NBER WORKING PAPER SERIES

RECONSIDERING RISK AVERSION

Daniel J. Benjamin
Mark Alan Fontana
Miles S. Kimball

Working Paper 28007
http://www.nber.org/papers/w28007

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2020,Revised September 2021

For financial support, we are grateful to NIH/NIA through grants R21-AG037741 to Cornell University, T32-AG00186 to the NBER, R01-AG020717 to the RAND Corporation, P01-AG026571 and R01-AG040787 to the University of Michigan, R01-AG051903 to UCLA, and R01-AG065364 to Hebrew University. We thank Mike Gideon for helpful early conversations, the RAND Corporation for access to MMIC and associated storage, Fudong Zhang for MATLAB guidance, and Bas Weerman for MMIC guidance. For helpful comments, we thank Zachary Breig, Jamie Druckman, Paul Feldman, João Ferreira, Alexia Gaudeul, Mike Gideon, Ori Heffetz, Eric Johnson, Ben Lockwood, Ted O'Donoghue, David Laibson, Kirby Nielsen, David Plunkett, Matthew Rabin, Daniel Reck, John Rehbeck, Valerie Reyna, Claudia Sahm, William Schulze, Bob Willis, participants at the Cornell Behavioral Economics Lab Meeting; seminar participants at the Consumer Financial Protection Bureau, Harvard University, Stanford University, University of Colorado Boulder, University of Michigan, University of Southern California, University of Texas at Dallas, University of Toronto, University of Warwick, and Yale School of Management; and conference participants at the Bounded Rationality In Economics Conference, China Greater Bay Area Experimental Economics Workshop, NBER Summer Institute, North American Economic Science Association Meeting, Normative Ethics and Welfare Economics Conference, 68° North Conference on Behavioral Economics, and the Stanford Institute for Theoretical Economics. We thank Tanner Bangerter, Shengmao Cao, Regina Chan, Kevin Coughlin, Christian Covington, Samantha Cunningham, Jakina Debnam, Alina Dvorovenko, Dan Golland, Hyun Gu, Hui Dong He, Duk Gyoo Kim, Sunsiree Kosindesha, Raymond Lee, Sichun Liu, Rebecca Royer, Michael West, Nancy Wang, and especially Xing Guo, Jordan Kimball, Derek Lougee, Tuan Nguyen, Yeo Yu (Leo) She, Andrew Sung, Fudong Zhang, and Jiannan Zhou for excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Reconsidering Risk Aversion
Daniel J. Benjamin, Mark Alan Fontana, and Miles S. Kimball
NBER Working Paper No. 28007
October 2020, Revised September 2021
JEL No. D63,D81,G11,H8

ABSTRACT

Risk aversion is typically inferred from real or hypothetical choices over risky lotteries, but such "untutored" choices may reflect mistakes rather than preferences. We develop a procedure to obtain a better measure of normatively relevant preferences: after eliciting untutored choices, we confront participants with their choices that are inconsistent with intertemporal-expected-utility axioms and allow them to reconsider their choices. We demonstrate this procedure via a survey about hypothetical retirement investment choices administered to 596 Cornell students. We find that, on average, reconsidered choices are more consistent with almost all axioms, with one exception related to a counterfactual reference point.

Daniel J. Benjamin
University of California Los Angeles
Anderson School of Management
and David Geffen School of Medicine
110 Westwood Plaza
Entrepreneurs Hall Suite C515
Los Angeles, CA 90095
and NBER
daniel.benjamin@anderson.ucla.edu

Miles S. Kimball
Department of Economics
University of Colorado
256 UCB
Boulder, CO 80309
and NBER
miles.kimball@colorado.edu

Mark Alan Fontana
Center for the Advancement of Value in Musculoskeletal Care
Hospital for Special Surgery
535 E. 70th Street
New York, NY 10021
and Weill Cornell Medical College
fontanam@nber.org

A data appendix is available at http://www.nber.org/data-appendix/w28007

I. Introduction

Policymakers, economists, and the popular media have long been worried that Americans may not be investing appropriately in preparation for retirement (e.g., Benartzi & Thaler 1999; Campbell 2006). 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 (e.g., Thaler & Sunstein 2009). 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 individuals' risk preferences. Economists' usual approach to measuring risk preferences is to infer them from real or hypothetical choices over risky lotteries.[1] But since risky decision making (even if simplified) is unfamiliar and cognitively challenging for most people, there is reason for concern that people's untutored choices may not accurately reflect their normatively relevant preferences. Moreover, people's choices often violate basic, seemingly compelling axioms of expected utility theory, such as the Reduction of Compound Lotteries axiom. While there are economic models that can accommodate such non-expected-utility 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 mistakes rather than normatively relevant preferences.

For both conceptual and practical reasons, reliably eliciting normatively relevant preferences is inherently challenging, and any approach for attempting to do so will have limitations. It is therefore valuable to develop a variety of approaches with different strengths and weaknesses. Taken together, different approaches can provide complementary evidence, with consistency in findings across approaches increasing confidence in the conclusions about preferences (for related discussion, see Beshears, Choi, Laibson, & Madrian 2008).

In this paper, we propose and develop a new, two-phase approach to measuring risk preferences that are more normatively relevant than untutored choices, and conduct a proof-of-concept implementation of the approach via a survey administered to a sample of 596 Cornell students. The first phase is the standard method of eliciting choices over risky lotteries. We used

[1] When financial advisors make their portfolio-allocation advice contingent on an individual's risk attitudes, they typically measure the individual's relative ranking in the population using assessment methods that have no clear relationship to the formal risk aversion measures in economic theory such as relative risk aversion. According to economic theory, however, what is needed is the numerical value of the individual's risk-preference parameters (or at least the distribution of these numerical values in the population, so that an individual's relative ranking can be interpreted numerically). These numerical values would need to be elicited using real or hypothetical choices over risky lotteries, as discussed here.

hypothetical choices about investing for retirement. We call participants' behavior in this phase their *untutored choices* (which we think of as the experimental analog to real-world investment choices that may be mistaken). Our innovation is the second phase: We confronted participants with *intransitivities* in subsets of their untutored choices and with other *inconsistencies* in their untutored choices: their different responses to choices framed differently that should be the same according to normative axioms of intertemporal expected utility theory. We asked participants whether their untutored choices were mistaken, and if so, how they would like to reconsider their choices. We call participants' revised choices their *reconsidered choices*. The central limitation of all approaches that aim to help participants understand how their choices violate axioms, including ours, is that they inherently introduce experimenter demand effects (e.g., Zizzo 2010; de Quidt, Haushofer, & Roth 2018). Below we discuss several of the ways we attempted to minimize or measure these effects.

The philosophical premise underlying our procedure is that, relative to the untutored choices, the reconsidered choices are closer to the preferences an individual would hold under idealized conditions of full deliberation and absence of cognitive error.[2] The mental states an individual would hold under idealized conditions such as these are commonly considered to be relevant in philosophical accounts of normative phenomena, including well-being (e.g., Railton 1986; Goodin 1986), the nature of value (e.g., Smith, Lewis, & Johnston 1989), and reasons for action (e.g., Smith 1994). Put in terms of Bernheim's (2016) framework for behavioral welfare economics (which develops and extends Bernheim & Rangel 2009), the idealized conditions are conducive to better information processing, and therefore choices made under those conditions are "decisions that merit deference." Because it is implausible that our experiment fully achieves the idealized conditions, our premise further relies on an assumption that preferences, when formulated under conditions *closer* to the idealized conditions, are *more* normatively relevant. As in the theory of the second best (Lipsey & Lancaster 1956), this assumption may not hold, but by iterating the *reconsideration procedure*, our experiment provides evidence consistent with the

[2] As Hausman (2016) argues, the "attribution of a latent, context-independent preference to the agent" is theoretically useful for economists but does not presuppose that they truly exist. Purified, inner preferences are "an account, which the agent can affirm or dispute, of what matters most to this flesh-and-blood individual" (p. 28). Hausman points out, however, that postulating such normatively relevant preferences is epistemologically problematic if there is no way of learning about them from actual behavior. Our reconsideration procedure is one proposal for how we may learn about them, and we discuss others in Section VI. See also Ferreira (2020) for a philosophical defense of "confirmed choices" as a proxy for welfare.

assumption: iteration increases deliberation and creates additional opportunities to correct cognitive errors, thereby moving actual conditions closer to idealized conditions, and we find no evidence of non-monotonic changes of behavior within an experimental session.

We conducted a proof-of-concept test of our reconsideration procedure in a sample of 596 Cornell students, 237 of whom returned to the lab 2-4 weeks later for a second wave of the experiment. In each wave, the first phase of our experiment elicited untutored choices. These were 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 screens, shown in Figure 1A, participants chose between two compound lotteries, a riskier lottery ("BDF") pictured on the right and a safer lottery ("BDE") pictured on the left. In another screen, shown in Figure 1B, participants again made a choice between two lotteries—the "same" two lotteries according to the Reduction of Compound Lotteries axiom but framed as reduced simple lotteries. (The speech balloons in each figure show what percentage of participants made each choice initially and at the end of the experiment, after all stages of our reconsideration procedure.) As in much prior work on framing effects and non-expected utility (e.g., Kahneman & Tversky 1979), we find that participants often make inconsistent choices across different frames. In the Reduction of Compound Lotteries example, among those who participated in both waves of our experiment, 26.1% made inconsistent choices in their initial choices.

In each wave, the second phase of our experiment is the *reconsideration procedure*: we confronted participants with intransitivities and (other) inconsistencies in their untutored choices. Continuing the example of an inconsistency, suppose that in their untutored choices, a participant had answered one of the survey screens with "BDE" and the other with "BDF." We 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, then we gave the participant the option of changing one or both decisions.

As noted above, the key concern with an experimental design like ours is possible experimenter demand effects. In our experiment, a participant may infer from the question itself that he or she *should* change the earlier choice or, relatedly, should be making consistent choices. Our experimental design includes three main 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 the demand effect weaker than if we had only offered the option to change one choice. Similarly, when we offered participants the opportunity to rank options for which pairwise choices led to an intransitive cycle, we always offered the option of not ranking the options.

Second, roughly half the time that we offered participants the opportunity to change their choices, we selected pairs of choices that were already *consistent* with the relevant normative axiom. By doing so, we further weakened the demand effect and also obtained a placebo measure of how often people change their choices when prompted to do so. While participants who participated in both waves revised their untutored choices on average 46.0% of the time when these choices were inconsistent, they did so only 1.9% of the time in the placebo cases when the choices were originally consistent.

Third, we formulated axioms of intertemporal expected utility theory in a way that made each axiom transparently simple. This is an important innovation of our design. It is crucial that participants understand the axioms, but explaining the axioms to participants could strengthen demand effects. We broke down the Independence Axiom, which is complex to someone untrained in economic theory, into six (mostly non-standard but) easy-to-understand subcomponents, one of which is the standard Reduction of Compound Lotteries. These six axioms plus transitivity, together with completeness and continuity (which we assume but do not test), imply intertemporal expected utility theory. By making the axioms easy to understand, we aimed to make it easier for participants to see the appeal or lack of appeal of the axiom to them. In addition, to minimize misunderstandings based on lack of familiarity with probabilities, we only used probabilities of 50% and 25%, and we provided participants with basic probability training.

We divide axiom violations into *intransitivities* (which involve cyclic choices within a frame) and *inconsistencies* (all other axiom violations, all of which involve inconsistent choices across frames). We find that, on average across the six axioms we examine, in their initial untutored choices, those who participated in both waves exhibit 22.4% of the inconsistencies that would be possible (for example, inconsistent choices in the two frames defining the Reduction of Compound Lotteries axiom, depicted in Figures 1A and 1B). The reconsideration procedure led

to substantial movement overall toward endorsing the axioms. By the end of the second wave of the experiment, the average inconsistency rate fell to 8.4%. Respondents similarly exhibited a substantial reduction in intransitivities, from 38.1% of potential intransitivities in initial untutored choices to 5.3% at the end of the second wave. While we do not know if we would see further declines with additional stages of reconsideration, we interpret our results as suggesting that for most participants and for all the axioms except one, violations of the axioms in untutored choices reflect mistakes rather than normative preferences. The one exception is an axiom we call "Irrelevance of Background Counterfactuals," which is related to reference-dependent preferences with a counterfactual reference point. For that axiom, we find no evidence of a reduction in inconsistencies across multiple stages of reconsideration in both waves of the experiment.

In addition to these descriptive analyses suggesting that most of the potential deviations from expected utility that we study are mistakes, in the three frames where we have sufficient data, we conduct a structural estimation; this structural estimation assumes that preferences fully satisfy expected-utility theory and, more specifically, constant relative risk aversion (CRRA). It is motivated by two questions. First, if deviations from CRRA are conceptualized as response errors—as is typical when economists analyze (untutored) risky-choice data—how much does the reconsideration procedure reduce response-error variance? Second, how does the reconsideration procedure affect the average level of risk aversion? We estimate that from initial choices to the end of the second wave of the experiment, response-error variance declined by 70%. While our estimates of the average level of risk aversion are noisy, we find no evidence of a systematic change in risk aversion resulting from our reconsideration procedure.

This paper contributes to a growing literature on behavioral welfare economics (see Bernheim & Taubinsky 2018 for a review) and methods of using observed choices to draw inferences about normatively relevant preferences despite the influence of systematic errors (e.g., Beshears et al. 2008; Benkert & Netzer 2018; Goldin & Reck 2020). It is most closely related to other work—both an older literature pioneered by MacCrimmon (1968) and several recent papers: Gaudeul ® Crosetto (2019), Breig & Feldman (2021), and Nielsen & Rehbeck (2020)—that also aims to distinguish preferences from mistakes by prompting experimental participants

to deliberate more, for example, by offering them an opportunity to revise their choices.[3] In Section VI, we draw together different literatures that, we argue, can helpfully be understood through a common philosophical lens. We review the approaches and findings across the experiments that have been conducted, offer a unifying perspective on these literatures, and place this paper in context.

The rest of the paper is organized as follows. In Section II, we describe the sample, the risky choices posed to participants, and the experimental design. Sections III and IV present descriptive results. In Section V, we report our structural estimation of CRRA utility. Section VI provides a review and synthesis of related experimental approaches drawn from different literatures, and Section VII concludes. The Survey Appendix contains screenshots from the experiment. The Web Appendixes contain robustness analyses and other additional analyses.

II. Experimental Design

II.A. Participants and Session Procedure

We recruited 596 experimental participants from the subject pools of Cornell's LEEDR and Business Simulation laboratories. Of the participants, 65% were female. Mean age was 20.9 years, with approximately 90% between the ages of 18 and 22.

Figure 2 shows a timeline of the experiment. We explain the details of the figure in this and the next two subsections.

We initially collected data on 321 participants from July to December 2013. We collected data on 275 additional participants in April 2014. We refer to these two groups as the "version 1" and "version 2" samples because we modified the experiment slightly in the later data collection. In particular, in version 1, we elicited open-ended responses to questions about why a person revised their choices as they did or why they did not revise their choices (these responses are

[3] Less directly related but in a similar spirit to the experiments reviewed here, Zhang & Abaluck (2016) interview Medicare Part D enrollees, walking them through the Medicare.gov online plan-finder tool to find the lowest-cost health insurance coverage, and then assessing their interest in switching away from their current plan. They found that few enrollees wanted to switch plans because their plan had been chosen by someone else (e.g., another family member) or because they were satisfied enough with their current plan and were afraid the new plan might be worse for an unanticipated reason. Our paper is also related to Ambuehl, Bernheim, & Lusardi (2017), who present experimental participants with different framings of the same asset, where the frames should be irrelevant according to mathematical principles (e.g., $10 invested for 15 days at 6% interest per day compounded daily versus $24 in 15 days). Ambuehl, Bernheim, & Lusardi use the difference in willingness-to-pay across frames for the assets with the same dollar value as a measure of participants' mistakes.

listed in Web Appendix A). In version 2, we instead elicited multiple-choice responses. Also, in version 2, as described below we elicited more information in one of the frames, Complete Contingent Action Plan. The experiment was otherwise identical.

We invited all of the version-2 sample back to the laboratory 2-4 weeks later, and 237 (86.2%) agreed. When we pool across the version-1 and version-2 samples and analyze data from only the first time participants came to the lab, we refer to the sample as "wave 1." The wave-1 sample thus includes all 596 participants. When we restrict the data to the version-2 participants who participated in both waves of the experiment, we refer to the sample as "wave 1+2." The wave 1+2 sample comprises the 237 who returned to the lab. Below we describe the differences between waves 1 and 2. The experiments were otherwise identical. Throughout the paper, in order to analyze data across both waves and keep the sample constant, we focus most discussion on the wave 1+2 sample. Web Appendix D reports complete results for the wave-1 sample.

Participants were paid $40.25 per wave. In wave 1, version-2 participants were also offered the option of participating in another, unrelated experiment that additionally paid them, and almost all participants agreed.[4] Experimental sessions were scheduled for 2 hours each. After initial introductions and setup that took roughly 10 minutes, participants filled out the online survey, which was programmed in the RAND Corporation's Multimode Interviewing Capability (MMIC) survey software. Mean completion time for the survey was 72.5 minutes for version 1, 63.8 minutes for version 2 wave 1, and 52.7 minutes for version 2 wave 2.

In designing our experiment, we decided to recruit undergraduate participants, and we had them make hypothetical investment choices. We recruited undergraduates because we could obtain a large sample at relatively low cost, and (relative to a web sample) we could better monitor them to minimize distractions during the long experiment. We focused on investment choices because such choices motivated our experiment, and hypothetical investment choices are

[4] This was a happiness-tracking experiment that had two components. First, participants received six text messages per day over a 4-week period, both before and after wave 2. Participants were paid $0.25 per text message that they responded to within 90 minutes, as long as they responded to at least 70% of the text messages. Those who reached the 70% threshold could earn between $35 and $50. Second, during the wave-2 experimental session, participants who signed up for wave 2 had an opportunity to win money from a coin flip. Each participant was randomly assigned to one of four amounts of money they could win: $1, $5, $25, and $125. The expected additional payment for wave-2 participation was therefore $19.50. The coin flip occurred immediately after the wave-2 elicitation of untutored preferences (described below). These participants also filled out a short questionnaire at the end of their wave-2 experimental session.

one way that economists estimate the relevant risk-preference parameters (Barsky et al. 1997). (Choices in small-stakes gambles, which can be more easily incentivized, primarily identify loss aversion rather than the large-stakes risk aversion relevant for retirement asset allocation decisions; Rabin & Thaler 2001). Our design decisions have disadvantages relative to studying an older sample for whom the choices would be more familiar and relevant. We view our experiment as a test of feasibility for our design, which could subsequently be refined and extended to other samples and choice contexts.

In wave 1, the experimental session had five parts:

1. Training Batteries: Explained the pictorial representation of the investment choices and described background assumptions we wanted subjects to hold while making their choices.[5] Participants could not continue until they correctly answered nearly all quiz questions about the training.

2. Elicitation of Untutored Choices: Choices between hypothetical investments.

3. Psychological and Cognitive Batteries: These included: the cognitive reflection task (Frederick 2005); a number series task (McArdle, Rodgers, & Willis 2015); a 10-question Big-Five personality battery (Gosling, Rentfrow, & Swann 2003); a probabilistic sophistication battery (developed by Miles Kimball); and the need for cognition scale (Cacioppo & Petty 1984). These batteries were interspersed between sets of questions from part 2 in a random order.

4. Reconsideration of Choices: Opportunities to revise untutored choices.

5. Post-Experimental Questionnaire and Demographics: Questions such as how much the experiment made them feel enjoyment, annoyance, stress, and frustration, and demographic questions.

[5] We randomized half the participants to also get a probability training and quiz (alongside the other training batteries). The other half of participants did not get the training and answered the quiz in Part 5 (immediately before the demographic questions). This randomization allowed us to test the effect of probability training on the number of inconsistencies in the untutored choices. Participants who received the training had somewhat fewer inconsistencies (mean = 5.6) than those who did not (6.0), but the *p*-value for the difference is 0.28 (Web Appendix Table C.12).

Parts 2 and 4 constitute the core of our experiment and are described in the next two sections. Other parts of the experiment are mentioned as needed throughout the paper.[6] Screenshots of the complete survey are in the Survey Appendix.

In wave 2, we omitted part 3, and we abbreviated parts 1 and 5 by dropping the quizzes and demographic questions. These returning participants saw re-randomized versions of parts 2 and 4.

II.B. Instructions

At the very beginning of the survey, before explaining the symbols used in the investment choices, we instructed participants to interpret the payoffs as representing annual consumption streams in retirement: "This amount of money is *how much you will be able to spend every year during retirement, from age 65 on*. It is the only money you will be able to spend each year. It must be used to cover rent, food, clothing, entertainment, etc. This amount is what you have to spend after paying income taxes." Later, among the training batteries in part 1, we further instructed participants to make some assumptions that help ensure that the monetary payoffs represent consumption streams:

> In addition to the instructions you've already seen, you should imagine the following situation. These things are meant to help make your decisions a little easier, by removing some uncertainties you might otherwise have considered in your decision making:
>
> The government provides free medical insurance, and you are in good health.
>
> The government **no longer** provides social security (i.e. monthly checks).
>
> There is **no** inflation.
>
> 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.
>
> When you retire at age 65, you plan to move into rental housing that will have a monthly payment.

[6] We also included a set of 12 text-based, binary hypothetical choices between a certain amount of consumption each year in retirement and a 50-50 gamble between two amounts. These were inspired by similar questions asked in the Health and Retirement Study (Barsky et al. 1997). The questions were randomized to appear during part 1 for half the participants and during part 5 for the other half. We initially designed these text-based questions because we intended to assign monetary amounts for the survey (as described in Section II.C below) to each participant based on which CRRA parameter was closest to what would be implied by their responses to the text-based questions. Ultimately, however, we decided instead to randomize the monetary levels and randomize the placement of the text-based questions, and we did not use the data from them.

> **Note that 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 quizzed participants about these instructions and reviewed them if participants got fewer than five out of six quiz questions correct.

II.C. The Master Decision Tree

All the risky gambles we posed were derived from a *master decision tree*, shown in Figure 3. (The speech balloons superimposed on this and other figures show what percentage of participants made each choice in initial choices and final, reconsidered choices.) There is an initial binary choice made at age 35 between A and B. A is a riskless choice, meant to correspond to a portfolio of safe assets, whereas B is a risky choice, meant to correspond to a portfolio of equities. Conditional on choosing B, there is a 50% chance of each of two subsequent binary choices to be made at age 50, meant to correspond to rebalancing a risky portfolio that may have gone up or down in value. Each of these choices, C versus D or E versus F, is between a safe and a risky option. Five contingent plans are possible: A, BCE, BCF, BDE, and BDF.

Several design decisions are apparent from the figure. For example, we labeled safe options as "conservative," we depicted probabilities with a shaded pie chart, and we only used familiar probabilities: 50% and (when we reduce compound lotteries) 25%. These and other design decisions were informed by feedback we received during pilot testing about how to make the risky gambles easier to understand.

Figure 3 depicts one of ten sets of monetary amounts into which participants were randomized. Web Appendix B lists all ten sets. Within each set, the payoff triples in each of the three risky choices in the master decision tree are a constant multiple of each other, except that we rounded payoff amounts to the nearest 1k. For example, for the monetary amounts in the figure, the triples (225k, 150k, 108k) (the payoffs in the E versus F choice), (108k, 72k, 52k) (the payoffs in C versus D), and (150k, 100k, 72k) (the payoffs in A versus B, assuming safe choices for B) are approximately constant multiples of each other. Given the roughly fixed payoff ratios, an agent with constant relative risk aversion (CRRA) equal to 1.576 would be roughly indifferent at every decision node. The next four sets of monetary amounts have payoffs 100k, 150k, and

225k for the same outcomes as in Figure 3 but adjust the other payoffs to correspond to CRRA indifference cutoffs of 2.958, 4.865, 12.113, and 17.967. We chose this range of CRRA indifference cutoffs to roughly correspond to the 10th and 90th percentiles of estimated CRRA parameter values from Kimball, Sahm, & Shapiro's (2008) study of Health and Retirement Study respondents' choices in hypothetical gambles, which were 2.5 and 16.0, respectively. The last five sets of monetary amounts are the same as the first five but with every payoff cut in half, and so the last five correspond to the same CRRA levels as the first five. Due to a bug in our code, all version-1 participants were randomized into one of the four sets corresponding to the CRRA levels of 12.113 and 17.967. To reduce the imbalance across all the monetary levels, we randomized all version-2 participants into one of the six sets corresponding to the CRRA levels of 1.576, 2.958, and 4.865.

Within a wave, a participant was always asked questions based on the same set of monetary amounts. Participants who returned for a second wave received a re-randomized version of the survey. Since all of these were version-2 participants, they had a 4/5 chance of being randomized into a different set of monetary amounts. Except where otherwise mentioned, all of our analyses pool data across the monetary amounts.

While Figure 3 and other figures in this paper depict the master decision tree with A at the top of the screen and the E vs. F choice at the bottom, we randomized whether participants saw their decision screens this way, which we call "rightside up," or saw instead an "upside down" orientation of all decision screens, in which A was at the bottom of the screen and E vs. F was at the top. This randomization allows us to test and correct for any tendency to choose options toward the top of the screen. We do not find such a tendency (comparing untutored choices in each of the 41 decision screens across the randomization, the *p*-value is less than 0.05 for only one screen; Web Appendix Table C.9). We thus pool these data in all analyses.

II.D. Frames, Normative Axioms, and Elicitation of Untutored Preferences

When eliciting untutored preferences, we posed a total of 37 risky choices across 36 separate screens. These are derived from the master decision tree by asking, within each of seven frames, all choices between the five contingent plans that make sense given the frame.

We formulated the frames and normative axioms we study to jointly satisfy a number of criteria. We aimed for frames that are standard ways of posing risky choices. Except for

transitivity, we wanted every axiom to be defined as stipulating that the "same" choice should be made in two of the frames. We wanted a set of axioms such that, when combined with completeness and continuity (which we assume but do not test), the axioms are necessary and sufficient for preferences to satisfy intertemporal expected utility theory. Finally, we wanted each axiom to be transparent enough that participants should be able to easily see the appeal or lack of appeal of the axiom to them. The goal of transparency is to make it possible for experimental participants to endorse or reject an axiom's logic without ever being required to understand a complex chain of reasoning. Transparency was part of our strategy for minimizing potential experimenter demand effects because it eliminated the need to explain the axioms to participants. Because the set of axioms in standard axiomatizations of intertemporal expected utility theory (e.g., Volij 1994) do not satisfy this transparency criterion—for example, anyone who has taught intermediate microeconomics can attest that people unfamiliar with the Independence Axiom find it difficult to understand—most of our axioms are *components* of the traditional axioms.

To understand the frames, it is helpful to accompany their descriptions with screenshots. Frames 6 and 7 are shown in Figure 1, and frame 4 is shown in Figure 3. The rest are shown in Figure 4. (The speech balloons are superimposed on the screenshots and were not seen by participants.) The seven frames are:

1. Single Action in Isolation (2 screens: C vs. D, E vs. F): A choice at a single node, with the rest of the tree not shown.

2. Single Action with Backdrop (2 screens: C vs. D, E vs. F): A choice at a single node, with the rest of the tree grayed out. Participants were instructed: "**These grayed-out parts of the picture are things that could have happened, but you know for sure did not happen.**"

3. Two Contingent Actions with Backdrop (1 screen: C vs. D and E vs. F): A choice at two nodes, with the relevant choice contingent on a 50-50 realization and with the rest of the tree grayed out.

4. Complete Contingent Action Plan (1 screen: master decision tree): First, a choice at the A vs. B node. Participants were instructed: "If you choose B, you also need to make two decisions that will lock in how you will invest at age 50." In version 1 of the experiment, participants also make a choice at the C vs. D and E vs. F nodes only if they choose B. In version 2 of the experiment, all participants make a choice at the C vs. D and E vs. F nodes; participants who choose A were instructed to imagine that that option was not available and to make a choice at the C vs. D and E vs. F nodes.

5. Pairwise Choices Between Complete Strategies (10 screens: all pairwise choices between the five contingent plans): Pairwise choice between two contingent plans, with the plans displayed in the master decision tree. Participants were instructed: "You need to make a **choice between two investment plans, Option 1 and Option 2.** Each has a set of choices locked in along the way (at age 35 and age 50), shown by circled letters…**Grayed out parts are used to show things that can't happen if you choose that investment plan.**"

6. Pairwise Choices Between Compound Lotteries (10 screens: all pairwise choices between the five contingent plans): Pairwise choice between two contingent plans, with the plans involving B displayed as compound lotteries.

7. Pairwise Choices Between Reduced Simple Lotteries (10 screens: all pairwise choices between the five contingent plans): Pairwise choice between two contingent plans, with the plans involving B displayed as reduced simple lotteries.

We call the first four *nodewise frames* because each question involves one or more actions at nodes of the master decision tree. We call the last three *pairwise frames* because each question involves a pairwise choice between contingent plans. We explained above in Section II.C how the nodewise frames were randomized to be rightside up or upside down. In all pairwise frames, in any given wave, *both* contingent plans were either shown in the orientation of the figures or in an upside-down configuration, whichever was consistent with how the participant was randomized for the nodewise frames. On each screen of a pairwise frame, we randomized which contingent plan was shown on the left and which on the right.

The frames are ordered such that every contiguous pair of frames defines a normative axiom, according to which the "same" choice should be made in both frames. For example, axiom 1 says that the "same" choice should be made in frames 1 and 2; axiom 2, in frames 2 and 3; and so on. Here are informal descriptions of what the resulting six normative axioms say, together with names that summarize their content:

1. Irrelevance of Background Counterfactuals: People should make the same choice at a node whether or not they are aware of the history that led to the node.

2. Simple Actions = State-Contingent Actions: People should make the same choice at a node when they know that node has occurred as when they do not yet know whether the node will occur.

3. Irrelevance of Counterfactual Choices: People should make the same choices at a node when they do not yet know whether that node will occur as when choosing a complete contingent action plan (which includes choices at prior nodes that may determine whether the node can occur).

4. Shift from Nodewise to Pairwise: People should make the same choices over complete contingent action plans when choosing the actions nodewise as when making pairwise choices between complete contingent action plans.

5. Complete Strategies = Implied Lotteries: (i) People's pairwise choices over complete contingent action plans should coincide with their pairwise choices over the temporal compound lotteries derived from those plans, *and* (ii) people are indifferent to the timing of the resolution of uncertainty in pairwise choices over temporal compound lotteries.[7]

[7] We consider (i) and (ii) to be distinct conceptually, but we combined them because, as a matter of survey design, (ii) alone—changing the timing of the resolution of uncertainty—boiled down to just a change in the age label on the decision tree, which seemed something whose full import would be easy for our participants to miss.

6. Reduction of Compound Lotteries: People should make the same choice in a compound lottery as in the reduced simple lottery derived from it.

Using the machinery for dynamic choice under uncertainty (Kreps & Porteus 1978), Web Appendix F formalizes these axioms (doing so requires introducing a lot of notation to distinguish between frames) and relates them to previous work linking the Independence Axiom to axioms of dynamic choice (Karni & Schmeidler 1991; Volij 1994). Aside from Reduction of Compound Lotteries, which is a standard axiom, the others are components of standard axioms, as noted above. In addition to these, we study the (standard) transitivity axiom.

7. Transitivity: Aside from indifference, if Option A is chosen over Option B, and Option B is chosen over Option C, then Option A is chosen over Option C.

In Web Appendix F, we prove the equivalence between this set of axioms (plus completeness and continuity) and intertemporal expected utility theory.

When eliciting untutored preferences, we randomized participants into one of three groups in which they saw the frames with equal probability: (i) order 1, 2, …, 7; (ii) the reverse order 7, 6, …, 1; and (iii) a random order. Within each frame, the ordering of questions was randomized.

II.E. Reconsideration Procedure

As mentioned in the Introduction, the goal of our reconsideration procedure is to move the conditions of choice closer to idealized conditions of full deliberation and absence of cognitive error. To prompt deliberation, our approach is to provide participants with different perspectives on a choice by showing them a choice problem framed in more than one way after they have previously made a choice in each frame separately. As described above, we formulated the frames and axioms such that when we face participants with a choice problem framed two ways, one of the axioms implies that the participant's choice should not depend on the frame. If participants see the appeal of making the same choice in both frames, then they can correct their cognitive error by revising either of their choices. If they instead continue to see different choices as attractive, then they can leave their choices unchanged (or revise both choices), and we can be

more confident that their axiom violation reflects their normatively relevant preferences rather than a mistake.

After participants completed the elicitation of untutored preferences, to introduce the reconsideration procedure, the instructions stated: "Our research project depends on understanding your choices in a deep way. Now, we're going to ask you about some of the choices you've made so far."

Our algorithm for the reconsideration procedure had four stages, in this order:

1. Inconsistencies: Participants are given the opportunity to reconsider every pair of untutored choices from adjacent frames that is inconsistent with the corresponding axiom (*inconsistencies*), as well as a randomly selected ¼ of the pairs of untutored choices from adjacent frames that are consistent (*placebos*). Because the placebos were randomly selected, the inconsistencies and placebos were often interspersed. Which choice from each pair was shown on the top versus bottom of the screen was randomized. On average across both waves, participants in the wave 1+2 sample faced 9.9 inconsistencies and 10.1 placebos at this stage.

2. Intransitivities: In the three pairwise frames (frames 5-7), intransitive (cyclic) choices are possible among the five complete contingent strategies. For example, consider frame 6. In the choice depicted in Figure 1A, a participant might choose BDF over BDE. In two other choices (not shown), the participant might choose BDE over BCF and BCF over BDF. If so, that is an intransitivity. For all intransitivities among the stage-1 choices, participants are given the opportunity to rank the options (in the example, BDF, BDE, and BCF).

Our algorithm for identifying intransitivities among the five strategies was as follows: (i) try to identify both a highest- and lowest-ranked strategy; (ii) if neither can be identified, then stop; (iii) if either or both can be identified, then eliminate that one or two strategies; (iii) if there are at least three strategies remaining, then return to step (i), and otherwise stop. If the algorithm stops with one or two strategies remaining, then the stage-1 choices are transitive. Otherwise, the number of strategies remaining determines whether we call it a 3-way, 4-way or 5-way intransitivity. On average across both waves, participants in the wave 1+2 sample faced 1.7 intransitivities at this stage.

3. Inconsistencies again: Stage 1 is repeated, except that all the inconsistencies from the stage-2 choices are presented. On average across both waves, participants in the wave 1+2 sample faced 6.8 inconsistencies and 11.2 placebos at this stage.

4. Intransitivities again: Stage 2 is repeated, except that all the intransitivities among the stage-3 choices from pairwise frames are presented. On average across both waves, participants in the wave 1+2 sample faced 0.8 intransitivities at this stage.

To determine the order in which a participant saw inconsistencies and intransitivities, the computer program walked through the participant's choices in the order they were made. The first inconsistency or intransitivity encountered was presented to the participant first, and so on. Figure 5 panels A and B show screenshots of an inconsistency reconsideration and an intransitivity reconsideration, respectively. We now discuss the stages in more detail.

II.E.i. Stages 1 and 3: Inconsistency and Placebo Reconsiderations

Inconsistencies and placebos proceeded identically, with only one difference. For inconsistencies, the top of the screen reads: "In one question you chose [Option 1] over [Option 2], but in another question you chose [Option 2] over [Option 1]." For placebos, the top of the screen reads: "In these two questions, you chose [Option 1] over [Option 2]." In both cases, the screen showed both of the participants' choices and then asked, "Do you think the two situations are different enough that it makes sense to have different choices, or should they be the same?" There were two possible answers, which triggered different follow-up questions. The sequence of follow-up questions, which we now describe, is also depicted as a flowchart in Figure 6.

One possible answer was "It makes sense to have different choices." In that case, there was one follow-up question: "Why do you want to make different choices in these two situations?" In version 1 of the experiment, we elicited open-ended responses to this question (listed in Web Appendix A). In version 2, we instead offered the following multiple-choice responses:

- 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.
- Other: [free-response box]

The other possible answer to the initial question, "It makes sense to have the same choice," triggered a longer set of follow-ups. First, we asked "Which better represents your preference: your choice of [Option 1] over [Option 2], or your choice of [2] over [1]?" To try to avoid communicating that we wanted or expected participants to make their choices consistent, participants could choose among all four logically possible responses: (i) "Option 1 over 2"; (ii) "Option 2 over 1"; (iii) "I changed my mind: I realized that it does make sense to have different choices in these two situations. I would like to keep my current choices"; and (iv) "I changed my mind: I realized that it does make sense to have different choices in these two situations. I would like to change *both* of my choices." If the participant chose (iii), then she was asked the follow-up question "Why do you want to make different choices in these two situations?" as above. If she chose (i), (ii), or (iv), then she was shown a screenshot of what it would look like to make that choice and was asked to confirm, "Is this what you wanted your choices to be changed to?" If the participant did not confirm, she was taken back to the earlier screen with response options (i)-(iv). If she did confirm, then she was asked the final follow-up question, "Why did you want to change your choices as you did?" In version 1 of the experiment, we elicited open-ended responses to this question (listed in Web Appendix A). In version 2, we instead offered the following multiple-choice responses:

- 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.
- Other: [free-response box]

II.E.ii. Stages 2 and 4: Intransitivity Reconsiderations

The instructions for intransitivity reconsiderations read:

> Sometimes it is possible to rank options. Given your choices so far, we could not figure out how to rank these [3, 4, or 5] options. Would you like to rank these options? If so, please label the option you like best 1, the option you like second best 2, etc. If there is a tie between two options, you can rank them in any order, but please enter 1 only once, 2 only once, etc. If you do not want to rank these options, please *only* check the box below.
>
> [BOX] I do not want to rank these options!

The 3, 4, or 5 options were displayed below this text. Figure 5 panel B shows an example of a 4-way intransitivity reconsideration. If participants ranked the options, then their pairwise choices for that frame were revised accordingly. If participants opted not to rank the choices, then their choices in that frame were left unchanged (and hence remained intransitive).[8]

In version 2 of the experiment (but not version 1), if a participant opted not to rank the choices, we asked them: "Why couldn't you rank the options on the previous slide?" We offered four response options:

- I couldn't rank the options because they are all equally good to me.
- I couldn't rank the options because I don't know which option I prefer.
- I should be able to rank the options, but it's extremely hard.
- I couldn't rank the options for another reason: [free-response box]

III. Main Results

This section contains the main results of our experiment. It describes how participants' choices changed over the course of the experiment from the initial untutored choices to the final reconsidered choices.

III.A. Untutored Choices

[8] In addition to having the option not to rank, as another attempt to avoid giving participants the impression that we disapprove of intransitivity, we asked a survey question about intransitivity immediately before part 2 (the elicitation of untutored preferences): "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?" In the wave 1+2 sample, 54.4% of participants in wave 1 and 55.3% in wave 2 answered "Yes, I can imagine feeling like Ian about some number of choices," rather than the other option "No, I cannot imagine feeling like Ian about some number of choices." The correlation between answering "Yes…" to this question with the number of intransitivities, when both are from stage 0 of wave 1, is 0.074.

For the example untutored choices shown in Figures 1, 3, and 4, the percentage of participants choosing each option is shown as the first number inside the speech balloon attached to the option. Analogous figures for the remaining untutored choices are in Web Appendix H. The inconsistencies across frames can be seen from these percentages. For example, when the choices were depicted as compound lotteries, Option 1 was chosen 25% of the time (Figure 1A) but 43% of the time when the choices were depicted as compound lotteries (Figure 1B). This difference implies that many participants violated the Reduction of Compound Lotteries axiom. We discuss the frequency of individual-level axiom violations below in Section III.D when we analyze how these frequencies change with reconsideration.

III.B. Reconsideration Placebos

Before examining how the reconsideration procedure affected the number of inconsistencies and intransitivities, we look at how participants reacted to the placebos—the cases when they were asked to reconsider choices that were already consistent. Reassuringly, participants virtually never changed their choices when facing the placebos: 98.1% of the time, wave 1+2 participants chose "It makes sense to have the same choice" (or chose "It makes sense to have different choices" but upon further reflection selected that they had changed their mind so ultimately remained consistent) (Web Appendix Figure C.6). This finding indicates that merely prompting participants to revise their choices does not lead them to do so. In contrast, when facing inconsistencies, wave 1+2 participants revised their choices toward consistency 46.0% of the time (Figure 6; we obtain this value by multiplying the proportion of participants who chose "It makes sense to have the same choice" by the proportion of participants who chose either "Option 1 over 2" or "Option 2 over 1") and 72.9% of the time when facing intransitivities (out of 591 intransitivities in total, Table 3 reports 160 unrevised).

III.C. Inconsistencies and Intransitivities Over the Course of the Experiment

Figure 7 panel A shows, for the wave 1+2 sample, for each stage of the experiment, a histogram of the number of inconsistencies in the current set of choices (we discuss the inconsistency rate by axiom in Section III.D). The *x*-axis in each histogram is the number of inconsistencies, from 0 to 20 (the maximum possible number is 19). The *y*-axis is the percentage of participants. In the first row of five histograms, the first histogram corresponds to the "stage-

0" untutored choices in wave 1; the second, to the stage-1 choices (after the first round of inconsistency reconsiderations); and so on, up to the fifth, which corresponds to the stage-4 choices (after both rounds of inconsistency and intransitivity reconsiderations). The second row of five histograms is analogous, except it shows the choices from wave 2.

Among the untutored choices in wave 1, the median number of inconsistencies was 6, and only 4.0% participants had zero inconsistencies. The numbers of inconsistencies declined in each stage of reconsideration. By the end of stage 4, the median number was 2, and 23.5% had zero. At the beginning of wave 2 several weeks later, there was partial "reset," but participants had fewer inconsistencies at the beginning of wave 2 than at the beginning of wave 1. This reduction from wave 1 to wave 2 is very unlikely to be due to participants remembering their earlier choices; it occurs even for the 4/5 of participants who faced gambles with different monetary payoffs across the two waves (Web Appendix Figure C.7). The reduction is more likely due to participants having learned from wave 1 to make more of an effort to avoid inconsistencies. By the end of wave 2, the median number of inconsistencies was 1, and 33.8% participants had zero inconsistencies.

If participants revised their inconsistent or intransitive choices in a random way, then their revised choices can create new inconsistencies, and we would not in general expect the total number of inconsistencies to fall on average. The decline in the number of inconsistencies over the course of the experiment therefore indicates that participants reconsidered their choices in a direction that led to a set of choices that were overall more consistent with the axioms.

Figure 7 panel B is analogous to panel A, except for intransitivities rather than inconsistencies. Each 3-way, 4-way, or 5-way intransitivity counts as one intransitivity, so the maximum possible is three (one for each of the three pairwise frames). As with the inconsistencies, the number of intransitivities declined over wave 1, reset partially at the beginning of wave 2, and then declined over wave 2. The mean number was 1.1 at the beginning of wave 1 and 0.1 at the end. For wave 2, the analogous numbers are 0.8 and 0.2.

III.D. Reconsidered Choices

We now turn from analyzing the effects of the reconsideration procedure to examining the properties of the set of choices that result from the procedure. It is clear from Figure 7 panel B that the reconsidered choices have far fewer intransitivities than the untutored choices. For

example, in the wave 1+2 sample shown in the figure, by the end of wave 2, the mean and median number of intransitivities are 0.2 and 0, respectively.

There are many more inconsistencies possible from the choices we posed to participants than intransitivities; Figure 7 panel A shows that inconsistencies too are dramatically reduced in the reconsidered choices relative to the untutored choices. Moreover, most of the remaining inconsistencies are driven by relatively few participants: in the wave 1+2 sample shown in the figure, by the end of wave 2, 33.8% of participants had zero inconsistencies, 54.0% have ≤ 1, and 67.1% had ≤ 2. Only 14.9% of participants had > 5.

To facilitate comparison across axioms, we calculate for each axiom the inconsistency rate: the number of inconsistencies divided by the number of possible inconsistencies. For the wave 1+2 sample, Table 5 shows the inconsistency rate for the untutored choices and reconsidered choices in each wave, separately by axiom and in aggregate. With one exception—the Irrelevance of Counterfactual Choices axiom, discussed below—the inconsistency rate fell substantially from the beginning of wave 1 to the end of wave 2 and was ≤ 11.0% by the end of wave 2. For example, the Reduction of Compound Lotteries axiom had an inconsistency rate at the beginning of wave 1 of 26.1%, which was the highest. By the end of wave 2, its inconsistency rate was 8.9%. In aggregate, inconsistency rates fell by almost 2/3, from 22.4% to 8.4%. We interpret the low inconsistency rates for the reconsidered choices as suggesting that many of the axiom violations in the untutored choices are math errors or other mistakes rather than reflections of normatively relevant preferences.

As noted above, the main exception to declining inconsistencies with reconsideration is the Irrelevance of Counterfactual Choices axiom. As Table 5 indicates, its inconsistency rate at the beginning of wave 1 was 13.9%, and, while it fluctuated over the course of the experiment, it ended up at 14.2% at the end of wave 2. Violations of the axiom might therefore reflect normatively relevant preferences for a non-trivial fraction of respondents. It can be seen from Figure 5 panel A that violations of this axiom involve making a different choice in C vs. D or E vs. F when the participant had chosen B over A than when the participant ended up on this branch of the decision tree without having chosen B over A. Most plausibly, such behavior may implicate reference-dependent risk preferences (as in Kőszegi & Rabin 2006; see also Loomes & Sugden 1982) with the counterfactual payoff from A influencing the reference point.

To explore whether our data may be consistent with one such model, we report additional analyses in Web Appendix G. We formally analyze a model of reference-dependent risk preferences in which the reference point is influenced by foregoing a sure payoff. In the frame Two Contingent Actions with Backdrop (frame 3), where the participant makes the C vs. D and E vs. F choices without having faced the A vs. B choice, we assume that the safe payoff—the payoff from C or from D—is the reference point in each choice. In the frame Complete Contingent Action Plan (frame 4), we assume that when a participant chooses B over A, foregoing the payoff from A shifts the reference point for both the C vs. D and E vs. F choices. We show that this shift in reference point leads to greater risk tolerance *regardless* of what the new reference point is. Therefore, the model predicts greater risk tolerance—i.e., more willingness to choose D and F—in frame 4 than in frame 3. However, in our data we find no evidence of this pattern in the reconsidered choices; in fact, we find roughly equal numbers of choices that become more risk averse and more risk tolerant (Web Appendix Table G.1). We conclude that the (in our view, plausible) form of reference-dependence we study does not predominantly explain violations of the axiom. Moreover, accounting for our data would require a theory with heterogeneity in the direction of the prediction.

IV. Additional Descriptive Results

In this section, we report or summarize analyses that aim to help us better understand how and why participants revised their choices. (In addition to what is below, in Appendix I.1, we report analyses in which we examine, and find little support for, the possibility that participants might have made their choices in accordance with one of several simple heuristics.)

IV.A. Why Participants Revised or Did Not Revise Choices

When faced with inconsistencies, when participants did *not* revise their choices, their survey responses indicated that it was usually because the two frames were considered "different situations" or because the participants were indifferent. Table 1 shows the percentage of responses to the question "Why do you want to make different choices in these two situations?" The rows of the table break down the results by axiom, and the bottom row shows the results aggregated across all axioms. In aggregate, participants selected "The two situations are different enough that I want different choices" 56.8% of the time. We interpret this response as consistent

with rejecting the axiom that implies that choices should be the same. Participants selected the indifference response option, "Some of the options are equally good to me, so it doesn't matter which one I choose," 24.9% of the time. The other response options were selected ≤ 6.2% of the time.

For inconsistencies, when participants revised their choices, their survey responses indicated that it was virtually always because they initially erred, they learned something from thinking through their choices, or they were indifferent. Table 2 is analogous to Table 1 but for the question "Why did you want to change your choices as you did?"[9] The response "I made a mistake when I first chose" was selected 45.7% of the time; "Answering all of these questions made me change what I want," 38.4%; and "Some of the options are equally good to me, so it doesn't matter which one I choose," 9.4%. We interpret the first of these as consistent with endorsing the axiom that implies that choices should be the same. We similarly interpret the second as endorsing the axiom because it suggests that the deliberation induced by the experiment led the participant to better understand their preference. The other response options were selected ≤ 2.9% of the time.[10]

When participants refused to rank intransitive choices, their survey responses indicated it was usually *not* because of indifference. As Table 3 shows, they selected "Options all equally good" only 15.0% of the time, whereas they selected "I don't know what I prefer" 45.6% of the time and "Too hard to rank" 13.8% of the time. The latter two responses do not distinguish between truly intransitive preferences, incomplete preferences, or insufficient effort to figure out the participant's preference ordering. Of the 5.6% who selected "Other," most complained about being tired or noted that the task was too hard.

IV.B. Are Reconsidered Choices Closer to Normatively Relevant Preferences?

[9] The row for the axiom Irrelevance of Counterfactual Choices is omitted from the table because, due to a programming error, this follow-up question was not asked for this one axiom.

[10] While we put little weight on participants' self-reports of experimenter demand effects, we note that these self-reports provide a bit of further evidence against experimenter demand effects driving participants' revisions. Specifically, recall that when we asked participants their reasons for revising or not revising their choices, we offered participants the option to select: "I chose how I thought the experimenters wanted me to choose." Although very rarely selected, it was selected more often when participants did not revise an inconsistency (3% of the time) than when they did (1%) (see Tables 1 and 2; the p-value for this comparison is <0.0001), the opposite of what might have been expected if participants were revising inconsistencies due to experimenter demand effects.

While we would like our reconsideration procedure to generate a more accurate elicitation of participants' normatively relevant preferences, an alternative possibility is that participants are averse to inconsistencies and intransitivities (or to being asked about them by us) and revise to eliminate them, without getting closer to normatively relevant preferences. We can obtain some germane evidence by examining how concordant choices at the end of wave 1 and at the end of wave 2 are with *each other*. The idea is that participants might be eliminating inconsistencies across frames within each wave but not converging toward the same set of choices. However, if they are revising toward *the same* preferences in each wave, then their choices should be more concordant across waves at the end of the waves than at the beginning of the waves. Table 4 shows an analysis that tests this hypothesis. Each row corresponds to a frame, except for the last row, which aggregates over all the data. Column (1) shows the number of potential concordances across waves, which is equal to the number of choices made in the frame. Column (2) shows the percentage of choices that are concordant across waves in the untutored choices (stage 0 in each wave) and the final choices (after stage 4 in each wave). For comparison, column (3) gives the expectation under random behavior. In aggregate, the amount of concordance increased from 68.1% at the beginning of the waves to 72.2% at the end of the waves. (If we restrict the analysis to participants who faced the same monetary amounts in waves 1 and 2, these percentages are 72.5% and 77.7%, respectively; see Web Appendix Table C.4.) Column (5) reports that the *p*-value for the null hypothesis of equality between these two percentages is < 0.0005. Column (6) reports that the *p*-value for the null hypothesis of equality between the amount of concordance at the end of the waves and random behavior is also < 0.0005. The table shows that we can similarly reject random behavior for each of the individual frames. The evidence for increasing concordance going from the beginning to the end of the waves is concentrated among the pairwise frames, where we have greater statistical power (due to the larger number of possible concordant choices). We conclude from this analysis that participants do appear to be moving toward the same set of preferences in the two waves.

IV.C. Do Participants Revise Toward Greater Risk Tolerance?

When participants revise their choices, do they revise in the direction of greater risk tolerance or greater risk aversion? In Section V, we examine this question with the help of structural assumptions about preferences, but here we aim to shed light on it with descriptive

data. Since D and F are the risk-tolerant choices regardless of utility's functional form, one simple approach is to focus only on the choices C vs. D and E vs. F. For such choices, the top and bottom panels of Figure 8 show, for the wave 1+2 sample, the percentage of participants who chose D and F, respectively, in each frame over the course of both waves. The standard error on each data point is relatively large (roughly 3 percentage points; see Web Appendix Table C.10) both because we are cutting the data by frame and because we are only using a subset of participants' choices. Nonetheless, both panels of the figure hint at some overall increase in the frequency of the more risk-tolerant choices. For a more formal test, we pool all the data underlying Figure 8, and we run an OLS regression of choice of D or F on stage of the experiment, with fixed effects for frame and wave and with standard errors clustered by participant. The regression results confirm the visual impression from the figures, with the coefficient on stage estimated to be 0.37 percentage points (SE = 0.11) (Web Appendix Table C.11 column 1).

As another descriptive approach to assessing whether participants revise in the direction of greater risk tolerance or risk aversion, we directly examine participants' revisions. This analysis aggregates across all revisions (not just those for C vs. D and E vs. F choices) but is slightly more complicated, so we relegate it to Web Appendix I.2. We find that, when participants revise their choices, on average they do so toward the riskier choice, which reinforces the conclusion from examining the C vs. D and E vs. F choices. A limitation of both of these analyses, however, is that they cannot disentangle whether reconsideration leads to a change in risk preferences or a change in unsystematic "response error"—a point we address in Section V.

V. Estimating CRRA Preferences

In our descriptive analyses in Section III.C, we found that inconsistencies with intertemporal expected utility theory declined over the course of the experiment (Figure 7). In this section, we assume that normatively relevant preferences satisfy expected utility theory and, in particular, as is common in studies of retirement investment, we assume constant relative risk aversion (CRRA) preferences. We structurally estimate this model within each of three frames that have sufficient data, as well as pooling across them, in order to address two questions: (i) if deviations from intertemporal expected utility and from CRRA are assumed to be response

errors, how much does the reconsideration procedure reduce response-error variance? and (ii) how does the reconsideration procedure affect the average level of risk aversion in the reconsidered choices? Question (ii) was addressed with descriptive analyses in Section IV.C, but the structural model in this section allows us to disentangle whether the change in choices is due to a change in the underlying preference parameter or to the change in response-error variance.

Following Barsky et al. (1997) and Kimball, Sahm, & Shapiro (2008) we estimate a random parameter model (as prescribed by Apesteguia & Ballester 2018). Specifically, our model is as follows. We separately estimate relative risk aversion in each of three frames (but omit the frame subscript from all variables). As before, in each wave, we label untutored choices as "stage 0" choices, the choices after the first set of inconsistency reconsiderations as "stages 1" choices, etc. We allow individual i's relative risk aversion to evolve across waves $w \in \{1,2\}$ and stages $s \in \{0,1,2,3,4\}$ and denote it by γ_{iws}. We work with log relative risk aversion: $x_{iws} \equiv \ln \gamma_{iws}$. In wave *w* and stage *s*, consider a question *q* eliciting a pairwise choice between Option 1 and Option 2. Let κ_q denote the level of log relative risk aversion at which an individual would be indifferent. We assume that the agent makes the safer choice if and only if

$$x_{iws} + \epsilon_{iwsq} \geq \kappa_q,$$

where $\epsilon_{iwsq} \sim N(0, e^{2z_{iws}})$ and is independently drawn for each question (we discuss below how the log standard deviation z_{iws} is determined). That is, experimental participants are assumed to respond as if their log relative risk aversion were their true log relative risk aversion x_{iws} plus a random error ϵ_{iwsq}. The response error captures deviations from intertemporal expected utility and from CRRA, including intransitivities. When we estimate the model using data from just one frame, inconsistencies across frames (which we focused on in Sections III.C and III.D) do not directly affect the structural estimates, but they matter indirectly when choices change in response to the reconsideration procedure. When we pool data across the three frames, inconsistencies across those frames are directly included in the response error.

Since there are a discrete number of contingent strategies in the master decision tree, any preference ordering over contingent strategies that can be rationalized by some relative risk aversion parameter value can be rationalized by a range of parameter values. Consequently, participant-specific parameters are not point-identified. To address this issue and to increase

statistical power, we model a participant's log relative risk aversion as a random effect whose mean is a function of wave, stage, and demographics:

$$x_{iws} = \mu_0 + \mu_w \cdot 1\{w = 2\} + \mu_s \cdot (s-1) + \mu_{ws} \cdot 1\{w = 2\} \cdot (s-1) + \boldsymbol{\mu}_X \boldsymbol{X}_{iws} + \eta_{1,i} \cdot 1\{w = 1\} + \eta_{2,i} \cdot 1\{w = 2\} + \nu_i, \quad (1)$$

where $\boldsymbol{X}_{iws}$ is a vector of controls (which we omit in most specifications), and $\eta_{1,i} \sim N\left(0, \sigma_{\eta_1}^2\right)$, $\eta_{2,i} \sim N\left(0, \sigma_{\eta_2}^2\right)$, and $\nu_i \sim N(0, \sigma_\nu^2)$ are mutually independent. According to this model, the distribution of log relative risk aversion parameters governing stage-0 choices have means $\mu_0 + \boldsymbol{\mu}_X \boldsymbol{X}_{iws}$ in wave 1 and $\mu_0 + \mu_w + \boldsymbol{\mu}_X \boldsymbol{X}_{iws}$ in wave 2. These means shift over stages of reconsideration, with slope μ_s in wave 1 and slope $\mu_s + \mu_{ws}$ in wave 2. The population distribution of log relative risk aversion parameters has variance $\sigma_\nu^2 + \sigma_{\eta_1}^2$ in wave 1, variance $\sigma_\nu^2 + \sigma_{\eta_2}^2$ in wave 2, and covariance σ_ν^2 across waves.

We model the log standard deviation of the random error in choice, z_{iws}, similarly:

$$z_{iws} = \tau_0 + \tau_w \cdot 1(w = 2) + \tau_s \cdot (s-1) + \tau_{ws} \cdot 1(w = 2) \cdot (s-1) + \boldsymbol{\tau}_Z \boldsymbol{Z}_{iws}, \quad (2)$$

where $\boldsymbol{Z}_{iws}$ is a vector of controls (which we omit in most specifications). That is, for the stage-0 choices, z_{iws} equals $\tau_0 + \boldsymbol{\tau}_Z \boldsymbol{Z}_{iws}$ in wave 1 and $\tau_0 + \tau_w + \boldsymbol{\tau}_Z \boldsymbol{Z}_{iws}$ in wave 2. The value of z_{iws} shifts over stages of reconsideration, with slope τ_s in wave 1 and slope $\tau_s + \tau_{ws}$ in wave 2.

Taken all together, the set of model parameters is $\left\{\mu_0, \mu_w, \mu_s, \mu_{ws}, \boldsymbol{\mu}_X, \sigma_{\eta_1}^2, \sigma_{\eta_2}^2, \sigma_\nu^2, \tau_0, \tau_w, \tau_s, \tau_{ws}, \boldsymbol{\tau}_Z\right\}$. We calculate joint maximum-likelihood estimates of the parameters. We describe our estimation procedure in Web Appendix E.

We focus our analysis on the three pairwise frames: Pairwise Choices Between Complete Strategies, Pairwise Choices Between Compound Lotteries, and Pairwise Choices Between Reduced Simple Lotteries. We have far greater statistical power in these frames, which elicit each participant's complete preference order over contingent plans, than in the nodewise frames, which only elicit partial preference orders. To keep the sample constant across waves, we use the wave 1+2 sample in our main analysis. We confirm robustness of our main findings in the wave 1 sample in Web Appendix Table D.8.

In our main analyses, we pool data across participants who faced different monetary levels in their choices. As a specification check, however, we estimate the model parameters separately for the subsamples who faced monetary amounts that differed by a factor of two. CRRA utility and the other assumptions of our model imply that we should estimate the same mean log risk aversion for these two subsamples. For the subsample that faced smaller monetary amounts, we see evidence of less risk aversion in wave 2 but not in wave 1 (see Web Appendix Table E.8e, panel D).

Table 6 shows the results from estimating the model. The three panels A-C correspond to the three pairwise frames, and the panel D corresponds to the three frames pooled. The results are broadly similar across the three frames and the pooled data, so for concreteness, we walk through the parameter estimates only for the pooled data (panel D). Columns 1, 3, 4, and 5 depict the estimates from equation (1) above. In column 1, the estimate $\hat{\mu}_0 = -0.407$ (SE = 0.183), implying risk aversion in the untutored choices that is in between log utility ($\mu_0 = 0$) and square-root utility ($\mu_0 = -0.693$). The low value of relative risk aversion we find compared to estimates from hypothetical choices of older Americans (Kimball, Sahm, and Shapiro 2008) may reflect the younger age of our experimental participants. The other estimated parameters in column 1, $\hat{\mu}_w$, $\hat{\mu}_s$, and $\hat{\mu}_{ws}$, are all small and not statistically distinguishable from zero, indicating that the mean log relative risk aversion among participants does not change systematically across stages of the reconsideration procedure or across waves. Below we discuss the apparent tension between this finding and the descriptive results pointing to greater risk tolerance in participants' revised choices.

The estimated variance parameters, $\hat{\sigma}^2_{\eta_1}$, $\hat{\sigma}^2_{\eta_2}$, and $\hat{\sigma}^2_{v}$, are in Columns 3-5. They all have relatively large standard errors, which makes us reluctant to use them to draw inferences about the change in variance of log relative risk aversion from wave 1 to wave 2 or the correlation across waves.

Column 2 depicts the estimates from equation (2) above. Continuing to focus on Panel D, the initial (wave 1, stage 0) log error-response standard deviation is estimated to be $\hat{\tau}_0 = 0.945$ (SE = 0.092), much larger than the estimated mean log relative risk aversion. The estimate $\hat{\tau}_s = -0.107$ (SE = 0.008) is negative, indicating that the log error-response standard deviation decreases over the stages of reconsideration. The estimate $\hat{\tau}_w = -0.442$ (SE = 0.132) is also negative, meaning that the log error-response standard deviation is smaller at the beginning of

wave 2 than wave 1. The estimate $\hat{\tau}_{ws} = 0.057$ (SE = 0.012) is positive, indicating that the rate of decrease over the stages of reconsideration is smaller in wave 2 than in wave 1.[11] On the whole, these estimates from our structural model are similar to our reduced-form analysis of intransitivities and other inconsistencies: declining from the beginning to the end of wave 1, then starting at the same level at the beginning of wave 2 as at the end of wave 1 (unlike in the raw choice data, there is no partial reset in the structural parameter estimates), then declining further from the beginning to the end of wave 2, though more slowly than in wave 1. Quantitatively, the point estimates imply that the response-error variance from the beginning of wave 1 to the end of wave 2 declined by roughly 70% ($\approx 1 - e^{2(-0.410+4(-0.106+0.057))}$).

What explains the apparent tension between the finding from our structural model that mean log risk aversion does not change with reconsideration and the findings from our descriptive analyses that risk tolerance increases? The structural results are consistent with the descriptive results given the maintained assumptions of the model because the parameter estimates imply the following behavior. Our structural estimate of mean log risk aversion near zero corresponds to a lower value of the CRRA parameter than all six of the CRRA "cutoff" parameter values we used to set the monetary levels in the gambles (see Section II.C). Therefore, absent response error, most participants would make risk tolerant choices in all of the gambles (regardless of which monetary levels the participants were assigned to). Response error would thus be responsible for the risk-averse choices we observe in earlier stages of the experiment, but as the amount of response error shrinks over the course of the experiment, participants revise their erroneous choices to better align with their (risk tolerant) preferences. While we believe our apparently conflicting observations can be reconciled in this way, we caution that the conclusion from our structural estimation that mean relative risk aversion does not change relies on our structural model being correctly specified.

In Web Appendix E, we report a number of robustness checks and additional analyses, which we briefly summarize here. As a robustness analysis, we show that our results are robust to dropping the quintile of the sample who completed the experiment the fastest, who we suspect

[11] While this finding could be evidence of convergence of the log error-response standard deviation toward a positive asymptote, we caution that other explanations are possible. For example, in wave 1, we asked more questions in between the elicitations of untutored and reconsidered choices. The greater time gap or additional questions in wave 1 may have caused participants to have a fresher perspective during the reconsideration procedure, making them more willing to revise their choices.

may have been paying less attention or deliberating less. We also test and confirm that our estimated parameters do not systematically vary across the subsamples randomized to different groups (e.g., the decision trees depicted upside down); while a few such comparisons reach $p < 0.05$, the estimates are not consistent across the three frames (Web Appendix Tables E.8c-g). In addition, to make our structural estimation sample the same as that used in our descriptive analysis from Section IV.C, we restrict to the subsamples involving C vs. D and E vs. F choices; the results are similar to those from our main analysis (Web Appendix Table E.8h). Finally, although not the focus of our paper, we use our data and model to investigate how log relative risk aversion varies by sex and psychological characteristics such as performance on cognitive tests, while also allowing these variables to be correlated with the log standard deviation of the response error. Among the most interesting results are that higher cognitive-performance participants are estimated to have lower risk aversion and less error (consistent with the literature reviewed in Dohmen et al. 2018).

VI. Related Literature Review and Synthesis

As mentioned in the Introduction, the reconsideration procedure we introduce in this paper is related to other experimental approaches that also aim to provide measures of preferences that are more normatively relevant than untutored choices. In this section, we offer a unifying perspective on four traditions of experimental approaches from different literatures. We discuss the merits and drawbacks of each tradition, briefly review the findings, and place our experiment and results in context.

VI.A. Revisiting An Earlier Choice

In the first tradition of experimental approaches, participants simply revisit an earlier choice. A number of papers present the same choice problems to respondents over several rounds (van de Kuilen & Wakker 2006; van de Kuilen 2009; Nicholls, Romm, & Zimper 2015; Birnbaum & Schmidt 2015). Hey (2001) motivates this work by asking if deviations from expected utility are due to response errors that are transitory and decay with repetition. Later papers refer to Plott's (1996) "discovered preference hypothesis," according to which "individuals have a consistent set of preferences over states, but such preferences only become known to the individual with thought and experience" (Myagkov & Plott 1997, p.821). These

experiments generally find that choices conform more to expected utility theory in later rounds of the experiment, but in some cases only if the participants experience the outcomes of the gambles, and typically substantial deviations from expected utility remain in the final choices.

More similar to our paper, Breig & Feldman (2021) faced participants with a set of 25 risky choices two consecutive times and then offer them the opportunity to revise a subset of those choices.[12] Whereas participants may prefer to randomize across repeated choices (as in Agranov & Ortoleva 2017), revised choices largely eliminate this motivation because the earlier choice is not paid if it is subsequently revised. Breig & Feldman find that participants revise toward making a set of choices that is more consistent with expected utility: the mean across participants of a version of Afriat's Index that measures consistency with expected utility increases from 0.676 to 0.761. This increase is only 26% of the way toward its maximum of 1, although this metric likely understates the increase in consistency with expected utility because participants were only offered the opportunity to revise a subset of choices.

Taken all together, the evidence from this experimental tradition indicates that revisiting choices increases consistency with expected utility but still leaves a great deal of inconsistency.

VI.B. Providing Reasons For A Choice

Like the first tradition, the second tradition aims to prompt experimental participants to deliberate about their choice without providing external input: participants are asked to write down reasons for their choice before making it. In the context of risky decision making, we are aware of only two papers that pursued this approach. Miller & Fagley (1991) studied experimental participants' choices in a risky lottery that was either framed in terms of "lives saved" or "lives lost." Sieck & Yates (1997) studied both that problem and a choice between options that were either framed as simple or compound lotteries. In their one experiment Miller & Fagley found that asking participants to write down reasons eliminated the difference in choice frequencies across frames, while Sieck & Yates found in all three of their experiments

[12] Along the same lines, Yu, Zhang, & Zuo (2021) allowed participants to revise their responses to multiple price lists, in order to eliminate non-monotonic responses. In an experiment where participants make initial choices but can update them in continuous time over 20 seconds, Gaudeul ® Crosetto (2019) find that initial choices, but not final choices, display the attraction effect (in which the addition of a dominated option to a choice set increases choice of the option that dominates it).

that the difference was reduced but not eliminated. Unfortunately, neither paper analyzed the reasons provided by participants.

While a large body of literature finds that justifying one's reasoning improves problem solving in some contexts, such as in typical classroom math problems, providing reasons can worsen performance on tasks that require non-verbal processing, such as "insight problems" and affective decision making (see, e.g., W. R. Sieck, Quinn, & Schooler 1999 for references). Since writing down reasons is known to increase some types of cognitive errors, it is unclear whether or when the risky choices participants make when writing down reasons should be considered more normatively relevant than untutored choices.

VI.C. Revising A Choice In Light of An Axiom

The third tradition builds directly on early theoretical work on subjective expected utility, which invoked introspection by decision theorists and their readers about whether they would revise a choice when it was pointed out that the choice conflicted with an axiom of expected utility theory (e.g., Savage 1954; Raiffa 1961). Experiments in this tradition present normative axioms explicitly and examine whether participants revise choices to align with them. In the pioneering paper, MacCrimmon (1968) held a free-form discussion of five postulates of decision theory with business executives for roughly 30 minutes after the executives had made initial choices and had read arguments both for and against the postulates. MacCrimmon then offered the executives the opportunity to revise their choices. Most choices were consistent with the postulates either initially or after the discussion. In one small part of a classic paper on transitivity, Tversky (1969) found that undergraduates often made intransitive choices but endorsed transitivity when their intransitivity was pointed out to them.

Subsequent work focused on the Independence Axiom (and Sure-Thing Principle). Like MacCrimmon, these experiments provided arguments for and against the axiom, to help participants understand the axiom and its appeal while minimizing an experimenter demand effect pushing participants to endorse the axiom. Unlike MacCrimmon, these experiments did not include a free-form discussion with participants (several of the papers expressed the concern that during the discussion, the experimenter may have inadvertently biased participants toward endorsing the postulates). The findings from experiments in this tradition have been equivocal.

For example, consider Slovic & Tversky's (1974) second of two experiments. Participants faced decisions in one Allais-paradox and one Ellsberg-paradox problem. They were presented with arguments in favor of each of two opposing choices in both problems, and then they made their decisions and rated the arguments. The results of this experiment were puzzling: participants rated the anti-Independence-Axiom argument as more compelling in both problems, even though in the Allais-paradox problem the majority of participants behaved in accordance with expected utility theory! Such results are typical of experiments in this tradition (Moskowitz 1974; MacCrimmon & Larsson 1979; Eli 2017), which sometimes find that choices move in the direction of expected utility but often find that participants rate arguments that favor intuitively appealing violations of the Independence Axiom as more persuasive.[13] A natural worry about all of these experiments is that the arguments against the axioms may not generate experimenter demand effects equal to those of the arguments in favor, and it is difficult to know which is larger.

Most recently, in an experiment involving incentivized choices over small-stakes lotteries, Nielsen & Rehbeck (2020) develop a clever experimental design that avoids presenting arguments: they directly elicited participants' preferences over axioms by presenting the axioms as algorithms, or "decision rules," that implement choices on behalf of participants as a function of earlier choices. Later in the experiment, they measured participants' willingness-to-pay to revise choices that conflict with the decision rules that the participants had selected. They studied six axioms, including transitivity and the Independence Axiom. Participants selected decision rules that implement axioms roughly 85% of the time, compared with roughly 10% for "control" decision rules that implement the opposite of an axiom. On average across the axioms, when confronted with an inconsistency between a choice the participant had made and the choice implied by an axiom the participant had selected: 47% of the time participants revised their

[13] Here is a brief summary of results from the other papers we are aware of. Using real-stakes decisions in Allais-paradox problems, Moskowitz (1974) found that participants changed their choices in the direction of expected utility theory after reading one argument for and one against the Independence Axiom, but the change was small. In an extensive set of experiments, MacCrimmon & Larsson (1979) studied 20 rules that reflect either normative principles or reasons commonly given for non-normative choices. The overall pattern of findings is similar to Slovic & Tversky's (1974): experimental participants typically rated non-normative rules higher than normative rules, yet their choices often contradicted their rule rankings. Eli (2017) trained experimental participants to understand (i) the Independence Axiom, (ii) a decision rule based on Allais' (1979) argument in favor of choices consistent with the Allais paradox (and hence inconsistent with the Independence Axiom), and (iii) an anti-Allais-paradox decision rule (also inconsistent with the Independence Axiom). Among participants understanding all three rules, the largest group of participants made choices consistent with the Allais paradox, not expected utility.

choice, 13% of the time participants unselected the axiom, and 37% of the time participants left inconsistency (the remaining 3% of the time, participants *both* revised their choice and unselected the axiom). Although the experimental design does not lend itself to comparisons across axioms, it is notable that the Independence Axiom receives relatively less support; for example, participants who had selected the Independence Axiom as a decision rule, when faced with an inconsistency with a choice they had made, revised their choice (and kept the axiom selected) only 34% of the time.

The weaker support for the Independence Axiom notwithstanding, Nielsen & Rehbeck find much clearer endorsement of the axioms than prior work in this tradition. We conjecture that this difference in results flows from the major design differences: Nielsen & Rehbeck's experiment did not present arguments in favor and against the axioms, but a decision rule, like an argument in favor, makes the axiom explicit to participants. When an axiom is presented explicitly, the logic and simplicity of the axiom can be seductive, while the reasons why a participant might want to violate the axiom when faced with a particular choice (e.g., anticipated regret for the Independence Axiom) are not made salient. In earlier experiments, the arguments against the axiom highlighted the reasons why a participant might want to violate the axiom, but there was no such countervailing force in Nielsen & Rehbeck's experiment. (The "control" decision rules do not serve this purpose because they implement the opposite of an axiom, which is not an appealing decision rule and does not provide participants with a good reason to reject an axiom.) Consequently, the design differences may have maintained an experimenter demand effect pushing participants to endorse the axioms while eliminating an effect in the opposite direction.

VI.D. A Choice Viewed Through Two Frames

Our paper builds on a fourth tradition: examining how experimental participants make choices when two framings of the "same" decision problem are presented together. Like the third tradition, the fourth tradition can study particular axioms, but it has two advantages relative to most experiments in the third tradition: it does not require presenting arguments for and against the axiom, *and* it does not present the axiom explicitly.

Both prior papers in this tradition, McNeil, Pauker, & Tversky (1988) and Druckman (2001), compared experimental participants' choices in a risky lottery framed in one of three

ways: in terms of "lives saved," "lives lost," and both together.[14] Both papers found that, when both frames are presented together, the percentage of participants choosing the risky option is in between the percentages in the two separate frames. While Druckman treats the mixed frame as providing a measure of "unadulterated preferences – that is, preferences unaffected by a particular frame," McNeil, Pauker, & Tversky offer a more guarded interpretation, arguing that the mixed frame "may be more helpful because it draws attention to both the positive and the negative aspects of the outcome." McNeil, Pauker, & Tversky conclude that if a given individual makes different decisions when a problem is framed differently, then the data are insufficient for drawing an inference about normatively relevant preferences. However, neither paper elicited the choices of a given participant in different frames.

Our experimental design develops this tradition in four main ways. First, our within-subject design elicits every participant's choice in every frame. Second, we have participants make a choice in each frame before facing both frames together. This ensures that participants think about which choice is attractive in each frame. Moreover, the reconsideration of earlier choices prompts more cognitive engagement and deliberation. Third, we apply the approach systematically to the normative axioms underlying expected utility theory. Fourth, we apply it iteratively over many rounds. As mentioned in the Introduction, this feature of the design allows us to check for non-monotonic changes of behavior after additional reconsideration.

VI.E. Complementarity of Traditions and Contributions of This Paper

We think it is helpful to view all four experimental traditions through the philosophical lens outlined in the Introduction: their results provide evidence about the choices participants make under conditions closer to idealized conditions of full deliberation and absence of cognitive error. We view the evidence from the different approaches as complementary, for three reasons. The first reason is philosophical: since there are many dimensions of "deliberation" and many

[14] In the context of time preferences, Frederick & Read (2021) used a similar approach (although not two framings of the *same* decision problem) to study the "magnitude effect": people discount smaller amounts of money more heavily than larger amounts of money. Making the common but questionable assumption that discounting over money corresponds to discounting over consumption (Frederick, Loewenstein, & O'Donoghue 2002), Frederick & Read argue that the magnitude effect violates the normative principle that discount rates should be independent of magnitude. They assessed experimental participants' discounting over receipt of $10 or $1,000 and replicated the typical finding of greater discounting for $10. They asked some participants *jointly* about both amounts and find no diminution of the magnitude effect for these participants. They conclude that participants either reject the normative principle that discount rates should be independent of magnitude or that participants fail to appreciate this principle.

types of "cognitive error," each type of experiment may be conceived as moving closer to the idealized conditions along a different path in a multidimensional space. The second reason is practical: in applications, some approaches may be more natural than others. For example, during financial advising, we conjecture that our reconsideration procedure would be more palatable to many clients than, for example, asking them to simply make the same choice again.

The third reason is related to an inherent tradeoff in experimental design that was expressed clearly by Slovic & Tversky (1974): designs that more fully ensure participants understand the axioms and their appeal generate stronger experimenter demand effects. The various traditions (and varying designs within a tradition) make this tradeoff differently. For example, since experiments in the first and second traditions make no attempt to convey to participants that their initial choices violated an axiom, the results from these experiments—some movement of choices toward satisfying expected-utility axioms—plausibly provides a lower bound on how much participants would endorse the axioms after more thorough deliberation (with the caveat, noted above, that the second tradition may increase some kinds of cognitive error). As another example, since our experiment likely has some experimenter demand effects pushing participants toward consistency with axioms, if our finding that reconsideration does *not* reduce violations of the Irrelevance of Counterfactual Choices proves to be robust, it would imply that counterfactual reference points are normatively relevant, at least for some participants. It should be a priority for future papers to systematically vary features of the experimental design that may differentially influence experimenter demand effects to map out how they affect the findings.

Because, as mentioned above, the fourth tradition of experiments has some advantages over the third tradition, developing this fourth tradition is a main contribution of our paper. Moreover, while there is no perfect experimental design, *improvements* in design are possible, in the sense of designs that are closer to the tradeoff frontier. We believe our breaking down the Independence Axiom into easier-to-understand axioms is one such improvement because (holding constant the experimental procedure for prompting participants to deliberate) it makes it easier for participants to see the appeal or lack of appeal of the axiom to them. Relative to prior work, our results much more clearly point to participants' endorsement of most components of the Independence Axiom, while at the same time pinpointing the possible exception of a particular component related to a counterfactual reference point.

VII. Conclusions

In this paper, we elicited a sequence of hypothetical investment choices from experimental participants, and then confronted participants with cases where their choices were intransitive or inconsistent with other normative axioms of expected utility theory, asking if they would like to reconsider their choices, and if so, how. In our data, this reconsideration procedure virtually eliminates intransitivities and substantially reduces the frequency of inconsistencies with other axioms. The remaining inconsistencies are concentrated among relatively few participants. The one axiom for which the frequency of inconsistencies does not decline suggests that a counterfactual reference point may cause normative preferences to deviate from expected utility theory, although for a minority of our experimental participants (given the choices we study), and in a way not consistent with a simple model of reference-dependent preferences (laid out in Web Appendix G).

Our results suggest, however, that other inconsistencies with expected utility are mainly mistakes, rather than normative preferences. For example, violations of the Reduction of Compound Lotteries axiom, which have received much attention from decision theorists and experimentalists (see, e.g., Halevy 2007; Shechter 2020), appear largely to be mistakes.

Three notes of caution are in order. First, while we see substantial reductions in inconsistencies over each of eight rounds of reconsideration (across two waves), we do not know what would happen after additional rounds of reconsideration. Our finding that eight rounds is insufficient for reconsidered choices to converge is itself of interest, but further work is needed to determine whether with additional rounds, inconsistency rates would stabilize or decline further. Moreover, although in our structural estimation we find no evidence of a change in risk aversion resulting from our reconsideration procedure, this could be due to the high initial level of risk tolerance in our student sample. We do not know whether reconsideration would lead to a systematic change in risk aversion in older populations.

Second, while we took many steps to minimize experimenter demand effects, in retrospect we can see several ways we might have gone even further. For example, the instructions for reconsidering inconsistencies stated: “In one question you chose [Option 1] over [Option 2], but in another question you chose [Option 2] over [Option 1].” The word “but” subtly suggests that the participant erred. For future experiments, we would recommend using the word

"and" instead. As another example, we labeled choices by their paths in the master decision tree (e.g., C and D). We did so to help make the axioms transparent to participants, but participants might have inferred that across two frames, they should choose the options that share the same label. For future experiments, we recommend exploring the alternative of labeling options differently in different frames, or even using the same label to refer to options that are different (according to a normative axiom) as a placebo treatment. Another placebo treatment would be to ask participants if they want to change a reconsidered choice back to an earlier, untutored choice.

Third, our procedure assumes that reconsidered choices are closer to normatively relevant preferences, but we have not validated that assumption. In principle, one could test whether after an experiment like ours, participants make real-world choices more in line with their reconsidered choices. If so, it would suggest that participants learned from the experiment and recognized that their reconsidered choices are closer to what they really prefer. Since we do not have such real-world data, the best we can do is test whether the change from untutored to reconsidered choices in wave 1 is reflected in untutored and reconsidered choices in wave 2. In Section IV.B, we report a relevant finding that is supportive: comparing wave 1 to wave 2, reconsidered choices are more concordant across waves than untutored choices are, suggesting that participants' choices are converging toward the same preferences in the two waves.

While our experiment was motivated by the applied problem of helping people make better retirement investment decisions, it is a proof of concept: our experimental procedure is likely too long and complicated to be used directly for financial advice. A useful next step would be to develop a simplified version with fewer questions that could be used as an input into financial advice or as an immediate precursor to having individuals choose their retirement plan asset allocation. Our results could help the simplification process (with the caveat that our findings for undergraduates may not generalize to the relevant population). For example, our results suggest that transitivity could be imposed on participants' choices, since virtually all of our participants were fully transitive in their reconsidered choices.

It would be helpful for practical applications if one of our frames elicited untutored choices that well approximate what the reconsidered choices in that frame will be: a "revelatory frame" in the terminology of Goldin (2015). In particular, one might hypothesize that participants revise their choices in the direction of what they chose in a "simpler" frame. In Web Appendix I.3, we report analyses that aim to test whether any of the frames have this relevatory

property. We find that none of our frames do. Moreover, surprisingly to us, we find no support for the hypothesis that people revise toward their choices in frames that would be naturally thought of as simpler.

While we developed the reconsideration procedure in the context of retirement investment choices, a similar procedure might be useful for helping to identify normatively relevant preferences in other types of choices. Within the realm of risk preferences, it would be interesting to apply the procedure to choices where loss aversion and probability weighting are known to influence untutored choices. Future work should also explore applying the procedure to other preference domains.

References

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

Allais, Maurice. 1979. "The So-Called Allais Paradox and Rational Decisions Under Uncertainty." In *Expected Utility Hypotheses and the Allais Paradox*, edited by Maurice Allais and Ole Hagen, 437–681. Dordrecht: D. Reidel.

Ambuehl, Sandro, B. Douglas Bernheim, and Annamaria Lusardi. 2017. "A Method for Evaluating the Quality of Financial Decision Making, with an Application to Financial Education." *NBER Working Paper Series* no. 20618.

Apesteguia, Jose, and Miguel A. Ballester. 2018. "Monotone Stochastic Choice Models: The Case of Risk and Time Preferences." *Journal of Political Economy* 126 (1): 74–106.

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): 537–79.

Benartzi, Shlomo, and Richard H. Thaler. 1999. "Risk Aversion or Myopia? Choices in Repeated Gambles and Retirement Investments." *Management Science* 45 (3): 364–381.

Benkert, Jean-Michel, and Nick Netzer. 2018. "Informational Requirements of Nudging." *Journal of Political Economy* 126 (6): 2323–55.

Bernheim, B. Douglas. 2016. "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics." *Journal of Benefit-Cost Analysis* 7 (1): 12–68.

Bernheim, B. Douglas, and Antonio Rangel. 2009. "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics." *Quarterly Journal of Economics* 124 (1): 51–104.

Bernheim, B. Douglas, and Dmitry Taubinsky. 2018. "Behavioral Public Economics." In *Handbook of Behavioral Economics: Applications and Foundations*, edited by B. Douglas Bernheim, Stefano Dellavigna, and David Laibson, 1:381–516. Amsterdam: North Holland.

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

Birnbaum, Michael H., and Ulrich Schmidt. 2015. "The Impact of Learning by Thought on Violations of Independence and Coalescing." *Decision Analysis* 12 (3): 144–52.

Breig, Zachary, and Paul Feldman. 2021. "Revealing Risky Mistakes through Revisions."

https://zacharybreig.com/papers/RMR.pdf.

Cacioppo, John T., and Richard E. Petty. 1984. "The Need for Cognition: Relationship to Attitudinal Processes." *Social Perception in Clinical and Counseling Psychology* 2: 113–40.

Campbell, John Y. 2006. "Household Finance." *The Journal of Finance* 61 (4): 1553–1604.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2018. "On the Relationship Between Cognitive Ability and Risk Preference." *Journal of Economic Perspectives* 32 (2): 115–34.

Druckman, James N. 2001. "Evaluating Framing Effects." *Journal of Economic Psychology* 22 (1): 91–101.

Eli, Vincent. 2017. "Essays in Normative and Descriptive Decision Theory." Paris-Saclay and HEC Paris.

Ferreira, João V. 2020. "On How to Disrespect Choice : A Comparison of Three Behavioural Proxies of Welfare." https://joaovferreira.weebly.com/uploads/9/5/0/5/95056108/how_to_disrespect_choice.pdf.

Frederick, Shane. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives* 19 (4): 25–42.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue. 2002. "Time Discounting and Time Preference: A Critical Review." *Journal of Economic Literature* 40 (2): 351–401.

Frederick, Shane, and Daniel Read. 2021. "Reflective Equilibrium & the Endorsement of 'Anomalous' Preferences: The Magnitude Effect as a Case Study."

Gaudeul, Alexia, and Paolo Crosetto. 2019. "Fast Then Slow: A Choice Process Explanation for the Attraction Effect." *Grenoble Applied Economics Laboratory Working Paper*. https://hal.archives-ouvertes.fr/hal-02408719.

Goldin, Jacob. 2015. "Which Way To Nudge? Uncovering Preferences in the Behavioral Age." *Yale Law Journal* 125: 226–70.

Goldin, Jacob, and Daniel Reck. 2020. "Revealed-Preference Analysis with Framing Effects." *Journal of Political Economy* 128 (7): 2759–95.

Goodin, Robert E. 1986. "Laundering Preferences." In *Foundations of Social Choice Theory*, edited by Jon Elster and Aanund Hylland, 75–101. Cambridge: Cambridge University Press.

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): 504–28.

Halevy, Yoram. 2007. "Ellsberg Revisited: An Experimental Study." *Econometrica* 75 (2): 503–36.

Hausman, Daniel M. 2016. "On the Econ Within." *Journal of Economic Methodology* 23 (1): 26–32.

Hey, John D. 2001. "Does Repetition Improve Consistency?" *Experimental Economics* 4 (1): 5–54.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (2): 263–91.

Karni, Edi, and David Schmeidler. 1991. "Atemporal Dynamic Consistency and Expected Utility Theory." *Journal of Economic Theory* 54: 401–8.

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): 1028–38.

Kőszegi, Botond, and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics* 121 (4): 1133–65.

Kreps, David M., and Evan L. Porteus. 1978. "Temporal Resolution of Uncertainty and Dynamic Choice Theory." *Econometrica* 46 (1): 185–200.

Kuilen, Gijs van de. 2009. "Subjective Probability Weighting and the Discovered Preference Hypothesis." *Theory and Decision* 67 (1): 1–22.

Kuilen, Gijs van de, and Peter P. Wakker. 2006. "Learning in the Allais Paradox." *Journal of Risk and Uncertainty* 33 (3): 155–64.

Lipsey, R. G., and Kelvin Lancaster. 1956. "The General Theory of Second Best." *Review of Economic Studies* 24 (1): 11–32.

Loomes, Graham, and Robert Sugden. 1982. "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty." *The Economic Journal* 92 (368): 805–24.

MacCrimmon, Kenneth R. 1968. "Descriptive and Normative Implications of the Decision-Theory Postulates." In *Risk and Uncertainty*, 3–32. London: Palgrave Macmillan.

MacCrimmon, Kenneth R., and Stig Larsson. 1979. "Utility Theory: Axioms Versus 'Paradoxes.'" In *Expected Utility Hypotheses and the Allais Paradox*, edited by Maurice Allais and Ole Hagen, 333–409. Dordrecht: Springer.

McArdle, John, Willard Rodgers, and Robert Willis. 2015. "Cognition and Aging in the USA

(CogUSA) 2007-2009." Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

McNeil, Barbara J., Stephen G. Pauker, and Amos Tversky. 1988. "On the Framing of Medical Decisions." In *Decision Making: Descriptive, Normative, and Prescriptive Interactions*, edited by David E. Bell, Howard Raiffa, and Amos Tversky, 562–68. Cambridge: Cambridge University Press.

Miller, Paul M., and N. S. Fagley. 1991. "The Effects of Framing, Problem Variations, and Providing Rationale on Choice." *Personality and Social Psychology Bulletin* 17 (5): 517–22.

Moskowitz, Herbert. 1974. "Effects of Problem Representation and Feedback on Rational Behavior in Allais and Morlat-Type Problems." *Decision Sciences* 5 (2): 225–42.

Myagkov, Mikhail, and Charles R. Plott. 1997. "Exchange Economies and Loss Exposure : Experiments Exploring Prospect Theory and Competitive Equilibria in Market Environments." *American Economic Review* 87 (5): 801–28.

Nicholls, Nicky, Aylit Tina Romm, and Alexander Zimper. 2015. "The Impact of Statistical Learning on Violations of the Sure-Thing Principle." *Journal of Risk and Uncertainty* 50 (2): 97–115.

Nielsen, Kirby, and John Rehbeck. 2020. "When Choices Are Mistakes." https://kirbyknielsen.com/wp-content/uploads/kirby/Mistakes.pdf.

Plott, Charles R. 1996. "Rational Individual Behavior in Markets and Social Choice Processes: The Discovered Preference Hypothesis." In *The Rational Foundations of Economic Behaviour*, edited by Kenneth Arrow, Christian Schmidt, Enrico Colombatto, and Mark Perlman, 225–50. London: McMillian.

Quidt, Jonathan de, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and Bounding Experimenter Demand." *American Economic Review* 108 (11): 3266–3302.

Rabin, Matthew and Richard H. Thaler. 2001. "Anomalies: Risk Aversion." *Journal of Economic Perspectives*, 15 (1): 219-232.

Raiffa, Howard. 1961. "Risk, Ambiguity, and the Savage Axioms: Comment." *Quarterly Journal of Economics* 75 (4): 690–94.

Railton, Peter. 1986. "Moral Realism." *The Philosophical Review* 95 (2): 163–207.

Savage, Leonard J. 1954. *The Foundations of Statistics*. New York: John Wiley & Sons.

Segal, Uzi. 1990. "Two-Stage Lotteries without the Reduction Axiom." *Econometrica* 58 (2): 349–77.

Shechter, Steven. 2020. "Deconstructing the Allais Paradox: The Reduction of Compound Lotteries vs. the Independence Axiom." *SSRN Electronic Journal*.

Sieck, Winston R., Clark N. Quinn, and Jonathan W. Schooler. 1999. "Justification Effects on the Judgment of Analogy." *Memory and Cognition* 27 (5): 844–55.

Sieck, Winston, and J. Frank Yates. 1997. "Exposition Effects on Decision Making: Choice and Confidence in Choice." *Organizational Behavior and Human Decision Processes* 70 (3): 207–19.

Slovic, Paul, and Amos Tversky. 1974. "Who Accepts Savage's Axiom?" *Behavioral Science* 19 (6): 368–73.

Smith, Michael. 1994. *The Moral Problem*. 1st ed. Wiley-Blackwell.

Smith, Michael, David Lewis, and Mark Johnston. 1989. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society, Supplementary Volumes* 63: 89–174.

Thaler, Richard H., and Cass R. Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New York: Penguin Books.

Tversky, Amos. 1969. "Intransitivity of Preferences." *Psychological Review* 76 (1): 31–48.

Volij, Oscar. 1994. "Dynamic Consistency, Consequentialism and Reduction of Compound Lotteries." *Economics Letters* 46: 121–29.

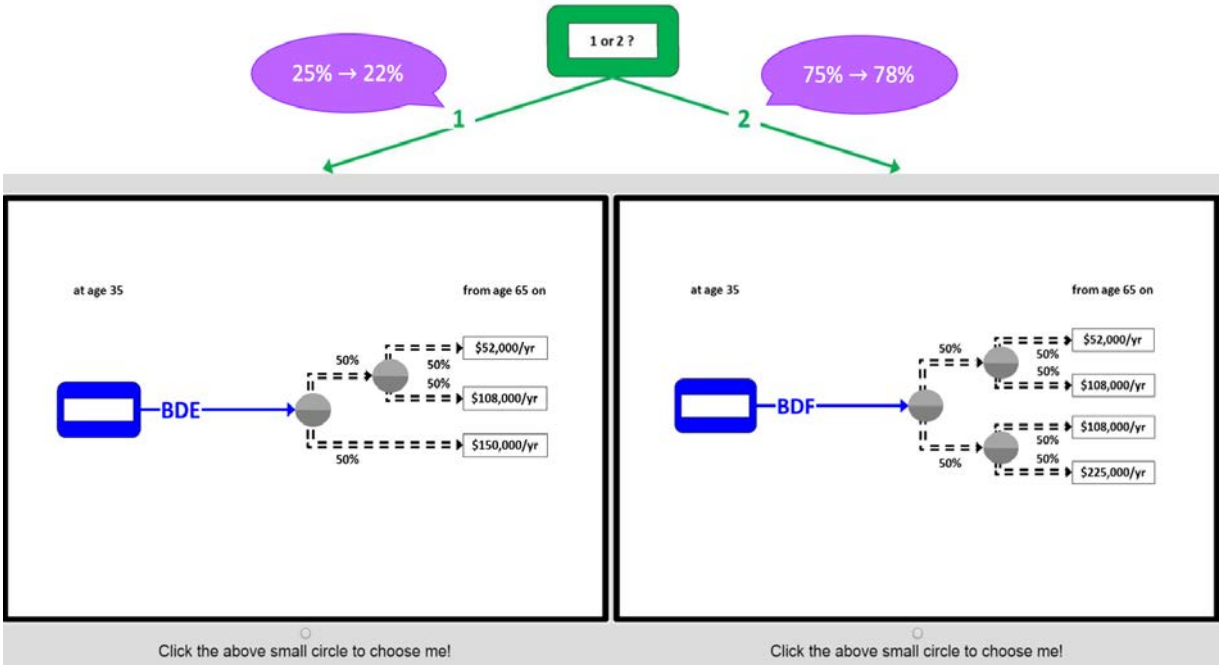
Yu, Chi Wai, Y. Jane Zhang, and Sharon Xuejing Zuo. 2021. "Multiple Switching and Data Quality in the Multiple Price List." *Review of Economics and Statistics* 103 (1): 136–50.

Zhang, Yalun, and Jason Abaluck. 2016. "Consumer Decision-Making for Prescription Drug Coverage and Choice Inconsistencies." Yale University.

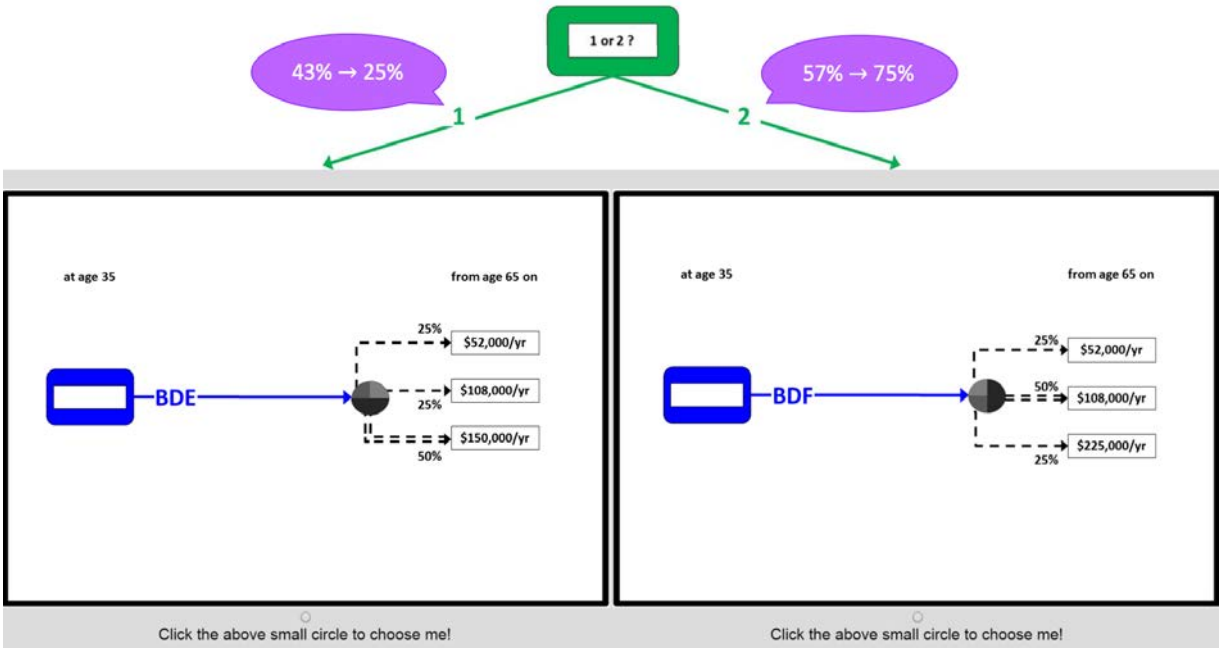
Zizzo, Daniel John. 2010. "Experimenter Demand Effects in Economic Experiments." *Experimental Economics* 13 (1): 75–98.

Figure 1: Examples of two frames

(A) Pairwise Choices Between Compound Lotteries (frame 6)



(B) Pairwise Choices Between Reduced Simple Lotteries (frame 7)



Note: Speech balloons are superimposed on screenshots. Speech balloons show, for the wave 1+2 sample, the percentages of participants choosing each possible action in the initial, untutored choices (wave 1, stage 0) and in the final choices (wave 2, stage 4). Percentages are aggregates over all the monetary levels (not just the particular monetary level depicted in the figure).

Figure 2: Timeline of experiment

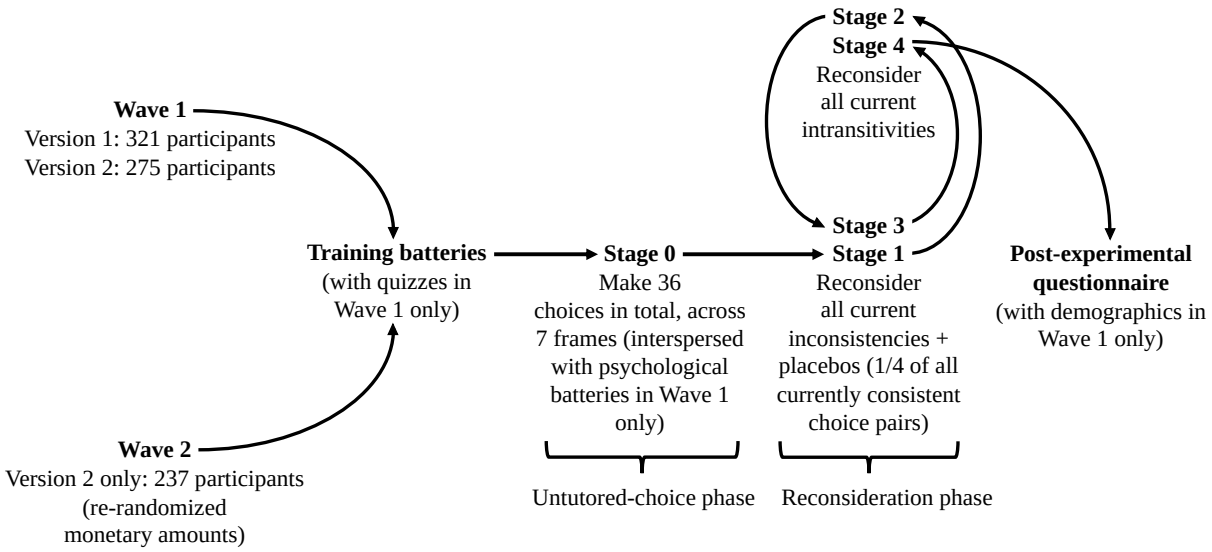
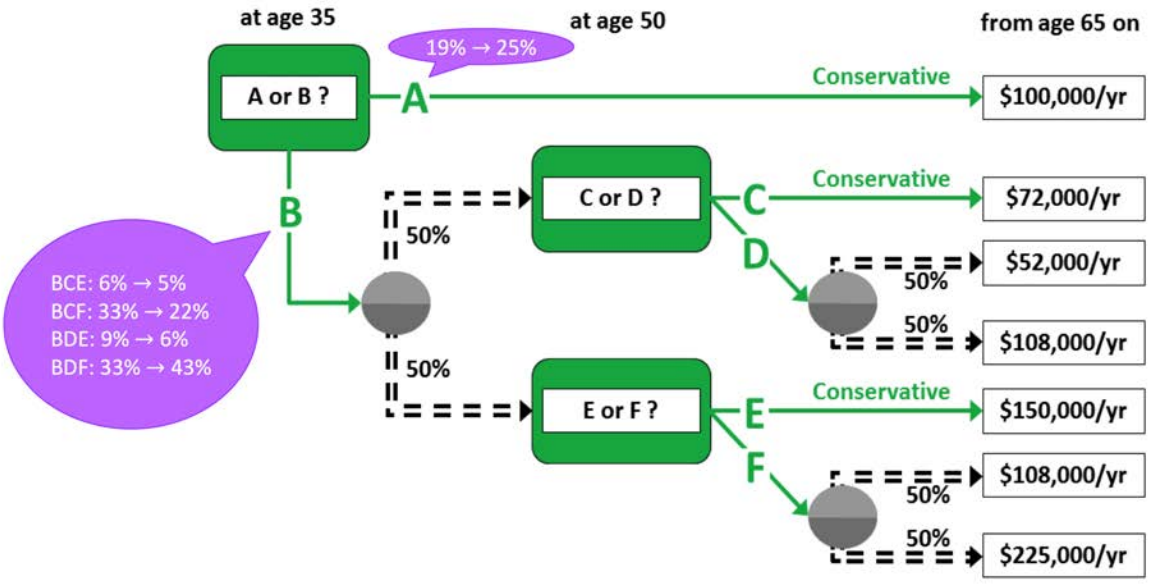


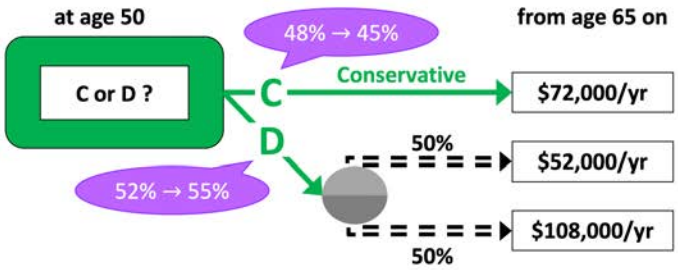
Figure 3: Master decision tree, which is also the frame Complete Contingent Action Plan (frame 4)



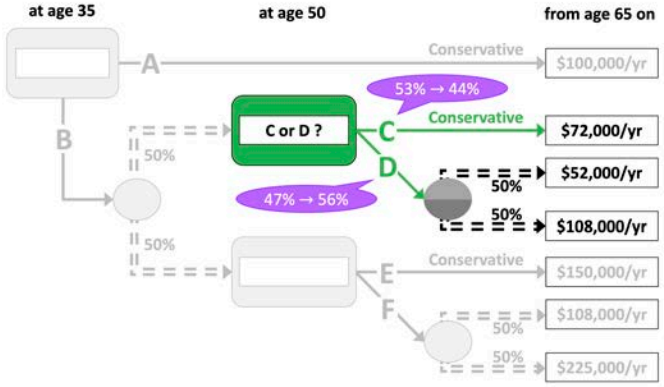
Note: Speech balloons are superimposed on screenshots. Speech balloons show, for the wave 1+2 sample, the percentages of participants choosing each possible action in the initial, untutored choices (wave 1, stage 0) and in the final choices (wave 2, stage 4). Percentages are aggregates over all the monetary levels (not just the particular monetary level depicted in the figure).

Figure 4: Examples of the remaining frames

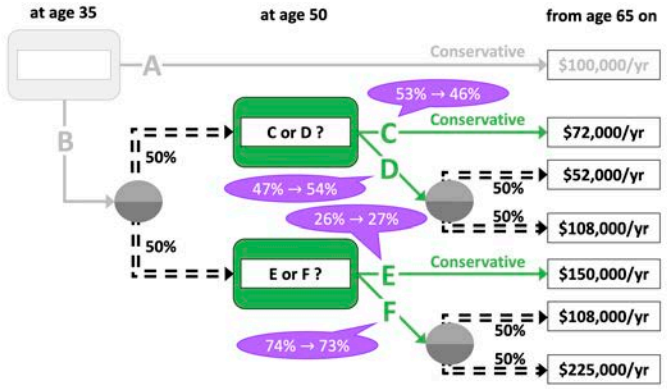
(A) Single Action in Isolation (frame 1)



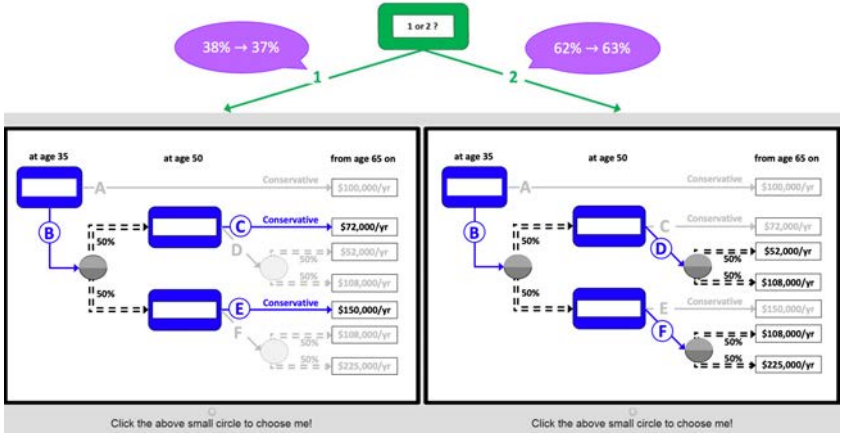
(B) Single Action with Backdrop (frame 2)



(C) Two Contingent Actions with Backdrop (frame 3)



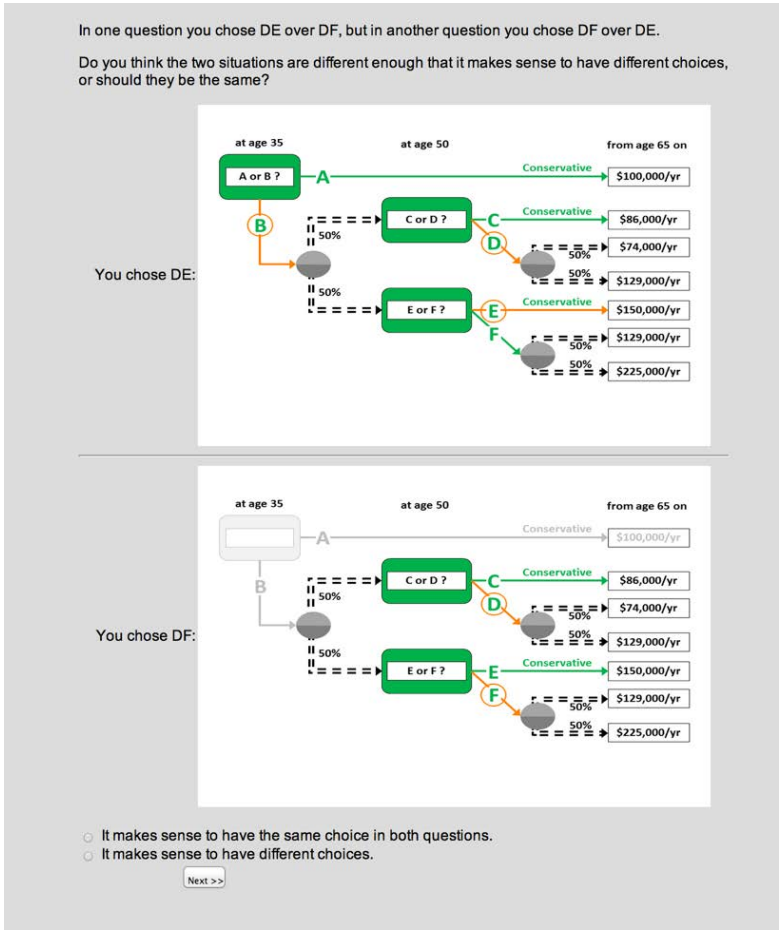
(D) Pairwise Choices Between Complete Strategies (frame 5)



Note: Speech balloons are superimposed on screenshots. Speech balloons show, for the wave 1+2 sample, the percentages of participants choosing each possible action in the initial, untutored choices (wave 1, stage 0) and in the final choices (wave 2, stage 4). Percentages are aggregates over all the monetary levels (not just the particular monetary level depicted in the figure).

Figure 5: Screenshots of inconsistency and intransitivity reconsiderations

(A) Inconsistency reconsideration



(B) Intransitivity reconsideration

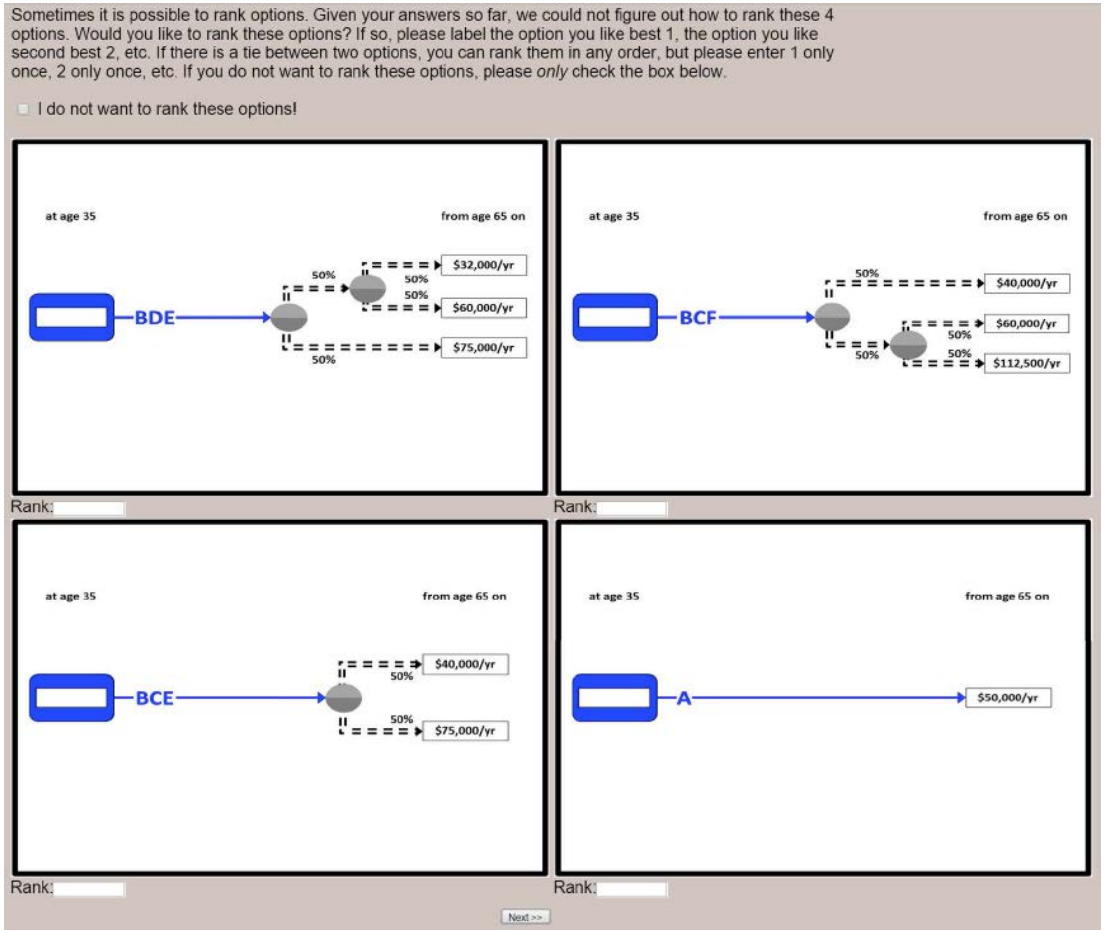
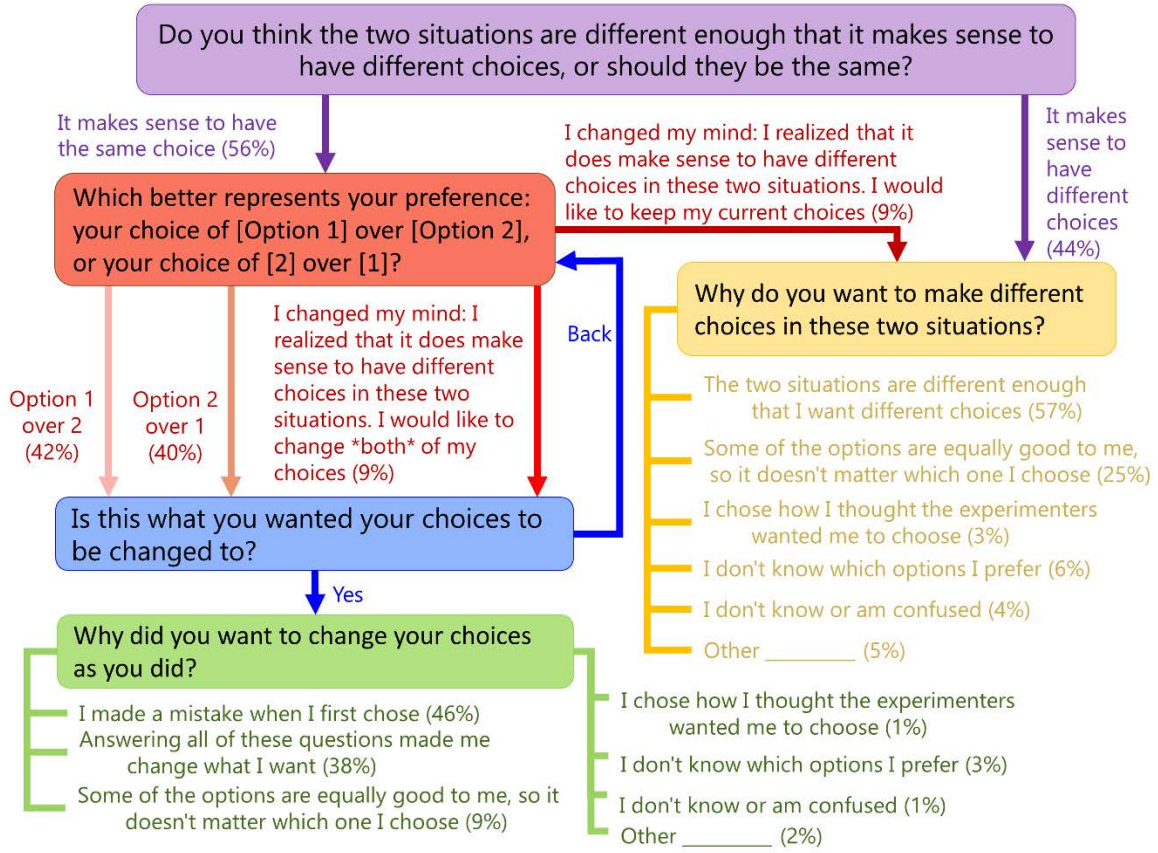


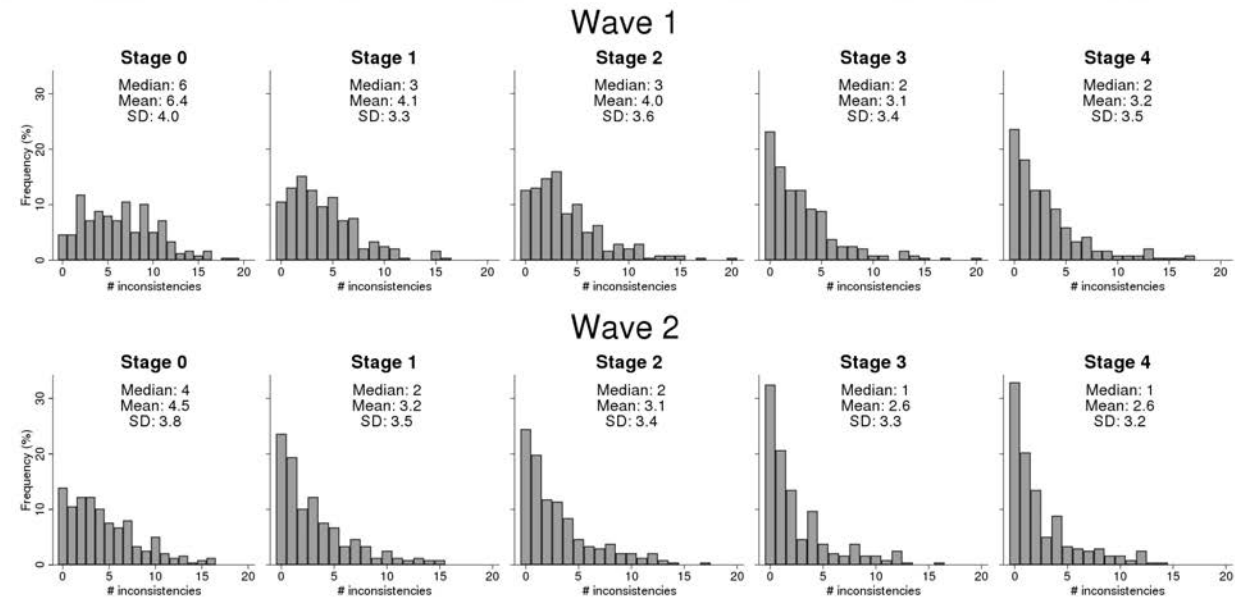
Figure 6: Flow chart for the inconsistency reconsideration procedure



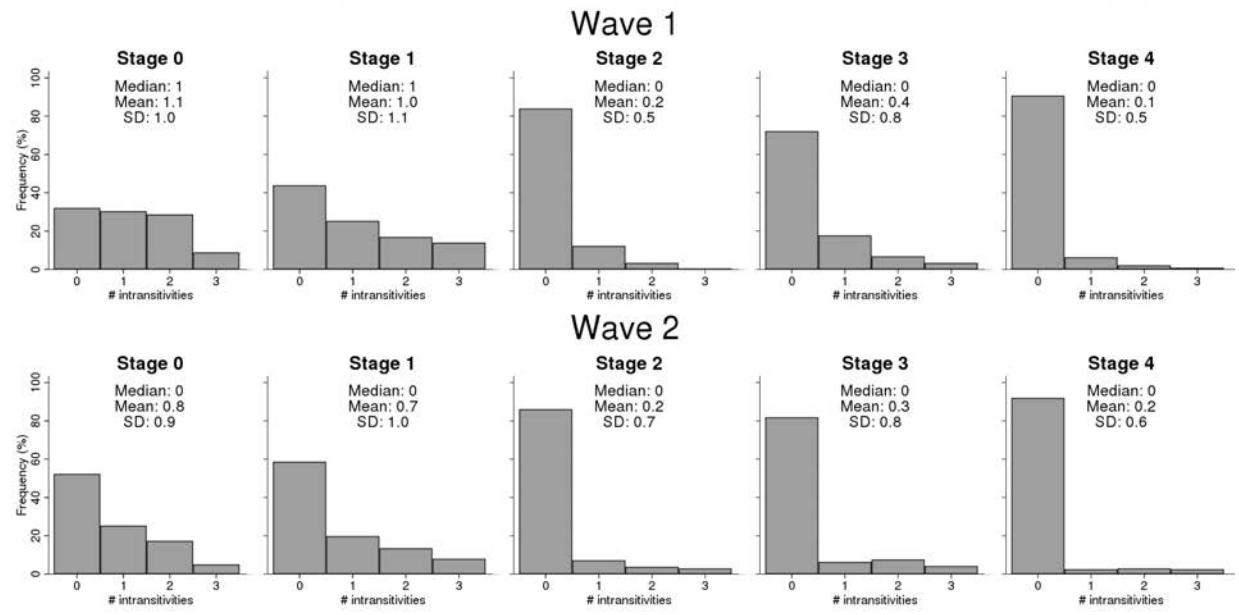
Note: The numbers in parentheses are frequencies of each choice across all instances of inconsistencies (not including placebos) in the Wave 1+2 sample. Percentages are rounded to the nearest 1% and therefore may not add up to 100%.

Figure 7: Histograms of number of inconsistencies and intransitivities

(A) Inconsistencies



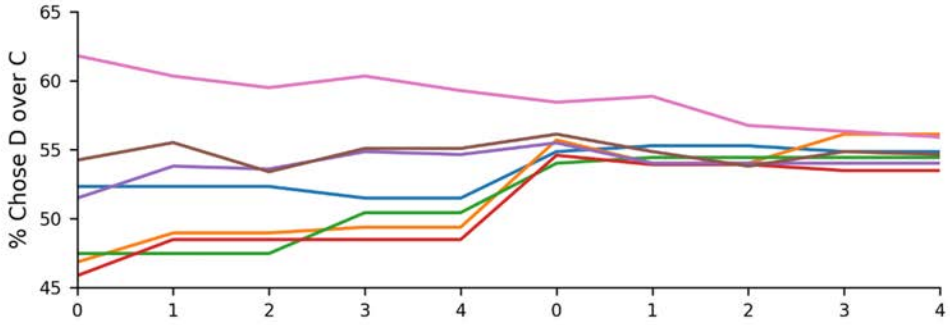
(B) Intransitivities



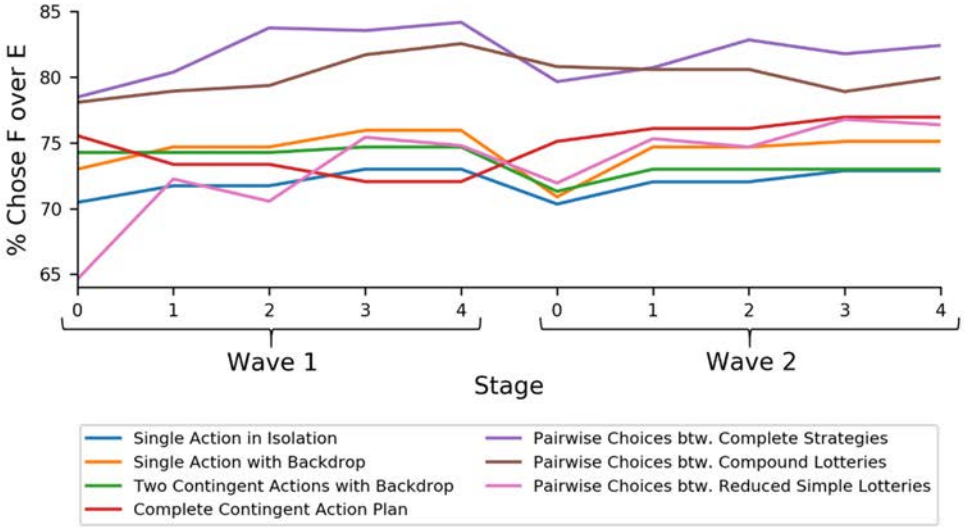
Note: Wave 1+2 sample (so that the sample is identical across waves).

Figure 8: Percentage of participants who make the risky choice in simple lotteries

(A) C versus D



(B) E versus F



Note: Wave 1+2 sample (so that the sample is identical across waves). For each of the three pairwise frames, the top panel reports the average from the two questions eliciting BCE vs. BDE and BCF vs. BDF, and bottom panel reports the average from the two questions eliciting BCE vs. BCF and BDE vs. BDF. Standard errors around each plotted point are roughly 2-3 percentage points (not shown to avoid cluttering the figure but reported in Web Appendix Table C.11).

Table 1: Responses after not revising an inconsistency

Axiom	Different Situation	Indiff	Expt'er Demand	IDK	Confused	Other	#Obs
Irrelevance of Background Counterfactuals	56.6%	22.4%	0.0%	9.2%	9.2%	2.6%	76
Simple Actions = State-Contingent Actions	74.8	15.9	1.9	1.9	4.7	0.9	107
Irrelevance of Counterfactual Choices	55.9	25.5	2.0	8.8	3.0	4.9	226
Fusion + Shift from Nodewise to Pairwise	63.3	20.4	1.8	7.1	3.5	4.0	226
Complete Strategies = Implied Lotteries	54.6	25.1	3.4	7.7	2.9	6.4	626
Reduction of Compound Lotteries	54.4	27.5	2.7	4.8	4.0	6.6	805
Overall	56.8	24.9	2.6	6.2	3.8	5.7	1942

Note: Wave 1+2 sample. Percentages are averages across all stages in both waves. The full text of the responses to the question "Why do you want to make different choices in these two situations?" after not revising an inconsistency are: "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". Percentages are rounded to the nearest 0.1% and therefore row percentages may not add up to 100%.

Table 2: Responses after revising an inconsistency

Axiom	Made Mistake	Learned	Indiff.	Expt'er Demand	IDK	Confused	Other	#Obs
Irrelevance of Background Counterfactuals	46.9%	38.3%	8.6%	1.2%	1.2%	1.2%	2.5%	81
Simple Actions = State-Contingent Actions	34.7	42.9	12.2	0.0	4.1	2.0	4.1	49
Fusion + Shift from Nodewise to Pairwise	51.2	31.7	8.5	1.2	3.7	3.7	0.0	82
Complete Strategies = Implied Lotteries	46.4	37.1	9.1	0.5	4.2	0.5	2.1	614
Reduction of Compound Lotteries	45.1	39.7	9.7	0.6	2.0	1.1	1.8	814
Overall	45.7	38.4	9.4	0.6	2.9	1.0	2.0	1640

Note: Wave 1+2 sample. Percentages are averages across all stages in both waves. The full text of the responses to the question "Why did you want to change your choices as you did?" after not revising an inconsistency are:
"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", and "Other".
Percentages are rounded to the nearest 0.1% and therefore row percentages may not add up to 100%.

Table 3: Responses after not revising an intransitivity

Frame	Indiff.	IDK	Real Intransitivity	Too Hard	Other	#Obs
Pairwise Choices between Complete Strategies	14.0%	37.2%	25.6%	9.3%	14.0%	43
Pairwise Choices Between Compound Lotteries	19.0	51.7	15.5	12.1	1.7	58
Pairwise Choices Between Reduced Simple Lotteries	11.9	45.8	20.3	18.6	3.4	59
Total	15.0	45.6	20.0	13.8	5.6	160

Note: Wave 1+2 sample. Percentages are averages across all stages in both waves. The full text of the responses to the question "Why couldn't you rank these options?" after not revising an intransitivity are: "I couldn't rank the options because they are all equally good to me", "I couldn't rank the options because I don't know which option I prefer", "I feel like Ian Trantivi on the game show. Remember Ian's story from earlier in the survey: he won a prize, and could choose between three piles of stuff, but he prefers the first pile to the second, the second pile to the third, and the third pile to the first", and "I couldn't rank the options for another reason". Among the three pairwise frames, when facing an intransitivity, the percentage of the time that participants did not revise was 21.9%, 31.5%, and 28.0%, respectively. Percentages are rounded to the nearest 0.1% and therefore row percentages may not add up to 100%.

Table 4: Concordance of choices across waves

Frame	(1) Max # Poss. Concord. Choices	(2) Wave 1 vs. 2 Concordance Rate		(3) Random Choice Concordance Rate	(4) *p*-value: Stage-0 Concordance Rate = Stage-4 Concordance Rate	(5) *p*-value: Stage-0 Concordance Rate = Random Concordance Rate	# Obs
		Stage 0	Stage 4				
Single Action in Isolation	2	67.4%	69.1%	50.0%	0.2286	<0.0005	236
Single Action with Backdrop	2	67.1	71.7	50.0	0.0050	<0.0005	237
Two Contingent Actions with Backdrop	2	68.6	67.4	50.0	0.3973	<0.0005	236
Complete Contingent Action Plan	1	43.6	44.1	20.0	0.8623	<0.0005	227
Pairwise Choices Between Complete Strategies	10	69.3	73.3	50.0	<0.0005	<0.0005	236
Pairwise Choices Between Compound Lotteries	10	70.4	74.5	50.0	<0.0005	<0.0005	234
Pairwise Choices Between Reduced Lotteries	10	68.0	72.8	50.0	<0.0005	<0.0005	234
Overall	37	68.1	72.2	49.2	<0.0005	<0.0005	221

Note: Wave 1+2 sample, restricted to participants who are not missing any data from either wave for that frame. Concordance rate = (total number of same choices)/(max. number of possible same choices). *P*-values are from two-sided *t*-tests. Percentages are rounded to the nearest 0.1%.

Table 5: Average inconsistency rates by axiom

Axiom	(1) Inconsistency Rate Wave 1	(2) Inconsistency Rate Wave 1	(3) Inconsistency Rate Wave 2	(4) Inconsistency Rate Wave 2	(5) *p*-value (1)-(2)	(6) *p*-value (1)-(3)	(7) *p*-value (3)-(4)	(8) *p*-value (1)-(4)	#Obs
	Stage 0	Stage 4	Stage 0	Stage 4					
Wave 1 sample									
Irrelevance of Background Counterfactuals	12.5%	5.7%			<0.0001				595
Simple Actions = State-Contingent Actions	11.9	8.1			0.0002				592
Irrelevance of Counterfactual Choices	12.5	15.0			0.0325				578
Fusion + Shift from Nodewise to Pairwise	23.4	13.4			<0.0001				579
Complete Strategies = Implied Lotteries	19.5	8.3			<0.0001				590
Reduction of Compound Lotteries	23.0	8.3			<0.0001				591
Overall	20.1	9.2			<0.0001				571
Wave 1+2 sample									
Irrelevance of Background Counterfactuals	12.9	5.9	8.3	3.8	<0.0001	0.0259	0.0012	<0.0001	236
Simple Actions = State-Contingent Actions	12.3	9.1	6.1	6.1	0.0218	0.0025	1.0000	0.0020	236
Irrelevance of Counterfactual Choices	13.9	17.2	11.8	14.2	0.0781	0.3713	0.1830	0.9045	221
Fusion + Shift from Nodewise to Pairwise	24.1	14.5	17.7	11.0	<0.0001	0.0052	<0.0001	<0.0001	226
Complete Strategies = Implied Lotteries	22.1	10.1	14.6	8.1	<0.0001	<0.0001	<0.0001	<0.0001	233
Reduction of Compound Lotteries	26.1	10.4	18.3	8.9	<0.0001	<0.0001	<0.0001	<0.0001	232
Overall	22.4	11.0	15.1	8.4	<0.0001	<0.0001	<0.0001	<0.0001	216

Note: Top panel: wave 1 sample. Bottom panel: wave 1+2 sample. Inconsistency rate = (total number of inconsistencies)/(number of potential inconsistencies). *P*-values are from two-sided *t*-tests.

Table 6: Results from structural estimation

A. Pairwise Choices Between Complete Strategies						**B. Pairwise Choices Between Compound Lotteries**				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
	ln(CRRA)	$\ln(\sigma_\epsilon)$	σ_ν^2	$\sigma_{\eta_1}^2$	$\sigma_{\eta_2}^2$	ln(CRRA)	$\ln(\sigma_\epsilon)$	σ_ν^2	$\sigma_{\eta_1}^2$	$\sigma_{\eta_2}^2$
wave2	0.173	-0.440				0.073	-0.414			
	(0.231)	(0.159)				(0.237)	(0.154)			
stage	-0.003	-0.095				0.021	-0.103			
	(0.021)	(0.015)				(0.021)	(0.015)			
wave2*stage	0.012	0.060				-0.002	0.060			
	(0.026)	(0.021)				(0.027)	(0.021)			
constant	-0.291	0.775	2.088	1.989	1.488	-0.394	0.788	2.436	1.446	1.478
	(0.193)	(0.110)	(0.439)	(0.668)	(0.639)	(0.201)	(0.110)	(0.480)	(0.599)	(0.658)

C. Pairwise Choices Between Reduced Simple Lotteries						**D. Pooled Data Across the Three Frames**				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
	ln(CRRA)	$\ln(\sigma_\epsilon)$	σ_ν^2	$\sigma_{\eta_1}^2$	$\sigma_{\eta_2}^2$	ln(CRRA)	$\ln(\sigma_\epsilon)$	σ_ν^2	$\sigma_{\eta_1}^2$	$\sigma_{\eta_2}^2$
wave2	0.060	-0.438				0.106	-0.442			
	(0.242)	(0.165)				(0.220)	(0.132)			
stage	-0.015	-0.111				0.005	-0.107			
	(0.025)	(0.015)				(0.014)	(0.008)			
wave2*stage	0.009	0.051				0.005	0.057			
	(0.030)	(0.020)				(0.017)	(0.012)			
constant	-0.275	0.979	2.461	1.330	1.656	-0.407	0.945	2.603	1.691	1.808
	(0.214)	(0.123)	(0.495)	(0.690)	(0.707)	(0.183)	(0.092)	(0.473)	(0.619)	(0.662)

Note: Wave 1+2 sample. #Obs is 21330 choices. Standard errors in parentheses. In panels A, B, C, and D, respectively, the difference between $\sigma_{\eta_1}^2$ (columns 4) and $\sigma_{\eta_2}^2$ (columns 5) is 0.305 (1.152), -0.011 (1.128), -0.203 (1.259), and 0.033 (1.150).