Reconsidering Risk Aversion*

Daniel J. Benjamin
University of Southern California

Mark Alan Fontana
University of Southern California

Miles Kimball
University of Michigan

This Draft: March 24, 2016

Abstract

Measuring risk preferences is challenging because people's choices over lotteries sometimes vary depending on how the lotteries are framed and violate expected-utility axioms. We develop a two-stage procedure to measure risk preferences, and we demonstrate it via a survey about hypothetical retirement investment choices administered to 601 Cornell students. The first stage is the standard method of eliciting choices over risky lotteries. In the second stage, we confront participants with their inconsistencies—their different responses to choices framed differently that should be the same according to expected-utility axioms—and allow them to update their choices. Our key assumption is that individuals' updated, "reasoned" choices more closely reflect their preferences than their original, "untutored" choices. We find that on average, participants update in the direction of consistency with expected-utility axioms, and their reasoned choices exhibit less risk aversion than their untutored choices. Our results suggest that deviations from the axioms usually reflect decision errors rather than non-expected-utility preferences. Our two-stage procedure may hold promise as a way to measure risk preferences for the purpose of setting optimal defaults or giving advice about portfolio allocation.

JEL Classification: D63, D81, G11, H8,

Keywords: risk preferences, retirement investing, framing effects, expected utility

---

* Benjamin: Center for Economic and Social Research, University of Southern California, 135 Downey Way, Los Angeles, CA 90089-3332, and NBER (email: djbenjam@usc.edu); Fontana: Center for Economic and Social Research, University of Southern California, 135 Downey Way, Los Angeles, CA 90089-3332 (email: markfont@usc.edu); Kimball: Department of Economics, University of Michigan, 312 Lorch Hall, Ann Arbor, MI 48109, and NBER (e-mail: mkimball@umich.edu). For financial support, we are grateful to NIH/NIA through grants R21-AG037741 to Cornell University, P01-AG026571 and R01-AG040787 to the University of Michigan, and T32-AG00186 to the NBER. We thank Mike Gideon for helpful early conversations, Fudong Zhang for MATLAB guidance, and Bas Weerman for MMIC guidance. We thank Shengmao Cao, Regina Chan, Kevin Coughlin, Christian Covington, Samantha Cunningham, Jakina Debnam, Alina Dvorovenko, Dan Golland, Hui Dong He, Duk Gyoo Kim, Sunsiree Kosindesha, Raymond Lee, Sichun Liu, Rebecca Royer, Michael West, Hyun Gu, and especially Yeo Yu (Leo) She, Xing Guo, Andrew Sung, and Jordan Kimball for excellent research assistance.

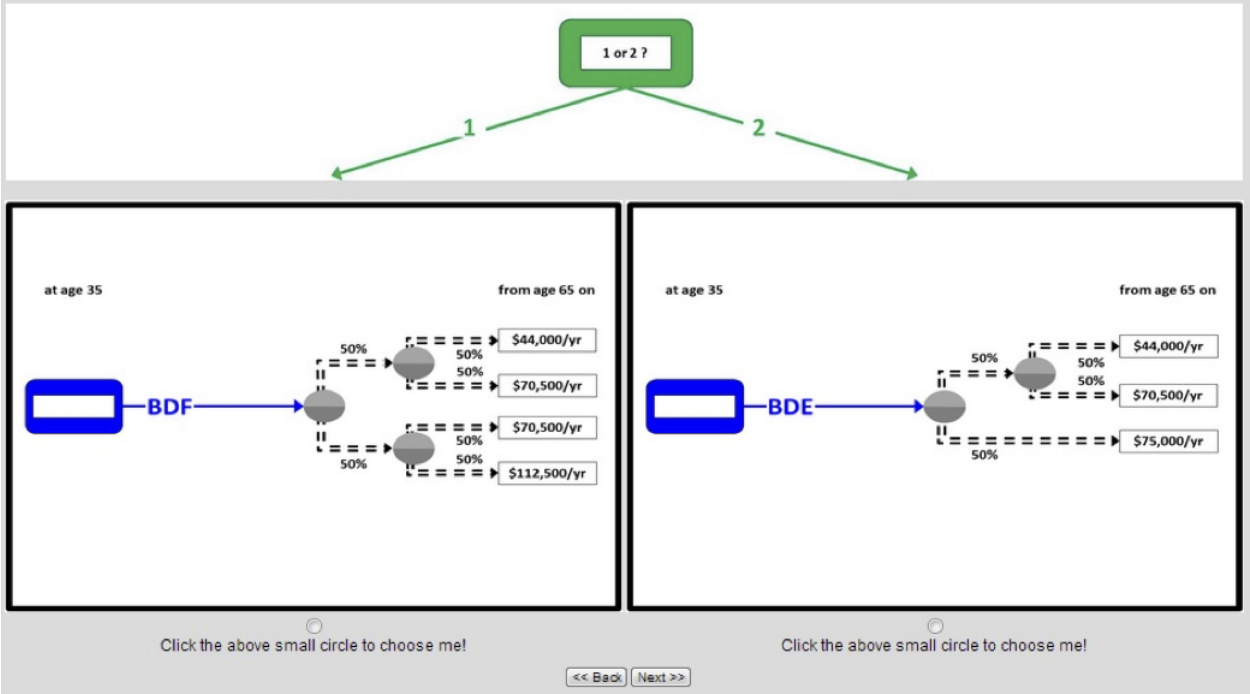
Policymakers, economists, and the popular media have long been worried that Americans may not be investing appropriately in preparation for retirement (e.g., Benartzi and Thaler, 1999).[1] It may therefore be valuable to give people advice about their asset allocation or to use policy tools, such as defaults, to nudge people toward better allocations. However, to identify what is a "good" asset allocation for an individual (or to set an optimal default allocation for a group), a key input is the individual's risk preferences. The usual approach to measuring risk preferences is to infer them from real or hypothetical choices over risky lotteries. But there is reason for concern that people's untutored choices may not accurately reflect their preferences, since for most people, risky decisionmaking (even if simplified) is unfamiliar and cognitively challenging. Moreover, people's choices often violate basic, seemingly compelling axioms of expected utility (EU) theory, such as the Reduction of Compound Lotteries axiom. While there are economic models that can accommodate such non-EU behavior (e.g., Segal, 1990), most are explicitly meant to be descriptive (rather than normative) and thus admit the interpretation that non-standard behaviors represent decision errors, rather than features of actual preferences (Beshears et al., 2008).

In this paper, we develop a two-stage procedure to measure risk preferences and implement it via a survey. The first stage is the standard method of eliciting choices over risky lotteries. We call participants' behavior in this stage their *untutored choices*. Our innovation is the second stage: We confront participants with *inconsistencies* in their untutored choices: their inconsistent responses to choices framed differently that should be the same according to normative axioms. We ask participants whether their untutored choices were mistaken, and if so, how they would like to reconsider their choices. We call participants' updated choices their *reasoned choices*.

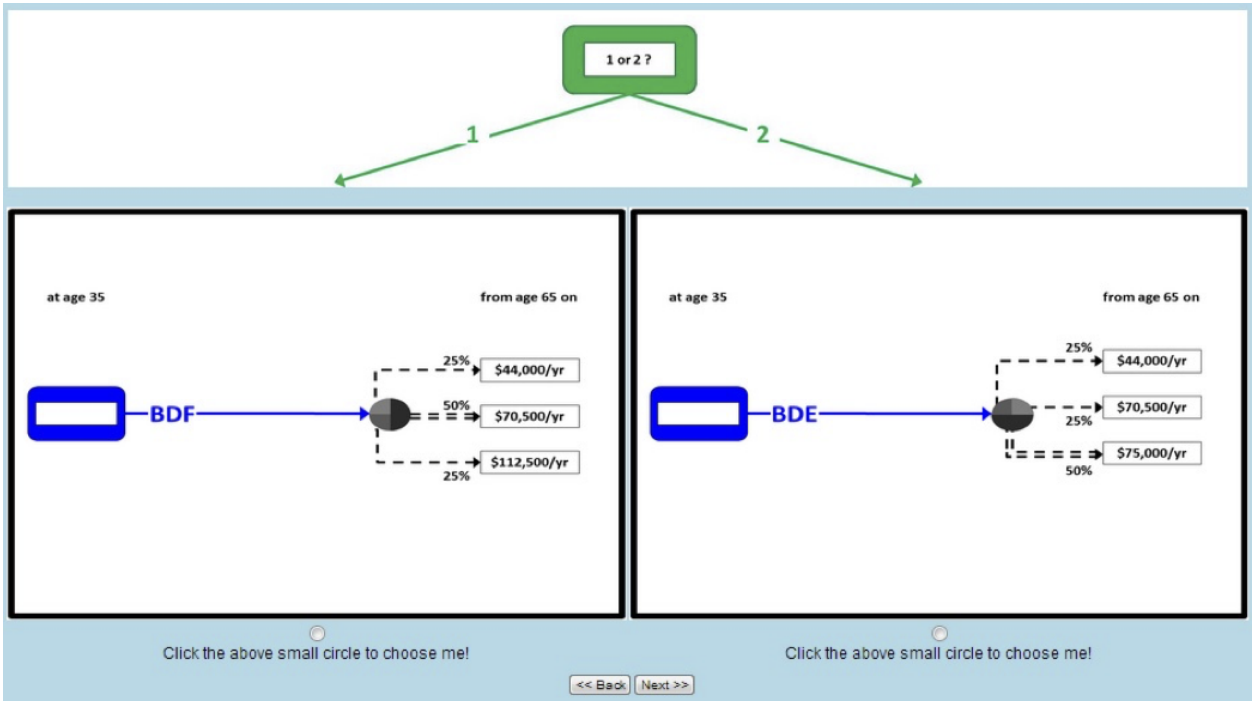
The key assumption underlying our procedure is that, relative to the untutored choices, the reasoned choices more closely reflect the individual's preferences. Given this assumption, we use our survey results to identify mistakes separately from deliberate violations of axioms that economists take to be normative, and to reduce the range of uncertainty about people's risk preferences.

[1] An example from the popular media is http://financialmentor.com/free-articles/investment-advice/alternative-investment-strategy/deadly-dozen-investment-mistakes.

We test our procedure in a sample of 601 Cornell students, 264 of whom returned to the lab 2-4 weeks later for a second survey wave. On each survey wave, the first stage of our survey elicited untutored risk choices, which we measured using hypothetical scenarios about investing for retirement, with monetary outcomes describing "how much you have to spend each year during retirement from age 65 on." For example, in one of the survey screens, participants chose between two compound lotteries, a riskier lottery ("BDF") pictured on the left and a safer lottery ("BDE") pictured on the right:



In another survey screen, participants again made a choice between two simple lotteries—the "same" two lotteries according to the Reduction of Compound Lotteries axiom, but framed differently:



As in much prior work on framing effects and non-expected utility (e.g., Kahneman and Tversky, 1979), we find that participants often make inconsistent choices across different frames; in the Reduction of Compound Lotteries example above, in our data 17% of participants made different choices.

In the second stage, we confront participants with inconsistencies in their untutored choices. Continuing the example from above, suppose that in the first stage, a participant had answered one of the survey screens with "BDE" and the other with "BDF." We would show both decisions on the same screen and ask the participant to endorse one of two statements: "It makes sense to have the same choice in both questions" or "It makes sense to have different choices." If a participant reported that it made sense to have different choices, we would ask a follow-up question about why, giving possibilities that included: the two situations are different, the participant is indifferent, the participant chose how he or she thought the experimenter wanted, and the participant is confused. If a participant reported that it made sense to have the same choice, then we gave the participant the option of changing one or both decision, and we asked a follow-up question about why, this time giving possibilities that included: the participant originally made a mistake, the participant learned about his or her own preferences from

answering other questions, the participant is indifferent, the participant chose how he or she thought the experimenter wanted, and the participant is confused.

A key concern with a survey design like this, in which participants are asked if they want to change choices they made earlier, is a possible "experimenter demand effect" (e.g., Fleming and Zizzo, 2015): a participant may infer from the question itself that he or she *should* change the earlier choice, or may infer that the experimenter wants him or her to do so. Our survey design included several features intended to minimize such effects or measure their impact. First, whenever we offered participants the opportunity to change one of their choices, we also offered the additional options of keeping both choices the same and of switching *both* choices, thereby making our intent less obvious than if we had urged participants to make their choices consistent. Second, roughly half the time that we offered participants the opportunity to change one of their choices, we selected pairs of choices that were already *consistent* with the relevant normative axiom. By doing so, we further masked our intentions and also obtained a placebo measure of how often people change their choices when prompted to do so. Finally, as already noted, we directly asked participants if their choices were driven by what they thought the experimenter wanted. Our measures suggest that experimenter demand effects on participants' choices were minimal. For example, while participants updated their untutored choices on average 40% of the time when these choices were inconsistent, they did so only 2% of the time in the placebo cases when the choices were originally consistent.

The first part of the survey, which elicited untutored choices, asked about choices between five investment plans framed seven different ways, for a total of 36 questions in total. We designed the set of choices so that the seven frames correspond to six normative axioms, each defined by the equivalence of adjacent frames. While most research on non-expected utility theory has focused on a traditional set of axioms, such as the Independence axiom, most of our "axioms" are sub-components of the Independence axiom. Indeed, a crucial feature of our experimental procedure is that no subject was ever asked to understand a complex chain of reasoning. Instead, each axiom was selected to correspond to a rather transparent equivalence between two frames.

We find that, on average across the six axioms we examine, in their initial untutored choices participants behave inconsistently 20.3% of the time. While a majority of participants maintain at least one inconsistent decision in their reasoned choices, there is substantial

movement overall toward endorsing the normative axioms, with the average inconsistency rate falling after the second stage to 9.0%. There is also substantial variation across axioms in both initial and final inconsistency rates. For the Reduction of Compound Lotteries axiom, the inconsistency rate begins at 23.1% and falls to 8.3%.

While most of our analyses allow for the possibility that participants may not accept one or more of the normative axioms, in order to quantify the degrees of risk aversion, we also conduct parametric analyses that assume not only expected-utility theory but constant relative risk aversion (CRRA) more specifically. Pooling choices across participants, we estimate the CRRA parameter using maximum likelihood from the untutored choices and from the reasoned choices, and from choices framed in different ways. Consistent with our non-parametric analyses, we find near-convergence of the CRRA parameter across frames as we move from untutored to reasoned choices, as well as a decrease in estimated choice-error variance. In our data, we also find that reasoned choices imply greater risk tolerance than untutored choices, with the CRRA parameter averaged across frames decreasing from roughly 0.5 to roughly 0 (i.e., log utility).

In additional analyses, we separately estimate the CRRA parameter for untutored and reasoned choices across demographic groups. While we find that female participants and those with lower cognitive ability have higher choice-error variance in untutored choices, the gender gap narrows in the reasoned choices, and in our data there is no significant difference in estimated risk aversion across groups for either untutored or reasoned choices.

As mentioned above, our two-stage procedure is premised on an assumption that reasoned choices are closer to preferences than untutored choices. This is a substantive psychological assumption that does not follow from the principle of revealed preference (Bernheim and Rangel, 2009). It seems most reasonable to us in settings where individuals are making abstract choices in a deliberative state of mind, conditions that are reasonably well approximated in our survey. Our procedure builds on an ancient tradition in moral philosophy, and a more recent tradition in the field of decision analysis (e.g., Raiffa, 1968), of having individuals attempt to resolve their internal inconsistencies and thereby discover their own preferences through introspective reasoning. Indeed, the argument that people would endorse the

EU axioms when it is pointed out that their choices violate the axioms has often been explicitly articulated in defense of EU as a normative theory.[2]

Several prior papers explore procedures with elements similar to ours. For example, McNeil, Pauker, and Tversky (1988) and Druckman (2001) compare experimental participants' choices in a risky lottery framed in one of three ways: in terms of "lives saved," "lives lost," and both together. Both papers find that behavior is intermediate between the two frames when both frames are presented together. Loewenstein and Sicherman (1991) find that experimental participants were more likely to maximize the expected discounted value of cash flows after they were provided with arguments for and against doing so. Our paper builds on such prior work by taking a systematic approach to evaluating normative axioms and comparing untutored with reasoned choices.

The rest of the paper is organized as follows. In Section 2, we briefly motivate the importance of giving people good advice about asset allocation by calibrating the expected welfare losses that result when a person's asset allocations are based on the wrong risk aversion parameter. In Section 3, we describe the structure of the experiment, the training provided to participants, and the sample. Section 4 lays out the risky choices posed to participants, as well as the results from both the untutored-choice and reasoned-choice stages of the survey. In Section 5, we report results from our maximum-likelihood estimations of the CRRA parameter from both untutored and reasoned choices. Section 6 concludes.

I. Importance of Setting Portfolio Defaults Appropriately

To help motivate why it is important to set portfolio defaults appropriately, we calculate the expected welfare losses that arise if an individual is defaulted into an allocation that is optimal for a *different* level of preference parameters than her own. Consider the continuous-time, 35-year-long investment-saving problem faced by a household with annual discount rate ρ and initial wealth $w_0 > 0$ (and no labor income), as in Merton (1969), except with Kreps-Porteus preferences that separate the elasticity of intertemporal substitution (EIS) s from relative risk aversion γ. In each instant, the household chooses what fraction of wealth to consume, and what

[2] For example, Morgenstern (1979) writes: "[EU theory] is normative in the sense that the theory is 'absolutely convincing' which implies that men will act accordingly. If they deviate from the theory, an explanation of the theory and of their deviation will cause them to readjust their behavior."

fraction to invest in a risky asset whose excess return is stochastic, with mean μ and standard deviation (per square-root of time) σ as opposed to a safe asset with rate of return r. Rather than assuming that the household behaves optimally, we suppose that the household's allocation is optimal with respect to the (possibly *wrong*) risk aversion, time preference, and EIS parameters: $\hat{\gamma}$, $\hat{\rho}$, and $\hat{s}$; while we focus below on the consequences of using the wrong risk aversion parameter, we allow all three preference parameters to vary for completeness.

Let $u^*(\gamma, \hat{\gamma}, \rho, \hat{\rho}, s, \hat{s}; w_0)$ denote the expected discounted utility, calculated using the household's true preference parameters γ, ρ, and s, of the state-contingent saving-investment path implied by solving the household's problem with risk aversion $\hat{\gamma}$, $\hat{\rho}$, and $\hat{s}$. Then let the real-valued function $c^*(\gamma, \hat{\gamma}, \rho, \hat{\rho}, s, \hat{s}; w_0)$ denote the certain, constant level of consumption that, if consumed at every moment from now on, would give expected discounted utility equal to $u^*(\gamma, \hat{\gamma}, \rho, \hat{\rho}, s, \hat{s}; w_0)$. We measure the welfare loss from choosing the suboptimal investment-saving path by

$$(1) \qquad \frac{c^*(\gamma, \hat{\gamma}, \rho, \hat{\rho}, s, \hat{s}; w_0)}{c^*(\gamma, \gamma, \rho, \rho, s, s; w_0)} = \left(\frac{1 - e^{-aT}}{a}\right)^{\frac{1}{1-s}} \left(\frac{1 - e^{-\hat{m}T}}{\hat{m}}\right)^{\frac{-s}{1-s}} \left(\frac{1 - e^{-\hat{a}T}}{\hat{a}}\right)^{-1},$$

where $a \equiv (1-s)\left(r + \frac{\mu^2}{2\gamma^2\sigma^2}\right) + s\rho$, $\hat{a} \equiv (1-\hat{s})\left(r + \frac{\mu^2}{2\hat{\gamma}^2\sigma^2}\right) + \hat{s}\hat{\rho}$, and $\hat{m} \equiv \rho + \left(\frac{1}{s} - 1\right)\left(r + \frac{\mu^2}{2\hat{\gamma}^2\sigma^2}(2\hat{\gamma} - \gamma) - \hat{a}\right)$. This ratio equals 1 if the household's portfolio choice is optimal for its preferences but will be less than 1 otherwise. Given the assumptions on preferences, the ratio is independent of w_0, and it can be interpreted as the proportional reduction in initial wealth (and hence consumption in all states) that is equivalent in terms of welfare to the mistaken behavior.

Table 1 presents illustrative calculations of (1), where we set $T = 35$, $r = 0.03$, $\mu = 0.03$, $\sigma = 0.15$, $T = 35$, $r = 0.03$, $\mu = 0.03$, $\sigma = 0.15$, $s = \hat{s} = 0.375$, $\rho = \hat{\rho} = 0.075$, and $\gamma, \hat{\gamma} \in \{1,2,4,8\}$.

Even though much existing literature has focused on the mistake of non-participation in financial markets (e.g., Haliassos and Bertaut, 1995), Table 1 shows that the more severe mistake

is behaving less risk-aversely than one actually is (because the extra risk is extremely costly). If a household with risk aversion $\gamma = 1$ behaves like $\hat{\gamma} = 2$, the welfare loss is equivalent to 6% of wealth. For the converse (the household takes on too much risk) then the welfare loss is 14%. If $\gamma = 8$ and $\hat{\gamma} = 1$, then the loss is 94%! These calculations imply it is important to choose a default allocation appropriately; a one-size-fits-all default may be particularly harmful for individuals who are more risk averse than the chosen default.

II. Experimental Setup

II.A. Experimental Procedure

The experiment has 6 parts: (1) Pre-Test, (2) Training Batteries, (3) Main Body Part 1: Elicitation of Untutored Preferences, (4) Psychological and Cognitive Batteries, (5) Main Body Part 2: Elicitation of Reasoned Preferences, and (6) Follow-Up and Demographic Batteries. We dedicate discussion of sections (3) and (5), which constitute the core components of our procedure and experiment, to their own sections later (during which we also discuss descriptive results). The software used to implement our procedure was designed using the RAND Corporation's Multimode Interviewing Capability technology. Subjects sat a computer terminal and were instructed to not speak nor interact with other participants nor use their phones while taking the survey. Proctors walked around the room and ensured these instructions were followed. Experimental sessions were scheduled for 2 hours each, but mean completion time was 68 minutes (not including initial introductions and interactions which last about 10 minutes). Subjects were paid $40 for 2 hours[3].

II.A.1. Pre-Test

The Pre-Test elicited preliminary risk aversion estimates for each subject using simple, text-based binary choices between hypothetical safe and risky assets. For more information on the questions asked, and for the details of analyses exploring the relationship these preliminary risk aversion measures and each of sex and cognitive performance, see Appendix: Pre-Test.

[3] Lack of direct incentives is often worrying in survey research. However, given our focus is on saving for retirement, fully incentivizing subjects with a large fraction of lifetime resources is well beyond our budget constraints (and would be even be well beyond our budget constraint if implemented in a developing country). Waiting many decades for payouts to realize is also more of an ambitious experiment than we hoped to undertake.

II.A.2. Training Batteries

The training sections: (1) reviewed basic facets of probability theory with examples using coin tosses and dice rolls, (2) taught subjects to interpret the symbols and figures used to convey the investment choices with which they would be faced, and (3) reviewed background assumptions we wanted subjects to hold while making their decisions. Each training module was followed by a short quiz meant to test subjects' understanding. Subjects could not continue to the next part of the survey until they got nearly all quiz questions correct. See the Appendix: Survey Walkthrough for more details on all the training sections.

II.A.3. Psychological and Cognitive Batteries

Interspersed between parts of Main Body Part 1: Elicitation of Untutored Preferences, we elicited several short psychological and cognitive batteries. These included: (1) cognitive reflex task (Frederick, 2005), (2) number series battery (adopted from the CogUSA, McArdle et al., 2015), (3) 10-question Big-Five personality battery (Gosling et al., 2003), (4) probabilistic sophistication battery (developed by Miles Kimball, see the Appendix: Survey Walkthrough for details), and (5) need for cognition battery (Cacioppo et al., 1984). Subjects alternated between answering all of the questions in a particular set of Main Body Part 1 questions and these psychological and cognitive batteries. The order in which these question batteries were presented was randomized.

II.A.4. Follow-Up and Demographic Batteries

After Main Body Part 1: Elicitation of Untutored Preferences and Main Body Part 2: Elicitation of Reasoned Preferences, we asked subjects to complete several follow-up question batteries. First, subjects were asked subjects how much they agreed that the survey made them feel each of the following emotions on a 6-point scale: enjoyment, annoyance, stress, and frustration (Fagerlin et al., 2007). Wave 1 responses are shown in Table 2. Subjects somewhat enjoyed the experience (44.8% respond 4/6 or higher); were somewhat annoyed by it (54.6% respond 4/6 or higher); were not overly stressed by it (66.3% respond 3/6 or lower), and were slightly frustrated by it (41.2% respond 4/6 or higher).

We next asked participants about the extent to which they preferred abstract descriptions of ways in which an investment decision could be framed, as well as how strongly they felt about those preferences. Overall, participants' preferences over abstract descriptions of frames are fairly unrelated to how they actually update their decisions in Main Body Part 2: Elicitation of Reasoned Preferences. However, our analyses are generally underpowered. See Appendix: Explicit Frame Preferences for more information.

We finally administered a basic demographics survey. Further details, as well as the exact wording of questions, can be found in the Appendix: Survey Walkthrough.

II.A.5. Subject Population and Timeline

We collected data at Cornell's LEEDR and Business Simulation laboratories on 601 people during 2013-2014 (almost all undergraduates)[4]. Mean subject age was 20.9 years. Approximately 90% of subjects were between the ages of 18 and 22 (i.e. college students). Subjects were disproportionately female, 64.7%.

See Table 3 for details of the experimental timeline. We first collected data on 317 subjects during July, November, and December 2013. These subjects were only invited to one experimental session, which we call wave 1. We next collected data on 284 subjects in April 2014. We invited all 284 of these subjects back to the laboratory for a follow-up session 2-4 weeks after the initial wave, which we call wave 2, and denote this subsample the wave 1+2 sample. Almost 87% (246 subjects) returned for the second wave, during which subjects only receive the Pre-Test, an abridged Training section, and both parts of the Main Body. Our motivation was to see how their decisions change over the course of several weeks, and to what extent any updating that took place during the first wave "stuck" with participants over the short-term. We should note that returning participants saw a re-randomized version of the survey (described in detail in the next section) and hence were usually faced with different main decisions—this mitigates concerns over participants simply remembering and repeating exactly what they did during the first wave.

[4] Our Cornell student population, although not representative, is perhaps non-representative in a way that makes up for the lack of direct, financial incentive. Cornell students are particularly motivated to perform well on difficult tasks, so much so that this (non-representative) internal motivation might be sufficient to overcome lack of direct, financial incentive. That being said, other non-representative characteristics of the particular student subject pool might sully the external validity of our findings, although we cannot do much more than caveat this potential difficultly.

II. B. Experimental Preference Elicitation and Descriptive Results

II.B.1. Main Body Part 1: Elicitation of Untutored Preferences

We begin by describing the single decision problem underlying all questions in the main parts of our experiment. Subjects are faced with a two-period investment horizon over 20+ years involving 3 decision nodes and gambles over "how much you have to spend each year during retirement from age 65 on." See Figure 3 for a visualization of the decision problem. It involves 3 binary choices between: A and B, C and D, and E and F. The first option in each pair always has a certain outcome, while the second option always has some degree of uncertainty associated with it. We exclusively use probabilities of 25% and 50%, which are more familiar to most people than the probabilities often used in tests of expected utility. The first binary choice, between A and B, is meant to be made at age 35. Conditional on choosing B, the subsequent 2 binary choices are meant to be made at age 50. Therefore, 5 investment plans are possible: A, BCE, BDF, BDE, and BDF (ordered from safest to riskiest).

We chose the context of retirement savings for the decision problem as a matter of practicality. Getting participants to internalize and understand a complex decision problem with multiple decision nodes and stochastic elements would be difficult if we had relied on an unfamiliar, abstract domain. While our subject pool is fairly young, they are all no doubt familiar with saving for retirement and that doing so involves multiple decisions and multiple sources of uncertainty. In other words, the retirement context makes our decision problem concrete and understandable, minimizing potential confusion about the decisions themselves. Although hypothetical, the fact that the amounts are over such large quantities and involve total resources during retirement brings a tone of seriousness to the questions that likely pushed participants into thinking more carefully about their decisions than they might otherwise have.

Another motivation behind the decision problem is the general observations that people are more risk averse over shorter horizons than longer horizons. Although not the focus of our analysis, this feature of human behavior motivated splitting the decision problem into decisions to be made at ages 30 and 45. As will become clear, it also partially motivated how we split this decision problem into different frames.

This single decision problem is used to form 7 different sets of axiomatically equivalent questions, or "frames," which elicit choices about the whole decision problem or specific subparts of it. Our frames vary whether irrelevant information is provided or not, whether independent choices are made sequentially or simultaneously, whether subjects are able directly to choose their favorite choice or can only choose between many pairs of filled-out plans, and whether lotteries are already reduced or not. Frames vary in length between a single question (i.e. the whole decision problem) and 10 questions (i.e. asking participants to choose between all "5 choose 2 = 10" pairs of investment plans). For frames involving more than one question, the ordering of questions was randomized. Each participant is asked all questions in each of these 7 frames, constituting their untutored preferences. We categorize the 7 frames into 2 groups: 4 Nodewise Action Choice Frames and 3 Pairwise Strategy Choice Frames, which we will describe in more detail momentarily.

Across all frames, participants were asked to hold the following assumptions while making decisions: (1) "The government provides free medical insurance, and you are in good health," (2) "The government no longer provides social security (i.e. monthly checks)," (3) "There is no inflation," (4) "Imagine that your friends and extended family outside of your household do not need financial help from you, and you cannot ask them for money," (5) "When you retire at age 65, you plan to move into rental housing that will have a monthly payment," and (6) "You have no other resources beyond the amounts specified by your decisions. For example, any money you get from selling your existing home has already been figured into the yearly spending you can afford."

We randomized participants into one of six different "monetary levels," i.e. sets of monetary amounts[5,6], used as payoffs in all 7 frames. The monetary levels vary the attractiveness of the risky options (see Table 4). Each is associated with a different level of constant relative risk aversion (CRRA): 1.576, 2.958, 4.865, 7.184, 12.113, or 17.967. The 3rd, 5th, and 6th

[5] Our goal in constructing these dollar amounts was to create gambles that would maximize initial inconsistencies. Subjects are faced with 3 binary choices. However, according to expected utility and CRRA, and given our monetary amounts, choosing the safer option in one pair of options usually implies one should also choose the safer option in the other pairs (and vice-versa for the choosing the risky option). We did not choose these amounts to maximize power in measuring cardinal risk aversion.

[6] Unfortunately, for our first 317 participants surveyed in summer 2013, our software suffered from an error: inadvertently, we were only randomizing participants into the two sets of monetary amounts that made the riskier investment plans the most attractive. Therefore, for the remaining 284 participants, we randomized participants into the three sets of monetary amounts that made the riskier investment plans the least attractive. One set of monetary amounts was never used.

amounts shown in Table 4 are always 100k, 150k, and 225k, respectively. The 2nd amount is the dollar value such that an expected utility maximizer with CRRA preferences (given the constant coefficient of relative risk aversion associated with that set of monetary amounts) would be indifferent between 100k for sure and a 50-50 chance of 150k and the unknown amount. Finally, the 1st amount is the 2nd amount squared and divided by 100. Within a wave, a participant was always asked questions drawing from the same monetary level (i.e. monetary amounts are held fixed for a given subject and wave). Subjects who returned for a second wave received a re-randomized version of the survey, thereby potentially receiving a different monetary level.

We further randomized approximately half of our participants into monetary amounts that are exactly half of those listed in Table 4. According to constant relative risk aversion (CRRA), this should not affect participants' choices. To be clear, we do not rely on this assumption in any of our descriptive results, as we seek to be agnostic about which facets of rationality or models of decisionmaking a participant use or ought to use. However, based on this randomization, our MLE analysis below verifies CRRA as a good modeling assumption.

Although we will explain our procedure for reaching reasoned preferences more thoroughly in the next section, Main Body Part 2: Elicitation of Reasoned Preferences, we note briefly that our procedure rests on asking subjects to reconcile any inconsistencies in their decisionmaking among the axiomatically-equivalent 7 frames. Of course, it would be extremely difficult to try to explain to a typical experimental subject why an entire set of choices among all frames is normatively problematic. Because we want to make sure that subjects *do* understand how combinations of choices violate normative axioms, we confront subjects *only* with instances where their risk preferences violate normative axioms *between similar frames*. We constructed the frames so that equivalences between these similar frames follow from *small steps of logical reasoning*. We do not take for granted adherence to any of the axioms underlying these small steps of logical reasoning. Now, let us turn to explaining the 7 frames themselves and the axioms underlying them. See Table 5 for the names of the axioms associated with each small step of logical reasoning and the names of the 2 associated frames.

The simplest frame is simply the entire single decision problem, which we call the Complete Contingent Action Plan. It is also 1 of 4 of the Nodewise Action Choice Frames. The remaining 6 frames are derived from the Complete Contingent Action Plan.

Let us first describe the remaining 3 Nodewise Action Choices Frames (or simply "nodewise frames"). First, we form Two Contingent Actions with Backdrop by disallowing a subject to choose between A and B (graying out those options, see Figure 4). She must make 2 simultaneous decisions to lock in how she will invest at age 50, between C and D as well as E and F. The subject is asked to pretend to be 35 years old when making this decision, told she has already chosen B, but does not yet know how this decision has turned out. We call the small step of logical reasoning between Two Contingent Actions with Backdrop and Complete Contingent Action Plan the axiom of *Irrelevance of Counterfactual Choices* (see Table 5)[7].

Second, we form Single Action with Backdrop from Two Contingent Actions with Backdrop by isolating and making sequential the 2 decisions between C and D as well as E and F. A subject is asked to pretend to be 50 years old and given information about how decisions made when she was 35 turned out that are beyond her control at this point, i.e. the 2 ways in which risky choice B could turn out. In each question, the other possibility and associated decision is grayed out; subjects are told those represent things that could have happened but they know for sure did not happen. See Figure 5 for how the E versus F question looked. The order of the 2 questions is randomized. We call the small step of logical reasoning between the 2 questions in Single Action with Backdrop and the single question in Two Contingent Actions with Backdrop the axiom of *Simple Actions = State-Contingent Actions* (see Table 5).

Third, we form Single Action in Isolation from each question in Two Contingent Actions with Backdrop. It consists of the same 2 sequential decisions with the grayed-out (irrelevant) features completely removed, again asking participants to pretend they are 50 years old. See Figure 6 for how the C versus D question looked. Again, the order of the 2 questions is randomized. We call the small step of logical reasoning between the 2 questions in Single Action in Isolation and the 2 questions in Single Action with Backdrop the axiom of *Irrelevance of Background Counterfactuals* (see Table \ref{fig:steps}).

Let us next describe the 3 Pairwise Strategy Choices Frames, or simply pairwise frames. Each consists of 10 questions, asking participants to choose between every possible pair of the 5

[7] For the 284 wave 1+2 subjects who were invited back for a second wave, during both their wave 1 and 2 surveys, if a subject chose A in Complete Contingent Action Plan, we followed-up by asking her to pretend she had missed the deadline to register for A, forcing her also to choose between each of C versus D and E versus F as a contingency plan for Complete Contingent Action Plan. This allows potential inconsistences with Two Contingent Actions with Backdrop, i.e. the axiom Irrelevance of Counterfactual Choices, even if a subject chose A in Complete Contingent Action Plan.

investment plans (5 choose 2 = 10). Each of the 10 questions presents the two options next to each other, labeled "Option 1" and "Option 2." The order of the 10 questions in each frame is randomized. Moreover, in each question, the orientation of which investment plan appears on the right and which appears on the left is randomized, thereby mitigating the tendency of English language readers to favor options on the left.

First, Pairwise Choices Between Complete Strategies asks a subject to choose between all possible pairs of already filled-out Complete Contingent Action Plans, with unselected options grayed-out. For example, Figure 7 shows the decision between A and BCE. We call the small step of logical reasoning between a subject's favorite investment plan given in Complete Contingent Action Plan and the 4 questions involving that same investment plan in Pairwise Choices Between Complete Strategies the axiom *Fusion + Shift from Nodewise to Pairwise* (see Table \ref{fig:stes}).

Second, Pairwise Choices Between Compound Lotteries simplifies things by stripping the questions in Pairwise Choices Between Complete Strategies of grayed-out (irrelevant) information and de-emphasizing the intertemporal aspect of learning about probabilistic outcomes at age 50. For example, Figure 2 shows the decision between BDE and BDF (also pictured in the Introduction). We call the small step of logical reasoning between each of the 10 questions in Pairwise Choices Between Complete Strategies and Pairwise Choices Between Compound Lotteries the axiom *Complete Strategies = Implied Lotteries* (see Table \ref{fig:steps}).

Third, Pairwise Choices Between Reduced Simple Lotteries reduces the compound lotteries presented in Pairwise Choices Between Compound Lotteries. For example, Figure 1 shows the decision again for BDE versus BDF (also pictured in the Introduction). The small step of logical reasoning between each of the 10 questions in Pairwise Choices Between Compound Lotteries and Pairwise Choices Between Reduced Simple Lotteries is simply the standard axiom of *Reduction of Compound Lotteries* (see Table \ref{fig:steps}).

We randomized the order in which the frames were asked during Main Body Part 1. This order was also used to determine the order in which subjects were allowed to update their choices in Main Body Part 2. For approximately 1/3 of subjects, the order was: Single Action in Isolation, Single Action with Backdrop, Two Contingent Actions with Backdrop, Complete Contingent Action Plan, Pairwise Choices Between Complete Strategies, Pairwise Choices

Between Compound Lotteries, and Pairwise Choices Between Reduced Simple Lotteries. For another 1/3, the order was reversed. A final 1/3 was given a random ordering[8].

Approximately half of our subjects received all of the survey questions oriented exactly as appear in the figures presented in this section. The remaining half were randomly assigned to a "down" randomization that flipped the orientation of all figures such that option A appeared on the bottom rather than at the top. This was meant to mitigate people's tendency to choose options toward the top of the screen. In practice, fortunately, orientation did not seem to matter for subjects' choices (see below MLE, although the "down" randomization was associated with slightly higher incidence of errors).

Table 6 details the participants untutored preferences in each pairwise frame for all 10 questions, i.e. the percentage of respondents initially favoring the relatively safe option in each of the 10 questions. The top panel combines all participants' wave 1 initial, untutored responses. The bottom panel restricts attention to wave 1+2 subjects' wave 2 untutored responses. Within each panel, and for each of the 10 questions, it is clear that there exist significant inconsistencies among the 3 pairwise frames.

Table 7 provides an overview of subjects' initial responses in each of the 7 frames. The top panel combine all participants' wave 1 initial responses. The bottom panel restricts attention to wave 1+2 subjects' wave 2 initial responses.

For the nodewise frames, we simply show the percentage of respondents having chosen each of the 4 or 5 investment plans. Subjects appear to particularly like BDF, although there exist clear inconsistencies among the nodewise frames.

Complete Contingent Action Plan is listed multiple times in each panel. Complete Contingent Action Plan (without any additional qualifiers) shows the percentage of respondents choosing each of the 5 investment plans, which are readily comparable to the pairwise frame entries (described below). The starred rows exclude A, showing the percentage of participants having chosen BCE, BCF, BDE, or BDF, re-normalized so those percentages sum to 100%. This allows easy comparison with the other nodewise frames. In bottom panel, Complete Contingent Action Plan (cont.) details how subjects who chose A answered the contingency question which forced them to pretend they had missed the deadline for registering for A. Both rows show

[8] If reasoned preferences truly reflect underlying preferences, then reasoned preferences should not depend on the order of reasoning. Unfortunately, we are not statistically well powered enough to analyze this randomization using our MLE procedure.

comparable percentages to the other nodewise frames, though there is significant evidence of respondents' inconsistencies here too.

It is not straightforward to compare these final 5 percentages, representing only participants' favorite choices in Complete Contingent Action Plan, to the 10 questions in each pairwise frame, representing a more thorough elicitation of preferences. To facilitate comparison to the pairwise frames, we calculate the percentage of respondents who clearly favor each investment plan among those who have an investment plan which is preferred strictly more times than any other investment plan. Participants again appear particularly to like BDF. The pairwise frames once again show signs of inconsistencies among themselves, as well as with Complete Contingent Action Plan.

II.B.2. Main Body Part 2: Elicitation of Reasoned Preferences

During Main Body Part 2, we elicited reasoned preferences by providing subjects the opportunity to systematically review and update their choices from each of the 7 frames answered in Main Body Part 1. The adaptive software first confronted subjects with one round of inconsistency checks, providing them the opportunity to update inconsistencies between frames highlighted as similar according to our axioms. Subjects were also given the opportunity to update a random subset of already-consistent choices, allowing us to rule out experimenter demand as the driving force behind updating (i.e. participants were not updating simply because we gave them the opportunity to do so). Next, the software confronted subjects with one round of intransitivity checks, which allowed them to explicitly rank options within pairwise frames among which there existed cyclical preferences. Finally, the software repeated one additional round of consistency checks (with placebos) and one additional round of intransitivity checks. In designing these questions, we tried to avoid being paternalistic, but instead strove to be deferentially light-handed. We consider a participant's choices after these checks to be their "reasoned preferences." By comparing participants' reasoned preferences with their untutored preferences, specifically detailing in which frames participants update and in which frames they do not, we can differentiate decision errors from genuine lack of agreement with our normative axioms.

II.B.3. Intransitivity Checks

We discuss intransitivity first because it constitutes more basic violations of rationality than inconsistencies implied by our axioms. For each pairwise frame, we confronted participants with intransitivities (i.e. instances of cyclical preferences) among their 10 choices. Subjects were presented with, and asked to explicitly rank, the choices among which they had an intransitivity (or note that they did not want to rank them). Their choices in that frame were then adjusted accordingly. Given 10 questions were necessary to elicit all pairwise preferences among 5 investment plans (i.e. 5 choose 2 = 10), intransitivities could be 5-way, 4-way, or 3-way. For 27.3% of intransitivity checks, participants answered that they did not want to rank the choices among which they were intransitive[9].

See Figure 8 for an example of a 4-way intransitivity among A, BCE, BCF, and BDE for Pairwise Choices Between Reduced Simple Lotteries.

Among the wave 1+2 sample, if a subject selected "I do not want to rank these options," they were given the opportunity to answer a follow-up question: "Why couldn't you rank the options on the previous slide?" Their options were: (1) "I couldn't rank the options because they are all equally good to me," (2) "I couldn't rank the options because I don't know which option I prefer," (3) "I should be able to rank the options, but it's extremely hard," or (4) "I couldn't rank the options for another reason." Subjects answered this question 90% of the time it was presented. Among instances of response, 16.86% responded "Options all equally good," 49.42% responded "I don't know what I prefer," 22.67% responded "Too hard to rank," and 11.05% responded "Other." Among those responding "Other," we further prompted for a free-response; most complained about being tired or noted that the task was too hard.

Table 8 shows the average number of intransitivities per participant at various points, or "rounds," during the survey among the 3 pairwise frames[10]. We restrict attention to the wave 1+2

[9] We also asked all participants how they felt about intransitivity in the abstract: "Ian Trantivi is facing a weird problem. He is on a game show, and has just won! As a prize, he can choose one of three piles of stuff. He says that he prefers the first pile to the second, the second pile to the third, and the third pile to the first! Do you think you could ever imagine feeling this way?" 58.3% of participants in the first wave and 56.1% of participants in the second wave answered "Yes, I can imagine feeling like Ian about some number of choices." (as opposed to "No, I cannot imagine feeling like Ian about some number of choices."). Comparing their answers to this question among the participants surveyed in both waves, 92 participants consistently answered "No," 120 participants consistently answered "Yes," and the remaining 34 participants swapped their answers between waves (18 initially said "Yes" and switched to "No").

[10] Pairwise Choices Between Reduced Simple Lotteries is slightly more likely to contain an intransitivity than the other two pairwise frames with respect to untutored preferences. However, for participants' reasoned preferences, the three pairwise frames have essentially identical propensities of intransitivities. With respect to the type of

sample who returned for wave 2. We further restrict attention to participants who answered all pairwise questions throughout both rounds. During the first wave, subjects averaged 1.183 intransitivities among their untutored preferences. The first round of inconsistency checks worked in the favor of also resolving intransitivities, leading to an average of 1.048 intransitivities. The first round of explicit intransitivity checks further decreased this to 0.210 intransitivities on average. The second round of intransitivities moved the point estimate in the other direction, indicating a slight trade-off between consistency and transitivity. The second round of intransitivity checks in the first wave lead to a final average of 0.135 intransitivities. During the second wave, these subjects averaged 0.742 and 0.140 intransitivities in their untutored and reasoned preferences, respectively, with round-by-round trends similar to the first wave. Indeed, comparing untutored preferences in both waves, there appears to be mild "stick" to our procedure with respect to transitivity. Table 9 shows the average number of intransitivities for *all* participants who answered all pairwise questions throughout wave 1. Results are similar to those above.

Previewing our maximum likelihood estimation (MLE) results, consider that intransitivities are one type of error contained in participants' preferences. Therefore, we expect within-frame measures of error to decline across the course of the survey. This is corroborated by our MLE. The fact that average the number of intransitivties increases between waves is also featured in our MLE results (i.e. within-frame measures of error increase between waves).

II.B.4. Inconsistency Checks

The procedure for a particular inconsistency check was as follows. First, when two choices subjects made violated some axiom in Table 5, they were asked whether they thought those two choices should be the same or different. If they said their choices should be the same, they were given the opportunity to revise their choices, and then were asked why they revised as they did. If they said their choices should be different, they were asked why. For the wave 1 sample, follow-up questions asking why participants updated or not were open-ended. We used these open-ended responses to generate multiple-choice responses used for the wave 1+2 sample. As we will describe in more detail, in the vast majority of instances where a subject revised, they

intransitivities: 3-way intransitivities were the most common, followed by 4-way intransitivities, and finally 5-way. Incidence of these types of intranstivities are rougthly constant across frames and rounds.

said it was because they either originally made a mistake or learned something new about their preferences during the procedure. Experimenter demand was infrequently cited as a reason for switching. Those who did not revise most often cited that they believed the situations were different enough to merit different answers. For more details on the exact wording of these questions, as well as a complete example, see Appendix: Flow of Inconsistency Checks.

We also presented participants with "placebo inconsistencies" to test more directly the influence of experimenter demand—subjects were given the opportunity to change already-consistent choices, using nearly identical text and questions as the normal inconsistency checks. For each pair of questions that could potentially trigger an inconsistency, but did not, a random 25% were pushed to the subject as placebo checks. These checks were randomly interspersed among the standard inconsistency checks. Fortunately, across all axioms, we find that people updated already-consistent choices less than 2% of the time. In contrast, people revised actual inconsistencies more than 40% of the time. This supports the limited influence of experimenter demand on our results. Of course, these placebo questions necessarily differed slightly from the normal inconsistency questions: when subjects said their choices should be the same, they were asked "why" without being given the opportunity to update; when subjects said their choices should be different, they were given the opportunity to update, and then asked why they did so. More than 90% of the time participants responded that the situations were similar enough to merit the same answers. For an example of one such placebo check and more details on the exact wording of questions, see Appendix: Flow of Inconsistency Checks.

The order in which inconsistent or placebo choices appeared within a check was randomized, so that half the time a choice from an "earlier" frame appeared first, and half the time a choice from that same "earlier" frame appeared second. This mitigates a natural tendency toward favoring options on the top.

While there were significant improvements toward consistency in most individual axioms, the majority of subjects still retain at least one inconsistency. For all participants having answered all questions during wave 1 (N = 565), 94.2% of subjects have untutored preferences with at least one inconsistency, and 68.0% of subjects have reasoned preferences with at least one inconsistency. For the wave 1+2 sample having actually returned for wave 2 and answered all questions throughout both waves (N = 214), 95.8% of subjects have wave 1 untutored

preferences with at least one inconsistency. The numbers for wave 1 reasoned, wave 2 untutored, and wave 2 reasoned preferences are 75.7%, 86.0%, and 64.4%.

We next describe *which* normative axioms subjects initially and subsequently endorsed, where they were most likely to update their choices, and toward which frames they revised. Overall, for any given axiom, most people endorsed consistency, although there was substantial heterogeneity across axioms. There were also always people who did not endorse consistency for a given axiom.

To hone in on resolution behavior at each axiom, we derive "inconsistency rates" for each axiom for subjects' untutored and reasoned preferences. Inconsistency rates control for the fact that axioms differ in the number of inconsistencies possible between the two associated frames, facilitating cross-axiom comparisons of inconsistency. An inconsistency rate is defined simply as the actual number of inconsistencies divided by the potential number of inconsistencies. We finally average these inconsistencies rates across participants independently for their untutored and reasoned preferences and for each axiom.

See the top panel of Table 10 for these inconsistencies rates among all subjects participating in wave 1 for their untutored and reasoned preferences. For each axiom, we restrict attention to those participants who are not missing any wave 1 data for the 2 frames constituting that axiom. We also report two-sided p-values associated with tests of differences of proportions for inconsistency rates between wave 1 untutored and reasoned preferences, giving us a sense of whether participants significantly updated toward consistency. Initially, subjects were particularly apt to endorse axioms involving the Nodewise Action Choices Frames relative to axioms involving the Pairwise Strategy Choices Frames. They were especially unlikely initially to endorse: (1) Fusion + Shift Between Nodewise and Pairwise and (2) Reduction of Compound Lotteries. However, Reduction of Compound Lotteries also saw the most movement toward consistency. Irrelevance of Background Counterfactuals had the lowest reasoned inconsistency rate. Overall, we see a movement toward consistency in all axioms except Irrelevance of Counterfactual Choices (a result we will explore more momentarily).

The bottom panel of Table 10 shows results for the subset of the wave 1+2 sample who actually showed up to both waves, detailing inconsistency rates for their untutored and reasoned preferences for both wave 1 and 2. For each axiom, we restrict attention to those participants who are not missing any wave 1 nor 2 data for the 2 frames constituting that axiom. We also

report 3 sets of proportion tests: between untutored and reasoned preferences in wave 1, between untutored preferences in waves 1 and 2, and between untutored and reasoned preferences in wave 2. Wave 1 results here are broadly consistent with those from the full sample: inconsistency rates for all axioms except Irrelevance of Counterfactual Choices significantly decline. Most inconsistency rates increased between reasoned wave 1 preferences and untutored wave 2 preferences. However, most are significantly lower at the beginning of wave 2 than the beginning of wave 1, i.e. our procedure has a positive short-term impact on consistency. Wave 2 sees further reductions in all inconsistency rates except Irrelevance of Counterfactual Choices and Simple Actions = State Contingent Actions.

Previewing our MLE results for the pairwise frames, given an overall trend of declining inconsistency rates among pairwise frames, we expect (1) a general decline in within-frame measures of error between untutored and reasoned preference in each wave, and (2) convergence in measures of population mean risk aversion among frames—both of which we indeed observe in the MLE results. Moreover, we expect an increase in within-frame measures of error between reasoned wave 1 preference and untutored wave 2 preference, which we also observe. The fact that reasoned preferences (in both waves) have strictly positive inconsistency rates is also corroborated; MLE results do not indicate complete convergence among the pairwise frames nor near-zero within-frame measures of error. Furthermore, note that Complete Strategies = Implied Lotteries has a lower untutored inconsistency rate than Reduction of Compound Lotteries. In the MLE, we therefore expect Pairwise Choices Between Reduced Simple Lotteries to have higher within-frame error than the other pairwise frames, which indeed holds. We also expect that Pairwise Choices Between Complete Strategies and Pairwise Choices Between Compound Lotteries to be more similar to each other than to Pairwise Choices Between Reduced Simple Lotteries with respect to mean risk aversion and within-frame error measures, again corroborated by our MLE results.

Table \ref{tab:updatenormperc} shows, for each axiom, conditional on having a particular normal (top panel) or placebo inconsistency (bottom panel), the percentage of the time subjects chose to update (or not), and, conditional on updating, how they updated. Data come from all wave 1+2 subjects and both waves. For each axiom “Frame 1” and “Frame 2” refer to the frames as labeled in Table 5. Note that updating toward frame 1 or 2 implies moving toward

consistency, while either not updating or swapping choices implies retaining inconsistent answers.

We calculate p-values denoting the significant of a T-test of equality between the percentage updating toward frame 1 and the percentage updating toward frame 2, indicating the extent to which there was a systematic preference for updating toward one frame or another. For the axiom with the most updating toward consistency, Reduction of Compound Lotteries, we see inconsistent subjects tend to update toward their choices in the non-reduced lotteries (i.e. from their initial decisions in Pairwise Choices Between Reduced Simple Lotteries toward their decisions in Pairwise Choices Between Compound Lotteries). The direction of updating is corroborated by our MLE results, which show mean risk aversion in the Reduced frame moving toward mean risk aversion in the Compound frame. For Complete Strategies = Implied Lotteries, conditional on an inconsistency, we inconsistent subjects tend to update toward their decisions in Pairwise Choices between Compound Lotteries. With respect to the nodewise frames, we see that subjects are apt to update *away* from Two Contingent Actions with Backdrop, both with respect to (1) Simple Actions = State-Contingent Actions and (2) Irrelevance of Counterfactual Choices.

Results for the placebo inconsistencies stand in stark contrast to the results from the normal inconsistencies (bottom panel of Table \ref{tab:updatenormperc}). Subjects rarely update choices that are already consistent. This confirms the limited influence of experimenter demand on subjects updating toward consistency.

Table 12 details why wave 1+2 subjects, combining data from both waves, said they did not revise. Results are similar across axioms. Overall, 57.1% of the time subjects said the situations were sufficiently different as to merit different answers; these subjects truly do not endorse the associated axiom. 25.0% of the time subjects said they were indifferent[11]. The remainder of the time, subjects cite either experimenter demand, not knowing what their preferences were, confusion, or "other."

Table 13 details why wave 1+2 subjects, combing data from both waves, said they revised one or more of their answers. Results are again similar across axioms. Overall, 45.3% of the time subjects cite that they previously made a mistake in making their choices. 36.0% of the

[11] Note that in these cases, we still consider subjects to be "inconsistent" despite specifically saying they were indifferent.

time that they learned something new about their preferences in answering questions. This indicates our procedure is in fact helping participants better express their underlying preferences. 11.7% of the time subjects said they were indifferent. The remainder of the time, they cite some combination of experimenter demand, not knowing their own preferences, confusion, or "Other." The axiom Simple Actions = State-Contingent Actions is a bit of an outlier: subjects are only 36.1% likely to say they initially made a mistake, 37.7% likely to say they learned something new, 18.0% likely to cite indifference (with nobody citing experimenter demand).

Table 14 details why wave 1+2 subjects said they did not update when presented with placebo inconsistencies (for both waves and all points of the survey). 89.7% of the time subjects cite that the situations are similar enough to merit the same answer. This corroborates that participants indeed endorse these axioms. 7.0% of the time subjects cite indifference. The remainder of the time, subjects cite some combination of experimenter demand (only 0.4%), that they don't know their own preferences, confusion, or "Other." Results are similar across axioms.

Irrelevance of Counterfactual Choices was the only axiom in which we did not observe a declining inconsistency rate. Indeed, this axiom is the step along the logical chain of reasoning that causes people the most trouble toward having completely consistent choices across the whole spectrum[12]. However, conditional on any particular inconsistency check for Irrelevance of Counterfactual Choices, subjects are about as likely to retain inconsistent answers as many of the other axioms (i.e. choosing "No Update" or "Swap"). This can be explained by the fact that, despite there only being 1 potential inconsistency possible, there are a high number of inconsistency checks for this axiom (see "Total" column of Table \ref{tab:updatenormperc}). In fact, here there are 253 / 1 = 253 inconsistency checks per potential inconsistency. For all other axioms involving nodewise frames, these numbers are: 187 / 2 = 93.5, 183 / 2 = 91.5, and 433 / 4 = 108.25, i.e. much lower. Note also that subjects update away from Two Contingent Actions with Backdrop both with respect to (1) Simple Actions = State-Contingent Actions and (2) Irrelevance of Counterfactual Choices (see Table \ref{tab:updatenormperc}). So, while there is indeed movement toward consistency conditional on any particular inconsistency check (on par with other axioms), subjects tend to undo these corrections by moving away from Two

[12] Unfortunately, this is the sole axiom in which the follow-up question asking why a subject wanted to update their answers was inadvertently not implemented. It is plausible they would have responded that the frames differed significantly enough that they wanted to make different choices (as in all other axioms), though we cannot say for sure.

Contingent Actions with Backdrop in Simple Actions = State-Contingent Actions, resulting in an overall high number of inconsistency checks and a high inconsistency rate.

Recall for the wave 1+2 sample, if a subject chose A in Complete Contingent Action Plan, they were prompted for their second-favorite choice as though they could not choose B. This contingency choice was then used to determine whether the subject had an inconsistency for Irrelevance of Counterfactual Choices. We therefore explicitly checked whether these subjects who choose A in Complete Contingent Action Plan are driving the results in Irrelevance of Counterfactual Choices. These people might have cared far less about their second-favorite options, and therefore been more willing to be inconsistent. First, considering only participants who were *not* given the opportunity to report their second-favorite choice conditional on choosing A in the Complete Contingent Action Plan, the inconsistency rate decreases slightly, but the drop is not statistically significant. Considering participants who *could* report a second-favorite choice, both (1) subjects who did *not* choose A (and therefore whose inconsistencies are driven by their favorite choice) and (2) subjects who *did* in fact choose A (and therefore whose inconsistencies are driven by their second-favorite choice) show no statistically significant change in the inconsistency rate in the waves 1 and 2. These facts suggest that people choosing A are *not* driving results in Irrelevance of Counterfactual Choices.

II.B.5. Summary of Descriptive Results

Overall, we find that subjects readily update toward consistency for all but 1 of the 6 axioms we explore (the exception being Irrelevance of Counterfactual Choices). For a subset of participants whom we bring back into the lab 2-4 weeks later for a second wave, we find that their wave 2 untutored preferences are more consistent and transitive than their wave 1 untutored preferences (although not as consistent as their wave 1 reasoned preferences), and their wave 2 reasoned preferences are the most consistent. Similar trends hold for intransitivities. However, a majority of participants still hold at least one inconsistent decision in their wave 1 and wave 2 reasoned preferences. When subjects do in fact update, they most readily cite either having made a mistake initially or having learned something about their own preferences, highlighting the effectiveness of our procedure. When subjects do not update, they most readily cite the situations being presented as fundamentally different, indicating genuine lack of agreement with those axioms. Fortunately for the effectiveness of our procedure, subjects rarely update already-

consistent preferences when presented with placebo inconsistencies. In so doing, they acknowledge the situations to be similar enough to merit the same choices.

II.C. Maximum Likelihood Estimation

We now quantify participants' risk aversion and incidence of decision errors using maximum likelihood estimation (MLE). In the above descriptive analyses, we were completely agnostic as to any particular model of decisionmaking or risk aversion, detailing the extent to which paricipants did or did not agree with a host of normative axioms. Here, we instead assume any deviations from normative axioms implied by expected utility theory can be attributed to errors in decisionmaking. In so doing, we are able to separately estimate variance due to these errors, variance due to true heterogeneity in risk preferences, as well as mean estimates of population risk aversion.

We wish to explore how estimates of risk aversion and incidence of decision errors vary by frame, as well as how they change between untutored and reasoned preferences. Only the Pairwise Complete Strategy Frames are well-powered for these analyses, owing to the fact that these choices elicited participants' complete preferences, whereas the nodewise frames only elicited participants' favorite choices (and sometimes, for the Complete Contingent Action Plan, their second-favorite choices). Among the pairwise frames, we find mild convergence in mean risk aversion between untutored and reasoned preferences, a well as increases in true risk aversion heterogeneity and decreases in variance due to within-frame errors.

We assume subjects are expected utility maximizers with utility functions that obey constant relative risk aversion (CRRA) of the canonical form, with (directly unobserved) risk aversion parameter γ_{ift}, for individual *i*, frame *f*, and round *t*. We use the randomization in which we presented approximately half of our participants with large-stakes monetary amounts and the other approximately half of our participants with small-stakes monetary amounts to verify the legitimacy of CRRA (see below for more details). For computational convenience we will focus on log risk aversion, $x_{ift} = \ln(\gamma_{ift})$. Like above, we use the term "round" to refer to a particular point in the survey. We denote round 1 as after elicitation of initial untutored preferences; round 2 as after the initial round of inconsistency checks; round 3 as after the initial round of intransitivity checks; round 4 as after the second round of inconsistency checks; and round 5 as

after the second round of intransitivity checks. Rounds 6-10 refer to the same points in the survey during wave 2. We specifically assume:

(2) $$x_{ift} \sim N(\mu_{ft}, \sigma^2_{x_{ft}})$$

Next, we define CE_k as the log certainty equivalent of choice k (where k is either A, BCE, BCF, BDE, or BDF) conditional on some known coefficient of relative risk aversion λ, as well as the (also known) monetary amounts on which the subject's payoffs are based (see Table 4, where MA_r denotes the monetary amount in row r for the specific column randomly assigned to a given subject):

(3) $$CE_A = \ln[MA_3]$$

(4) $$CE_{BCE} = ln\left[((1-\lambda)(\frac{1}{2}\frac{{MA_2}^{1-\lambda}}{1-\lambda} + \frac{1}{2}\frac{{MA_5}^{1-\lambda}}{1-\lambda}))^{\frac{1}{1-\lambda}}\right]$$

(5) $$CE_{BCF} = ln\left[((1-\lambda)(\frac{1}{2}\frac{{MA_2}^{1-\lambda}}{1-\lambda} + \frac{1}{4}\frac{{MA_4}^{1-\lambda}}{1-\lambda} + \frac{1}{4}\frac{{MA_6}^{1-\lambda}}{1-\lambda}))^{\frac{1}{1-\lambda}}\right]$$

(6) $$CE_{BDE} = ln\left[((1-\lambda)(\frac{1}{4}\frac{{MA_1}^{1-\lambda}}{1-\lambda} + \frac{1}{4}\frac{{MA_4}^{1-\lambda}}{1-\lambda} + \frac{1}{2}\frac{{MA_5}^{1-\lambda}}{1-\lambda}))^{\frac{1}{1-\lambda}}\right]$$

(7) $$CE_{BDF} = ln\left[((1-\lambda)(\frac{1}{4}\frac{{MA_1}^{1-\lambda}}{1-\lambda} + \frac{1}{2}\frac{{MA_4}^{1-\lambda}}{1-\lambda} + \frac{1}{4}\frac{{MA_6}^{1-\lambda}}{1-\lambda}))^{\frac{1}{1-\lambda}}\right]$$

Because of individual response error, ϵ_{ift}, which is assumed to be the same across questions within a round and frame, we cannot directly observe latent risk aversion. A gamble is evaluated as $CE_k + \epsilon_k$, i.e. certainty equivalents are evaluated with error, obfuscating measurement of latent risk aversion. Specifically, we assume:

(8) $$\epsilon_{ift} \sim N(0, \sigma^2_{\epsilon_{ft}})$$

Our maximum likelihood procedure estimates the vector of parameters: $(\mu_{ft}, \sigma_{x_{ft}}, \sigma_{\epsilon_{ft}})$ Again, μ_{ft} is the mean of the distribution of log risk aversion for a particular frame and round, $\sigma_{x_{ft}}$ is the standard deviation of the distribution of log risk aversion for a particular frame and round (i.e. a measure of true risk aversion heterogeneity), and $\sigma_{\epsilon_{ft}}$ represents within-frame decision error for a particular frame and round. Estimation is carried out using the Berndt-Hall-Hall-Hausman (BHHH) numerical optimization algorithm, which uses the outer product of scores to approximate the hessian (see Berndt et al., 1974). We use a MATLAB implementation programmed by Thomas Jørgensen[13], which also computes robust standard errors. Point estimates using BHHH match those found using global search optimization (but runs more than an order of magnitude faster).

II.C.1. Pairwise Complete Strategy Frames

Recall that subjects choose between 10 pairs of options in the Pairwise Complete Strategy Frames. Let q index each of these 10 pairs, and furthermore assume that for a given q, a subject chooses option $q1$ over $q2$ when the former certainty equivalent is greater than the latter (after including error on each term so a gamble is evaluated as $CE_k + \epsilon_k$). Utilizing 10 different logit discrete choice functions (each between 2 options), the MLE can be written as follows:

$$\max_{\mu,\sigma_x,\sigma_\epsilon} \sum_{i=1}^{n} \ln \left(\int_{-\infty}^{\infty} \frac{1}{\sigma_x \sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma_x^2}} \prod_q f\left[\frac{CE_{q1}}{\sigma_\epsilon}, \frac{CE_{q2}}{\sigma_\epsilon} \right] dx \right) \tag{9}$$

$$f\left(\frac{CE_{q1}}{\sigma_\epsilon}, \frac{CE_{q2}}{\sigma_\epsilon} \right) = \frac{1}{1 + e^{\frac{CE_{q2} - CE_{q1}}{\sigma_\epsilon}}} \tag{10}$$

Results by frame and round can be seen in Figure 9. Diagnostic tests indicate stable likelihood functions and sufficient statistical power (see Appendix: MLE Diagnostic Tests). Note that all available data are used for each round and frame, meaning the first 5 rounds (i.e. wave 1) include about twice as much data as the latter 5 (i.e. wave 2). Results are broadly consistent with our non-MLE analyses above.

[13] http://www.econ.ku.dk/phdstudent/jorgensen/code.htm

First focusing attention on the leftmost pane detailing μ, our measure of mean population risk aversion: we cannot reject complete convergence in mean risk aversion across the frames. Untutored preferences for both Pairwise Choices Between Complete Strategies and Pairwise Choices Between Compound Lotteries start with relatively low μ, increasing slightly over the course of the survey, but remain indistinguishable from each other throughout. Untutored preferences for Pairwise Choices Between Reduced Simple Lotteries start with relatively high μ, declining over the course of the survey toward the other 2 pairwise frames. This convergence corroborates the declining inconsistency rates we described above. The fact that inconsistency rates for Complete Strategies = Implied Lotteries are weakly lower than those for Reduction of Compound Lotteries is consistent with the fact that, in the MLE, Pairwise Choices Between Complete Strategies and Pairwise Choices Between Compound Lotteries are more similar to each other than to Pairwise Choices Between Reduced Simple Lotteries.

Wave 1 untutored preferences' mean risk aversion point estimates range fairly widely, i.e. $\mu = \ln(\gamma) = [-0.76, 0.52]$, $\gamma = [0.47, 1.68]$. Interestingly, by the end of wave 2, our estimates appear to be converging on log risk aversion (i.e. $\mu = \ln(\gamma) = 0$, $\gamma = 1$). The range of point estimates significantly narrows: $\mu = [-0.32, 0.05]$, $\gamma = [0.73, 1.05]$. Our welfare calculations in Table 1 highlight the potentially large average welfare losses stemming from using decisions not adjusted by our procedure. [CALCULATE IMPLIED WELFARE LOSSES?]

Our point estimates of μ are fairly low. However, given our young population, this is perhaps unsurprising. Beyond the fact that younger individuals tend to be more risk-taking, it is likely that choices focused on retirement savings are particularly abstract for our student population, contributing to their overall risk tolerance.

Focusing next on the middle pane detailing $\sigma_{x_{ft}}$, our measure of true risk aversion heterogeneity: Pairwise Choices Between Reduced Simple Lotteries are associated with overall lower true heterogeneity in risk aversion compared to the other two pairwise frames. For all frames, this true heterogeneity increases over the course of the survey.

Finally, focusing on the rightmost pane detailing our measure of within-frame decision errors, $\sigma_{\epsilon_{ft}}$: all pairwise frames see a significant overall decrease in the error response variance, except between waves, when it jumps up slightly before again declining. This is completely consistent with the fact that our inconsistency rates are generally declining, except between

waves. Pairwise Choices Between Reduced Simple Lotteries are, overall, associated with lower error, as might be expected, given these questions are already reduced (and hence simpler). However, this result and our intuition contrasts with our prior analyses, which found Reduction of Compound Lotteries had weakly higher inconsistency rates compared to Complete Strategies = Implied Lotteries.

II.C.2. Nodewise Action Choice Frames

For details of the MLE for the nodewise frames, including exact functional forms, see Appendix: MLE Nodewise Action Choice Frames. Unfortunately, results from the nodewise frames are statistically underpowered. This is driven by the fact that subjects in these frames are only asked for their favorite investment plan[14]. This is in contrast with the Pairwise Complete Strategy Choice Frames, which are indeed well-powered, owing to the fact that we elicit participants' preferences over every possible pair of choices, i.e. we have much more information in the pairwise frames. We conduct diagnostic tests to get a sense of the behavior of the likelihood function when our estimated standard errors are large. We do this by manually varying one of the three parameters while fully optimizing over the remaining two. We next plot the log likelihood as a function of the manually varied parameter. We repeat this separately for each of the three parameters, finding that our likelihood functions break down and behave erratically (e.g. severe non-monotonicity). More details of these tests can be found in Appendix: MLE Diagnostic Tests.

II.C.3. Gentzkow-Shapiro Method for Explaining MLE Estimates

We employ a general method developed by Gentzkow and Shapiro (2014) to estimate sufficiency and sensitivity statistics which describe the extent to which summary statistics explain relevant parameter estimates. We apply this procedure to the Pairwise Complete Strategy Choices Frames. We employ their method to measure the extent to which 3 vectors of crude measures of risk aversion explain μ.

This exercise serves 3 functions: (1) act as sanity checks on our other analyses, i.e. provide an explicit bridge between raw features of our data and our MLE estimates of mean risk

[14] Respondents answering "A" in the Complete Contingent Action plan among the wave 1+2 sample were also asked for their second-favorite choices.

aversion, (2) corroborate that our procedure purifies measurements of risk aversion, and (3) hints at which choices are the most important determinants of our mean risk aversion estimates. The 3 vectors of crude measures are 1-dimensional, 2-dimensional, and 10-dimensional, in order of increasing complexity. Except for the second element of the 2-d measure, all other elements of these vectors are constructed so higher values are associated with greater risk *aversion*. For details on how these are constructed, see Appendix: Gentzkow-Shapiro Method for Explaining MLE Estimates.

Briefly, toward function (1): the 1-dimensional vector captures the leftmost pane of 9 showing μ, as do the two elements of the 2-dimensional vector, see Figures \ref{fig:crudera_1d} and \ref{fig:crudera_2d}. Toward function (2): we point to the sufficiency statistics outlined in Tables \ref{tab:gs_suff_and_sens_1_2} and \ref{tab:gs_suff_and_sens_10}. If our crude measures of risk aversion, which are simply derived from participants 10 choices in each pairwise frame, are initially composed of both true signal about risk preferences and error, and we suppose that our procedure generally increases this signal-to-noise ratio (i.e. corrects for errors), we should expect these 10 questions to better capture risk preferences at the end of the survey than at the beginning—indeed corroborated by the increasing sufficiency statistics. Toward (3): we can compare the sensitivities among the elements the 10-d summary statistics, noting those with higher absolute values are generally more consequential: A versus BDF is, perhaps unsurprisingly, particularly informative across the pairwise frames.

II.C.4. Pairwise Complete Strategy Frames, Stratified by Covariates

We also investigate how our MLE estimates differ by certain fundamental covariates such as cognitive performance, sex, whether monetary amounts were large- or small-stakes (recall half of participants were randomly assigned monetary amounts that were all cut in half), and how all of the choice images were oriented (recall half of participants saw a "down" randomization where everything was flipped vertically). Although the latter two randomizations should, on average, wash out any associated effects, these supplemental analyses allow us to explicitly test whether these factors mattered for individual decision making. For details of these analyses, see Appendix: Pairwise Complete Strategy Frames, Stratified by Covariates.

Although both low and high cognitive performance groups see a sizeable decline in the error response variance, high cognitive performance is associated with robustly lower error

response variance across all rounds, with no convergence between the low and high cognition groups. Estimates from the other two parameters do not significantly differ between the two cohorts. Although our procedure indeed helps both relatively low and high cognitive functioning individuals, it does not close the gap between these groups.

With respect to sex, error response variance is initially lower for men than for women, but converges by the second wave. This is consistent with men being more engaged in the survey initially. It is also consistent with lower financial sophistication among women, a phenomenon for which our procedure corrects.

Next, we compared MLE estimates for two randomized groups, one receiving large-stake monetary amounts, and the other receiving low-stake monetary amounts (i.e. exactly one-half of the large-stakes amounts). Estimates of mean risk aversion are fortunately indistinguishable between the two cohorts, lending credence to CRRA as a modeling assumption. True heterogeneity is slightly higher for the large-stakes randomization. Error response variance is robustly higher for the large-stakes randomization, consistent with participants being more careless when given bets involving ubiquitously larger amounts of money.

Finally, we compare MLE estimates for two other randomized groups, one receiving images oriented as appear throughout this text, and the other receiving images with the orientation of choices flipped vertically. While randomizing between these groups should result in any differences washing out in the final estimation, fortunately, for all three parameters, the groups are indistinguishable.

III. Conclusions

Quantitative measures of individual risk aversion, which themselves are necessary inputs for portfolio calibration and the selection default asset allocations in pension plans, vary depending on a variety of factors that should not matter according to standard normative axioms—notably how the problem is framed. Our main contribution is the development of a surveying procedure which attempts to correct for framing biases in measuring risk aversion.

We test our procedure on a sample of 601 mostly student subjects. We first elicit "untutored" risk preferences for each of 7 axiomatically-equivalent ways of framing the same fundamental decision problem, involving gambles over "how much you have to spend each year during retirement, from age 65 on.". We then lead subjects through a reconciliation phase,

allowing them to update their choices where they have contradicted normative axioms of rational choice (i.e. choosing inconsistently among frames or intransitively within a frame), finally reaching "reasoned" preferences, which we define as those measured after reconciliation. We also allow subjects to update already-consistent preferences to test whether our procedure's results are driven by experimenter demand.

Subjects readily become more consistent and transitive and rarely move in the other direction. However, few subjects completely endorse *all* normative axioms implied by differences between frames, and there is substantial heterogeneity across frames in both number of initial consistencies and propensity to update. Initially, subjects are particularly apt toward consistency in Nodewise Action Choice Frames (as opposed to Pairwise Strategy Choice Frames). They are particularly unlikely initially to endorse Reduction of Compound Lotteries, although this axiom also sees the most resolution toward endorsement after sufficient reasoning. Subjects readily update toward consistency for all but 1 of the 6 axioms we explore (the exception being Irrelevance of Counterfactual Choices).

For a subset of participants whom we bring back into the lab 2-4 weeks later for a second wave, we find that their wave 2 untutored preferences are more consistent and transitive than their wave 1 untutored preferences (although not as consistent as their wave 1 reasoned preferences), and their wave 2 reasoned preferences are the most consistent. Similar trends hold for intransitivities.

When subjects do in fact update, they most readily cite either having made a mistake initially or having learned something about their own preferences, highlighting the effectiveness of our procedure. When subjects do not update, they most readily cite the situations being presented as fundamentally different, indicating genuine lack of agreement with those axioms. Fortunately for the effectiveness of our procedure, subjects rarely update already-consistent preferences when presented with placebo inconsistencies. In so doing, they acknowledge the situations to be similar enough to merit the same choices.

Using maximum likelihood estimation, we quantify risk aversion and incidence of decision errors among frames and over the course of our procedure. We see convergence in risk aversion among pairwise frames and an overall reduction in decision errors. Subjects tend to be more risk averse in their responses for Pairwise Choices Between Reduced Simple Lotteries, and have a higher incidence of errors, as compared to the other pairwise frames. Our wave 1 mean

risk aversion point estimates are statistically distinguishable and have a fairly wide range: $\mu = \ln(\gamma) = [-0.76, 0.52]$, $\gamma = [0.47, 1.68]$. By the end of wave 2, these point estimates are statistically indistinguishable and have a narrower range: $\mu = [-0.32, 0.05]$, $\gamma = [0.73, 1.05]$. Interestingly, our estimates appear to be converging on log risk aversion (i.e. $\mu = \ln(\gamma) = 0$, $\gamma = 1$). [IMPLIED WELFARE LOSSES HERE.]

Overall, we have demonstrated a method that makes progress in reducing the range of uncertainty surrounding risk aversion. However, the method is still imperfect. More data collection and further tweaks to our procedure are necessary for us to state definitely "the" level of risk aversion across frames (i.e. to provide default asset allocation recommendations for pension plans or 401(k)'s).

We have several ideas moving forward. First, our procedure focused on being as light-handed as possible in terms of encouraging resolution; we gave subjects ample opportunity to do what they wanted, and our results strongly reject experimenter demand as a driving force. However, heavy-handed promotion could help close the gap between measures of risk aversion derived from different frames (e.g. by changing the wording of the inconsistency checks to more strongly encourage resolution). Perhaps this extra push could facilitate even more updating and hence greater convergence.

Second, we would like to elicit second-, third-, and fourth-favorite choices among the Nodewise Action Choice Frames. This will help boost the statistical power of our MLE procedure and provide a more comprehensive picture of subjects' preferences. It would also facilitate greater comparability between nodewise and pairwise frames.

Third, it would be interesting to bring subjects back into a lab for third or even fourth waves. We saw substantial evidence of persistence of wave 1 reasoned preferences into wave 2; it would be interesting to evaluate how preferences persist over greater periods of time. Moreover, subjects had the greatest amount of consistency, and lowest within-frame error responses, after the second wave. It would be interesting see whether this trend continues into further waves (or to what extent there are declining marginal returns to further waves).

Fourth, it would be useful to start collecting data on a more nationally representative sample. The fact that our estimates of μ are unexpected low is likely driven by our fairly young, student sample. Indeed, external validity is a concern that we can only really address by

broadening our subject pool. This is especially important given our goal of identifying "the" level of risk aversion in the population across frames.

Once we are confident we have validated a procedure that addresses these concerns and more accurately assesses reasoned risk preferences, our next step will be to find a short survey that successfully approximates the results of the more thorough procedure. Such a survey could be implemented in real time on actual investors to identify individual-specific optimal asset allocations.

Since the question of determining optimal policy when choices violate normative axioms is central to behavioral welfare economics (Bernheim and Rangel, 2008), we believe that the idea of eliciting reasoned preferences could have quite general implications for economics and policy. We envision that applying the kind of procedure we develop here could help distinguish decision errors from reasoned preferences in a wide range of economic settings aside from investment decision making and for other kinds of preferences besides risk preferences (e.g. time preferences, elasticity of intertemporal substitution). Evidence on time preferences suggests equally important inconsistencies exist among differently-framed questions. For example, individuals are more patient between two future dates than between today and a future date (Strotz, 1955; Thaler, 1981). Estimates of the elasticity of intertemporal substitution range widely from about zero (Hall, 1988; Dynan, 1993) to well over 1.0 (Mulligan, 2002; Gruber, 2006).

References

Arrow, Kenneth J. 1983. "Behavior under uncertainty and its implications for policy." In Stigum, B.P., and F. Wenstøp (eds.), Foundations of Utility and Risk Theory with Applications. The Netherlands, Reidel.

Barsky, Robert B., F. Thomas Juster, Miles S. Kimball, and Matthew D. Shapiro 1997. "Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study." *Quarterly Journal of Economics*, 112(2): 537-579.

Benartzi, Shlomo, and Richard H. Thaler. 1999. "Risk aversion or myopia? Choices in repeated gambles and retirement investments." *Management Science*, 45(3): 364-381.

Benjamin, Daniel J., Sebastian A. Brown, and Jesse M. Shapiro. 2006. "Who is 'behavioral'? Cognitive ability and anomalous preferences." Harvard University mimeo, May.

Bernheim, B. Douglas, and Antonio Rangel. 2008. "Beyond revealed preference: choice theoretic foundations for behavioral welfare economics." No. w13737. National Bureau of Economic Research, 2008.

Berndt, Ernst R., et al. 1974. "Estimation and inference in nonlinear structural models." Annals of Economic and Social Measurement, NBER, 3(4): 653-665.

Beshears, John, James J. Choi, David I. Laibson, and Brigitte C. Madrian. 2008a. "The importance of default options for retirement saving outcomes: Evidence from the United States." In Kay, Stephen J., and Tapen Sinha (eds.), Lessons from Pension Reform in the Americas.

Beshears, John, James J. Choi, David I. Laibson, and Brigitte C. Madrian. 2008b. "How are preferences revealed?" *Journal of Public Economics*, 92: 1787-1794.

Beshears, John, James J. Choi, David I. Laibson, and Brigitte C. Madrian. 2009. "Can psychological aggregation manipulations affect risk-taking? Evidence from a framed field experiment." Harvard University mimeo, February.

Burks, Stephen, Jeffrey Carpenter, Lorenz Goette, and Aldo Rustichini. 2008. "Cognitive skills explain economic preferences, strategic behavior and job attachment." IZA Discussion Paper No. 3609, July.

Cacioppo, John T., and Richard E. Petty. 1984. "The need for cognition: Relationship to attitudinal processes." Social perception in clinical and counseling psychology 2 (1984): 113-140.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde. 2008. "Are risk aversion and impatience related to cognitive ability?" IZA Discussion Paper No. 2735, April.

Druckman, James N. 2001. "Evaluating framing effects." *Journal of Economic Psychology* 22.1: 91-101.

Dynan, Karen E. 1993. "How Prudent Are Consumers?" *Journal of Political Economy*, 101(6): 1104-1113.

Fagerlin, Angela, Brian J. Zikmund-Fisher, Peter A. Ubel, Aleksandra Jankovic, Holly A. Derry, and Dylan M. Smith. 2007. "Measuring numeracy without a math test:

Development of the Subjective Numeracy Scale." Medical Decision Making, 27: 672-680.

Frederick, Shane. 2005. "Cognitive reflection and decision making." *Journal of Economic perspectives* (2005): 25-42.

Fuchs, Victor R. 1982. "Time preference and health: An exploratory study." In Victor R.Fuchs (ed.), Economic aspects of health, chapter 3, pp. 93-120. The University of Chicago Press, Chicago.

Fudenberg, Drew, and David Levine. 2006. "A Dual Self Model of Impulse Control," *American Economic Review*, 96: 1449-1476.

Gentzkow, Matthew, and Jesse M. Shapiro. 2014. "Measuring the sensitivity of parameter estimates to sample statistics." No. w20673. National Bureau of Economic Research.

Gosling, Samuel D., Peter J. Rentfrow, and William B. Swann. 2003. "A very brief measure of the Big-Five personality domains." *Journal of Research in Personality* 37.6 (2003): 504-528.

Gruber, Jonathan. 2006. "A Tax-Based Estimate of the Elasticity of Intertemporal Substitution," NBER Working Paper No. 11945.

Gul, Faruk, and Wolfgang Pesendorfer. 2005. "The Revealed Preference Theory of Changing Tastes." *Review of Economic Studies*, 72: 429-448.

Haliassos, Michalis and Carol Bertaut. 1995. "Why Do So Few Hold Stocks?" *The Economic Journal*, 195: 1110-29.

Hall, Robert. 1988. "Intertemporal Substitution in Consumption." *Journal of Political Economy*, 96(2): 339-357.

Huck, Steffan, and Wieland Müller. 2008. "Allais for all: Revisiting the paradox." University College London mimeo, November 3.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect theory: An analysis of decision under risk." *Econometrica*, 47: 263-291.

Kimball, Miles S., Claudia R. Sahm, and Matthew D. Shapiro. 2008. "Imputing risk tolerance from survey responses." *Journal of the American Statistical Association*, 103.483 (2008): 1028-1038.

Koszegi, Botond, and Matthew Rabin. 2006. "A model of reference-dependent preferences." *Quarterly Journal of Economics,* 121(4): 1133-1166.

Laibson, David I. 1997. "Golden eggs and hyperbolic discounting," *Quarterly Journal of Economics*, 112(2): 443-477.

Loewenstein, George, and Nachum Sicherman. 1991. "Do workers prefer increasing wage profiles?" *Journal of Labor Economics*, 9(1): 67-84.

Loewenstein, George, and Drazen Prelec. 1992. "Anomalies in intertemporal choice: Evidence and an interpretation," *Quarterly Journal of Economics*, 107(2): 573-597.

Loewenstein, George, and Ted O'Donoghue. 2007. "The Heat of the Moment: Modeling Interactions Between Affect and Deliberation," Cornell University mimeo, June.

McArdle, John, Willard Rodgers, and Robert Willis. 2015. "Cognition and Aging in the USA (CogUSA) 2007-2009." ICPSR36053-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2015-04-16.

McNeil, Barbara J., Stephen G. Pauker, and Amos Tversky. 1998. "On the framing of medical decisions."

Merton, Robert C. 1969. "Lifetime portfolio selection under uncertainty: the continuous time case." *Review of Economics and Statistics*, 51: 247-257.

Mulligan, Casey B. 2002. "Capital, Interest, and Aggregate Intertemporal Substitution," NBER Working Paper No. 9373.

Poterba, James M., Joshua Rauh, Steven F. Venti, and David A. Wise. 2005. "Utility Evaluation of Risk in Retirement Saving Accounts," in Wise, David A. (ed.), Analyses in the Economics of Aging. Chicago: University of Chicago Press.

Rabin, Matthew. 2000. "Risk aversion and expected-utility theory: A calibration theorem." *Econometrica*, 68(5): 1281-92.

Raiffa, Howard. 1968. Decision Analysis. Reading, MA: Addison-Wesley.

Strotz, Robert. 1955. "Myopia and inconsistency in dynamic utility maximization," *Review of Economic Studies*, 23(3): 165-180.

Thaler, Richard H. 1981. "Some empirical evidence on dynamic inconsistency," Economic Letters, 8: 201-207.

Thaler, Richard H., and Hersh M. Shefrin. 1981. "An economic theory of self-control," *Journal of Political Economy*, 89(2): 392-410.

TIAA-CREF website as of August 4, 2008.

Tversky, Amos, and Daniel Kahneman. 1981. "The framing of decisions and the psychology of choice." *Science*, 211: 453-58.

Wiseman, David B., and Irwin P. Levin. 1996. "Comparing risky decision making under conditions of real and hypothetical consequences." *Organizational Behavior and Human Decision Processes*, 66(3): 241-250.

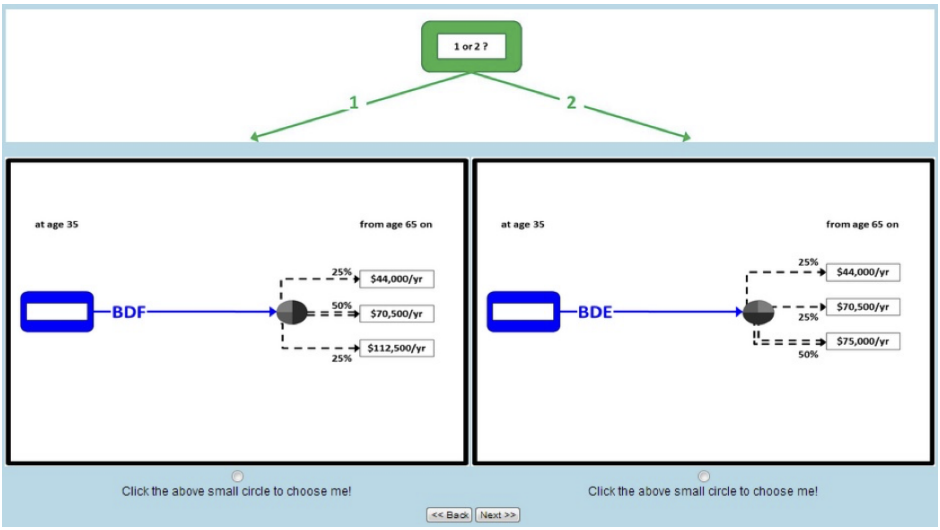


Figure 1: Example question from Pairwise Choices between Compound Lotteries (10 questions total). BDE versus BDF shown here. 9 other questions not pictured.

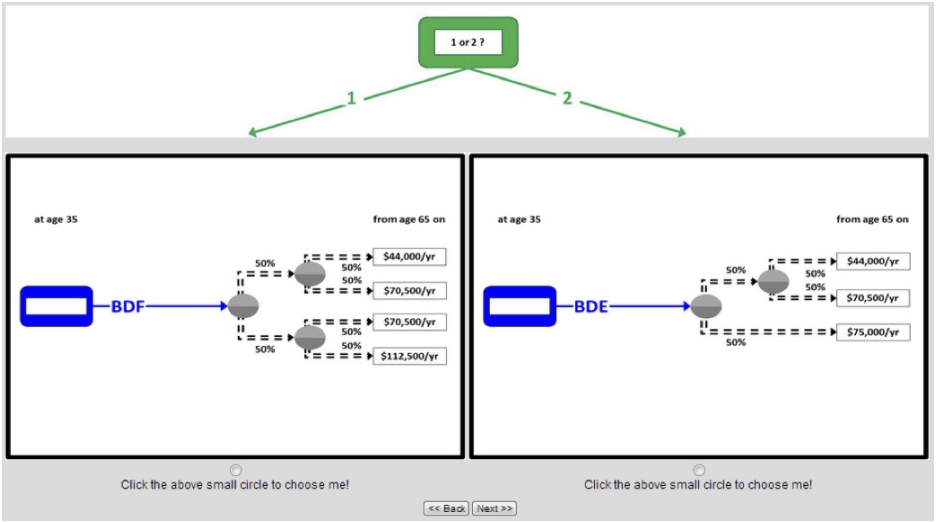


Figure 2: Example question from Pairwise Choices between Reduced Simple Lotteries (10 questions total). BDE versus BDF shown here. 9 other questions not pictured.

Portfolio Risk Aversion \ True Risk Aversion	$\gamma = 1$	$\gamma = 2$	$\gamma = 4$	$\gamma = 8$
$\hat{\gamma} = 1$	0%	14%	56%	94%
$\hat{\gamma} = 2$	6%	0%	07%	29%
$\hat{\gamma} = 4$	13%	3%	0%	3%
$\hat{\gamma} = 8$	17%	7%	2%	0%

Table 1: Welfare cost of investing with the "wrong" risk preferences. Wealth loss (in percent) that would generate the same loss in expected utility as the mistaken behavior. True risk aversion, γ, is along the columns. Risk aversion used for portfolio choice, $\hat{\gamma}$, is along the rows.

Emotion	**1 (least)**	**2**	**3**	**4**	**5**	**6 (most)**
Enjoyment	11.7%	18.2%	25.4%	32.6%	9.0%	3.2%
Annoyance	4.5%	17.0%	23.9%	23.7%	18.4%	12.5%
Stress	23.7%	22.2%	20.4%	19.4%	10.7%	3.7%
Frustration	15.2%	21.2%	22.4%	24.0%	10.7%	6.5%

Table 2: After wave 1, percentage of respondents answering on a scale of 1-6: "To what extent do you agree with the following statement": "I enjoyed thinking through these choices," "Thinking through these choices was annoying," "Thinking through these choices made me feel stressed," "Thinking through these choices was frustrating."

Sample	Dates	N	Details
Wave 1 Sample	July-December 2013	317	Subjects were only invited to a single experimental session (wave 1 only). Questions about why a person updated choices as they did were open-ended.
Wave 1+2 Sample	April-May 2014	284	Subjects were initially invited to one session (wave 1), then all invited back for a second session (wave 2). 246 returned for the second wave. Questions about why a person updated as they did were multiple-choice.

Table 3: Timeline of data collection, including dates when the samples were collected, sample size (N), and details of the associated surveys.

	ML_1	ML_2	ML_3	ML_4	ML_5	ML_6
MA_1	52K	64K	74K	81K	88K	92K
MA_2	72K	80K	86K	90K	94K	96K
MA_3	100K	100K	100K	100K	100K	100K
MA_4	108K	120K	129K	135K	141K	144K
MA_5	150K	150K	150K	150K	150K	150K
MA_6	225K	225K	225K	225K	225K	225K

Table 4: Columns denote 6 possible monetary levels (ML), each consisting of 6 monetary amounts (MA). Sets are used as lottery payoffs throughout the experiment and vary the attractiveness of the risky options. Each ML is associated with a different level of constant relative risk aversion (CRRA): 1.576, 2.958, 4.865, 7.184, 12.113, or 17.967. The 3rd, 5th, and 6th amounts are always 100k, 150k, and 225k, respectively. The 2nd amount is the dollar value such that an expected utility maximizer with CRRA preferences (given the constant coefficient of relative risk aversion associated with that ML) would be indifferent between 100k for sure and a 50-50 chance of 150k and the unknown amount. Finally, the 1st amount is the 2nd amount squared and divided by 100. Within a wave, a participant was always asked questions drawing from the same ML (i.e. monetary amounts are held fixed for a given subject and wave). Subjects who returned for a second wave received a re-randomized version of the survey, thereby potentially receiving a different ML.

(Normative) Axiom	**Frame 1**	**Frame 2**
Irrelevance of Background Counterfactuals	Single Action in Isolation	Single Action with Backdrop
Simple Actions = State-Contingent Actions	Single Action with Backdrop	Two Contingent Actions with Backdrop
Irrelevance of Counterfactual Choices	Two Contingent Actions with Backdrop	Complete Contingent Action Plan
Fusion + Shift from Node-wise to Pairwise	Complete Contingent Action Plan	Pairwise Choices between Complete Strategies
Complete Strategies = Implied Lotteries	Pairwise Choices between Complete Strategies	Pairwise Choices Between Compound Lotteries
Reduction of Compound Lotteries	Pairwise Choices Between Compound Lotteries	Pairwise Choices Between Reduced Simple Lotteries

Table 5: Name of each (normative) axiom and the two frames that are equivalent according to it.

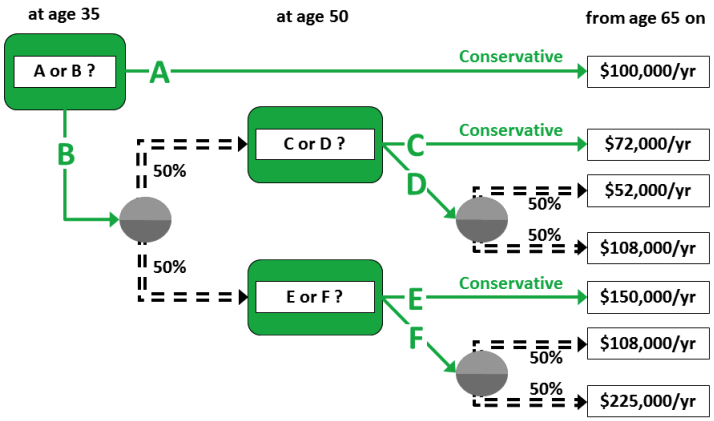


Figure 3: Single decision problem underlying all 7 frames, also called Complete Contingent Action Plan. It constitutes our most basic frame. Subjects are prompted with instructions: "Imagine you are currently 35 years old. You need to make three decisions. The decision between A and B takes place now. If you choose B, you also need to make two decisions that will lock in how you will invest at age 50. If you choose A, you do not need to make any more decisions. All decisions will affect how much money you will be able to spend each year during retirement (from age 65 on). In each decision, you will choose between two strategies: risky or conservative. Each conservative strategy will guarantee you a fixed amount to spend each year during retirement. Each risky strategy allows for possibly higher amounts. You do not need to choose the same kind of strategy in each decision."

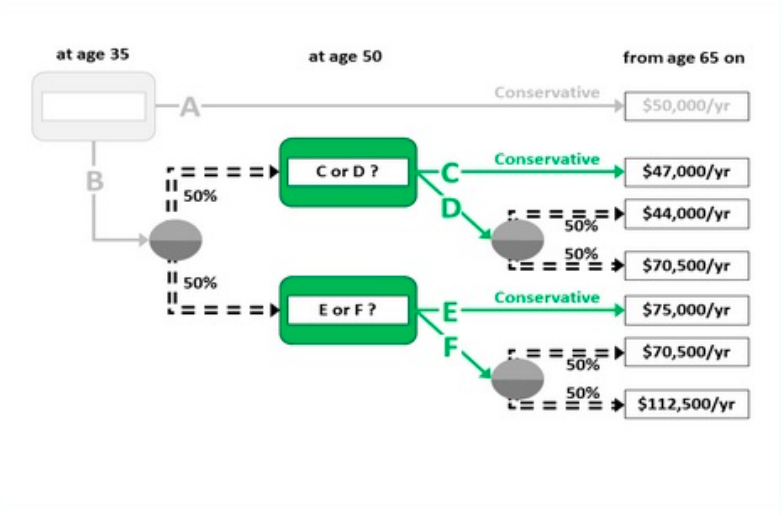


Figure 4: Two Contingent Actions with Backdrop (1 question). Subjects are prompted with the same instructions as Complete Contingent Action Plan except for the following initially: "Imagine you are currently 35 years old, and have chosen risky decision B. You do not yet know how this decision has turned out. So, you need to make two decisions that will lock-in how you will invest at age 50."

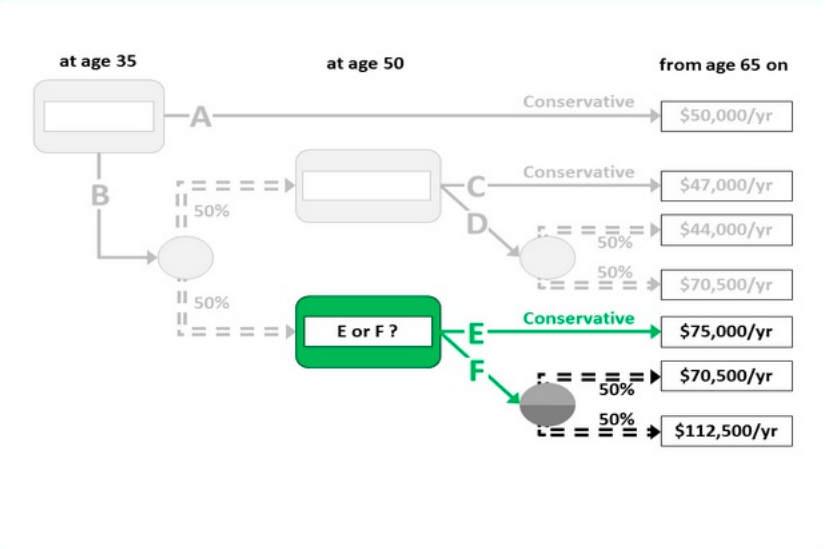


Figure 5: Single Action with Backdrop (2 questions). E versus F shown here. C versus D is the other question. Subjects are prompted with the same instructions as Two Contingent Actions with Backdrop except for the following: "[Imagine] that you are currently 50 years old. You will also be given information about how decisions you made when you were 35 turned out, that are beyond your control at this point. These grayed-out parts of the picture are things that could have happened, but you know for sure did not happen."

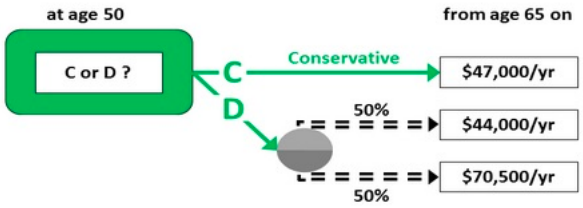


Figure 6: Single Action in Isolation (2 questions). C versus D shown here. E versus F is the other question.

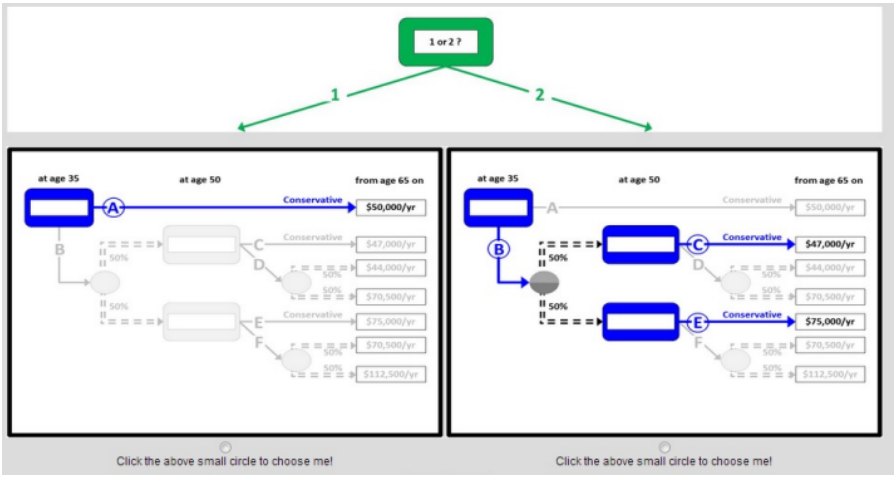


Figure 7: Example question from Pairwise Choices Between Complete Strategies (10 questions total). A versus BCE shown here. 9 other questions not pictured.

Frame	Round	A > BCE	A > BCF	A > BDE	A > BDF	BCE > BCF	BCE > BDE	BCE > BDF	BCF > BDE	BCF > BDF	BDE > BDF	N
Complete	U1	23%	16%	25%	17%	16%	40%	33%	68%	40%	21%	591
Compound	U1	22%	18%	24%	16%	14%	40%	31%	67%	36%	20%	590
Reduced	U1	21%	27%	23%	22%	24%	37%	34%	43%	24%	37%	590
Complete	U2	32%	24%	36%	30%	19%	45%	36%	77%	46%	24%	244
Compound	U2	28%	22%	35%	30%	20%	45%	36%	77%	44%	20%	246
Reduced	U2	28%	30%	33%	34%	23%	43%	38%	59%	39%	36%	245

Table 6: Untutored preferences for pairwise frames. "Complete" refers to Pairwise Choices between Complete Strategies; "Compound" refers to Pairwise Choices Between Compound Lotteries; "Reduced" refers to Pairwise Choices Between Reduced Simple Lotteries. Entries are the percentage choosing the *first* option in the pair (e.g. for "A > BCE" the percentage choosing "A"). Top panel: all participants' wave 1 untutored responses (round "U1"). Bottom panel: wave 1+2 participants' wave 2 untutored responses (round "U2").

Frame	A	BCE	BCF	BDE	BDF	N
Single Action in Isolation	-	12%	24%	10%	54%	592
Single Action with Backdrop	-	14%	27%	7%	52%	592
Two Contingent Actions with Backdrop	-	12%	29%	9%	50%	589
Complete Contingent Action Plan*	-	6%*	30%*	9%*	55%*	575
Complete Contingent Action Plan	16%	5%	25%	8%	46%	575
Pairwise Choices B/w Complete Strats.	6%	5%	29%	8%	52%	476
Pairwise Choices B/w Compound Lotts.	8%	3%	28%	8%	53%	485
Pairwise Choices B/w Reduced Lotts.	9%	7%	16%	21%	47%	471
Single Action in Isolation	-	22%	25%	9%	44%	245
Single Action with Backdrop	-	23%	23%	8%	46%	246
Two Contingent Actions with Backdrop	-	20%	27%	9%	44%	245
Complete Contingent Action Plan*	-	7%*	26%*	7%*	60%*	242
Complete Contingent Action Plan (cont.)	-	44%	34%	11%	11%	64
Complete Contingent Action Plan	28%	5%	19%	5%	43%	242
Pairwise Choices B/w Complete Strats.	19%	3%	30%	5%	43%	220
Pairwise Choices B/w Compound Lotts.	17%	2%	28%	6%	47%	219
Pairwise Choices B/w Reduced Lotts.	19%	4%	27%	14%	36%	207

Table 7: Untutored preferences: nodewise frames and winningest responses for pairwise frames. Top panel: all participants' wave 1 untutored responses. Bottom panel: wave 1+2 participants' wave 2 untutored responses. Top panel has two rows for Complete Contingent Action Plan: (1) unstarred row details percentage choosing each of the 5 investment plans (i.e. favorite choices), and (2) starred row details percentages among investment plans BCE, BCF, BDE, and BDF (also starred) re-normalized to sum to 100% (i.e. excluding A, for comparison with other nodewise frames). Bottom panel has an additional row: Completely Contingent Action plan (cont.), which details contingency responses for those wave 1+2 participants who chose "A" and who were then prompted for their 2nd favorite (i.e. contingency) choice among the remaining 4 investment plans. Pairwise frames detail the percentage of respondents who clearly favor each investment plan among those participants who have an investment plan which is preferred strictly more times than any other investment plan (i.e. winningest plans).

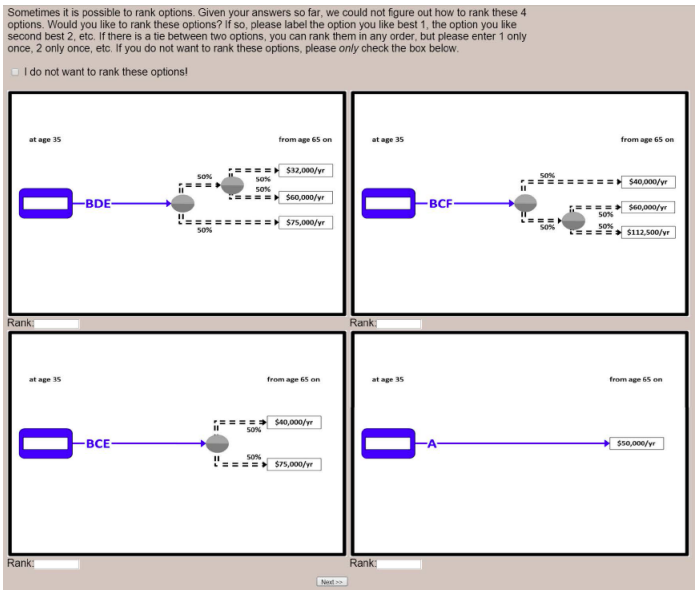


Figure 8: Example of an intransitivity resolution for the frame Pairwise Choices Between Reduced Simple Lotteries. Pictured here is a 4-way intransitivity among A, BCE, BCF, and BDE, i.e. the participant had cyclical preferences among these investment plans. The participant may either select "I do not want to rank these options!" or explicitly rank them by entering 1 for their favorite choice, 2 for their second favorite choice, etc.

Stage of Survey	Intransitivities	N
Wave 1, untutored	1.183	229
Wave 1, after 1st inconsistency check	1.048	229
Wave 1, after 1st intransitivity check	0.210	229
Wave 1, after 2nd inconsistency check	0.432	229
Wave 1, after 2nd intransitivity check	0.135	229
Wave 2, untutored	0.742	229
Wave 2, after 1st inconsistency check	0.690	229
Wave 2, after 1st intransitivity check	0.218	229
Wave 2, after 2nd inconsistency check	0.332	229
Wave 2, after 2nd intransitivity check	0.140	229

Table 8: Average number of intransitivities among the 3 pairwise frames. Only includes participants in the wave 1+2 sample who actually return for both waves. N denotes sample size of these participants, including only those having answered all pairwise questions throughout the survey.

Stage of Survey	Intransitivities	N
Untutored	1.070	584
After 1st inconsistency check	0.875	584
After 1st intransitivity check	0.228	584
After 2nd inconsistency check	0.361	584
After 2nd intransitivity check	0.166	584

Table 9: Average number of intransitivities among the 3 pairwise frames. Includes all participants' data from wave 1 (i.e. combining the wave 1 sample and the wave 1+2 sample). N denotes sample size of these participants, including only those having answered all pairwise questions throughout the survey.

Axiom	# Pot. Incon.	U1	R1	U2	R2	P-value U1-R1	P-value U1-U2	P-value U2-R2	N
Irrelevance of Background Counterfactuals	2	12.5%	5.8%	-	-	<0.0005	-	-	592
Simple Actions = State-Contingent Actions	2	12.0%	8.2%	-	-	0.003	-	-	589
Irrelevance of Counterfactual Choices	1	18.7%	19.2%	-	-	0.749	-	-	572
Fusion + Shift from Nodewise to Pairwise	4	23.4%	12.3%	-	-	<0.0005	-	-	573
Complete Strategies = Implied Lotteries	10	19.6%	8.3%	-	-	<0.0005	-	-	586
Reduction of Compound Lotteries	10	23.1%	8.3%	-	-	<0.0005	-	-	586
Irrelevance of Background Counterfactuals	2	12.9%	5.9%	8.5%	3.8%	0.0002	0.027	0.003	236
Simple Actions = State-Contingent Actions	2	12.3%	9.1%	6.1%	6.1%	0.114	0.001	1.000	236
Irrelevance of Counterfactual Choices	1	20.8%	25.8%	17.6%	21.7%	0.080	0.232	0.128	221
Fusion + Shift from Nodewise to Pairwise	4	24.6%	14.6%	17.8%	11.0%	<0.0005	0.0004	<0.0005	225
Complete Strategies = Implied Lotteries	10	22.2%	10.1%	14.4%	8.1%	<0.0005	<0.0005	<0.0005	232
Reduction of Compound Lotteries	10	26.3%	10.6%	18.2%	8.7%	<0.0005	<0.0005	<0.0005	231

Table 10: Average inconsistency rates by axiom = (average number of inconsistencies) / (number of potential inconsistencies) at a given point in the survey. First 6 lines correspond to all participants (where N only includes those participants who are not missing any wave 1 data for the two frames constituting each axiom). Second 6 lines correspond only to the wave 1+2 sample who return for both waves (where N only includes those participants who are not missing any data from either wave within each axiom). P-values are from two-sided tests for differences in proportions between wave 1 untutored ("U1") and reasoned ("R1") inconsistency rates, wave 1 untutored ("U1") and wave 2 untutored ("U2") inconsistency rates, and wave 2 untutored ("U2") and reasoned ("R2") inconsistency rates.

Axiom	Toward Frame 1	Toward Frame 2	P-Value 1 = 2	No Update	Swap or Both	Total
Irrelevance of Background Counterfactuals	24.1%	24.6%	0.917	48.7%	2.7%	187
Simple Actions = State-Contingent Actions	20.8%	12.0%	0.039	62.8%	4.4%	183
Irrelevance of Counterfactual Choices	20.6%	28.9%	0.060	43.9%	6.7%	253
Fusion + Shift from Nodewise to Pairwise	21.5%	16.9%	0.121	56.0%	5.5%	433
Complete Strategies = Implied Lotteries	21.2%	26.1%	0.008	48.1%	4.5%	1429
Reduction of Compound Lotteries	27.8%	20.7%	<0.0001	45.9%	5.6%	1898
Irrelevance of Background Counterfactuals	0%	0%	-	100.0%	0%	482
Simple Actions = State-Contingent Actions	0%	0%	-	100.0%	0%	427
Irrelevance of Counterfactual Choices	0%	0.61%	0.319	99.39%	0%	164
Fusion + Shift from Nodewise to Pairwise	1.20%	3.30%	0.071	94.89%	0.60%	333
Complete Strategies = Implied Lotteries	0.50%	0.85%	0.178	98.11%	0.55%	2007
Reduction of Compound Lotteries	0.89%	0.47%	0.095	97.80%	0.84%	2134

Table 11: How wave 1+2 subjects (across both waves) updated when presented with normal or placebo inconsistency checks (first and last 6 rows, respectively). Percentage of the time that subjects (1) updated toward frame 1, (2) updated toward frame 2, (3) did not update, or (4) swapped their choices (for normal inconsistencies) or updated both choices (for placebo inconsistencies). "P-value 1=2" denotes the significant of a T-test of equality between the percentage updating toward frame 1 and the percentage updating toward frame 2. Total denotes total such opportunities to update for all wave 1+2 subjects, including both waves and across all points of the survey.

Axiom	Different situations	Indiff.	Expt'er Demand	IDK	Confused	Other	Total
Irrelevance of Background Counterfactuals	57.1%	20.9%	2.2%	8.8%	8.8%	2.2%	91
Simple Actions = State-Contingent Actions	74.1%	17.2%	1.7%	1.7%	4.3%	0.9%	116
Irrelevance of Counterfactual Choices	56.8%	24.3%	1.8%	9.0%	3.6%	4.5%	111
Fusion + Shift from Nodewise to Pairwise	62.7%	20.5%	1.6%	7.4%	4.1%	3.7%	244
Complete Strategies = Implied Lotteries	55.8%	24.8%	3.3%	7.2%	2.9%	5.9%	690
Reduction of Compound Lotteries	54.5%	27.8%	2.9%	4.7%	3.9%	6.2%	873
Total	57.1%	25.0%	2.7%	6.1%	3.8%	5.3%	2130

Table 12: Percentage of the time wave 1+2 subjects (across both waves and all points of the survey) that subjects gave each of the following responses after not updating an inconsistency: "Why do you want to make different choices in these two situations?": "The two situations are different enough that I want different choices", "Some of the options are equally good to me, so it doesn't matter which one I choose", "I chose how I thought the experimenters wanted me to choose", "I don't know which options I prefer", "I don't know or am confused", or "Other." Total denotes total such opportunities to update.

Axiom	Mistake	Learned	Indiff.	Expt'er Demand	IDK	Confused	Other	Total
Irrelevance of Background Counterfactuals	48.4%	34.1%	12.1%	1.1%	1.1%	1.1%	2.2%	91
Simple Actions = State-Contingent Actions	36.1%	37.7%	18.0%	0.0%	3.3%	1.6%	3.3%	61
Fusion + Shift from Nodewise to Pairwise	47.0%	31.9%	13.3%	1.2%	4.2%	2.4%	0.0%	166
Complete Strategies = Implied Lotteries	46.2%	35.4%	10.2%	0.7%	4.7%	0.9%	1.9%	678
Reduction of Compound Lotteries	44.6%	37.3%	12.0%	0.8%	2.4%	1.3%	1.7%	926
Total	45.3%	36.0%	11.7%	0.8%	3.3%	1.2%	1.7%	1922

Table 13: Percentage of the time wave 1+2 subjects (across both waves and all points of the survey) that subjects gave each of the following responses after updating an inconsistency: "Why did you want to change your choices as you did?": "I made a mistake when I first chose", "Answering all of these questions made me change what I want", "Some of the options are equally good to me, so it doesn't matter which one I choose", "I chose how I thought the experimenters wanted me to choose", "I don't know which options I prefer", "I don't know or am confused", or "Other." Total denotes total such opportunities to update.

Axiom	Same Situations	Indiff.	Expt'er Demand	IDK	Confused	Other	Total
Irrelevance of Background Counterfactuals	90.3%	7.5%	0.2%	0.9%	0.6%	0.4%	466
Simple Actions = State-Contingent Actions	85.6%	10.5%	0.2%	1.9%	1.0%	0.7%	411
Irrelevance of Counterfactual Choices	84.7%	9.8%	0.0%	0.0%	1.8%	3.7%	163
Fusion + Shift from Nodewise to Pairwise	91.5%	7.3%	0.0%	0.6%	0.3%	0.3%	316
Complete Strategies = Implied Lotteries	89.4%	6.4%	0.5%	1.2%	1.2%	1.3%1970	
Reduction of Compound Lotteries	90.7%	6.6%	0.3%	0.8%	1.0%	0.6%	2084
Total	89.7%	7.0%	0.4%	1.0%	1.0%	0.9%	5410

Table 14: Among wave 1+2 subjects, percentage of the time across both waves that subjects gave each of the following responses after not revising a placebo inconsistency: "Why do you want to make the same choices in these two situations?": "The two situations are similar enough that I want to make the same choices", "Some of the options are equally good to me, so it doesn't matter which one I choose", "I chose how I thought the experimenters wanted me to choose", "I don't know which options I prefer", "I don't know or am confused", or "Other." Total denotes total such opportunities to update (across all points of the survey and waves).

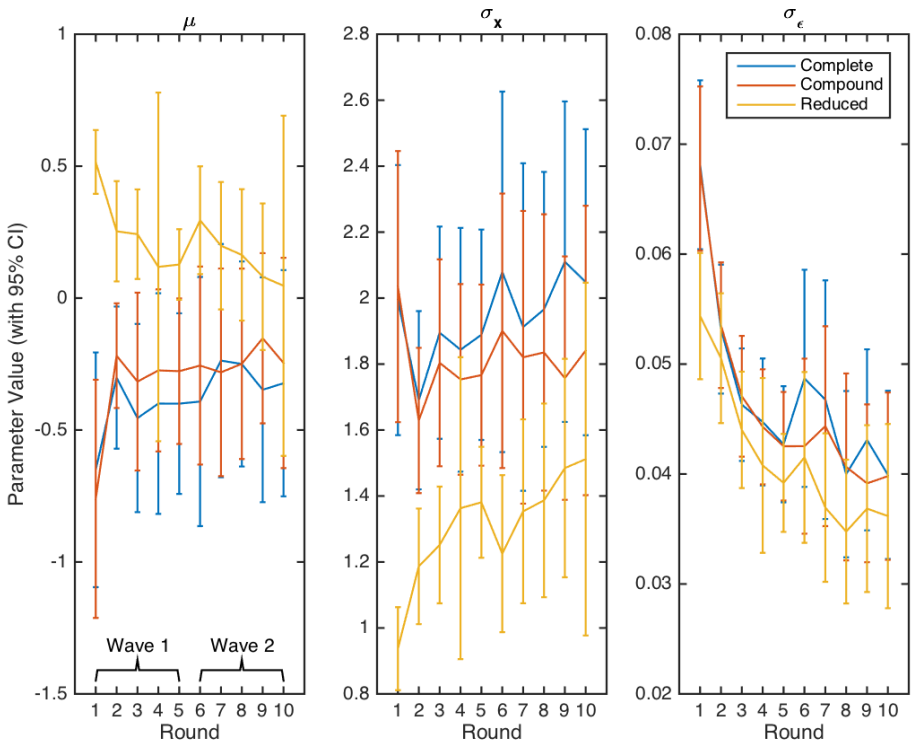


Figure 9: MLE results for Pairwise Strategy Choices Frames. Wave 1 consists of rounds 1-5 and wave 2 consists of rounds 6-10. All available data are used for each round and frame. Error bars denote 95% confidence intervals. "Complete" refers to Pairwise Choices Between Complete Strategies; "Compound" refers to Pairwise Choices Between Compound Lotteries; and "Reduced" refers to Pairwise Choices Between Reduced Simple Lotteries.