CHARACTERIZATION FAILURE, CONFIDENCE, AND CHOICE: A BEHAVIORAL WELFARE ANALYSIS OF DECISIONS UNDER RISK AND CERTAINTY

GARRETT HALL, MITCHELL LINEGAR, ALDO LUCIA, KIRBY NIELSEN, CHARLES D. SPRENGER

1. Motivation

Both general and behavioral welfare analysis of choice is preceded by the definition of a welfare-relevant domain: a potential subset of a body of decisions within which choices are understood to plausibly be the product of maximization of true preferences. Choices within the welfare-relevant domain are understood to have a normative interpretation. They plausibly represent maximized preferences and so can be used as the basis both for understanding and estimating the nature of true preferences, and as a basis for welfare calculations. Choices outside of the welfare-relevant domain may serve to meaningfully estimate some bias, but such choices merit no weight in policy guidance.

Given the fundamental importance of defining the welfare-relevant domain, researchers have developed general approaches for choice exclusion. In traditional welfare economics, a welfare criterion is often proposed that permits exclusion of a class of decisions. For example, under the famous long-run criterion in intertemporal choice, decisions involving the present are excluded from the welfare-relevant domain, though such decisions may be useful for identifying any present bias in decision-making. Decisions taken without some elements of information on the objects of choice may similarly be excluded both from the welfare-relevant domain and any estimation of true preferences (for discussion, see Bernheim and Rangel, 2005).

Behavioral welfare analysis often departs from general welfare analysis in the conservatism of choice exclusion. Bernheim and Rangel (2009) argue that no specific criteria contradicting choice should be imposed without demonstrating that said choice suffers from characterization failure. That is, unless the researcher can demonstrate that an individual fundamentally mischaracterized their decision and so could not have implemented their true preferences, there is no grounds for exclusion.[1]

Characterization failure is notably challenging to demonstrate or assess. From choice data alone, how is a researcher to know if the choice is mischaracterized? One technique used to plausibly induce characterization failure is to design experimental conditions that provide limited information on the choice options. Choice in such conditions is rendered welfare irrelevant by assumption due to the obfuscation of relevant choice-related details

[1]For the purposes of having a working definition in the context of risk, we consider accurate characterization as correctly perceiving the mapping from states to outcomes.

(Allcott and Taubinsky, 2015). Another is to induce values for objects and study if individuals make choices that are consistent with such values. If induced values are not reliably maximized in such definitive choice environments, the individual plausibly mischaracterized the task and the corresponding choices can be presumed welfare irrelevant. In one prominent example, Cason and Plott (2014) induce values for a token of \$2 and show that subjects facing a Becker DeGroot Marschak (BDM) mechanism with limited instructions do not reliably reveal a value of \$2. Such mistaken BDM choices can plausibly be excluded from the welfare-relevant domain.[2]

Recently, a further tool has arisen which could potentially be used for identifying decisions to be excluded from the welfare-relevant domain. Oprea (2024) develops an approach based on pattern matching. As in Cason and Plott (2014), the researcher constructs a decision environment where the decision-maker should have a specific induced value, and then documents a pattern of definitively mistaken choice which renders the choices excludable from the welfare-relevant domain. Then, the researcher considers a similar environment without induced values. If the data patterns look similar to the definitively mistaken choice patterns, then this second set of decisions may also be excludable from the welfare-relevant domain. Oprea (2024) demonstrates this pattern matching exercise in the domain of risk: He documents that individuals exhibit similar deviations from expected value maximization when presented with risky lotteries as when presented with riskless "mirrors" that are presented in a way to resemble lotteries but represent deterministic payments. If deviations from expected value maximization over mirrors are definitively excludable from the welfare-relevant domain—and if the same patterns exist with risk—then it stands to reason that the patterns revealed in these risky choices are also excludable from the welfare-relevant domain. On the basis of this pattern matching, Oprea (2024) questions the normative relevance of non-expected utility models and, by extension, the relevance of the underlying risky choice data often used to identify and estimate such models. Similarly, Enke et al. (2023) use the pattern matching approach to question the normative relevance on non-exponential intertemporal choice models and experimental discounting choices.

The results obtained from these pattern matching exercises pose serious challenges for the interpretation of choices revealed through experiments in these environments. If the same data patterns are observed with and without risk or with and without time, then the central economic dimension of interest—preferences under uncertainty or intertemporal preferences—may be largely a second-order concern. The first-order object of interest must be the underlying process of perception or judgement that guides decisions. Oprea (2024) reasons in precisely this direction by arguing the decisions under risk should be reformulated as decisions under complexity.

Framed as a behavioral welfare problem, when the analyst is considering removing choices from the welfare-relevant domain, they need to identify characterization failures.

[2]Indeed, Cason and Plott (2014) go further arguing that even more than these isolated induced-value BDM choices are excludable and suggesting that other phenomena studied with the BDM mechanism may actually be mistakes. This extrapolative step requires that the characterization failures in the definitive environment are informative of characterization failures in other environments using the BDM mechanism.

In this way, the logic of the pattern matching approach is that the characterization failures in the definitive environment, say without risk, are informative of the nature of the characterization failures in the target environment, say with risk.[3] However, this relies on the critical assumption that individuals correctly perceive the definitive environment as removing all preference-relevant features of the target environment. In other words: What if the characterization failure in the risk-free environment was to misconstrue the risk-free questions as actually risky? In such a case, similar data patterns could emerge in risky and risk-free environments simply because the risk-free environment is perceived as risky. If this is the case, though, the risky choice data would not be excludable from the welfare-relevant domain and the interpretation of choice patterns reflecting complexity preferences rather than risk preferences would not prevail. It is therefore critical to know and measure the nature of characterization failures and understand their similarities and differences across decision contexts. Oprea (2024) argues away this challenge by stating that subjects had to take a pre-test prior to starting the study to ensure that they reported example risk-free questions to be risk-free and example risky questions to be risky. However, no measurement of characterization is made for each type of task.

We implement a simple tool to measure characterization failures. In a risky and a risk-free choice environment, we ask subjects to report the mapping from the objects of choice to the distribution of outcomes that would obtain. We argue that this mapping is precisely the information required to correctly characterize a decision. These characterization questions take the form of filling out a contingency table. Subjects provide these characterization tables for every decision, allowing us to measure the characterization failure at the decision level. This decision-level characterization is required for explicit exclusion from the welfare-relevant domain under the behavioral welfare tradition. Across experimental conditions, the characterizations are either provided before or after choice; and either with or without explicit incentives for accurate representation.

As an alternative to pattern matching as a basis for exclusion, our project considers one additional recent technique for identifying error-prone choices: the use of ex-post confidence statements. In effect, individuals themselves provide a statement of whether they view their choice as the implementation of their true preferences by which they can confidently stand. Ex-post statements of confidence can be related to patterns in choice in order to identify potentially excludable decisions, or bodies of decisions. Enke and Graeber (2023) show that greater deviations from expected value maximization in risky choice and changing risk tolerance as probabilities increase are both correlated with lower confidence measures. If non-maximizing choices primarily arise when subjects are not confident it stands to reason that such choices are associated with characterization failure, and so can be excluded from welfare relevance. Correspondingly, any utility model estimated on such

[3]Alternatively, one could imagine thinking of all choices as correctly characterized and failure of maximization therefore reflects the behavioral response to the complexity of a correctly-characterized choice problem, and therefore from the behavioral welfare approach, neither the environment with risk nor the environment without risk would be excludable from the welfare-relevant domain on the basis of characterization failure. However, there may be other paths to exclusion, such as the obfuscation of relevant choice-related details produced by complexity.

data cannot be understood as a model of true preferences. As with Oprea (2024), patterns in risky choice data are re-interpreted as reflecting something deeper than risk; in this case cognitive imprecision on optimal actions. Within our study, we will also explore confidence measures, which will be collected in each condition directly after choice.

2. Experimental Design

We analyze binary choices between lotteries and binary choices between objects, the values of which must be determined via expected value computation. We represent lotteries as "risky bags" containing 100 differently colored tickets, each color corresponding to a distinct value. Each lottery permits the drawing of one ticket from the corresponding bag, and the drawer receives the value indicated on the ticket. To replicate the state-to-value mapping found in lotteries, but without the associated risk, we introduce "safe bags." These bags also contain 100 tickets of varying colors, but all tickets share the same value. This uniform value is conveyed to the subjects as an expected value calculation. We refer to choice tasks involving risky bags as RISKY-tasks, and to those that involve safe bags as EV-tasks.

We denote a risky bag, B, which contains n tickets with a payoff of $\$L$ and $100-n$ tickets with a payoff of $\$H$, as $(n, \$L; 100-n, \$H)$. For each risky bag, we devise a corresponding deterministic bag, M_B, which contains the same number of each color ticket as in the risky bag B. However, each ticket in M_B is worth the expected value of lottery B. In other words, each ticket is valued at $[\$L \times n + \$H \times (100-n)]/100$. Figure 1 illustrates an example of a risky bag containing 10 purple tickets that pay \$30 and 90 gold tickets that pay \$0, alongside its corresponding safe bag.

We construct risky and corresponding safe bags to investigate whether the common ratio effect, a phenomenon widely documented in the context of binary choices over lotteries, also applies to riskless options. Here, we outline the binary choice tasks employed to examine the common ratio effect with risky bags. In our experiment, each of these tasks has a corresponding choice task, where the risky bags are substituted by their related safe bags.

To examine the common ratio effect, we consider two types of binary choice tasks:

Unmixed: $A_n = (\$M_n, 100)$ vs. $B_n = (\$0, n; \$H, 100-n)$.
Mixed: $C_n = (\$0, 80; \$M_n, 20)$ vs. $D_n = (\$L, 100-(100-n)\times 0.2; \$H, (100-n)\times 0.2)$.

We consider five different potential values for n: 10, 20, 50, 80, and 90. For each possible value of n, we consider two potential values for M_n: the expected value of lottery B_n subtracted by one, and added by one. The rationale for this design choice is that, when examining the unmixed and mixed choice tasks for the associated safe bags, any decision-maker with monotone preferences over money should strictly prefer one of the two bags. In total, we have two versions of each task (risky and safe bags), two types of choice tasks (unmixed and mixed), five possible values for n, and two possible values of M_n for each n. This gives us a total of 40 possible choice tasks. Each subject in the experiment encounters

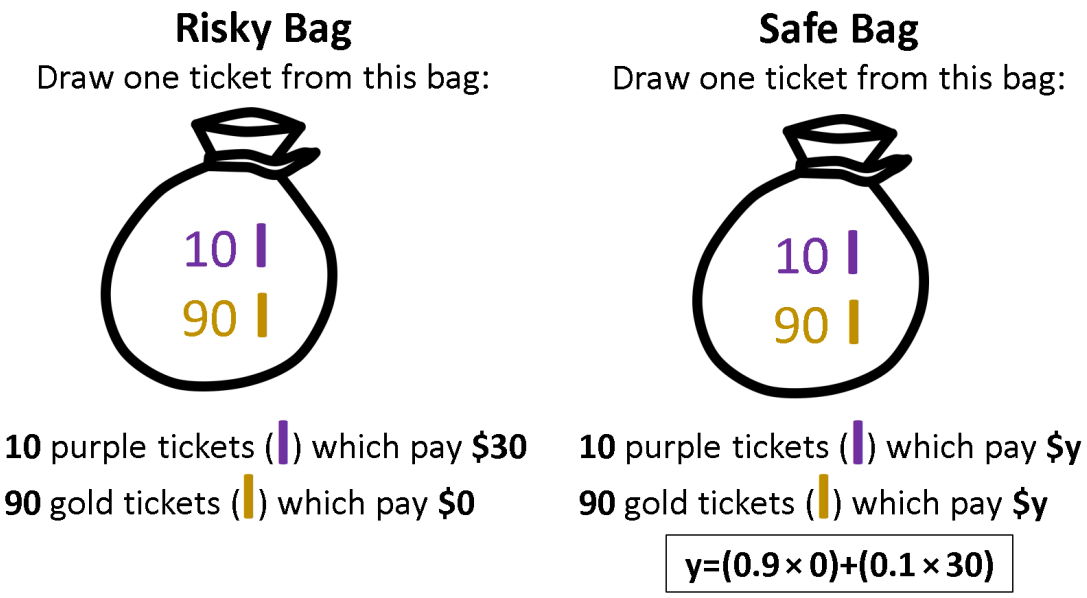

FIGURE 1. Risky and associated safe bags.

20 out of these 40 possible tasks. In each choice task, subjects are asked to select their preferred bag and indicate their confidence in their choices on a scale from 0 to 100.

If the common ratio effect can be observed both with and without risk, then we may infer that the presence of risk is not a prerequisite for the emergence of the common ratio effect. This conclusion hinges on the assumption that subjects perceive the riskless 'safe bags' as truly free of risk. If this condition is not met, then there's no basis to question the normative relevance of risk preferences as the key driver of the common ratio effect.

To assess subjects' characterization of the tasks, we ask subjects to complete some contingency tables in three different conditions: Incentivized, Unincentivized, and Control. The contingency tables remain the same across all three conditions. We ask subjects to denote the probabilities and the values of the different tickets for each of the two bags in a choice task. Figure 2 presents an example of the tables that subjects are required to fill in. Initially, they are asked to write the probabilities of drawing each possible ticket from each bag. Subsequently, they are asked to specify the values associated with all tickets that, according to their previous answers, have a non-zero probability of being drawn.

The three conditions—Incentivized, Unincentivized, and Control—differ in the timing of when subjects complete the tables and the specifics of the incentive schemes. In both the Incentivized and Unincentivized conditions, subjects are asked to fill in the tables for each of the 20 choice tasks before indicating their preferred option. They can review their responses in the contingency tables when they are asked to choose their preferred bag. Each

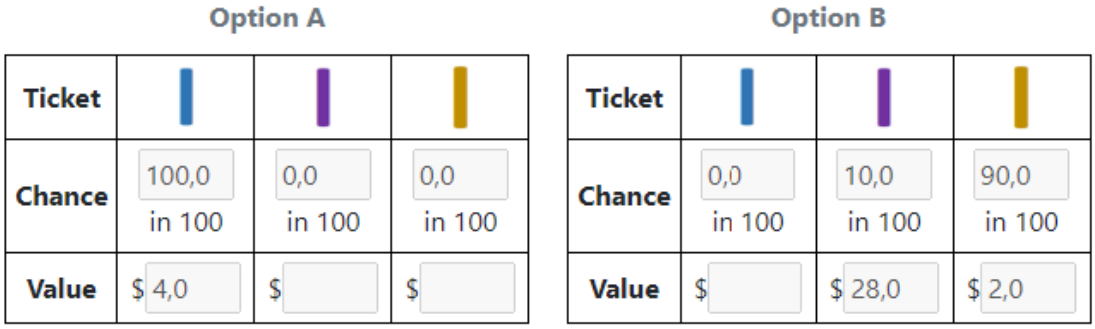

FIGURE 2. Contingency tables.

subject in these conditions has an 20 in 100 chance of qualifying for a bonus payment. For those who are eligible, one of the 20 tasks they have undertaken during the experiment is randomly selected.

In the Incentivized condition, subjects receive a bonus only if they have correctly completed the contingency table in the task chosen for payment. In contrast, the Unincentivized condition does not tie the bonus payment to the accuracy of the responses given in the contingency tables. The Control treatment is somewhat different; here, subjects are asked to complete contingency tables about each task only after they have made their choices for all tasks. Similar to the Unincentivized condition, the Control condition doesn't tie the subjects' responses in the contingency tables to their bonus payments. The key details of the three conditions in the experiment are summarized in Table 1. We do not have a treatment with incentivized characterization after all choices because this would be difficult to explain to subjects while they are making their decisions.

TABLE 1. Design Summary.

	Incentivized	**Unincentivized**	**Control**
Characterization Table Timing	Before each choice	Before each choice	After all choices
Characterization Incentives	Incentivized	Unincentivized	Unincentivized
Number of tasks	20	20	20

2.1. **Comprehension Questions.** We assess subjects' understanding of the experimental instructions with a series of comprehension questions. In the Incentivized and Unincentivized conditions, we introduce participants to two training tasks designed to assess their understanding of risky and safe bags. These tasks involve fictitious choice scenarios: one with risky bags and another with safe bags. Participants are asked to fill out contingency

tables for each scenario. If errors are made, explanations are provided to clarify mistakes. Participants must accurately complete these comprehension tasks before proceeding with the experiment. After completing each comprehension task, participants are also asked to indicate their preferred choice and confidence level, facilitating familiarity with the experimental process.

In the Control condition, participants engage in the same fictitious choice tasks prior to beginnning the experiment without the initial requirement of completing contingency tables. After making all their choices, they are introduced to the contingency tables and asked to correctly fill them out for the two fictitious tasks they previously encountered. As in the Incentivized and Unincentivized conditions, feedback is provided for mistakes. To conclude the experiment, participants must provide correct responses in these comprehension tasks.

In all three conditions, we assess participants' understanding of how bonus payments are awarded. Using a fictitious example, we pose the following four multiple-choice questions to participants:

(1) How do we determine your potential bonus payment?
(2) What is your potential bonus payment if you preferred Option B and a purple ticket is drawn?
(3) Under what condition will you receive the potential bonus payment at the end of the experiment?
(4) What is the chance that you are eligible for a potential bonus payment?

Subjects are required to select one out of four possible answers for each question, and explanations are provided in case of mistakes. All subjects must correctly answer all questions to begin the experiment. In addition to these comprehension checks, at the end of the study we ask the subject if they made use of any decision aids such as a calculator.

3. Analysis

We say that a choice task suffers from a *characterization failure* if we find any mistakes in the contingency tables. Among the types of possible characterization failures, we are particularly interested in instances where subjects wrongly perceive an EV-task as carrying risk, or a RISK-task as carrying no risk. We will refer to characterization failures of this type as *recognition failures*. Finally, we say that an EV-task suffers from a *maximization failure* when a subject fails to choose the option with the higher value.

The first step in our analysis will be to describe behavior in RISKY-tasks and EV-tasks across the three conditions of the experiment. We will

- Document the sensitivity of choice probabilities in RISKY-tasks and EV-tasks to two factors: the proportion of high-value tickets involved in each risky or safe bag, and the value of different tickets.
- Document the eventual emergence of the common ratio effect in RISKY-tasks and EV-tasks.
- Document the relationship between the distribution of choices in RISKY-tasks and EV-tasks.

We will calculate the extent of the common ratio effect as the difference between the proportion of subjects choosing the "safer" option in Unmixed tasks and the proportion of subjects choosing the "safer" option in the associated Mixed tasks.[4]

After documenting behavior in RISKY-tasks and EV-tasks, we will proceed to evaluate our main analysis. Our main analysis proceeds in four steps.

(1) Control Condition Analyses: The first step focuses on our Control condition and has three components.
   (a) Pattern Matching: We will assess the extent of pattern matching between RISKY-tasks and EV-tasks in terms of choice probabilities and the emergence of common ratio effects.
   (b) Characterization Failure and Maximization Failure: Within EV-tasks we will assess the connection between maximization failure and characterization failure in subsequent characterization tables.
   (c) Characterization Failure and Recognition Failure: Within EV-tasks characterization tables, we will assess what proportion of characterization failures derive from failing to accurately recognize these tasks as risk free.

(2) Incentivized and Unincentivized Conditions: The second step focuses on our conditions where subjects complete characterization tables prior to completing the tasks. This step has four components.
   (a) Effect of Prior Tables on Characterization and Maximization Rates: We will compare correct characterization and maximization rates with EV-tasks across the control group and the prior characterization groups.
   (b) Effect of Prior Tables on Pattern Matching: We will assess the extent of pattern matching between RISKY-tasks and EV-tasks in terms of choice probabilities and the emergence of common ratio effects. The correlation in behavior across tasks will be compared to that in the control condition. The magnitude of the CRE in each task type will be compared to that in the control condition.
   (c) Pattern Matching Conditional on Recognition. We will assess the extent of pattern matching between RISKY-tasks and EV-tasks in terms of choice probabilities and the emergence of common ratio effects, conditional on accurate recognition of the task type in the prior characterization tables.

(3) Analysis of Self-Reported Confidence. As an alternative to pattern matching, low self-reported confidence has been considered a basis for choice exclusion. This analysis has two components.
   (a) Confidence, Characterization and Maximization in EV-tasks. We will assess the relationship between Confidence and correct characterization and maximization in EV-tasks. Analysis will be conducted within each condition. Analysis will indicate whether removing observations on the basis of confidence also removes characterization and maximization failures leading to more correct choices and less anomalous CRE behavior in EV-tasks.

[4]Note that options in EV-tasks carry no risk. We consider one safe bag A to be "safer" than another safe bag B if the risky bag associated with A is safer than the risky bag associated with B.

(b) Confidence, Characterization, and Maximization in RISKY-tasks. We will assess the relationship between confidence and correct characterization and behavior in RISKY-tasks. Analysis will be conducted within each condition. Analysis will indicate whether removing observations on the basis of confidence also removes characterization and expected value maximization failures leading to less CRE behavior in RISKY-tasks.

(4) The Effect of Incentives versus Prior Characterization Alone: Analysis will be provided comparing the additional effects of incentives beyond the prior characterization tables for increasing characterization rates, recognition rates, increasing confidence, and value or expected value maximization rates in both EV and RISKY-tasks. This analysis will be conducted separately for mixed and unmixed problems as such problems may differ in the challenge of accurate characterization.

3.1. **Sample Size and Power Calculations.** Given the four blocks of analysis considered above, we propose to collect a sample of roughly 1000 subjects, each completing 20 choice tasks. These subjects will be split equally between the Incentivized, Unincentivized, and Control Conditions. Our sample size selection is guided by two principle power calculations:

(1) Powering identification of CRE within each probability, payment values, and task type. Within our pilot data's control condition the average choice probability for the safe or simple option in unmixed questions was approximately 0.75. In order to detect a CRE of 0.15 or greater (in absolute value) with 80% power, approximately 150 observations are required in each cell (mixed and unmixed). 333 subjects assigned to the control condition assures an expectation of 166 observations in each randomly assigned cell as each subject provides half of a full data set in expectation. Within the prior literature on the CRE summarized by Blavatskyy et al. (2023), the median number of observations is 76, derived primarily from within-subject designs. Given our primarily between-subject design comparing choices of subjects to randomly assigned conditions, we consider this expansion of the average sample size appropriate.

(2) Powering treatment effects on correct characterization and maximization rates in EV-tasks. Within our pilot data's control condition EV-tasks, the average correct characterization rate was approximately 0.7 and the average correct maximization rate was approximately 0.74. In order to detect a treatment effect of 0.05 or greater (in absolute value) with 80% power, 1251 observations are required for characterization rates and 1128 are required for maximization rates in each cell. Our design provides an expectation of $333 \times 20 = 6660$ observations in each cell if each observation is treated as independent.[5]

[5]A more conservative approach would be to use only one observation per subject in which case we would be powered to detect treatment effects of 0.10 in each case.

References

Allcott, H. and Taubinsky, D. (2015). Evaluating behaviorally motivated policy: Experimental evidence from the lightbulb market. *American Economic Review*, 105(8):2501–2538.

Bernheim, B. D. and Rangel, A. (2005). Behavioral public economics: Welfare and policy analysis with non-standard decision-makers. Technical report.

Bernheim, B. D. and Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124(1):51–104.

Blavatskyy, P., Panchenko, V., and Ortmann, A. (2023). How common is the common-ratio effect? *Experimental Economics*, 26(2):253–272.

Cason, T. N. and Plott, C. R. (2014). Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. *Journal of Political Economy*, 122(6):1235–1270.

Enke, B. and Graeber, T. (2023). Cognitive uncertainty. *Quarterly Journal of Economics*, 138(4):2021–2067.

Enke, B., Graeber, T., and Oprea, R. (2023). Complexity and time. Technical report.

Oprea, R. (2024). Simplicity equivalents. Technical report.