RISK ATTITUDES, SAMPLE SELECTION, AND ATTRITION IN A LONGITUDINAL FIELD EXPERIMENT

Glenn W. Harrison, Morten I. Lau, and Hong Il Yoo*

Abstract—We evaluate the temporal stability of risk preferences using a remarkable data set that combines sociodemographic information from the Danish Civil Registry with information on risk attitudes from a longitudinal field experiment. Our econometric model accounts for endogenous sample selection and attrition processes that may confound inferences about temporal stability. Our experimental design builds in randomization on the incentives for participation that facilitates empirical identification of the model. In general, we find evidence consistent with temporal stability after correcting for the effects of selection and attrition. When neglected, these effects change our inferences in an economically and statistically significant manner.

I. Introduction

ANY longitudinal survey or experimental design raises concerns about sample selection and attrition, and response rates may vary dramatically depending on the nature of the study and incentives provided in the design. Controlling for endogenous effects of sample selection requires some background information on subjects who did not select into the survey or experiment, so that one can estimate a latent selection process and its correlation with the primary outcome of interest. This information is often missing, and most longitudinal studies are concerned just with attrition effects. For nonparticipants, attrition outcomes are also missing, and strictly speaking one cannot control for attrition effects without addressing endogenous selection first. Without controlling for selection effects, the estimates of a latent attrition process may be subject to selection bias even when there is no effect of selection on the primary outcome in the initial wave of the study.

Using a structural model of risky choices that allows for endogenous sample selection and panel attrition, we analyze data from a longitudinal field experiment with a stratified sample of the adult Danish population. The data are linked to administrative data from the Civil Registry in Denmark, allowing us to observe background information on nonparticipants. We illustrate the importance of controlling for within-wave and between-wave effects of sample selection in the evaluation of individual risk attitudes at different points in time.

Temporal stability of risk preferences is a common assumption in evaluations of economic behavior.[1] When the potential benefits of any social insurance policy are evaluated, for example, one must know the risk preferences of the beneficiaries of the policy in order to calculate expected individual welfare (Harrison & Ng, 2016). If preferences are unstable, then what might be a socially attractive policy today could become an unattractive policy in the future. When nudges or boosts are provided to improve decision making over risky portfolios, to take another example, one must also condition these on knowledge of the risk preferences of the target population in order to ensure that they are welfare enhancing (Harrison & Ross, 2018). If those preferences are unstable over time, what might seem like a welfare-enhancing nudge today could again become a welfare-reducing nudge in the future. Behavioral welfare economics requires that we not only identify risk preferences but check their stability over time as policies that are contingent on those preferences take effect.

Testing the assumption of temporal stability of risk preferences with the same individuals requires, of course, that one address problems of sample selection and attrition. We design and evaluate a longitudinal field experiment with a nationally representative sample of Danish adults between 19 and 75 years of age to address this question. The sample is randomly drawn from the Civil Registry and stratified with respect to population size in each county. Our design builds in explicit randomization on the incentives for participation, an idea suggested by the theoretical literature on sample selection models and easy to implement in the sampling process and subsequent experiment.

The classic problem of sample selection refers to possible recruitment biases, such that individuals with certain types of characteristics are more likely to be in the observed sample. The statistical problem is that there may be some unobserved characteristics that simultaneously affect someone's chance of being in the sample as well as affecting other outcomes that the analyst is interested in. In any longitudinal study, there is also an inherent scope for post-recruitment selection bias due to panel attrition, which occurs as some subjects may leave the panel.[2] We build on the direct likelihood approach of Heckman (1976), Hausman and Wise (1979), and Diggle and Kenward (1994) and use maximum simulated likelihood

Received for publication August 4, 2017. Revision accepted for publication April 10, 2019. Editor: Shachar Kariv.

*Harrison: Georgia State University; Lau: Copenhagen Business School and Durham University Business School, Durham University; Yoo: Durham University Business School, Durham University.

We thank the Danish Social Science Research Council for research support under project 275-08-0289 and 12-130950; the Carlsberg Foundation under grant 2008-01-0410; and John DiNardo, Justin McCrary, Bill Schworm, and three referees for helpful comments. Harrison is also affiliated with the University of Cape Town.

Appendixes A through F are available in the online supplement at http://www.mitpressjournals.org/doi/suppl/10.1162/rest_a_00845.

[1]The term *stability* can mean unconditional stability, or it can mean stable preferences conditional on a given set of covariates. In the latter case, the question is whether preferences are a stable (and known) function of those covariates (Andersen et al., 2008b). We consider both forms of stability.

[2]The attrition problem is not the same as the dropout problem. As Heckman, Smith, and Taber (1998) stressed, the latter refers to subjects that leave some randomized program or intervention but remain in the sample. The attrition problem concerns subjects that completely drop out of the sample.

The Review of Economics and Statistics, July 2020, 102(3): 552–568

https://doi.org/10.1162/rest_a_00845

to estimate unique probit-kernel models that consider the full longitudinal design of the experiment. Our models control for the effects of selection and attrition on risk preferences inferred from both waves of the experiment, as well as addressing unobserved heterogeneity in risk preferences of the underlying population.

We consider a structural analysis of two theories of decision making under risk, expected utility theory (EUT) and rank dependent utility (RDU), where the latter is a highly influential alternative to EUT that relaxes the independence axiom under EUT.[3] Each theory has a set of structural parameters that characterize risk preferences. Previous analyses of temporal stability do not control for recruitment bias and focus on either population averages of the structural parameters or individual-level estimates that have no structural link to the population distribution of risk preferences. In contrast, our analysis controls for endogenous sample selection and attrition and captures unobserved heterogeneity around the population averages by modeling all structural parameters as individual-level random coefficients that follow a population distribution. We allow the population distribution to vary over time and the random coefficients to be correlated with the error terms in the selection and attrition equations.

This estimation approach allows us to consider temporal stability of risk attitudes at two levels, with and without controls for endogenous sample selection and attrition: (a) the population level, by comparing the population distributions of structural parameters over time, and (b) the individual level, by considering the correlation between individual-specific random coefficients over time. Our direct likelihood approach is inspired by the trivariate probit model of Capellari and Jenkins (2004), which includes two types of selection equations, but their primary outcome equation is the linear index probit model and their selection equations do not address selection bias in the sense of recruitment bias. We are not aware of past statistical models that capture unobserved heterogeneity in latent structural parameters with controls for recruitment bias and/or attrition bias in longitudinal studies, or empirical studies that use the panel correlations of preference parameters to measure individual temporal stability.

No existing studies test temporal stability of risk attitudes in the context of a model that addresses unobserved preference heterogeneity across the population. Glöckner and Pachur (2012) and Zeisberger, Vrecko, and Langer (2012) are so far the only studies that test temporal stability of risk preferences at the individual level. But they do not consider temporal stability at the population level and do not control for sample selection or attrition bias.

Existing studies on temporal stability of risk attitudes do not control for selection bias or attrition bias.[4] In fact, most studies do not even make a passing reference to "sample selection" and, perhaps more remarkably, "attrition" or "retention."[5] Dasgupta et al. (2017) report a significant difference in the sample average risk attitudes of the attrited and the retained, but do not undertake statistical correction for attrition bias based on unobservables and do not mention selection bias.

We draw several conclusions from our statistical analysis. First, we find evidence that the use of different fixed recruitment fees can affect the decision to participate in our experiment.[6] When we used a relatively substantial recruitment fee of 500 kroner, which is about US$100, 24.1% of invitees accepted the invitation to the initial wave of our experiment. The initial acceptance rate fell to 18.1% when we instead used 300 kroner. Of course, this is just a "law-of-demand" effect from paying more money for people to participate, but it demonstrates that there are indeed deliberate decisions being made about participation. The second wave of our experiment paid the same recruitment fee of 300 kroner to every person, and there was no significant difference in the retention rates of subjects who were initially recruited with the high fee (48.4%) and subjects who were initially recruited with the low fee (54.7%).

Second, we find evidence that correcting for endogenous sample selection and panel attrition changes our inferences about risk preferences in an economically and statistically significant manner. The results suggest that one should not discount the potential effects of selection and attrition a priori, even when a self-selected sample and an underlying population of interest look more or less similar in terms of observed characteristics. Subjects participating in each wave of our experiments have demographic characteristics that are comparable to the adult population in Denmark, but without correcting for endogenous selection and attrition, our EUT specification would have overestimated the average Dane's relative risk aversion in the first wave by a factor of about 2. Under RDU, nonlinear probability weighting, capturing pessimism or optimism in relation to objective probabilities, may generate a positive or negative risk premium even when the individual has a linear utility function. Without correction for endogenous selection and attrition, our RDU specification

[3]Considerable experimental evidence points to violations of the independence axiom under EUT, at least for some individuals. Several earlier alternatives to EUT relaxed the independence axiom in ways that maintained the linearity of indifference curves in the Marschak-Machina triangle representation, but experimental evidence quickly rejected those alternatives in favor of models that had nonlinear indifference curves. RDU has emerged as the most popular alternative in the literature that allows for these types of violations of the independence axiom in the gain domain. Starmer (2000) provides an excellent review of these developments.

[4]Andersen et al. (2008b) is a hybrid, viewing the sample in their first wave as the population that is then selected into later waves, and model the sample selection into later waves.

[5]Smidts (1997), Goldstein, Johnson, and Sharpe (2008), Baucellis and Villasís (2010), Glöckner and Pachur (2012), and Zeisberger et al. (2012).

[6]Paying no fixed recruitment fee is not a panacea for the sample selection issues we consider; it just masks it and makes it impossible to evaluate since there is no variation in those fees. There are other sensible reasons why one should avoid zero show-up fees, since that could generate altogether different, and nasty, biases in sample selection documented by Kagel, Battalio, and Walker (1979) and Eckel and Grossman (2000).

would have substantially underestimated the population share of individuals who have an inverse-S probability weighting function that captures optimism for small probabilities and pessimism for large probabilities.

Finally, we draw several conclusions on temporal stability of risk preferences that depend on which aspect of temporal stability one is interested in. The range of results reflect the strengths of our empirical specifications that allow us to define and test temporal stability in several ways. For example, consider risk aversion in the EUT sense of a concave utility function. Under both EUT and RDU, we find that the average Dane is risk averse in this sense, and this conclusion is robust over time. But we still find some instability in the population distribution of risk aversion under RDU as there is a decline in the extent of unobserved preference heterogeneity around the average. When focusing on the within-individual autocorrelation of risk aversion, we find estimates of 0.36 under EUT and 0.69 under RDU, which lie between the two extreme cases of completely unrelated and completely stable preferences. Of course, under RDU, risk preferences are also characterized by the probability weighting function. We find more evidence on the stability of the probability weighting function than for the utility function, at both the population and individual levels. Overall, we find evidence consistent with temporal stability under EUT and RDU at the aggregate population level.

Our use of exogenously varied recruitment fees demonstrates how one can constructively design features of a survey, or experiment to facilitate empirical identification of sample selection effects. Building on Heckman (1976, 1979), the emphasis in the literature has been on the discovery of some exclusion restrictions, referring to variables that affect the probability of selection but not the primary outcome of interest.[7] The collection of these variables could be designed by the surveyor or experimenter but often were not.[8] In most cases, analysts simply have to live with the existing set of variables in a survey or experiment and search for exclusion restrictions on an a priori basis. The later theoretical literature, typified by Das, Newey, and Vella (2003), stresses the value of direct controls over the probability of selection rather than relying on some variables selected on an a priori basis.

II. Data

A. *Field Sampling Procedures*

Between September 28 and October 22, 2009, we conducted an artifactual field experiment[9] with 413 Danes.[10] The sample was drawn to be representative of the adult population as of January 1, 2009, using sampling procedures that are virtually identical to those documented in Andersen et al. (2008a). We received a random sample of the population between ages 18 and 75, inclusive, from the Danish Civil Registration Office, stratified the sample by geographic area, and sent out 1,996 invitations. We drew this sample of 1,996 invitees from a random sample of 50,000 adult Danes obtained from the Danish Civil Registration Office, which includes information on sex, age, residential location, marital status, and whether the individual is an immigrant. Thus, we are in the fortunate, and rare, position of knowing some basic demographic characteristics of the individuals who did not agree to participate in our experiment.[11]

At a broad level, our final sample is representative of the population: the sample of 50,000 subjects had an average age of 49.8, 50.1% of them were married, and 50.7% were women; our final sample of 413 subjects had an average age of 48.7, 56.5% of them were married, and 48.2% were women. We stress this comparison because it is often made to assuage concerns about sample selection: check if the final sample is similar to the population for a few observed characteristics, and then assume it is representative in all characteristics, including those that are latent and unobserved. In the absence of the type of data we have access to in Denmark, this may appear to be a reasonable second-best procedure, but our results show that it may be an inadequate check on endogenous sample selection effects.

The initial recruitment letter for the experiment explained the purpose and that it was being conducted by Copenhagen Business School. The letter clearly identified that there would be fixed and stochastic earnings from participating in the survey. In translation, the uncertainty was explained as follows:

> **You can win a significant amount**
> To cover travel costs, you will receive 500 kroner at the end of the meeting. Moreover, each participant will have a 10% chance of receiving

[7] Without such exclusion restrictions, identification of sample selection models has to rely on the validity of functional form assumptions alone, such as the bivariate normality of the error terms in the maximum likelihood estimation of the standard Heckman model. Identification in this instance is formally achieved, but is known to be "weak" (Meng & Schmidt, 1985, and Keane, 1992). Exclusion restrictions are formally required for identification when semiparametric specifications are used (Lee, 1995).

[8] We know of only two applications of the constructive approach to building exclusion restrictions into the experimental design. Appendix B provides a review of the related studies. It is folklore in survey research that information is often retained on how many calls were made to a subject, how hard that person was to contact in other ways, or which interviewer conducted the survey. Although not the object of randomization, information of this kind might be used as an instrument to model the probability of selection.

[9] An artifactual field experiment is defined by Harrison and List (2004) as involving the use of artifactual instructions, task, and environment with a field subject pool.

[10] The negative effects of the global financial crisis of 2007 and 2008 were largely in place by the time of our experiments, between September 2009 and October 2010. The European sovereign debt crisis was just starting to manifest when our experiments began, and Denmark was about to begin a fiscal budgetary crisis in 2010 that persisted for several years. Kickert (2013) provides a detailed account of Denmark's responses to these crises.

[11] It is possible to extend this list of characteristics by taking our experimental data to Statistics Denmark, which stores the same data that we obtained from the Civil Registration Office, and merging it with the entire set of data that is available on all of the invited subjects. One can then undertake the same statistical analyses but with a larger set of covariates to explain sample selection.

> an amount between 50 and 4,500 kroner in one part of the survey. In another part of the survey, each participant will have a 10% chance of receiving at least 1,500 kroner. Some of these amounts will also be paid out at the end of the meeting, and some amounts will be paid out in the future. A random choice will decide who wins the money in the different parts of the survey.

The fixed amount is 500 kroner in the treatment that this text comes from and 300 kroner in another treatment. Subjects were randomly assigned to one of these two recruitment treatments. The stochastic earnings referred to in the recruitment letter were for a risk aversion task and separate tasks eliciting time preferences.[12] Thus, the subjects should have anticipated the use of randomization in the experiment.

The experiments were conducted in hotel meeting rooms around Denmark, so that travel logistics for the invited sample would be minimized. The average home-to-hotel distance was slightly larger for the 1,583 nonparticipants than the 413 participants (10.2 miles versus 8.1 miles), suggesting that distance might have had some influence on their participation decisions.[13] Various times of day were also offered to subjects to facilitate a broad mix of attendance. The largest session had fifteenth subjects, but most had fewer. The procedures were standard. Appendix A documents an English translation of the instructions and shows a typical screen display for the risk aversion task. Subjects were given written instructions that were read out and then made choices in a trainer task for small, nonmonetary rewards. The trainer task was played out and illustrated the procedures in the experiment. All decisions were made on computers. After all choices had been made, the subject was asked a series of standard sociodemographic questions.

There were 40 risk attitude choices and 40 discounting choices, and each subject had a 10% chance of being paid for one choice in each block of 40 choices.[14] The risk attitude choices preceded the discounting choices in one treatment and vice versa in another treatment. Average payments for the risk attitude choices were 242 kroner, and average payments for the discounting choices were 201 kroner (although some were for deferred receipt), for a combined average of 443 kroner. The exchange rate at the time was close to 5 kroner per U.S. dollar, so expected earnings from these tasks combined were $91. The subjects were also paid a 300 kroner or 500 kroner fixed show-up fee, plus earnings from subsequent tasks.[15]

Between April 2010 and October 2010, we repeated the risk aversion and discounting tasks with 182 of the 413 subjects who participated in the first experiment.[16] Each subject was interviewed in private in the new experiment, and the meeting was conducted at a convenient location for them (e.g., their private residence or the hotel where the first experiment took place). All subjects were paid a fixed fee of 300 kroner for their participation in the second experiment.[17]

Table 1 provides the sample response in each panel wave and definitions of the explanatory variables used in the statistical analysis and summary statistics. We observe a significant

TABLE 1.—SAMPLE SIZES AND DESCRIPTIVE STATISTICS

A. Sample Sizes

Recruitment	Variable	Wave 1	Wave 2	All
High fixed fee	Invited	865	184	1,049
	Accepted	208	89	297
	Percent accept	24.1%	48.4%	28.3%
Low fixed fee	Invited	1,131	170	1,301
	Accepted	205	93	298
	Percent accept	18.1%	54.7%	22.9%

B. Descriptive Statistics for Participants

Variable	Definition	Mean Wave 1	Mean Wave 2
female	Female	0.48	0.45
young	Aged less than 30	0.16	0.13
middle	Aged between 40 and 50	0.23	0.21
old	Aged over 50	0.49	0.53
IncLow	Lower level income	0.22	0.23
IncHigh	Higher level income	0.47	0.45
Number of subjects		413	182

Most variables have self-evident definitions. The omitted age group is 30 to 39. Lower income is defined in the variable "IncLow" by a household income in 2008 below 300,000 kroner. Higher incomes is defined in the variable "IncHigh" by a household income of 500,000 kroner or more.

[12] Results from the discounting task are reported in Andersen et al. (2013, 2014), and results from the correlation aversion task are reported in Andersen et al. (2018).

[13] The 2.1-mile difference, albeit small, is statistically significant with a two-sided p-value <0.001. To derive distances, we downloaded geographical coordinates of relevant locations from Google Maps and applied software due to Picard (2010) that measures the length of the shortest curve between two locations over an estimated surface of the Earth.

[14] The number of subjects in each session varied between 3 and 15, which is independent of the 10% probability of being paid for one of the 40 risk attitude choices. Harrison, Lau, and Williams (2002) randomly selected one subject in each session of their Danish field experiment to pay out their discounting choices, and find a small, positive, but statistically insignificant effect of group size on elicited discount rates.

[15] An extra show-up fee of 200 kroner was paid to 24 subjects who had received invitations stating 300 kroner, but then received a final reminder that accidentally stated 500 kroner. The additional tasks earned subjects an average of 659 kroner, so total earnings from choices made in the session averaged 1,102 kroner, or roughly $221, in addition to the fixed fee of $60 or $100. These 24 subjects were treated in the analysis as if they were 300 kroner subjects, since that was the incentive in the original invitation. Treating them as 500 kroner subjects does not change the results.

[16] There were four steps in the construction of this subsample. First, we divided the country into five regions, and each region was divided into subregions. Each subregion was assigned 1 or 2 numbers, in rough proportionality to the population of the subregion. In total, we assigned 24 numbers. Second, although Denmark is a relatively small country, it was necessary to consider logistical constraints, and we randomly picked 12 of the 24 numbers for the experiment in April 2010 and the remaining 12 numbers for the experiment in October 2010. Third, we picked the first 50% of the randomly sorted records within each subregion. This provided a subsample of 100 subjects for each experiment. Fourth, we contacted subjects by phone and invited them to participate again in the experiments.

[17] We did not vary the recruitment fee in the second experiment because we offered to interview the subjects at home or the hotel where the first experiment was conducted. The experiments were time-consuming and expensive to conduct, and we paid subjects the low recruitment fee of 300 kroner in the second experiments to keep costs down. We certainly see value from varying recruitment fees in the second stage as well.

difference in sample response with the high recruitment fee compared to the low recruitment fee. The drop from 24.1% to 18.1% in the first wave is statistically significant according to Fisher's exact test, with a p-value less than 0.001. After participating in the first wave, the sample response to recruitment into the second wave was slightly lower for those recruited into the first wave with the high recruitment fee compared to those recruited with the low fee. The sample response rates were 48.4% and 54.7% in the second wave, and are not statistically different according to Fisher's exact test with a two-sided p-value of 0.24. One might infer from these statistics that the effects of attrition on elicited risk attitudes are not significant, but of course that depends on who responded, which can only be assessed with an appropriate statistical model.

B. *Experiments to Infer Risk Attitudes*

Risk attitudes were evaluated from data in which subjects made a series of binary lottery choices. For example, lottery A might give the individual a 50-50 chance of receiving 1,600 kroner or 2,000 kroner to be paid today, and lottery B might have a 50-50 chance of receiving 3,850 kroner or 100 kroner today. The subject picks A or B. We used the procedures of Hey and Orme (1994) and presented each binary choice to the subject as a pie chart showing prizes and probabilities.[18] We gave each subject the same set of 40 choices, in four sets of ten choices with the same prizes. The prize sets employed are [A1: 2,000 and 1,600; B1: 3,850 and 100], [A2: 1,125 and 750; B2: 2,000 and 250], [A3: 1,000 and 875; B3: 2,000 and 75], and [A4: 2,250 and 1,000; B4: 4,500 and 50]. The order of these four prize sets was randomized for each subject, with the probabilities varying within each set of ten choices.[19] We refer to the first and last of these four prize sets as the high-stakes lotteries compared to the low-stakes lotteries in the second and third sets. These four treatments with different prize sets were administered within subjects.

We asked each subject to respond to all 40 risk aversion tasks and then randomly decided which one to play out using numbered dice. The large incentives and budget constraints precluded us from paying all subjects, so each subject was given a 10% chance to actually receive the payment associated with his decision. The typical findings from lottery choice experiments of this kind are that individuals are generally averse to risk, and that there is considerable heterogeneity in risk attitudes across subjects (see Harrison & Rutström, 2008, for an extensive review).

[18]The use of pie charts is common in experimental elicitation of risk preferences but should not be viewed as the only way that one might present lottery choices. Arguably, probabilities appear more salient than prizes in a pie chart, since probabilities are displayed both graphically (as pie slices) and numerically, whereas prizes are only displayed numerically. Harrison and Rutström (2008; appendix A) review alternative ways of presenting lotteries in the literature, none of which has emerged as obviously superior for all purposes.

[19]Within each prize set, the ten choices were presented one at a time in an ordered manner, with the probability of the high prize starting at 0.1 and increasing by 0.1 until the last choice is between two certain amounts of money.

III. Identification of Risk Preferences

We first write out a structural model to estimate risk attitudes assuming EUT, to focus on essentials. We then discuss how the likelihood function changes to account for sample selection and attrition, and then finally discuss the extension from EUT to the more general RDU model.

A. *Baseline EUT Specification*

Consider the estimation of risk preferences in the simplest possible model of decision-making under risk, EUT, without worrying about sample selection or attrition. In our experiment, each decision task presented a choice between two lotteries, and each lottery had two potential outcomes. Let M_{ij} be the jth outcome of lottery i, where $i =$ A,B and $j = 1,2$. Assume that the utility of an outcome is given by the constant relative risk aversion (CRRA) specification,

$$U(M_{ij}) = M_{ij}^{(1-r)}/(1-r), \tag{1}$$

for $r \neq 1$, where r is the CRRA coefficient. Then, under EUT, $r = 0$ denotes risk neutral behavior, $r > 0$ denotes risk aversion, and $r < 0$ denotes risk loving behavior.

EUT predicts that the observed choice is lottery B when it gives the larger expected utility (EU) than lottery A and vice versa. Probabilities for each outcome, $p(M_{ij})$, are those that are induced by the experimenter, so the EU of lottery i is simply the probability-weighted average of its outcome utilities,

$$EU_i = p(M_{i1}) \times U(M_{i1}) + p(M_{i2}) \times U(M_{i2}), \tag{2}$$

where $p(M_{i2}) = 1 - p(M_{i1})$. Let y denote a binary indicator of whether the observed choice is lottery B ($y = 1$) or lottery A ($y = 0$). Using the indicator function $\mathbf{I}[.]$, the observed choice under EUT can be compactly written as $y = \mathbf{I}[(EU_B - EU_A) > 0]$.

To allow observed choices to deviate from deterministic theoretical predictions, the EUT model is combined with a stochastic behavioral error term. Specifically, assume that the choice depends not only on the EU difference, but also on a random error term ε such that $y = \mathbf{I}[(EU_B - EU_A) + \upsilon \times \varepsilon > 0]$, or equivalently $y = \mathbf{I}[(EU_B - EU_A)/\upsilon + \varepsilon > 0]$, where υ is a positive scale factor that we will parameterize shortly. Assume further that ε is normally distributed with the standard deviation of μ, $\varepsilon \sim N(0, \mu^2)$. The choice probability of lottery B is then $\Phi(\nabla EU)$ where $\Phi(.)$ is the standard normal cumulative density function (CDF), and the index ∇EU is given by

$$\nabla EU = [EU_B - EU_A)/\upsilon]/\mu. \tag{3}$$

It follows that the likelihood function for each choice observation takes the form

$$P(r, \mu) = \Phi(\nabla EU)^y \times (1 - \Phi(\nabla EU))^{(1-y)}. \tag{4}$$

As the noise parameter μ approaches 0, this stochastic EUT specification collapses to the deterministic EUT model; conversely, as μ gets arbitrarily large, it converges to an uninformative model that predicts a 50-50 chance regardless of the underlying EU difference.

We complete the behavioral error specification by adopting the contextual utility model of Wilcox (2011): υ is set to $(U_{\max} - U_{\min})$, where $U_{\max}$ and $U_{\min}$ are the maximum and minimum of the four potential outcome utilities, $U(M_{A1})$, $U(M_{A2})$, $U(M_{B1})$, and $U(M_{B2})$. Supposing that lottery B is riskier than lottery A, it is arguably desirable to have a statistical model that predicts a smaller probability of choosing B for a more risk averse person with a larger r. The traditional Fechner error model ($\upsilon = 1$) leads to choice probabilities that do not vary monotonically with r in this manner, an issue identified by Wilcox (2011) and reiterated by Apesteguia and Ballester (2018).[20] The contextual utility model addresses this potential drawback.

To clarify our econometric methods, more notation is needed than one would typically see in the context of nonlinear models for panel data. We subscript the choice-level likelihood function in equation (4) as $P_{ntw}(r_{nw}, \mu)$, henceforth, to emphasize that it describes subject n's choice in decision task t of panel wave w.[21] The CRRA coefficient r_{nw} is indexed by subject n and wave w for two reasons. First, to capture unobserved preference heterogeneity across individuals, we model the CRRA coefficient as an individual-specific random coefficient drawn from a population distribution of risk preferences. Second, to test temporal stability, we allow the underlying population distribution, as well as the CRRA coefficient drawn from it, to vary freely across waves. We use $f(r_{n1}, r_{n2}; \theta)$ to denote the joint density function for the random CRRA coefficients, where θ is a set of parameters that characterize their joint distribution.

It is possible to estimate the set of parameters θ directly and draw inferences about the population distribution of risk preferences once the joint density $f(r_{n1}, r_{n2}; \theta)$ is fully specified. Assume that r_{n1} and r_{n2} are jointly normal so that $\theta = (\bar{r}_1, \bar{r}_2, \sigma_{r1}, \sigma_{r2}, \sigma_{r1r2})$, where $\bar{r}_w$ and σ_{rw} are the population mean and standard deviation of the CRRA coefficient r_{nw}, and σ_{r1r2} is the covariance between r_{n1} and r_{n2}. Conditional on a particular pair of CRRA coefficient draws, the likelihood of observing a series of 40 or 80 choices made by subject n can be specified as

$$\begin{aligned} &CL_n(r_{n1}, r_{n2}, \mu) \\ &\quad = \Pi_t P_{nt1}(r_{n1}, \mu) && \text{if } s_{n2} = 0 \\ &\quad = \Pi_t P_{nt1}(r_{n1}, \mu) \times \Pi_t P_{nt2}(r_{n2}, \mu) && \text{if } s_{n2} = 1 \end{aligned} \tag{5}$$

where s_{n2} is an indicator of whether subject n participated in only the first panel wave ($s_{n2} = 0$) or both panel waves ($s_{n2} = 1$). Since r_{n1} and r_{n2} are modeled as random coefficients, the "unconditional" (Train, 2009) or actual likelihood of subject n's choices is then obtained by taking the expected value of $CL_n(r_{n1}, r_{n2}, \mu)$ over the joint density $f(r_{n1}, r_{n2}; \theta)$

$$\begin{aligned} &L_n(\bar{r}_1, \bar{r}_2, \sigma_{r1}, \sigma_{r2,} \sigma_{r1r2}, \mu) = L_n(\theta, \mu) \\ &\quad = \iint CL_n(r_{n1}, r_{n2}, \mu) f(r_{n1}, r_{n2}; \theta) dr_{n1} dr_{n2}. \end{aligned} \tag{6}$$

Unobserved heterogeneity is similarly integrated out from many textbook models for panel data, such as random-effects probit (Wooldridge, 2010).[22] Our application is distinctive because unobserved heterogeneity enters the index function ∇EU_{ntw} nonlinearly via the CRRA coefficient and varies across two wave-specific blocks of observations instead of being time invariant.[23] The unconditional likelihood function $L_n(\theta, \mu)$ does not have a closed-form expression but can be approximated using simulation methods (Train, 2009). We compute maximum simulated likelihood (MSL) estimates of risk preference parameters θ and the behavioral noise parameter μ by maximizing a simulated analogue to the sample log-likelihood function $\Sigma_n \ln(L_n(\theta, \mu))$. The estimation sample is 413 subjects who participated in the first experiment or both experiments.

Our modeling framework offers several ways to define and analyze temporal stability of risk attitudes. One can test if the entire population distribution of risk preferences is stable, which can be expressed as a joint hypothesis H_0: $\bar{r}_1 = \bar{r}_2$ and $\sigma_{r1} = \sigma_{r2}$. Alternatively, one can test the temporal stability of the average person's risk attitude (H_0: $\bar{r}_1 = \bar{r}_2$), or test the temporal stability of unobserved preference heterogeneity (H_0: $\sigma_{r1} = \sigma_{r2}$). We can also accommodate observed heterogeneity by writing $\bar{r}_1$ and $\bar{r}_2$ as linear functions of the subject's characteristics, such as age, gender, and income.[24] It is then possible to consider the question of which

[20] In appendix F, we reestimate our main models assuming the Fechner error specification.

[21] We repeated the same set of experiments across two panel waves, and within each wave, the subject completed a series of decision tasks over 40 lottery pairs. The outcomes and probabilities associated with lottery pairs vary from task to task, and the same subject may make different choices across tasks and waves. Each lottery outcome and its probability are then M_{ijntw} and $p(M_{ijntw})$, leading to the expected utilities EU_{intw} and the index function ∇EU_{ntw}. The indicator y_{ntw} is 1 (0) if subject n chooses lottery B (lottery A) in decision task t of the experiment in wave w.

[22] Much as one finds with a random-effect probit model, our random-coefficient model allows for panel correlation across repeated observations on the same individual. Although equation (5) is a product formula akin to the pooled probit model, it is only one building block for the actual likelihood function in equation (6) that integrates such formulas. The log of this likelihood function does not simplify into a sum of observation-level log-likelihood functions, so our statistical approach does not rely on the independence of choice observations within individuals.

[23] Methods for estimating nonlinear random coefficients models of risk aversion were developed by Andersen et al. (2012).

[24] For illustration, we analyze a model of male-female differences in risk attitudes in appendix E.

demographic groups tend to be more risk averse and examine if the answer to that question is temporally stable.

The questions so far pertain to temporal stability at the population level, but the analysis can focus on temporal stability at the individual level as well. By normalizing the scale of covariance σ_{r1r2}, one can derive a coefficient $\varrho_{r1r2} = \sigma_{r1r2}/(\sigma_{r1} \times \sigma_{r2})$ that directly measures the within-individual correlation of the CRRA coefficient over time. Andersen et al. (2008b) elicit risk preferences using multiple price list formats and compute this type of correlation based on the midpoints of CRRA intervals that predict observed behavior under EUT. The approach we take here is far more general because it allows for behavioral errors and can be applied with any elicitation format as long as the statistical model incorporates a random coefficient specification similar to ours. Moreover, as reported below, one can estimate the within-individual correlations of structural parameters in an analogous manner after correcting for selection and attrition biases, as well as in the context of RDU models.

B. *EUT Specification with Endogenous Sample Selection and Panel Attrition*

The experimental design allows us to correct for sample selection into both panel waves of the experiment.[25] Estimates of risk aversion could be sensitive to the sample selection and attrition process in any longitudinal setting, and the estimated coefficients in the behavioral model may be significantly biased if subjects condition their participation on unobservable characteristics that correlate with their latent risk preferences. It is not obvious a priori that individuals with stable preferences are more likely to self-select into the early or later stages of our experiment. Since the decision to participate in the experiment may be correlated with individual risk preferences, it is appropriate to account for possible sample selection and attrition effects in the statistical model.

To control for sample selection bias, we take the initial pool of 1,996 invited subjects as a random sample from the population and model the initial selection process that led to 413 subjects in the first experiment. From this sample, 354 subjects were invited to the second experiment. To control for panel attrition bias, we take those 354 subjects as a random sample from the subpopulation who self-selected into the first experiment and model the attrition process that led to 182 subjects in the second experiment. This general strategy is consistent with our experimental design, under which the experimenter exogenously determines whether someone is invited to the first experiment and which subjects in the first experiment get invited to the second experiment.

We first describe a system of binary response models that describes sample selection and attrition. Let s_{nw} be an indicator of whether subject n accepted the invitation to the experiment in wave w ($s_{nw} = 1$) or not ($s_{nw} = 0$). For those who were not invited to the second experiment, we set $s_{n2} = -1$. Assume that each observed outcome s_{nw} is determined by a latent propensity S_{nw}, such that $s_{n1} = \mathbf{I}[S_{n1} > 0]$, and $s_{n2} = \mathbf{I}[S_{n1} > 0 \cap S_{n2} > 0]$ if subject n was invited to the second experiment. The latent propensities are specified as

$$S_{n1} = X_{n1}\beta_1 + u_{n1} = X_{n1}\beta_1 + (a_{n1} + e_{n1}), \tag{7}$$

$$S_{n2} = X_{n2}\beta_2 + u_{n2} = X_{n2}\beta_2 + (a_{n2} + e_{n2}), \tag{8}$$

where X_{nw} is a vector of explanatory variables including a constant, β_w is a conformable vector of coefficients to estimate, and u_{nw} is a random disturbance. We decompose u_{nw} further into a_{nw} and e_{nw}, which are orthogonal to each other. The term a_{nw} captures unobserved characteristics that are potentially correlated with risk attitudes, and across selection and attrition processes. In contrast, e_{nw} captures purely idiosyncratic errors.

Assume that the correlated components a_{n1} and a_{n2} are bivariate normal and that each idiosyncratic error e_{nw} is independently normal. Under this assumption, the composite errors u_{n1} and u_{n2} are also bivariate normal. When viewed in isolation from the random coefficient EUT model, the system of equations (7) and (8) is analogous to the probit model with sample selection (Van de Ven & Van Praag, 1981), which views the sample retention indicator s_{n2} as the primary outcome of interest. It is common to normalize this type of model by setting $\mathrm{Var}(u_{n1}) = \mathrm{Var}(u_{n2}) = 1$, and identify β_1, β_2, and $\varrho_{s1s2} = \mathrm{Corr}(u_{n1}, u_{n2}) = \mathrm{Cov}(a_{n1}, a_{n2})$. We could follow the same convention but prefer to normalize the system by setting $\mathrm{Var}(u_{n1}) = 2$ and $\mathrm{Var}(u_{n2}) = 2 + \mathrm{Cov}(a_{n1}, a_{n2})$, and identify β_1, β_2, and $\sigma_{s1s2} = \mathrm{Cov}(u_{n1}, u_{n2}) = \mathrm{Cov}(a_{n1}, a_{n2})$. This scheme allows us to assume $\mathrm{Var}(a_{n1}) = \mathrm{Var}(e_{n1}) = \mathrm{Var}(e_{n2}) = 1$ and $\mathrm{Var}(a_{n2}) = 1 + \sigma_{s1s2}$ without loss of generality; then equations (7) and (8) can more easily be combined with the random coefficient EUT model by attaching probit probabilities to equation (5), as shown below.

Let $g(a_{n1}, a_{n2}, r_{n1}, r_{n2}; \Theta)$ denote a density function for the joint distribution of risk attitudes and relevant selection/attrition errors, which is characterized by parameters in Θ. Let σ_{s1rw} and σ_{s2rw} denote $\mathrm{Cov}(a_{n1}, r_{nw})$ and $\mathrm{Cov}(a_{n2}, r_{nw})$, respectively. We allow for the full set of correlations among the four random components. Given the earlier assumptions, $g(.; \Theta)$ is then multivariate normal and $\Theta = (\theta, \Sigma)$, where $\theta = (\bar{r}_1, \bar{r}_2, \sigma_{r1}, \sigma_{r1}, \sigma_{r1r2})$ characterizes the population distribution of the CRRA coefficients and $\Sigma = (\sigma_{s1s2}, \sigma_{s1r1}, \sigma_{s1r2}, \sigma_{s2r2}, \sigma_{s2r2})$ collects covariance parameters that may induce selection and attrition biases. For example, a positive σ_{s1r1} means that those with relatively large CRRA coefficients in wave 1 are more likely to participate in the first experiment, and a positive σ_{s2r1} means that such subjects with high CRRA coefficients in wave 1 are also more likely to participate in the second experiment. Without correction for selection and attrition, one would overestimate the initial degree of risk aversion in the population. While σ_{s1s2} does not address risk attitudes directly, this parameter

[25] Vella (1998) surveys alternative specifications for modeling sample selection, including semiparametric methods.

corrects the attrition process for initial selection bias since the attrition outcomes are observed only for the self-selected sample of participants in the first experiment. If σ_{s1s2} is falsely constrained to 0, the resulting correction for attrition bias becomes invalid.

We now turn to a likelihood function that augments the baseline EUT specification with controls for selection and attrition biases. Conditional on a particular set of a_{n1}, a_{n2}, r_{n1}, and r_{n2}, the joint likelihood of subject n's selection/attrition outcomes and risky choices can be specified as

$$\begin{aligned} &CL_n(a_{n1}, a_{n2}, r_{n1}, r_{n2}, \beta_1, \beta_2, \mu) \\ &\quad = 1 - \Phi(\tau_{n1}) && \text{if } s_{n1} = 0 \\ &\quad = \Phi(\tau_{n1}) \times \Pi_t P_{nt1}(r_{n1}, \mu) && \text{if } s_{n1} = 1, s_{n2} = -1 \\ &\quad = \Phi(\tau_{n1}) \times (1 - \Phi(\tau_{n2})) && \text{if } s_{n1} = 1, s_{n2} = 0 \\ &\qquad \times \Pi_t P_{nt1}(r_{n1}, \mu) \\ &\quad = \Phi(\tau_{n1}) \times \Phi(\tau_{n2}) && \text{if } s_{n1} = 1, s_{n2} = 1 \\ &\qquad \times \Pi_t P_{nt1}(r_{n1}, \mu) \\ &\qquad \times \Pi_t P_{nt2}(r_{n2}, \mu) \end{aligned} \tag{9}$$

where $\tau_{nw} = X_{nw}\beta_w + a_{nw}$, $\Phi(.)$ is the standard normal CDF, and $P_{ntw}(.)$ is the choice-level likelihood under the baseline EUT model. The exact form of the conditional likelihood function thus varies for those who rejected the first invitation ($s_{n1} = 0$), those who participated in the first experiment but did not receive the second invitation ($s_{n1} = 1, s_{n2} = -1$), those who participated in the first experiment but rejected the second invitation ($s_{n1} = 1, s_{n2} = 0$), and finally those who participated in both experiments ($s_{n1} = s_{n2} = 1$). The unconditional likelihood function for subject n can be obtained by taking the expected value of $CL_n(a_{n1}, a_{n2}, r_{n1}, r_{n2}, \beta_1, \beta_2, \mu)$ over the joint distribution of the four random components

$$\begin{aligned} &L_n(\Theta, \beta_1, \beta_2, \mu) \\ &\quad = \iiiint CL_n(a_{n1}, a_{n2}, r_{n1}, r_{n2}, \beta_1, \beta_2, \mu) \\ &\qquad \times g(a_{n1}, a_{n2}, r_{n1}, r_{n2}; \Theta) da_{n1} da_{n2} dr_{n1} dr_{n2}, \end{aligned} \tag{10}$$

where $\Theta = (\bar{r}_1, \bar{r}_2, \sigma_{r1}, \sigma_{r2}, \sigma_{r1r2}, \sigma_{s1s2}, \sigma_{s1r1}, \sigma_{s1r2}, \sigma_{s2r1}, \sigma_{s2r2})$ in full. Since equation (10) does not have a closed-form expression, we compute the MSL estimates of Θ, β_1, β_2, and μ by maximizing a simulated analogue to the sample log-likelihood function $\Sigma_n \ln(L_n(\Theta, \beta_1, \beta_2, \mu))$. The estimation sample is all 1,996 subjects who were invited to the first experiment.

Parametric models with selection and attrition such as ours are theoretically identified without the aid of cross-equation exclusion restrictions. Nevertheless, our experimental design provides natural candidates for such restrictions that we use to assist empirical identification. The initial invitation letter randomized subjects to different recruitment fees, and the longitudinal design allows us to observe each subject's additional earnings from the first experiment.[26] Before coming to the first experiment, subjects did not know anything about the 40 lottery pairs used and, during the first experiment, everyone faced the same set of 40 lottery pairs. We assume that the recruitment fees affect the initial decision to accept the first invitation but do not affect the decision to accept the second invitation once we control for additional earnings from the first experiment.[27] We maintain the usual hypothesis that the recruitment fees and prior earnings do not affect the subject's evaluation of lottery pairs directly. Finally, subjects had to travel to hotel meeting rooms to participate in the first experiment, whereas each subject chose his or her own preferred venue for the second experiment.

The preceding discussion motivates us to include the recruitment fees only in X_{n1} for the selection equation, the actual earnings from the first experiment only in X_{n2} for the attrition equation, and the lottery payoffs and probabilities only in ∇EU_{ntw} for the structural model of risky choices. In addition, we augment X_{n1} with each subject's home-to-hotel distance (in miles) and its square.[28] Both X_{n1} and X_{n2} also include the subject's age and gender, and X_{n2} additionally includes self-reported income that is available only for those who participated in the first experiment.

To see the flexibility of our extended specification, one may compare it with several special cases. Consider first a naive approach, in which each panel wave is evaluated separately, using equation (7) to correct for selection into the first wave and equation (8) to correct for selection into the second wave. This approach is naive in the sense that it fails to recognize the longitudinal nature of the experiments and requires $\sigma_{s1s2} = \sigma_{s1r2} = \sigma_{s2r1} = 0$. However, even when these restrictions are valid, the approach cannot identify σ_{r1r2}, and hence ϱ_{r1r2}, that measures the temporal stability of risk preferences within individuals. Two special cases arise if both waves are analyzed jointly, but they correct for only selection bias or attrition bias. With correction for selection bias only, one can estimate all structural parameters consistently when $\sigma_{s2r1} = \sigma_{s2r2} = 0$. The other special case ignores selection bias and requires $\sigma_{s1s2} = \sigma_{s1r1} = \sigma_{s1r2} = 0$. The latter case is perhaps more interesting, considering that it resembles what

[26]Since the recruitment fee is an observed characteristic and the model is theoretically identified without using this as an exclusion restriction, it is possible to test whether the use of different recruitment fees results in recruitment of subjects with systematically different risk attitudes. For instance, as shown in tables C5 and C6 of appendix C, we can condition the mean of each structural parameter (r_{n1} and r_{n2} under EUT, and r_{n1}, r_{n2}, φ_{n1}, and φ_{n2} under RDU, which we will describe shortly) on the recruitment fee indicator and study whether the estimated coefficient on that indicator is significant. The results support our intended use of the recruitment fee as an exclusion restriction to assist empirical identification. The recruitment fee has an insignificant effect on the mean of r_{n1} and r_{n2} under EUT with p-values of 0.173 and 0.447, and under RDU with p-values of 0.191 and 0.246. Similarly the recruitment fee has an insignificant effect on the mean of φ_{n1}, and φ_{n2} under RDU, with p-values of 0.997 and 0.295.

[27]Additional earnings in the first experiment include payments for choices in three sets of decision tasks that elicit individual risk attitudes, discount rates, and correlation aversion, respectively.

[28]How closely the home-to-hotel distance approximates the actual inconvenience involved in traveling is an open question. The validity of our statistical corrections for endogenous selection and attrition does not rely on any precise interpretation that one might place on the distance variable. As usual, the selection equation in our framework is a reduced-form index model, and its coefficients need not have any causal interpretation.

one would do in typical longitudinal studies that observe no information on those who did not participate in the first wave.

Our modeling strategy provides a general framework for the structural estimation of risk preferences with correction for endogenous selection and attrition. While we parameterize the statistical model using multivariate normal densities and probit kernels, with a few notational changes, the likelihood functions above can incorporate other joint distributions of $\{a_{n1}, a_{n2}, r_{n1}, r_{n2}\}$ and kernel CDFs. We focus on the multivariate normal-probit kernel specification primarily to reach a wider audience; the workhorse sample selection models in the empirical literature assume either the bivariate normality of selection and structural errors in a maximum likelihood framework or the marginal normality of selection errors in Heckman's two-step procedure. In many longitudinal studies, the researcher may apply correction for panel attrition but not for initial selection due to the lack of information on nonparticipants. Our econometric approach can be adapted to such settings to specify a structural model with endogenous attrition by omitting the selection equation and renormalizing the standard deviation of the attrition error.[29] As usual, the resulting correction for attrition bias would be a second-best solution that presumes the absence of selection bias.

C. Rank Dependent Utility Theory Specifications

RDU is a popular generalization of EUT, due to Quiggin (1982), that allows the decision maker to transform the objective probabilities presented in lotteries and use these weighted probabilities to determine decision weights when evaluating lotteries. If $w(p)$ is the probability weighting function assumed and each lottery has only two prizes such that $M_{i1} > M_{i2}$, then we have

$$RDEU_i = [w(p(M_{i1})) \times U(M_{i1})] + [(1 - w(p(M_{i1}))) \times U(M_{i2})], \tag{2'}$$

where $RDEU_i$ refers to rank dependent expected utility of lottery i and the remaining notation is as defined in the context of equation (2).

The logic behind our econometric specifications extends naturally to RDU once we replace EU_i with $RDEU_i$. Of course, one has to specify the functional form for $w(p)$ and estimate additional parameters. Prelec (1998) offers a two-parameter probability weighting function that exhibits considerable flexibility. This function is

$$w(p) = \exp\{-\eta(-\ln p)^{\varphi}\}, \tag{12}$$

and is defined for $0 < p < 1$, $\eta > 0$, and $\varphi > 0$. We use its one-parameter special case that assumes $\eta = 1$ and model φ as a log-normally distributed random coefficient φ_{nw} that varies across individuals and panel waves. The resulting one-parameter function exhibits inverse-S probability weighting (optimism for small p, and pessimism for large p) for $\varphi < 1$, S-shaped probability weighting (pessimism for small p, and optimism for large p) for $\varphi > 1$, and linear probability weighting that reduces RDU to EUT when $\varphi = 1$. It rules out the cases of globally concave (optimism for all p) or globally convex (pessimism for all p) probability weighting a priori, and also implies that the fixed point where $w(p) = p$ occurs at $p = 0.368$ for any value of φ. The two-parameter function can admit concave and convex cases, and also inverse-S or S-shaped probability weighting with other fixed points. But allowing for the unrestricted joint distribution of random coefficients and selection or attrition errors leads to several extra parameters, making the use of the two-parameter function less practical for our purposes.

One implication of the RDU model is that risk preferences are characterized by more than the concavity of the utility function. The risk premium is a complex function of all of the parameters that define the utility function, as well as the probability weighting function. Indeed, a concave utility function might be mitigated by probability "optimism" such that the net effect is risk neutrality or even risk loving. We simply have to examine all parameters to characterize risk preferences in the case of RDU: r and φ.[30]

IV. Results

We are interested in testing several hypotheses. First, is the distribution of risk attitudes in the general adult Danish population temporally stable over the one-year period we consider in the experiment? Second, are risk attitudes temporally stable at the individual level? Third, does the possibility of nonrandom sample selection and attrition change our inferences about the temporal stability of risk attitudes?

We use MSL to estimate the full statistical model that captures unobserved preference heterogeneity, endogenous selection into the first experiment, and endogenous panel attrition between the two experiments. Train (2009) provides details on MSL estimation of heterogeneous preference models without selection. Cappellari and Jenkins (2004) show how one can control for endogenous selection and attrition using MSL in the context of models without unobserved preference heterogeneity. By modeling the joint likelihood of observing the entire series of responses by each subject and adjusting standard errors for clustering at the subject level,

[29]The conditional likelihood function under this endogenous attrition model is algebraically equivalent to the special case of equation (9) that assumes $s_{n1} = 1$ and $\Phi(\tau_{n1}) = 1$ for every n. Since the covariance between the selection and attrition errors is no longer identified, the scale of the attrition error should be renormalized—for example, by setting $\mathrm{Var}(u_{n2}) = 2$.

[30]The EUT model retains some descriptive value, however. The EUT and RDU models assume the same overall risk premium, even if they explain it differently. It is sometimes useful to focus on the parameter r in the EUT model as a summary statistic on the overall risk premium, even if the RDU model may provide the correct structural decomposition into aversion to outcome variability (the r parameter) and probability weighting (the φ parameter).

our statistical specification allows for clustered responses by the same subject. Panel-robust Wald statistics are used to test various hypotheses with respect to the estimated coefficients. The statistical model also allows for heteroskedasticity in the behavioral error term by conditioning the noise parameter on binary variables for each treatment in the experimental design; one variable captures the order of risk aversion and discounting tasks, and the other variable captures our use of high and low stakes in the risk aversion tasks. We also condition the population mean coefficients of latent risk preference parameters on these two treatment variables.

We transform several estimates into alternative forms that are easier to interpret and report correlation coefficients instead of covariance parameters. For the log-normal random coefficient φ in the RDU model, all results are for φ itself instead of $\ln(\varphi)$.[31] Finally, we divide selection and attrition equation coefficients by the normalized standard deviation of each equation so that they can be interpreted in the same manner as familiar probit coefficients.

A. *Temporal Stability of Risk Attitudes*

We find evidence of temporal stability for inferred risk attitudes under EUT when the model fully corrects for endogenous sample selection and attrition bias. Table C1 of appendix C contains detailed estimates. Single-hypothesis tests show that the mean CRRA parameter $\bar{r}_w$ for each treatment group is the same over time. For example, the estimated mean coefficient of relative risk aversion for the baseline case of our econometric model (when $RAfirst = RAhigh = 0$) is equal to 0.413 in wave 1 and equal to 0.594 in wave 2; the estimated difference in the two mean population coefficients is equal to 0.180, which is not significantly different from 0 with a p-value of 0.236.[32] The estimated population mean coefficient is also larger in wave 2 relative to wave 1 when we control for the high-stakes treatment; the estimated difference between the two coefficients is 0.151, which is insignificant with a p-value of 0.294. We also find that the estimated population standard deviation of relative risk aversion is temporally stable; the estimated standard deviation of the r parameter, σ_r, drops from 0.856 in wave 1 to 0.787 in wave 2, and the estimated difference between the two coefficients is not significantly different (p-value of 0.637). A joint test of estimated mean population coefficients and standard deviation coefficients across the two waves allows us to evaluate whether the entire population distribution is temporally stable. The $\chi^2(4)$ test statistic has a p-value of 0.480, so we cannot reject the hypothesis of temporal stability.[33] Although the estimated population mean is higher in wave 2 compared to wave 1 for low- and high-stakes treatments, we find statistical evidence of temporal stability for the entire population distribution of relative risk aversion.

The upper panel in figure 1 shows the estimated population distributions of relative risk aversion across the two waves and two monetary treatments, with controls for nonrandom selection and attrition bias. The population distributions of relative risk aversion for both monetary treatments shift to the right in wave 2 compared to wave 1, but the apparent increase in risk aversion is not statistically significant, as noted above.[34] The marginal effect of the high-stakes treatment on the estimated population mean is positive and the population distribution shifts to the right in both waves. The estimated coefficient of the high-stakes treatment is equal to 0.088 with a p-value of 0.017 in wave 1 and equal to 0.059 with a p-value of 0.260 in wave 2. We thus observe a significant effect of the high-stakes treatment on relative risk aversion in wave 1 and an insignificant effect in wave 2.

We next consider temporal stability at the individual level. The estimated correlation coefficient between relative risk aversion in wave 1 and 2, ϱ_{r1r2}, is equal to 0.360, which is significantly different from 0 (p-value < 0.001). The significant positive correlation suggests that risk preferences are temporally stable at the individual level, in the sense that someone with an above-average r parameter in wave 1 also tends to have an above-average r parameter in wave 2, and thus we reject the hypothesis that the two population distributions are independent.

Turning to the results for RDU, reported in detail in table C2 of appendix C, we draw mixed conclusions that depend on which aspect of temporal stability that one is interested in. Under RDU, risk preferences are characterized by the r parameter as well as the weighting parameter, φ, which is log-normally distributed. The entire population distribution of risk preferences may be said to be stable when the joint distribution of r and φ is stable. More formally, this joint

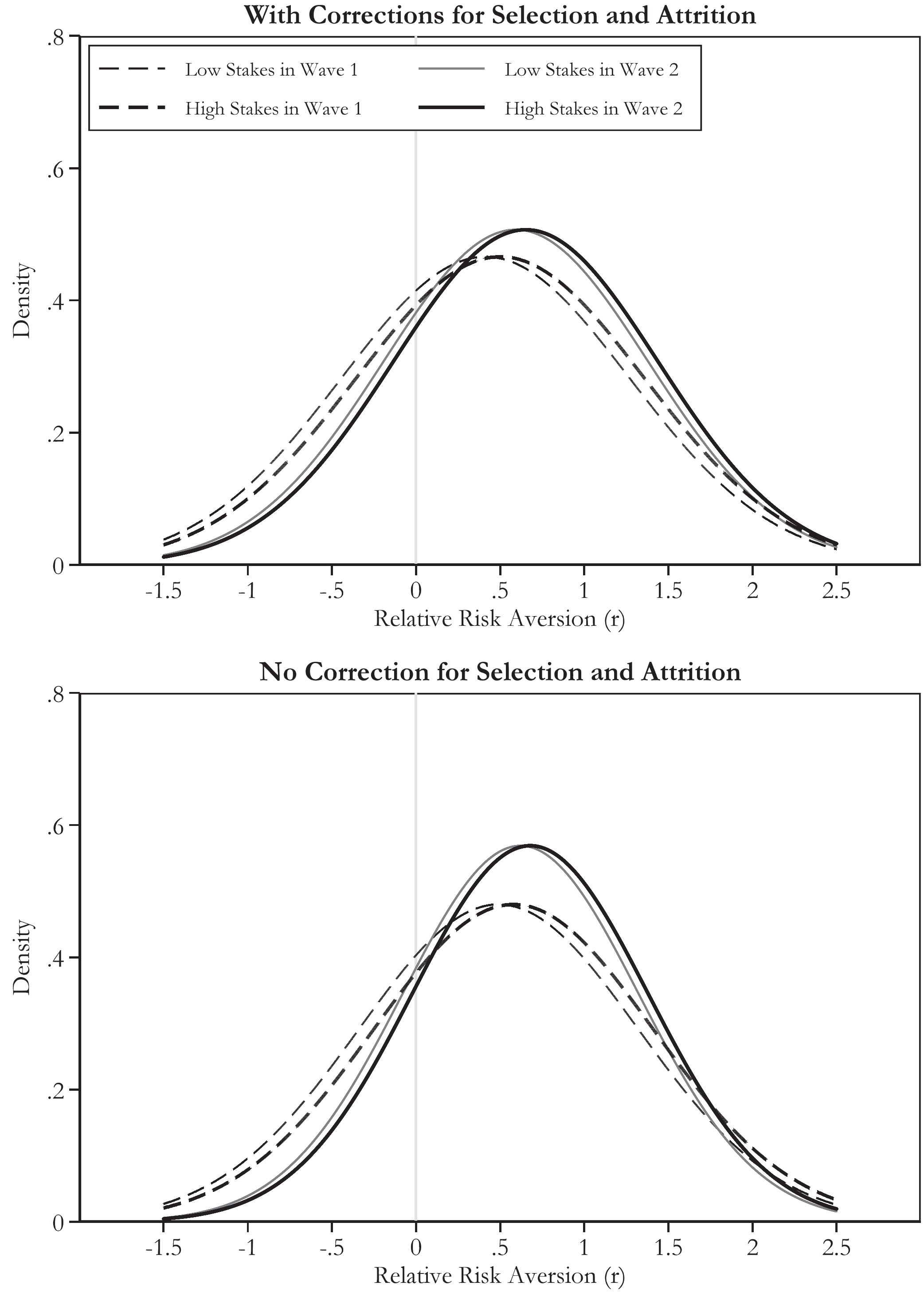
[31] Specifically, we report the mean and median of φ for the base group (constant), along with the marginal effect of each observed characteristic on the mean and median of φ for the base group. The standard deviation of φ is evaluated at the sample average characteristics. The within-individual correlation of φ is computed by applying the usual formula for the correlation coefficient of bivariate log-normal random variables. Other correlations involving φ present cases where we compute the correlation between a log-normal random variable and a normal random variable. Garvey, Book, and Covert (2015, theorem B.1) provide a closed-form formula that can be applied to these cases.

[32] Our risk aversion experiment was part of a larger experiment that involved a discounting choice tasks and correlation aversion tasks. The order of risk aversion and discounting tasks was randomized on a between-subject basis; half of the subjects faced risk aversion tasks first ($RAfirst = 1$) and the remaining half faced discounting tasks first ($RAfirst = 0$). The correlation aversion tasks always followed the risk and discounting tasks. In each wave, each subject completed twenty risk aversion tasks that we classify as low stake ($RAhigh = 0$) and twenty decision tasks that we classify as high stake ($RAhigh = 1$). Our model allows for systematic variations in risk preferences across the order and stake treatments. To avoid potential clutter, our figures focus on comparisons across the stake treatments, since the order treatment effect is not statistically significant at the 5% level in any of our estimation results.

[33] Since the mean of the r parameter has been conditioned on two treatment variables, in each wave there are three estimates associated with the mean (constant, $RAfirst$, $RAhigh$). Temporal stability of the population distribution therefore entails four between-wave equality restrictions, comprising three restrictions on the mean and one restriction on the standard deviation.

[34] Figure 1 is generated from the point estimates of the population mean and population standard deviation of the relative risk aversion parameter. It does not reflect the standard errors around those point estimates or the covariance between them. Our statistical tests do take these into account.

FIGURE 1.—POPULATION DISTRIBUTIONS OF RISK AVERSION UNDER EUT

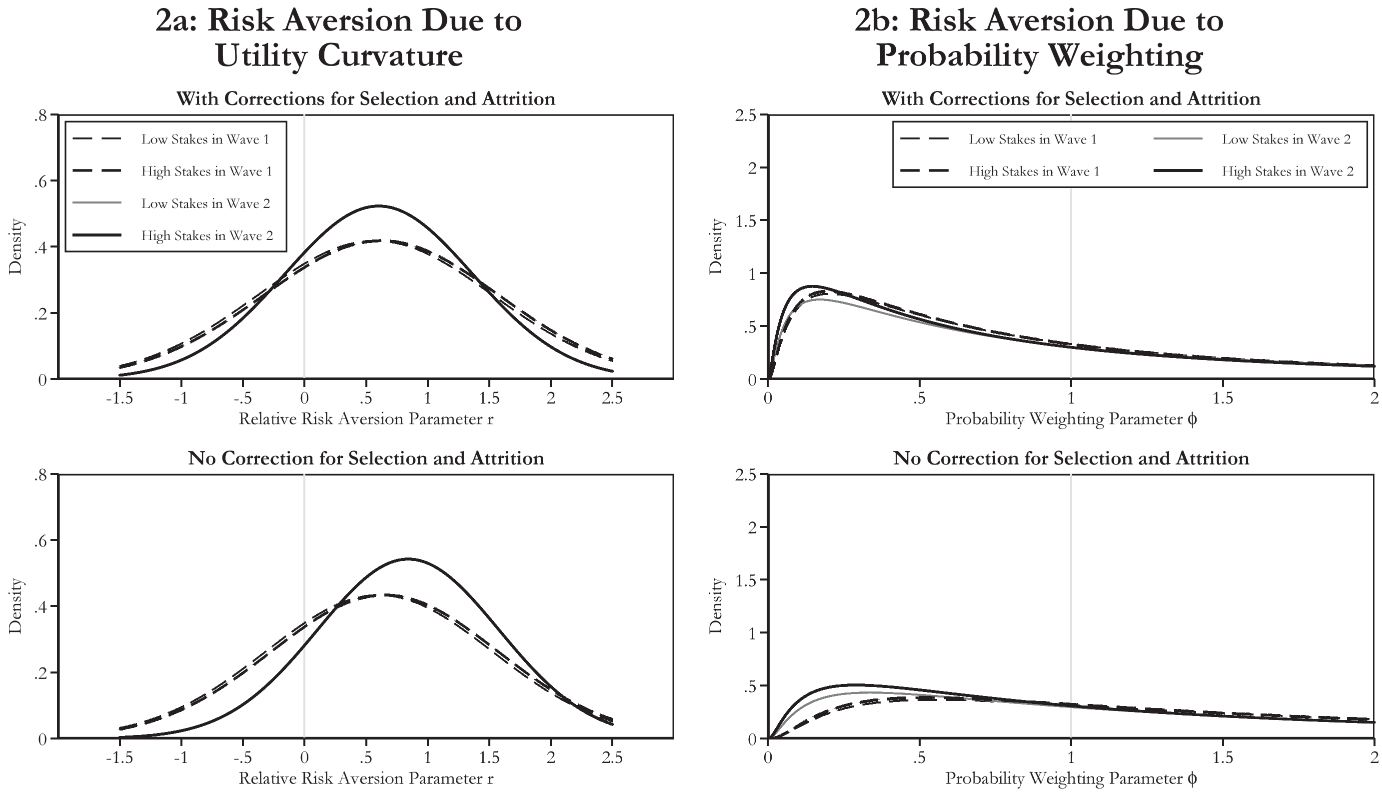


hypothesis requires stability in the estimated population means of the r and φ parameters, the estimated population standard deviations of r and φ, and the estimated correlation between r and φ. We cannot reject this type of temporal stability; the associated $\chi^2(9)$ test statistic has a p-value of 0.303.[35]

Figure 2a displays the estimated population distributions of relative risk aversion for each wave and monetary treatment. The estimated distributions in the upper panel

[35]The stable marginal distribution of the r parameter entails four restrictions. Similarly, the stable marginal distribution of the φ parameter entails another set of four restrictions. In total, temporal stability in the joint distribution of r and φ parameters entails nine between-wave equality restrictions: eight restrictions on the marginal distributions and one restriction on the correlation coefficient between the two parameters.

FIGURE 2.—POPULATION DISTRIBUTIONS OF RISK AVERSION UNDER RDU



control for selection and attrition bias, and we observe that the estimated population means of the *r* parameter are almost identical across the two waves. The estimated between-wave difference in the population mean is 0.031 for the low-stakes treatment and 0.022 for the high-stakes treatment, and neither estimate is statistically significant. We also observe that the population distributions in wave 2 have a smaller standard deviation than the distributions in wave 1; the estimated standard deviation is 0.955 in wave 1 and 0.763 in wave 2, and we reject the null hypothesis that the estimated difference in the two coefficients is equal to 0 at the 5% significance level (*p*-value of 0.042). Hence, we find temporal stability with respect to population mean and temporal instability with respect to the standard deviation of the *r* parameter. The estimated correlation coefficient between the population distributions of the *r* parameter over time, ϱ_{r1r2}, is equal to 0.689, which is somewhat higher than the estimated coefficient under EUT, and we reject the hypothesis that the two population distributions are independent.

The estimated population distributions of the probability weighting parameter φ are displayed in figure 2b. The distributions in the upper panel control for selection and attrition bias, and we observe insignificant differences in the estimated population distributions of the φ parameter between the two waves. We cannot reject the hypothesis that the population distribution of the φ parameter is temporally stable; the $\chi^2(4)$ test statistic has a *p*-value of 0.306. The estimated difference in the population mean between the two waves is statistically insignificant across each monetary treatment, and we also find that the standard deviation of the population distribution is temporally stable. The estimated standard deviation is higher in wave 2 compared to wave 1, but the estimated difference in the standard deviation is statistically insignificant (*p*-value = 0.326). Finally, we find that the estimated between-wave correlation of the φ parameter is 0.662 with a standard error of 0.159, which suggests a strong degree of temporal stability at the individual level.

In summary, we contribute to the literature by modeling risk preferences in a nonlinear, structural manner, allowing for unobserved heterogeneity across the population and endogenous selection and attrition. The use of panel correlations in structural parameters to test individual-level stability is also a unique feature of our analysis. The ability to analyze temporal stability at both the population and individual level in a single econometric model demonstrates the coherence and flexibility of our econometric modeling approach. Appendix D reviews related previous literature.

B. *Effects of Sample Selection and Attrition on Risk Attitudes under EUT*

We observe significant evidence of exogenous and endogenous selection and attrition effects on the estimated coefficients reported in table C1. We find a positive and significant effect of the higher recruitment fee on the propensity to self-select into the first wave of our experiment. In effect, the law of demand applies to participation in the experiments, and response rates increase significantly when the recruitment fee is raised from 300 kroner to 500 kroner for participation in wave 1. We also find a statistically significant and U-shaped association between the self-selection index and the home-to-hotel distance, suggesting a negative and diminishing marginal effect of the distance up to a turning point at 34.22 miles. In other words, as one may expect, people who live farther away from the session venues are less likely to participate, and people who live closer are more sensitive to a small increase in the distance. Of course, the sign of the marginal effect changes after the turning point, but this is more or less an artifact of the quadratic specification that is of limited economic significance, since only six out of the 1996 invitees lived outside a 34.22 mile radius from a venue.[36] Looking at observable characteristics, middle-aged and older subjects were more likely to select into the first wave compared to omitted age group. It is generally difficult to explain panel retention rates in terms of observed characteristics, although the results do suggest that young and high-income subjects are less likely to select into the second wave than otherwise.

Turning to endogenous effects of sample selection and attrition, we find enough statistical evidence to reject the hypotheses of no selection and attrition bias, respectively. The hypothesis of no endogenous sample selection bias is evaluated using the joint test of H_0: $\varrho_{s1s2} = \varrho_{s1r1} = \varrho_{s1r2} = 0$. This hypothesis is rejected, with a p-value less than 0.001. The hypothesis of no endogenous attrition bias can be tested by H_0: $\varrho_{s2r1} = \varrho_{s2r2} = 0$, which again is rejected, with a p-value less than 0.001. The estimated correlation coefficient between the error terms in the selection and attrition equations, ϱ_{s1s2}, is equal to -0.340 with a standard error of 0.125, which means that one cannot take the naive approach of correcting for each source of sampling bias separately.

We can see the overall effects of controlling for selection and attrition bias on the estimated population distributions of relative risk aversion in figure 1. The lower panel shows the estimated distributions with no correction for sample selection and attrition bias. Despite the significant statistical evidence of sample selection and attrition bias, we draw qualitatively similar conclusions about temporal stability. We observe that the population mean increases over time and the population distribution becomes tighter around the mean.[37] Although the estimated population mean is higher in wave 2 compared to wave 1 for both monetary treatments, there is statistically significant evidence of temporal stability with respect to relative risk aversion at the population level. We also find temporal stability at the individual level. The estimated correlation coefficient between relative risk aversion in waves 1 and 2 is equal to 0.537, which is significantly different from 0 (p-value < 0.001).

Correcting for endogenous attrition is often easier than correcting for endogenous selection, since in the case of attrition, one potentially knows a lot about the subjects who did not attend later waves from their participation in the very first wave. It would then be possible to correct for attrition bias under the assumption of no selection bias, as in Andersen et al. (2008b). When the maintained assumption fails, as in the analysis, this may lead to a sharply different conclusion from the full approach that corrects for both types of biases. For example, only correcting for attrition bias would have led us to reject temporal stability in the population mean and standard deviation of relative risk aversion, with a p-value of 0.007.[38]

We do not claim that correcting for attrition bias under the assumption of no selection bias is less desirable than making no correction at all. This is an empirical issue that must be evaluated on a case-by-case basis.[39] Characterizing situations in which endogenous selection has substantive effects is an inherently difficult task, since it is correlation in unobservables that drives selection bias. The constructive implication of our analysis is that one can identify the effects of selection directly by adopting an experimental design that exogenously varies show-up fees and avoid speculating on the presence and magnitude of selection bias.

C. *Effects of Sample Selection and Attrition on Risk Attitudes under RDU*

We continue to observe significant selection and attrition bias under RDU. The hypothesis test of no sample selection bias now involves the correlation coefficients between the error term in the selection equation and the five other random components (the error term in the attrition equation, two r parameters, and two φ parameters). This hypothesis is rejected at all conventional levels, since the p-value is less than 0.001. The hypothesis test of no attrition bias involves the correlation coefficients between the error term in the attrition equation and four structural parameters (two r parameters, and two φ parameters), and we again reject the null hypothesis of no attrition bias (p-value < 0.001). The estimated correlation coefficient between the error terms in the selection and attrition equations, ϱ_{s1s2}, is equal to -0.416 with a standard error of 0.162, so we can again reject the naive approach of correcting for each source of sampling bias separately.

Figure 2b displays the overall effects of controlling for selection and attrition bias on the estimated population

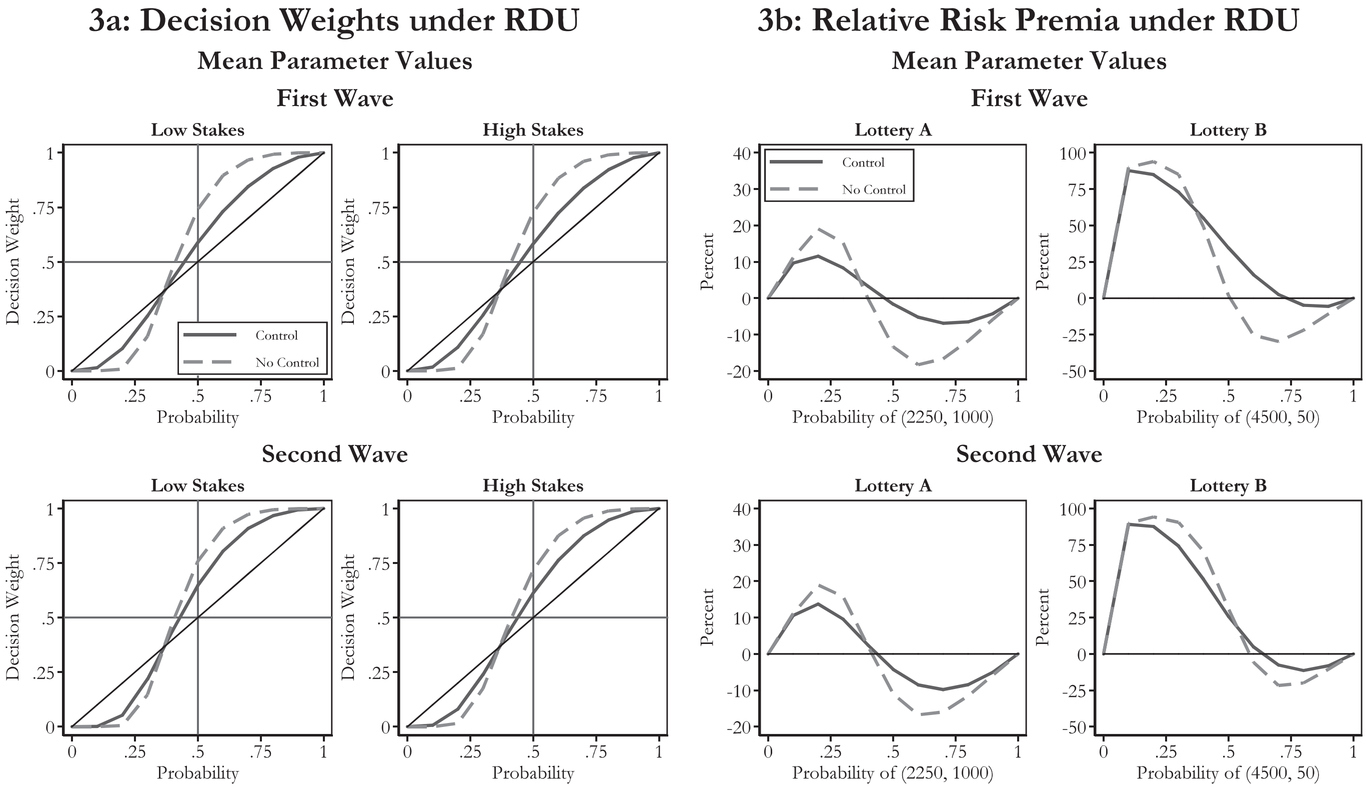
[36] All but one of the 1996 invitees lived within a 36.2 mile radius from a venue. The exception was one subject who lived in Copenhagen but participated in the experiment in Århus.

[37] Table C3 in appendix C reports the estimated parameters for the EUT model with no correction for selection and attrition bias.

[38] Table C7 in appendix C reports the estimated parameters for this EUT model with corrections for attrition bias and no corrections for selection bias.

[39] Whether correcting for only one type of bias worsens the overall bias depends on the interplay of all correlation coefficients pertaining to selection and attrition errors (in our case, ϱ_{s1s2}, ϱ_{s1r1}, ϱ_{s1r2}, ϱ_{s2r1}, and ϱ_{s2r2}). There is no analytic formula, or even reliable intuition, that can provide a guide. This issue may be best addressed by a Monte Carlo study of misspecification biases under systematically varied patterns of correlations.

FIGURE 3.—DECISION WEIGHTS AND RELATIVE RISK PREMIA UNDER RDU



distributions of the probability weighting parameter. The lower panel shows the estimated distributions with no correction for sample selection and attrition bias, and here we find statistical evidence of temporal stability.[40] More specifically, without corrections for nonrandom selection and attrition bias, we cannot reject the null hypothesis that the population distribution of the φ parameter is temporally stable (the $\chi^2(4)$ test statistic has a *p*-value of 0.304). Viewed another way, the uncorrected estimates of the probability weighting parameter seem relatively stable around biased base levels. We also observe that the shape of the population distribution for the weighting parameter changes when we correct for selection and attrition bias. Figure 2b shows that the population distribution of the φ parameter is more skewed to the right in the upper panel with corrections compared to the lower panel without corrections. A larger fraction of subjects can be classified by an inverse-S shaped probability weighting function when we correct for selection and attrition bias compared to the noncorrected estimates.

We can look more closely at the effect of adding controls for sample selection and attrition on risk attitudes under RDU. The effects on the mean of the *r* parameter are modest: estimates of concavity slightly decline in both waves 1 and 2 when we control for selection and attrition bias, so the risk premium derived from utility concavity, ceteris paribus, is lower. The effects on the mean of the φ parameter are shown in figure 3a. The top (bottom) panel refers to the first (second) wave, and the left-right-panel refers to the low (high) stakes treatment. There are two outcomes in each lottery, and the probability weighting functions displayed in figure 3a are identical to the implied decision weights on the highest outcome. Based on figure 3a we can infer the effect of probability weighting on risk attitudes evaluated at the mean of φ. The S-shape of the probability weighting function leads to a negative (positive) risk premium for lotteries with a relatively high (low) probability of the highest outcome, ceteris paribus. We see similar S-shaped probability weighting across the two waves. While corrections for selection and attrition bias do not change our qualitative inferences regarding the shapes of the probability weighting functions, they lead to smaller mean estimates in both waves, making the extent of probability distortion less pronounced. This finding on S-shaped probability weighting at mean values does not contradict the upper panel of figure 2b that classifies a large fraction of the population as inverse-S instead: φ follows a right-skewed distribution, and the mean is sensitive to a long right tail.

We can again assess the potential error in assuming away selection bias and just correcting for attrition bias. As with EUT, this second-best approach again leads to incorrect inferences.[41] Under RDU this approach would lead one to reject the hypothesis that the population mean and standard deviation of *r* and φ was temporally stable, with a two-sided

[40] The estimated parameters are reported in table C4 in appendix C.

[41] Table C8 in appendix C reports the estimated parameters for the RDU model with corrections for attrition bias and no corrections for selection bias.

p-value of 0.07.[42] This is again sharply different from the conclusion when correcting for both selection and attrition.

We can derive certainty equivalents for each lottery in option A and option B of the 40 decision tasks and then evaluate the risk premia associated with different prize sets. Figure 3b displays the estimated risk premium in percent as a function of the probability of the highest outcome in lottery A with 2,250 kroner and 1,000 kroner and lottery B with 4,500 kroner and 50 kroner. Lottery pairs like these were presented in decision tasks that involved the largest stake within our experiment. The solid line is based on the estimated parameter values for r and η with corrections for selection and attrition bias, and the dashed line is based on the model without correction for endogenous selection and attrition. The results show that endogenous selection and attrition bias can have a substantive effect on the estimated risk premium. For example, the upper-right panel shows that the risk premium for lottery B with a 50-50 chance of 4,500 kroner and 50 kroner is 1.7% of the expected value in the model with corrections for endogenous selection and attrition bias, and is equal to 34.6% in the model with no control for selection and attrition bias.

V. Conclusion

Heckman and Smith (1995, 99) noted, "Surprisingly, little is known about the empirical importance of randomization bias." Aggregate data on participation rates from job training experiments by Hotz (1992) and clinical trials by Kramer and Shapiro (1984) suggest that the bias due to endogenous participation decisions might be significant, but we know of no study that directly evaluates the hypothesis.[43] We do not a priori know the direction of randomization bias in economics experiments, and whether the use of recruitment fees mitigates the effects of randomization bias on elicited risk attitudes. Given the importance of randomized control trials in policy experiments in economics and concerns with inferences drawn from such designs (Harrison, 2011a, 2011b, 2013), there is surely some urgency to understand if randomization per se affects the latent characteristics of subjects.

We find evidence of temporal stability for inferred risk attitudes under EUT when the model fully corrects for endogenous sample selection and attrition bias. A joint test of the estimated mean population coefficients for relative risk aversion and standard deviation coefficients for relative risk aversion, across the two waves, allows us to demonstrate that the entire population distribution of relative risk aversion is temporally stable. Furthermore, the estimated mean and estimated standard deviation of the population relative risk aversion are each temporally stable. Finally, the correlation of the population distribution of relative risk aversion is positive and statistically significant between waves, consistent with temporal stability at the individual level.

We obtain similar aggregate results for temporal stability under RDU, but with one difference. Under RDU, the risk premium depends on utility curvature and probability weighting. When we consider the joint distribution of all parameters characterizing utility curvature and probability weighting, we cannot reject the hypothesis of temporal stability. This is what one would expect from the EUT results, since the two must agree in terms of the aggregate risk premium. But we find that there is temporal stability of the mean of the utility curvature parameter and temporal instability of the standard deviation of the utility curvature parameter. The parameter characterizing probability weighting demonstrates temporal stability. We again observe correlations between parameters over time, consistent with individual-level temporal stability.

These results are encouraging, in the sense that temporal stability allows policymakers to have some sense of confidence when designing policies that affect risky outcomes over time, such as social insurance. But the results are particularly striking because we also find statistically significant evidence of endogenous sample selection and attrition. One might find temporal stability without making a correction for selection and attrition because the raw data are literally the same from wave to wave, or even the inferred risk preferences are literally the same from wave to wave. We conclude that one must make that correction and that it results in changes in the averages and standard deviations of risk preference parameters: compare the top and bottom panels of figure 1 under EUT and figure 2a under RDU, and the two sets of probability weights in each panel in figure 2b under RDU. These changes in risk preferences translate into economically significant changes in risk premia as shown in figure 3b. Although we find evidence consistent with temporal stability with no corrections for selection and attrition, this is temporal stability with respect to biased estimates of risk preferences.

The effects of selection and attrition also accord with intuition. For example, we find a positive and significant effect of the higher recruitment fee on the propensity to self-select into the first wave of our experiment. People who live farther away from the session venues are less likely to participate, and people who live closer are more sensitive to a small increase in the distance.

Our results therefore show that randomization bias can have significant effects on inferences about risk attitudes. Neglecting endogenous sample selection and attrition could lead one to draw erroneous conclusions about risk attitudes at a point in time (e.g., the average Dane's relative risk aversion now), as well as stability in risk attitudes over time (e.g., whether the average Dane's relative risk aversion has changed over time). In fact, we find that neglecting selection and attrition leads to the first type of erroneous conclusion but not,

[42] Under EUT (RDU) the instability comes from the estimated mean (standard deviation) of the population parameter r.

[43] Many other hypotheses about the effects of sample selection and attrition in longitudinal studies have been evaluated, of course. In the case of clinical trials, for instance, Beunckens, Molenberghs, and Kenward (2005) compare the effects of obvious ad hoc methods (such as assuming that the last observed case for some subject who does not participate in later sessions is the observation that the subject would have provided, or only using subsamples that participate in all sessions), methods based on imputation and corrections for the imprecision of the imputation, and direct-likelihood methods such as those used here.

in general, to the second type of erroneous conclusion. These results hold whether one uses an EUT or RDU characterization of risk attitudes, although the way in which sample selection and attrition affect the analysis is different across the two decision theories as well as alternative measures of temporal stability that one may consider.

These effects of randomization bias on risk attitudes are clear in our design because of the exogenous variation in recruitment fees. We do not claim that our findings generalize beyond the adult Danish population, the specific recruitment fees we employed, or the battery of lotteries we employed. On the other hand, our sample is wide and representative of the adult Danish population, and our recruitment fees and lottery parameters fall well within common practice in field experiments. The constructive implication for future experimental design is to exogenously vary show-up fees and evaluate the effects on a case-by-case basis. Access to administrative data such as the Danish Civil Registry is not a prerequisite for operationalizing the proposed design. Recruiting experimental subjects from an existing household survey sample (Tanaka, Camerer, & Nguyen, 2010) is an example of an alternative study design that allows one to obtain background information on nonparticipants. Of course, in the latter case, correcting for the effects of selection would lead to inferences that pertain to a survey population instead of a general population.

The need for corrections to mitigate randomization bias is bad news from our results, because it requires renewed attention to ex ante sample design or ex post statistical corrections. It also raises deep concerns with experimental designs that rely on randomization to infer causal effects and that only check for consistency of observables over time. However, the good news is that even after making such corrections, there are still many quantitative and qualitative aspects of risk attitudes that remain temporally stable, at least for this population and the time frame evaluated in our experiments.

Why is it that we observe such stability of risk preferences in Denmark, during a period in which all major industrialized countries experienced various macroeconomic disruptions? One hypothesis might be that the extensive social network of consumer protections in Denmark mitigated the effect of changes in these background risks on the foreground risk aversion our experiments measured. There is also evidence that Danes view the foreground risks of experiments as distinct from their extra-experimental wealth (Andersen, Cox et al., 2018). The methodology we develop can be applied to different populations to evaluate the extent to which they exhibit the same temporal stability of risk preferences.

REFERENCES

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström, "Eliciting Risk and Time Preferences," *Econometrica* 76 (2008a), 583–618.

——— "Lost in State Space: Are Preferences Stable?" *International Economic Review* 49 (2008b), 1091–1112.

——— "Discounting Behavior and the Magnitude Effect: Evidence from a Field Experiment in Denmark," *Economica* 80 (2013), 670–697.

——— "Discounting Behavior: A Reconsideration," *European Economic Review* 71:1 (2014), 15–33.

——— "Multiattribute Utility Theory, Intertemporal Utility and Correlation Aversion," *International Economic Review* 59 (2018), 537–555.

Andersen, Steffen, Glenn W. Harrison, Arne Risa Hole, Morten I. Lau, and E. Elisabet Rutström, "Non-Linear Mixed Logit," *Theory and Decision* 73 (2012), 77–96.

Andersen, Steffen, James C. Cox, Glenn W. Harrison, Morten Lau, E. Elisabet Rutström, and Vjollca Sadiraj, "Asset Integration and Attitudes to Risk: Theory and Evidence," this REVIEW, 100 (2018), 816–830.

Apesteguia, Jose, and Miguel A. Ballester, "Monotone Stochastic Choice Models: The Case of Risk and Time Preferences," *Journal of Political Economy* 126 (2018), 74–106.

Baucells, Manel, and Antonio Villasís, "Stability of Risk Preferences and the Reflection Effect of Prospect Theory," *Theory and Decision* 68 (2010), 193–211.

Beunckens, Caroline, Geert Molenberghs, and Michael G. Kenward, "Direct Likelihood Analysis versus Simple Forms of Imputation for Missing Data in Randomized Clinical Trials," *Clinical Trials* 2 (2005), 379–386.

Capellari, Lorenzo, and Stephen P. Jenkins, "Modeling Low Income Transitions," *Journal of Applied Econometrics* 19 (2004), 593–610.

Das, Mitali, Whitney Newey, and Francis Vella, "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies* 70:1 (2003), 33–58.

Dasgupta, Utteeyo, Lata Gangadharan, Pushkar Maitra, and Subha Mani, "Searching for Preference Stability in a State Dependent World," *Journal of Economic Psychology* 62 (2017), 17–32.

Diggle, P., and M. G. Kenward, "Informative Drop-Out in Longitudinal Data Analysis," *Applied Statistics* 43:1 (1994), 49–93.

Eckel, Catherine, and Philip J. Grossman, "Volunteers and Pseudo-Volunteers: The Effect of Recruitment Method on Dictator Experiments," *Experimental Economics* 2 (2000), 107–120.

Garvey, Paul R., Stephen A. Book, and Raymond P. Covert, *Probability Methods for Cost Uncertainty Analysis: A Systems Engineering Perspective*, 2nd ed. (Boca Raton, FL: CRC Press, 2015).

Glöckner, Andreas, and Thorsten Pachur, "Cognitive Models of Risky Choice: Parameter Stability and Predictive Accuracy of Prospect Theory," *Cognition* 123 (2012), 21–32.

Goldstein, Daniel G., Eric J. Johnson, and William F. Sharpe, "Choosing Outcomes versus Choosing Products: Consumer-Focused Retirement Investment Advice," *Journal of Consumer Research* 35 (2008), 440–456.

Harrison, Glenn W., "Randomisation and Its Discontents," *Journal of African Economies* 20 (2011a), 626–652.

——— "Experimental Methods and the Welfare Evaluation of Policy Lotteries," *European Review of Agricultural Economics* 38 (2011b), 335–360.

——— "Field Experiments and Methodological Intolerance," *Journal of Economic Methodology* 20 (2013), 103–117.

Harrison, Glenn W., Morten I. Lau, and Melonie B. Williams, "Estimating Individual Discount Rates in Denmark: A Field Experiment," *American Economic Review* 92 (2002), 1606–1617.

Harrison, Glenn W., and John A. List, "Field Experiments," *Journal of Economic Literature* 42 (2004), 1009–1055.

Harrison, Glenn W., and Jia Min Ng, "Evaluating the Expected Welfare Gain from Insurance," *Journal of Risk and Insurance* 83 (2016), 91–120.

Harrison, Glenn W., and Don Ross, "Varieties of Paternalism and the Heterogeneity of Utility Structures," *Journal of Economic Methodology* 25 (2018), 42–67.

Harrison, Glenn W., and E. Elisabet Rutström, "Risk Aversion in the Laboratory," in J. C. Cox and G. W. Harrison, eds., *Risk Aversion in Experiments* (Bingley, UK: Emerald, 2008).

Hausman, Jerry A., and David A. Wise, "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica* 47 (1979), 455–473.

Heckman, James J, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5 (1976), 475–492.

——— "Sample Selection Bias as a Specification Error," *Econometrica* 47 (1979), 153–162.

Heckman, James J., and Jeffrey A. Smith, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives* 9 (1995), 85–110.
Heckman, James J., Jeffrey A. Smith, and Christopher Taber, "Accounting for Dropouts in Evaluations of Social Programs," this REVIEW 80 (1998), 1–14.
Hey, John D., and Chris Orme, "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica* 62 (1994), 1291–1326.
Hotz, V. Joseph, "Designing an Evaluation of JTPA," in C. Manski and I. Garfinkel, eds., *Evaluating Welfare and Training Programs* (Cambridge, MA: Harvard University Press, 1992).
Kagel, John H., Raymond C. Battalio, and James M. Walker, "Volunteer Artifacts in Experiments in Economics: Specification of the Problem and Some Initial Data from a Small-Scale Field Experiment," in V. L. Smith ed., *Research in Experimental Economics* (Greenwich, CT: JAI Press, 1979).
Keane, Michael P., "A Note on Identification in the Multinomial Probit Model," *Journal of Business and Economic Statistics* 10 (1992), 193–200.
Kickert, Walter, "How the Danish Government Responded to Financial Crises," *Public Money and Management* 33:1 (2013), 55–62.
Kramer, Michael, and Stanley Shapiro, "Scientific Challenges in the Application of Randomized Trials," *Journal of the American Medical Association* 252 (1984), 2739–2745.
Lee, Lung-Fei, "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics* 65 (1995), 381–428.
Meng, Chun-Lo, and Peter Schmidt, "On the Cost of Partial Observability in the Bivariate Probit Model," *International Economic Review* 26 (1985), 71–85.
Picard, Robert, "GEODIST: Stata Module to Compute Geodetic Distances," Statistical Software Components S457147, Boston College Department of Economics (2010).
Prelec, Drazen, "The Probability Weighting Function," *Econometrica* 66 (1998), 497–527.
Quiggin, John, "A Theory of Anticipated Utility," *Journal of Economic Behavior and Organization* 3 (1982), 323–343.
Smidts, Ale, "The Relationship between Risk Attitude and Strength of Preference: A Test of Intrinsic Risk Attitude," *Management Science* 43 (1997), 357–370.
Starmer, Chris, "Developments in Non-expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk," *Journal of Economic Literature* 38 (2000), 332–382.
Tanaka, Tomomi, Colin F. Camerer, and Quang Ngyuen, "Risk and Time Preferences: Linking Experimental and Household Survey Data from Vietnam," *American Economic Review* 100:1 (2010), 557–571.
Train, Kenneth, *Discrete Choice Models with Simulation*, 2nd ed. (Cambridge: Cambridge University Press, 2009).
Van de Ven, Wynand P. M. M., and Bernard M. S. Van Praag, "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection," *Journal of Econometrics* 17 (1981), 229–252.
Vella, Francis, "Estimating Models with Sample Selection Bias: A Survey," *Journal of Human Resources* 33:1 (1998), 127–169.
Wilcox, Nathaniel T., "'Stochastically More Risk Averse': A Contextual Theory of Stochastic Discrete Choice under Risk," *Journal of Econometrics* 162 (2011), 89–104.
Wooldridge, Jeffrey, *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. (Cambridge, MA: MIT Press, 2010).
Zeisberger, Stefan, Dennis Vrecko, and Thomas Langer, "Measuring the Time Stability of Prospect Theory Preferences," *Theory and Decision* 72 (2012), 359–386.