD α	17.07	54.25	134.02
	UPD θ	10.44	10.75	8.14
	PRD α	0.03	0.02	0.02
	PRD θ	0.04	0.09	0.23

Note. MPD = multivariate parameter discrimination; PRD = percent reduced discrimination; UPD = univariate parameter discrimination.

Table A2. Results of utility standardization parameter recovery for risk preference α

	Random choice set		No-dominance choice set		Switching choice set	
Variable	α = 0.8	α = 1.2	α = 0.8	α = 1.2	α = 0.8	α = 1.2
Mean estimate α	0.87 [0.00, 2.33]	1.31 [0.17, 3.00]	0.81 [0.44, 1.20]	1.22 [0.79, 1.78]	0.84 [0.35, 1.48]	1.2 [0.45, 1.99]
Relative bias	8% [−100%, 191%]	9% [−86%, 150%]	1.2% [−45%, 50%]	1% [−34%, 49%]	5% [−55%, 85%]	−0.2% [−62%, 66%]
Bias |± 0.2|	68%	74%	30%	33%	49%	49%
Absolute deviation	0.42 [0.02, 1.53]	0.53 [0.02, 1.80]	0.15 [0.01, 0.44]	0.19 [0.01, 0.60]	0.23 [0.01, 0.68]	0.28 [0.01, 0.86]
Relative absolute deviation	53% [2%, 191%]	44% [2%, 150%]	19% [1%, 55%]	15% [1%, 50%]	29% [1%, 86%]	23% [1%, 72%]

Note. Reported are mean values and empirical 95% quantile ranges (in brackets) for the estimation based on 10,000 simulations. The corresponding θs and their estimates can be found in Table A3.

Table A3. Results of utility standardization parameter recovery for choice consistency θ

	Random choice set		No-dominance choice set		Switching choice set	
Variable	θ = 6.1	θ = 5.8	θ = 14.5	θ = 13.5	θ = 10.4	θ = 9.5
Mean estimate θ	7.18 [4.00, 13.21]	6.75 [3.81, 12.03]	16.7 [9.23, 30.81]	15.7 [8.81, 29.34]	11.62 [6.76, 19.91]	10.81 [5.90, 18.91]
Relative bias	18% [−34%, 117%]	16% [−34%, 107%]	15% [−36%, 112%]	16% [−35%, 117%]	12% [−34%, 91%]	14% [−34%, 96%]
Absolute deviation	1.78 [0.06, 7.11]	1.62 [0.05, 6.23]	4.15 [0.13, 16.31]	3.89 [0.12, 15.84]	2.58 [0.08, 9.51]	2.43 [0.08, 9.17]
Relative absolute deviation	29% [1%, 117%]	28% [1%, 107%]	29% [1%, 112%]	29% [1%, 117%]	25% [1%, 91%]	26% [1%, 96%]

Note. Reported are mean values and empirical 95% quantile ranges (in brackets) for the estimation based on 10,000 simulations. The corresponding αs were 0.8 for the first and 1.2 for the second column in each choice set.

Table A4. Results of outcome standardization parameter recovery for risk preference α

	Random choice set		No-dominance choice set		Switching choice set	
Variable	α = 0.8	α = 1.2	α = 0.8	α = 1.2	α = 0.8	α = 1.2
Mean estimate α	0.86 [0.04, 2.27]	1.28 [0.18, 3.00]	0.8 [0.41, 1.26]	1.24 [0.75, 1.96]	0.8 [0.54, 1.11]	1.21 [0.86, 1.59]
Relative bias	8% [−95%, 184%]	7% [−85%, 150%]	0% [−49%, 58%]	3% [−38%, 63%]	0.4% [−32%, 39%]	1% [−27%, 32%]
Bias |± 0.2|	67%	74%	33%	47%	17%	28%
Absolute deviation	0.41 [0.01, 1.47]	0.52 [0.02, 1.80]	0.17 [0.01, 0.52]	0.24 [0.01, 0.76]	0.12 [0.00, 0.33]	0.15 [0.01, 0.42]
Relative absolute deviation	51% [2%, 184%]	43% [2%, 150%]	21% [1%, 64%]	20% [1%, 64%]	15% [1%, 41%]	12% [0%, 35%]

Note. Reported are mean values and empirical 95% quantile ranges (in brackets) for the estimation based on 10,000 simulations. The corresponding θs and their estimates can be found in Table A5.

Table A5. Results of outcome standardization parameter recovery for choice consistency θ

	Random choice set		No-dominance choice set		Switching choice set	
Variable	θ = 6.7	θ = 6.2	θ = 16	θ = 14	θ = 13.9	θ = 13.4
Mean estimate θ	12.15 [4.36, 74.84]	8.82 [4.46, 19.58]	18.6 [10.02, 34.72]	15.9 [9.05, 28.37]	15.6 [8.90, 26.62]	14.91 [8.63, 25.7]
Relative bias	81% [−35%, 1,017%]	42% [−33%, 216%]	16% [−37%, 117%]	13% [−35%, 103%]	12% [−36%, 92%]	11% [−35%, 92%]
Absolute deviation	6.17 [0.07, 68.12]	3.26 [0.06, 13.38]	4.76 [0.15, 18.72]	3.74 [0.13, 14.37]	3.49 [0.11, 12.72]	3.31 [0.12, 12.30]
Relative absolute deviation	92% [1%, 1,017%]	53% [1%, 216%]	30% [1%, 117%]	27% [1%, 103%]	25% [1%, 92%]	25% [1%, 92%]

Note. Reported are mean values and empirical 95% quantile ranges (in brackets) for the estimation based on 10,000 simulations. The corresponding αs were 0.8 for the first and 1.2 for the second column in each choice set.

Table A6. Results of monetary equivalence standardization parameter recovery for risk preference α

	Random choice set		No-dominance choice set		Switching choice set	
Variable	α = 0.8	α = 1.2	α = 0.8	α = 1.2	α = 0.8	α = 1.2
Mean estimate α	0.78 [−0.72, 3.00]	1.252 [−0.36, 3.00]	0.79 [0.36, 1.25]	1.25 [0.72, 2.09]	0.81 [0.56, 1.09]	1.21 [0.91, 1.53]
Relative bias	−2% [−190%, 275%]	4% [−130%, 150%]	−1% [−55%, 56%]	4% [−40%, 74%]	0.8% [−30%, 36%]	0.5% [−24%, 28%]
Bias |± 0.2|	77%	80%	34%	50%	14%	20%
Absolute deviation	0.68 [0.02, 2.20]	0.70 [0.02, 1.80]	0.17 [0.01, 0.54]	0.26 [0.01, 0.9]	0.11 [0.00, 0.31]	0.13 [0.00, 0.36]
Relative absolute deviation	85% [3%, 275%]	59% [2%, 150%]	21% [1%, 67%]	22% [1%, 75%]	14% [0%, 39%]	11% [0%, 30%]

Note. Reported are mean values and empirical 95% quantile ranges (in brackets) for the estimation based on 10,000 simulations. The corresponding θs and their estimates can be found in Table A7.

Table A7. Results of monetary equivalence standardization parameter recovery for choice consistency θ

	Random choice set		No-dominance choice set		Switching choice set	
Variable	θ = 0.07	θ = 0.07	θ = 0.15	θ = 0.14	θ = 0.15	θ = 0.16
Mean estimate θ	0.08 [0.05, 0.14]	0.08 [0.04, 0.13]	0.17 [0.09, 0.31]	0.15 [0.09, 0.27]	0.17 [0.09, 0.29]	0.18 [0.10, 0.31]
Relative bias	13% [−35%, 93%]	11% [−36%, 90%]	12% [−38%, 106%]	11% [−37%, 94%]	12% [−37%, 93%]	12% [−35%, 92%]
Absolute deviation	0.02 [0.00, 0.07]	0.02 [0.00, 0.06]	0.04 [0.00, 0.16]	0.04 [0.00, 0.13]	0.04 [0.00, 0.14]	0.04 [0.00, 0.15]
Relative absolute deviation	26% [1%, 93%]	25% [1%, 90%]	27% [1%, 106%]	25% [1%, 94%]	25% [1%, 93%]	25% [1%, 92%]

Note. Reported are mean values and empirical 95% quantile ranges (in brackets) for the estimation based on 10,000 simulations. The corresponding αs were 0.8 for the first and 1.2 for the second column in each choice set.

Table A8. Results of variance standardization parameter recovery for risk preference α

	Random choice set		No-dominance choice set		Switching choice set	
Variable	α = 0.8	α = 1.2	α = 0.8	α = 1.2	α = 0.8	α = 1.2
Mean estimate α	0.95 [0.00, 3.00]	1.31 [0.00, 3.00]	0.84 [0.25, 1.66]	1.26 [0.54, 2.36]	0.80 [0.56, 1.07]	1.20 [0.91, 1.53]
Relative bias	19% [−100%, 275%]	9% [−100%, 150%]	5.5% [−68%, 108%]	5.1% [−55%, 96%]	0.1% [−30%, 34%]	0.3% [−24%, 27%]
Bias |± 0.2|	79%	82%	53%	64%	12%	20%
Absolute deviation	0.62 [0.02, 2.20]	0.70 [0.03, 1.80]	0.27 [0.01, 0.86]	0.36 [0.01, 1.16]	0.1 [0.00, 0.29]	0.12 [0.00, 0.35]
Relative absolute deviation	77% [3%, 275%]	58% [2%, 150%]	38% [1%, 108%]	30% [1%, 97%]	13% [1%, 37%]	10% [0%, 29%]

Note. Reported are mean values and empirical 95% quantile ranges (in brackets) for the estimation based on 10,000 simulations. The corresponding θs and their estimates can be found in Table A9.

Table A9. Results of variance standardization parameter recovery for choice consistency θ

	Random choice set		No-dominance choice set		Switching choice set	
Variable	θ = 1.27	θ = 1.29	θ = 3.3	θ = 3.1	θ = 5.9	θ = 5.9
Mean estimate θ	1.45 [0.79, 2.58]	1.46 [0.79, 2.69]	3.68 [2.04, 6.60]	3.48 [1.91, 6.32]	6.51 [3.70, 11.15]	6.48 [3.74, 11.03]
Relative bias	14% [−38%, 103%]	15% [−39%, 108%]	12% [−38%, 100%]	12% [−39%, 104%]	10% [−37%, 89%]	10% [−37%, 87%]
Absolute deviation	0.36 [0.01, 1.31]	0.36 [0.01, 1.40]	0.88 [0.03, 3.30]	0.83 [0.03, 3.22]	1.44 [0.04, 5.25]	1.40 [0.05, 5.13]
Relative absolute deviation	28% [1%, 103%]	29% [1%, 108%]	27% [1%, 100%]	27% [1%, 104%]	24% [1%, 89%]	24% [1%, 87%]

Note. Reported are mean values and empirical 95% quantile ranges (in brackets) for the estimation based on 10,000 simulations. The corresponding αs were 0.8 for the first and 1.2 for the second column in each choice set.

Table A10. Results of baseline parameter recovery for α based on a random subset of 15 choice situations repeated four times

	Random choice set		No-dominance choice set		Switching choice set	
Variable	α = 0.8	α = 1.2	α = 0.8	α = 1.2	α = 0.8	α = 1.2
Mean estimate α	0.91 [0.24, 2.66]	1.32 [0.55, 2.99]	0.81 [0.31, 1.44]	1.25 [0.63, 2.21]	0.8 [0.55, 1.09]	1.21 [0.88, 1.56]
Relative bias	13% [−69%, 232%]	10% [−53%, 150%]	1.4% [−61%, 80%]	4.5% [−47%, 848%]	0.2% [−31%, 36%]	0.6% [−26%, 31%]
Bias |± 0.2|	72%	75%	28%	54%	14%	23%
Absolute deviation	0.45 [0.02, 1.86]	0.5 [0.02, 1.50]	0.20 [0.00, 0.64]	0.29 [0.01, 1.01]	0.11 [0.00, 0.31]	0.14 [0.00, 0.39]
Relative absolute deviation	56% [2%, 236%]	42% [1%, 150%]	25% [0.8%, 80%]	24% [1%, 84%]	14% [0.5%, 40%]	11% [1%, 33%]

Note. Reported are means and empirical 95% quantile ranges (in brackets) for the parameter estimation based on 10,000 simulations.

Table A11. Results of baseline parameter recovery for θ based on a random subset of 15 choice situations repeated four times

	Random choice set		No-dominance choice set		Switching choice set	
Variable	θ = 0.18	θ = 0.03	θ = 0.41	θ = 0.05	θ = 0.4	θ = 0.06
Mean estimate θ	0.99 [0, 3.6]	0.15 [0, 0.59]	1.07 [0.01, 8.19]	0.14 [0, 99]	0.58 [0.1, 1.07]	0.09 [0.01, 0.28]
Relative bias	449% [−99%, 1,899%]	424% [−99%, 1,899%]	160% [−96%, 1,899%]	179% [−99%, 1,899%]	44% [−75%, 373%]	50% [−82%, 439%]
Absolute deviation	0.94 [0.02, 3.42]	0.15 [0.03, 0.57]	0.85 [0.01, 7.78]	0.12 [0.01, 0.95]	0.32 [0.01, 1.49]	0.05 [0.00, 0.26]
Relative absolute deviation	520% [1%, 1,899%]	498% [1%, 1,899%]	208% [3, 1,899%]	240% [5%, 1,899%]	79% [2%, 373%]	90% [2%, 439%]

Note. Reported are means and empirical 95% quantile ranges (in brackets) for the parameter estimation based on 10,000 simulations.