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

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 Paul Feldman, Alexia Gaudeul, Mike Gideon, Ori Heffetz, Eric Johnson, Ben Lockwood, Ted O'Donoghue, David Laibson, Kirby Nielsen, Matthew Rabin, 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 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, and Fudong Zhang 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
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 disentangle preferences from mistakes: after eliciting untutored choices, we confront participants with their choices that are inconsistent with expected-utility axioms (broken down enough to be self-evident) 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 expected-utility axioms, with one exception related to regret.

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 and 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 and 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 there is reason for concern that people's untutored choices may not accurately reflect their preferences, since for most people, risky decisionmaking (even if simplified) is unfamiliar and cognitively challenging. Moreover, people's choices often violate basic, seemingly compelling axioms of expected utility (EU) theory, such as the Reduction of Compound Lotteries axiom. While there are economic models that can accommodate such non-EU behavior (e.g., Segal 1990), most are explicitly meant to be descriptive (rather than normative) and thus admit the interpretation that non-standard behaviors represent mistakes, rather than features of actual preferences.

In this paper, we develop a two-stage procedure to measure risk preferences and implement it via a survey administered to a sample of 596 Cornell students. The first stage is the standard method of eliciting choices over risky lotteries. We use hypothetical choices about investing for retirement. We call participants' behavior in this stage 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 stage: We confront participants with *inconsistencies* in their untutored choices: their inconsistent responses to choices framed differently that should be the same according to normative axioms of expected utility theory. We also confront participants with *intransitivities* in their untutored choices. We ask 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 major potential concern with such a

[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, e.g., using a qualitative scale. 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.

reconsideration procedure is experimenter demand effects (e.g., Zizzo 2010; de Quidt, Haushofer, and Roth 2018), as we discuss in more detail below.

The key assumption underlying our procedure is that, relative to the untutored choices, the reconsidered choices are a better indicator of the individual's *normative preferences* (which could be used to give people advice or set policy). By normative preferences, we mean the choices an individual would make after full deliberation and absent cognitive error (as in typical philosophical definitions of well-being: e.g., Railton 1986; Hausman 2011).[2] Our approach can be viewed as an implementation of the "behavioral revealed preference" program (Bernheim and Rangel 2009; Bernheim 2016), where we identify a particular frame—namely, the frame generated by understanding of the axiom and full deliberation (which we aim to approximate with our reconsideration procedure)—as the welfare-relevant domain, in which choices are identified with preferences. Our assumption builds on an ancient tradition in moral philosophy, in which individuals aim to achieve a state of harmony between particular judgments and general principles, often called the "method of reflective equilibrium" (Rawls 1971). Decision analysis (e.g., Raiffa 1968) typically draws on a similar method, in which individuals formulate their preferences by resolving internal inconsistencies between initial choices. Our assumption is also related to a traditional defense of the expected utility axioms as normative: the claim that when it is pointed out that people's choices violate the axioms, people would endorse the axioms and reject their choices (e.g., Savage 1954; Raiffa 1961; Morgenstern 1979). To be clear, we do not claim that revised choices are closer to normative preferences in all contexts. For example, time-inconsistent preferences may also cause people to revise choices, typically *away* from normative preferences as a tempting option becomes closer in time (as in Giné et al. 2018). We believe our assumption is most reasonable for abstract choices made in a deliberative state of mind—a context well approximated in our experimental setting. Given our assumption, we can use our

[2] As Hausman (2016) argues, the attribution of what we are calling normative (i.e., latent, error-free) preferences to an individual is theoretically useful for economists but does not presuppose that they truly exist. Normative 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 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 normative preferences. See also Ferreira (2018) for a philosophical defense of "confirmed choices" as a proxy for welfare.

reconsideration procedure to separate mistakes from deliberate axiom violations, and to obtain a better measure of risk preferences for normative purposes.[3]

We conduct 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 stage of our experiment elicited untutored risk 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 left and a safer lottery ("BDE") pictured on the right. 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. As in much prior work on framing effects and non-expected utility (e.g., Kahneman and Tversky 1979), we find that participants often make inconsistent choices across different frames. In the Reduction of Compound Lotteries example, among those who participated in both waves of our experiment, 26.1% made different choices in their initial, wave-1 untutored choices.

In the second stage, we confront participants with intransitivities and inconsistencies in their untutored choices. Continuing the example, suppose that in the first stage, 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 the same choice, then we give the participant the option of changing one or both decisions.

As noted above, the key concern with an experimental design like this, in which participants are asked if they want to change choices they made earlier, is a possible "experimenter demand effect": 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

[3] In our context of retirement investment choices, other approaches to inferring normative preferences (as in Beshears et al. 2008) are difficult to implement. For example, we typically cannot observe what choices people would make with repetition and feedback because retirement investment choices are only made once, and we cannot rely on the choices made by experienced or expert decision-makers if there is meaningful heterogeneity in risk preferences.

experimental design included 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 our intent less obvious than if we had urged participants to make their choices consistent. 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 one of their choices, we selected pairs of choices that were already *consistent* with the relevant normative axiom. By doing so, we further masked our intentions and also obtained a placebo measure of how often people change their choices when prompted to do so. 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 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 easy-to-understand subcomponents, one of which is Reduction of Compound Lotteries. These six axioms plus transitivity, together with completeness and continuity (which we assume but do not test), imply expected utility theory. When we confronted experimental participants with choices inconsistent with an axiom, it was thus self-evident why someone might want the choices to be consistent, and we left it up to the participants to decide if they thought the situations were sufficiently different to warrant different choices. 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" and "inconsistencies" (all other axiom violations). 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 (relative to a benchmark of 50% for random choice). The reconsideration procedure leads to substantial movement overall toward endorsing the normative axioms. By the end of the second wave of the experiment, the average inconsistency rate fell to 8.4%. Respondents similarly exhibit 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 anticipated regret or 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.

To complement our descriptive analyses, we conduct a structural estimation. While our descriptive analyses allow for the possibility that participants' preferences may violate one or more expected-utility axioms, our structural estimation assumes not only expected-utility theory but, more specifically, constant relative risk aversion (CRRA). Within each of three frames, pooling choices across participants, we estimate participants' relative risk aversion and response-error variance. In the structural model, intransitivities and any other within-frame inconsistencies are attributed to response error. We estimate that response-error variance declined with each round of reconsideration. This result largely reflects the finding from our descriptive analysis that intransitivities declined over the course of the experiment. We estimate a high level of risk tolerance in our sample (close to log utility), and 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 and Taubinsky 2018 for a review) and methods of using observed choices to draw inferences about normative preferences (e.g., Benkert and Netzer 2018; Goldin and Reck forthcoming). It is most closely related to other work that also aims to learn about normative preferences by offering experimental participants an opportunity to revise their choices (both an older literature pioneered by MacCrimmon 1968 and several recent papers: Gaudeul and Crosetto 2019; Breig and Feldman 2020; Nielsen and Rehbeck 2020).[4] In Section V, we briefly

[4] Less directly related but in a similar spirit to the experiments reviewed here, Zhang and 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

review that body of work, categorize the experiments that have been conducted, and place this paper in the context of that literature.

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. Section III presents descriptive results. In Section IV, we report results from our structural estimation of CRRA utility. Section V discusses related literature, and Section VI concludes. The Survey Appendix contains screenshots from the experiment. The Web Appendixes contain additional and robustness 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.

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 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

member) or because they were satisfied enough with their current plan and were afraid the new plan may be worse for an unanticipated reason. Our paper is also related to Ambuehl, Bernheim, and 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, and Lusardi use the difference in willingness-to-pay across frames for the assets with the same dollar value as a measure of participants' mistakes.

who participated in both waves of the experiment, we refer to the sample as "wave 1+2." The wave 1+2 sample includes the 237 who came back 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.[5] 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 decided to have 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. Our design decisions have disadvantages relative to studying an older sample for whom the choices would be more familiar and relevant, and/or using incentivized small-stakes gambles. 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:

[5] 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.

1. **Training Batteries:** Explained the symbols used in the investment choices and described background assumptions we wanted subjects to hold while making their choices.[6] Participants could not continue until they correctly answered nearly all quiz questions about the training.
2. **Elicitation of Untutored Preferences:** Choices between hypothetical investments.
3. **Psychological and Cognitive Batteries:** These included: the cognitive reflection task (Frederick 2005); a number series task (McArdle, Rodgers, and Willis 2015); a 10-question Big-Five personality battery (Gosling, Rentfrow, and Swann 2003); a probabilistic sophistication battery (developed by Miles Kimball); and the need for cognition scale (Cacioppo and Petty 1984). These batteries were interspersed between sets of questions from part 2 in a random order.
4. **Reconsideration of Preferences:** Opportunities to revise untutored preferences.
5. **Post-Experimental Questionnaire and Demographics:** Questions such as how much the experiment made them feel enjoyment, annoyance, stress, and frustration, and demographic questions.

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.[7] 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 a re-randomized version of parts 2 and 4.

II.B. Instructions

[6] 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 (5.6) than those who did not (6.0), but the *p*-value for the difference is 0.28 (Web Appendix Table C.12).

[7] 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.

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.
>
> **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 2. 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 2 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 2 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, and 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 2 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 36 risky choices, each on a separate screen. 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. After describing the frames and the normative axioms implicitly defined by them, we explain our rationale for studying these frames.

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 2. The rest are shown in Figure 3. 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 are randomized to be rightside up or upside down. In all pairwise frames, in any given wave, *both* contingent plans are either shown in the orientation of the figures or in an upside-down configuration, whichever is consistent with how the participant was randomized for the nodewise frames. On each screen of a pairwise frame, we randomize which contingent plan is 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. We call the resulting six normative axioms:

1. Irrelevance of Background Counterfactuals
2. Simple Actions = State-Contingent Actions
3. Irrelevance of Counterfactual Choices
4. Shift from Nodewise to Pairwise
5. Complete Strategies = Implied Lotteries
6. Reduction of Compound Lotteries

Using the machinery for dynamic choice under uncertainty (Kreps and 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 and Schmeidler 1991; Volij 1994). 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.

We chose the frames such that the implied normative axioms 1-6 would satisfy two criteria: transparency and yielding expected utility theory. The expected utility theory criterion is that, together with completeness and continuity (which we assume but do not test) and the transitivity axiom, axioms 1-6 are necessary and sufficient for preferences to satisfy expected utility theory. The transparency criterion is that the argument for making the "same" choice across adjacent frames would be self-evident to experimental participants. In Web Appendix F, we prove the equivalence between the axioms and expected-utility theory. Aside from Reduction of Compound Lotteries, which is a standard axiom, the others are implicit in the setup of expected utility theory or are components of the well-known Independence Axiom. We broke down the Independence Axiom into "baby step" components because the Independence Axiom taken as a whole would fail the transparency criterion, as anyone who has taught intermediate microeconomics can attest. Because the axioms satisfy transparency, it was possible for experimental participants to endorse the logic of expected utility theory 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. We also chose the frames we did because we believe that axioms 1-6 are interesting in themselves as elements of normative economic reasoning.

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. With each frame, the ordering of questions was randomized.

II.E. Procedure for Reconsideration

After participants completed the elicitation of untutored preferences, 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 phase of the experiment 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*). The inconsistencies and placebos were presented in a random (therefore often interspersed) order, and 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: For all intransitivities among the stage-1 choices in pairwise frames, participants are given the opportunity to rank the options. Our algorithm for identifying intransitivities among the five strategies was as follows: (i) try to identify the highest- and lowest-ranked strategies; (ii) if we can identify both, then eliminate both and return to step (i); (iii) if we cannot identify both a highest- and lowest-ranked strategy, then we have an intransitivity. The number of strategies remaining determines whether we call it a 3-way, 4-way or 5-way intransitivity. (If this algorithm eliminates all strategies without identifying an intransitivity, then the stage-1 choices are transitive.) 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 4 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, are also depicted as a flowchart in Figure 5.

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]

We interpret the first of these options as suggesting that the participant rejects the axiom that implies that the choices should be the same. We interpret the second as suggesting indifference.

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 (also 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]

Assuming participants changed their choices to be consistent, we interpret the first two of these options as suggesting that the participant endorses the axiom that implies that the choices should be the same. We interpret the third as suggesting indifference.

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. Figure 4 panel B shows an example of a 4-way intransitivity reconsideration.

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]

We interpret the first response as suggesting indifference. The second and third responses do not distinguish between truly intransitive preferences, incomplete preferences, or insufficient effort to figure out the participant's truly complete, transitive preference order.[8]

III. Descriptive Results

III.A. Reconsideration Procedure

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.5). 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 5; 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).

Figure 6 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 frame later in this section and by axiom in Section III.B). 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

[8] As an additional attempt to avoid giving participants the impression that we disapprove of intransitivity, we asked a survey question about intransitivity in the abstract 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.

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.6). Thus, the reduction in inconsistencies from wave 1 to wave 2 suggests that participants had learned from the wave-1 procedure. 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 6 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 number 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 of intransitivities 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.

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 two of these as consistent with endorsing the axiom that implies that choices should be the same. 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. Of the 5.6% who selected "Other," most complained about being tired or noted that the task was too hard.

While we would like our reconsideration procedure to generate a more accurate elicitation of participants' normative preferences, an alternative possibility is that participants are averse to inconsistencies and intransitivities and revise to eliminate them, without getting closer to their normative preferences. We can obtain some relevant evidence by examining how consistent 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 within each wave but not converging

[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 have little confidence in 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.

toward the same set of choices. However, if they are revising toward *the same* preferences in each wave, then their choices should be more consistent across waves at the end of the waves than at the beginning of the waves. Table 4 shows an analysis that aims to test this hypothesis. Each row corresponds to a frame, except for the last row, which aggregates over all the data. Columns (1)-(4) show the number of potential inconsistencies and three percentages of *consistent* choices: untutored choices from waves 1 and 2, final choices (after stage 4 of the reconsideration procedure) from waves 1 and 2, and the expectation under random behavior. In aggregate, the amount of consistency 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 consistency 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 consistency 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 potential inconsistencies). We conclude from this analysis that participants do appear to be moving toward the same set of preferences in the two waves.

One way that participants might revise their choices that would eliminate inconsistencies and intransitivities but *not* necessarily represent their normative preferences is if they followed a heuristic. To obtain some relevant evidence, we considered four heuristics that would be simple to follow in this experiment: choose the option that maximizes expected value (EV), choose the option that minimizes EV, choose the option shown on the top of the screen, and choose the option shown on the bottom of the screen. In our data, behaviors consistent with these heuristics are confounded: for participants randomized to the rightside-up orientation, the top option is always the safe option that minimizes EV, and vice-versa for participants randomized to the upside-down orientation. We break these correlations by restricting the sample to the 109 participants who were in both waves but were randomized to opposite rightside-up/upside-down orientations in the two waves, and we pool together their choices from the end of wave 1 and the end of wave 2. The top panel of Table 5 shows the percentage of participants whose choices are consistent with each heuristic and with none of the heuristics. The choices of 93.6% of the sample do not fit any of the four heuristics. The bottom panel of Table 5 examines a less stringent criterion: at least 60 of the 68 choices (34 per wave) are consistent with the heuristic. According to this criterion, 80.7% of the sample do not fit any of the heuristics. We interpret this evidence as casting doubt on the possibility that our results are driven by participants behaving according to the heuristics we examine.

When participants revise their choices, do they revise in the direction of greater risk tolerance or greater risk aversion? In Section IV, we examine this question with the help of structural assumptions about preferences, but here we aim to shed light on it with descriptive data and minimal assumptions. 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 7 show, for the wave 1+2 sample, the percentage of participants choosing 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 7, 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. For each reconsideration in the wave 1+2 sample, and for each axiom $j = 1,2, \dots ,6$ as well as overall, the top panel of Table 6 shows how often participants revised toward their frame-j or frame-$(j + 1)$ choice in cases where their current frame-j choice was riskier. The middle panel shows the revision frequencies in cases where participants' current frame-$(j + 1)$ choice was riskier. Although not relevant for our purposes, for completeness the bottom panel shows the cases where the current frame-j and frame-$(j + 1)$ choices are not risk-ranked (e.g., CF vs. DE) (the first two rows have no observations because this is not possible for the choices in frames 1 and

2). As an example of how to read the table, consider the first row of the top panel, which shows results for Axiom 1 (Irrelevance of Background Counterfactuals). In cases of inconsistencies with Axiom 1 where participants' choice in the frame 1 (Single Action in Isolation) was riskier than in frame 2 (Single Action with Backdrop), the first column shows that participants revised to make both choices consistent with their (riskier) frame-1 choice 33.3% of the time. The second column shows that they revised to make both choices consistent with their (less risky) frame-2 choice 20.5% of the time. The third column gives the *p*-value for the null hypothesis that these percentages are equal, which is 0.1236. (The fourth and fifth columns show how often participants did not revise either choice and swapped their choices, and the sixth column gives the number of observations.) To facilitate reading the table, in each row we have bolded whichever frame-j or frame-$(j+1)$ number is larger. The main result from Table 6 can be seen from the "overall" rows of both panels: on average, when participants revise their choices, they do so toward the riskier choice. This tendency is also seen for most axioms (albeit sometimes with large *p*-values). This finding from examining revision behavior thus reinforces the conclusion from examining the C vs. D and E vs. F choices.

As a final question about the effects of the reconsideration procedure, we ask: is there a particular frame (or frames) toward which participants revise their choices? And in particular, do participants revise their choices in the direction of what they chose in a "simpler" frame? We conjecture that the cognitively simplest nodewise frame is Single Action in Isolation (frame 1), and the cognitively simplest pairwise frame is Pairwise Choices Between Reduced Simple Lotteries (frame 7). From Table 6 we find no evidence that participants revise toward their choices in these frames; on average, participants revise toward their choices in those frames if those were the riskier choices and away from them otherwise. There is only one frame where we see some evidence for participants revising toward their choice in that frame: Complete Contingent Action Plan (frame 4). Specifically, participants revised in the direction of their choice in frame 4 even when that was the less risky choice (see row 3 in the top panel and row 4 in the bottom panel). Although intriguing, we do not draw any confident conclusion about this frame because the *p*-values are large in both relevant rows.

III.B. 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 6 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, but Figure 6 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 have zero inconsistencies, 54.0% have ≤ 1, and 67.1% have ≤ 2. Only 14.9% of participants have > 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 7 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 rates fall substantially from the beginning of wave 1 to the end of wave 2 and are ≤ 11.0% by the end of wave 2. For example, the Reduction of Compound Lotteries axiom has an inconsistency rate at the beginning of wave 1 of 26.1%, which is the highest. By the end of wave 2, its inconsistency rate is 8.9%. In aggregate, inconsistency rates fall 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 normative preferences.

As noted above, the main exception is the Irrelevance of Counterfactual Choices axiom. As Table 7 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 normative preferences. It can be seen from Figure 4 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 anticipated regret (e.g., Loomes and Sugden 1982) or reference-dependent risk preferences (as in Kőszegi and Rabin 2006) 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. We conclude that the (in our view, plausible) form of reference-dependence we study does not predominantly explain violations of the axiom.

IV. Estimating CRRA Preferences

In our descriptive analyses in Section III.A, we found that intransitivities (as well as inconsistencies with other normative axioms) declined over the course of the experiment (Figure 6). When we restricted the data to C vs. D and E vs. F choices (Figure 7) and examined the frequency of revisions toward the choices made in particular frames (Table 6), we found some evidence that when participants revised their inconsistencies, they tended to do so in the direction of more risk tolerant choices. As a complementary approach to addressing these questions, in this section we estimate a structural model for risk aversion within each frame and examine how the risk aversion parameter and the variance in response error vary over the course of the experiment. The structural estimation has the advantages that it allows us to use more of the data when measuring risk aversion (not just C vs. D and E vs. F choices), control for covariates, model heterogeneity in risk preferences, incorporate response error into the analysis, and measure risk aversion in economically meaningful units. The disadvantages are that we assume preferences satisfy expected utility theory and impose a particular structure on those preferences and on the response error. Because we assume expected utility theory, intransitivities in participants' choices are attributed entirely to response error. Inconsistencies do not directly affect the structural estimates because we estimate the model separately within each frame, but they matter indirectly because changes in risk aversion occur in response to the reconsideration procedure. As is common in studies of retirement investment, we assume constant relative risk aversion (CRRA) preferences. Following Barsky et al. (1997) and Kimball, Sahm, and Shapiro (2008), we estimate a random parameter model (as prescribed by Apesteguia and Ballester 2018).

Specifically, our model is as follows. We separately estimate relative risk aversion in each frame (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}. As discussed more below, 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}.

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 could 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:

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

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. That is, 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:

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

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 estimate the parameters with maximum likelihood, using a Stata implementation of adaptive quadrature. For more details, see 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. Indeed, we cannot statistically distinguish the estimates of mean log risk aversion parameters $\{\mu_0, \mu_w, \mu_s, \mu_{ws}\}$ across these subsamples (although at $p < 0.05$, there is a difference in the estimate of the log standard deviation of the error parameter, τ_0, for one of the three frames; see Web Appendix Table E.8e).

Table 8 shows the results from estimating the model. The three panels A-C correspond to the three pairwise frames. The results are broadly similar across the three frames, so for concreteness, we walk through the parameter estimates only for Pairwise Choices Between Complete Strategies (panel A). Columns 1, 3, 4, and 5 depict the estimates from equation (1) above. In column 1, the estimate $\hat{\mu}_0 = -1.312$ (SE = 1.103) is not statistically distinguishable from zero. This finding is intriguing because $\mu_0 = 0$ would correspond to log utility in the untutored choices. However, 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 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 $\hat{\sigma}^2_{\nu}$, are in Columns 3-5. They all have 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 A, the initial (wave 1, stage 0) log error-response standard deviation is estimated to be $\hat{\tau}_0 = 1.300$ (SE = 0.455), similar in magnitude to the estimated mean log relative risk aversion. The estimate $\hat{\tau}_s = -0.0953$ (SE = 0.0151) is negative, indicating that the log error-response standard deviation decreases over the stages of reconsideration. The estimate $\hat{\tau}_w = -0.735$ (SE = 0.550) is also negative, which may suggest that the log error-response standard deviation is smaller at the beginning of wave 2 than wave 1, but the estimate is not statistically distinguishable from zero. The estimate $\hat{\tau}_{ws} = 0.0589$ (SE = 0.0208) 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 mirror our reduced-form analysis of intransitivities: declining from the beginning to the end of wave 1, then partial reset at the beginning of wave 2, and then further declines from the beginning to the end of wave 2. This is because, as noted above, intransitivities provide much of the information that identifies response-error variances.

What explains the 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 may have been paying less attention or deliberating less. Indeed, in this reduced sample, the standard errors are smaller and our results generally strengthen. 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 $p <$

[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.

0.05, the estimates are not consistent across the three frames (Web Appendix Tables E.8e-g). In addition, to make our structural estimation sample the same as that used in our descriptive analysis from Section III.A, we restrict to the subsamples involving C vs. D and E vs. F choices; despite larger standard errors, 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, Falk, Huffman, and Sunde 2018), and more extraverted participants are estimated to have greater risk aversion (the opposite direction to that found by Becker, Deckers, Dohmen, Falk, and Kosse 2012).

V. Related Literature

Classic early work on subjective expected utility invoked introspection by decision theorists and their readers about whether they would revise choices that conflicted with axioms of expected utility theory (e.g., Savage 1954; Raiffa 1961). MacCrimmon (1968) pioneered a small literature using laboratory experiments to collect evidence on whether people endorse the axioms (e.g., Moskowitz 1974; Slovic and Tversky 1974; MacCrimmon and Larsson 1979). Recently, motivated by the goal of distinguishing preferences from mistakes (which is crucial for behavioral welfare economics), a few papers have returned to the topic of whether people endorse the axioms (e.g., Gaudeul ⓡ Crosetto 2019; Breig and Feldman 2020; Nielsen and Rehbeck 2020). In this section, we review the literature and place our paper in context.

The fundamental problem facing all work that aims to test if people endorse or reject axioms (including ours) is distinguishing between well-considered rejection of an axiom and failure to understand the axiom or apply it in practice. Addressing this problem introduces a tradeoff: a heavier-handed effort to explain the axiom to participants generates a stronger demand effect. Experiments in the literature lie on a continuum of heavy-handedness but can be roughly categorized into three approaches.

The first approach is the most light-touch: participants are simply asked if they want to revise earlier choices, without any reference to axioms or inconsistencies in choice. Some such research has allowed participants to revise their responses to multiple price lists, in order to eliminate non-monotonic responses (e.g., Yu, Zhang, and Zuo forthcoming). 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). More closely related to our paper, Breig and Feldman (2020) face participants with a set of 25 risky choices two consecutive times and then offer them the opportunity to revise a subset of those choices. Breig and Feldman find that participants revise toward consistency with rational preferences (as measured by Afriat's Index) and stationarity (i.e., the same choices both times). A small related literature presents the same choice problems to respondents over several rounds, and generally finds that choices conform more to expected utility theory in later rounds (Hey 2001; van de Kuilen and Wakker 2006; van de Kuilen 2009; Nicholls, Romm, and Zimper 2015; Birnbaum and Schmidt 2015).[12] Since this lightest-touch approach makes no attempt to convey to participants that their initial choices violated normative axioms, it likely understates their well-considered endorsement of the axioms.

The second approach is the most heavy-handed, presenting normative axioms explicitly and examining whether participants revise choices to align with them. In the pioneering paper, MacCrimmon (1968) discussed 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 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 using the second approach has focused on the Independence Axiom (or Sure-Thing Principle) and well-known patterns of choices that violate it, such as the Allais (or Ellsberg) paradox. Many of these papers expressed the concern that MacCrimmon's discussion with participants may have biased participants toward endorsing the postulates. None of the experiments included such a discussion, and perhaps for this reason, the findings have been

[12] Following Agranov and Ortoleva (2017), a related line of work poses the same risky choice problem multiple times to study whether participants are intentionally randomizing their responses to the problem.

much more equivocal than MacCrimmon's. With students making real-stakes decisions over course grades 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.[13] Slovic and Tversky (1974) reported two experiments. In the first, participants made a decision in one Allais-paradox and one Ellsberg-paradox problem, then read an argument for making the opposite decision to their own in the Allais-paradox problem, and then made both decisions again. Most participants violated expected utility theory in both problems, and few changed their choices. In the second experiment, participants were presented arguments in favor of each of the two opposing choices in both problems before making their decisions and rating 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! In an extensive set of experiments, MacCrimmon and Larsson (1979) studied 20 rules that reflect either normative principles or reasons commonly given for non-normative choices (e.g., that spell out the reasoning for Allais-paradox choices). The paper reports a rich set of results in many decision problems, but the overall pattern of findings is similar to Slovic and Tversky's: experimental participants typically rated non-normative rules higher than normative rules, yet their choices often contradicted their rule rankings. In another experiment leaning against MacCrimmon's conclusion, Eli (2017) trained, incentivized, and tested experimental participants for their understanding of (i) the Independence Axiom, (ii) a decision rule based on Allais's (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.

Most recently, in an experiment involving incentivized choices over small-stakes lotteries, Nielsen and Rehbeck (2020) developed a lighter-touch version of the second approach that avoids presenting arguments: they directly elicited participants' preferences over axioms by

[13] Using a similar methodology but outside the context of expected utility theory, Loewenstein and Sicherman (1991) found that experimental participants were more likely to maximize the expected discounted value of cash flows after they were provided with arguments for and against doing so.

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 the choice made by a decision rule a participant had selected and a choice made directly by the participant, participants revised their choice (and kept the decision rule selected) 47% of the time, unselected the decision rule 13% of the time, kept both choices inconsistent 37% of the time, and changed both 3% of the time. Nielsen and Rehbeck’s evidence, like ours, suggests that participants largely find the normative axioms compelling. The Independence Axiom, however, is among the axioms that 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 decision rule selected) only 34% of the time.

Because it explicitly presents axioms to participants, the second approach is heavier-handed than other approaches in three ways. First, the logic and simplicity of an axiom can be seductive, obscuring reasons why a participant might want to violate the axiom when faced with particular choices (e.g., anticipated regret for the Independence Axiom). Second, participants (especially students) may be predisposed to think an axiom ought to be appealing merely because it is a general rule. Third, it is difficult to design placebo axioms or arguments that have demand effects equal to those of the normative axioms or arguments.

Our paper uses the third and least explored approach, which aims to be as light-touch as possible (avoiding any explicit statement of an axiom) while still making clear to participants that their original choices may violate a normative axiom: examining how experimental participants make choices when two framings of the “same” decision problem are presented together. Prior to our paper, McNeil, Pauker, and 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. Both papers found that behavior is intermediate between the two frames when both frames are presented together.[14]

While we find the third approach to be particularly appealing (and therefore pursue it in this paper), we believe that the three approaches provide complementary evidence about reconsidered choices because they engage different deliberations. For example, the second approach gets participants to think about whether they find an axiom’s logic compelling regardless of the specific choices involved. In contrast, the third approach gets participants to think about whether a difference across frames warrants a difference in choices. Moreover, while lighter-touch approaches may underestimate participants’ endorsement of normative axioms and heavier-handed approaches may overestimate it, taken together they should provide useful bounds.

Our paper differs from prior work in three main ways. First, we develop the third approach far beyond McNeil, Pauker, and Tversky (1988) and Druckman (2001), applying it systematically and iteratively over many rounds to the normative axioms underlying expected utility theory. Moreover, we have participants make decisions in each frame before facing both frames together, which may prompt more cognitive engagement and deliberation. To the extent that participants deliberate more deeply, we expect that the reconsidered choices are more informative about preferences.

Second, while prior work studied the Independence Axiom as a whole, we instead break it down into components that are easy to understand (even without explicitly stating the component axioms). This is important because, as Slovic and Tversky (1974) emphasized, it is hard to be confident that participants understand the Independence Axiom, and if they do not, the evidence on whether they endorse or reject it is difficult to interpret. Indeed, participants’ failures to understand the axiom may be responsible for the mixed and somewhat contradictory evidence reviewed above. Our results much more clearly point to participants’ endorsement of most

[14] In the context of time preferences, Frederick and Read (revise and resubmit) use 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, and O’Donoghue 2002), Frederick and Read argue that the magnitude effect violates the normative principle that discount rates should be independent of magnitude. They assess experimental participants’ discounting over receipt of $10 or $1,000 and replicate the typical finding of greater discounting for $10. They ask 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.

components of the Independence Axiom, while at the same time pinpointing the possible exception of a particular component related to anticipated regret or counterfactual reference points.

Finally, our experiment is grounded in a particular applied problem: how to measure risk aversion in retirement saving portfolio-allocation decisions for the purpose of giving advice or nudging behavior appropriately. As we discuss briefly in Section VI, we view our experiment as a proof of concept for a procedure that might, after further development and simplification, be adopted for practical use.

VI. Conclusions

In this paper, we elicited a sequence of hypothetical investment choices from experimental participants, and then we confronted participants with cases where their choices were intransitive or inconsistent with other normative axioms of expected utility theory. We asked 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 normative axioms. The remaining inconsistencies are concentrated among relatively few participants. The exception of one axiom for which the frequency of inconsistencies does not decline suggests that anticipated regret or counterfactual reference points 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 to be largely 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 normative preferences, but we have not validated that assumption. In principle, it could be tested 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 III.A, we report two relevant findings, both supportive. First, the untutored choices in wave 2 have fewer inconsistencies than the untutored choices in wave 1, suggesting that participants had learned from the wave-1 procedure. Second, comparing wave 1 to wave 2, reconsidered choices are more consistent with each other 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 no doubt 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 particular helpful for practical applications if one of the frames elicited untutored choices that well approximated the reconsidered choices (a "revelatory frame" in the terminology of Goldin 2015). Looking at Figure 6, the frame Pairwise Choices Between Compound Lotteries is the only candidate for having this property in both the C vs. D and E vs. F choices. However, given the large standard errors and incomplete convergence of choices across frames by the end of the experiment, we do not draw a confident conclusion from this observation but instead view it as suggesting a hypothesis to be tested in future work.

Finally, while we developed the reconsideration procedure in the context of retirement investment choices, a similar procedure might be useful for helping to identify normative 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. It would also be of interest to apply 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 M. Allais and O. Hagen (eds.), *Expected Utility Hypotheses and the Allais Paradox*. Dordrecht, Holland: Reidel, pp. 437–681.

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 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): 537–579.

Becker, Anke, Thomas Deckers, Thomas Dohmen, Armin Falk, and Fabian Kosse. 2012. “The Relationship Between Economic Preferences and Psychological Personality Measures.” *Annual Review of Economics* 4: 453–478.

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): 2323–2355.

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.” *The 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 1*, vol. 1, edited by B. Douglas Bernheim, Stefano Dellavigna, and David Laibson, 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): 144–152.

Breig, Zachary and Paul Feldman. 2020. "Revealing Risky Mistakes through Revisions." https://zacharybreig.com/papers/RMR.pdf. Accessed on 2020-10-01.

Cacioppo, John T., and Richard E. Petty. 1984. "The Need for Cognition: Relationship to Attitudinal Processes." *Social Perception in Clinical and Counseling Psychology* 2: 113–140.

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

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

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): 115–134.

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." Doctoral dissertation, Paris-Saclay and HEC Paris.

Ferreira, Joao V. 2018. "Would you choose it again? On confirmed choices as a proxy of welfare." https://joaovferreira.weebly.com/uploads/9/5/0/5/95056108/on_confirmed_choices.pdf. Accessed on 2020-10-01.

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. "Reflective equilibrium & the endorsement of "anomalous" preferences: The magnitude effect as a case study." Revise and resubmit at *Journal of Behavioral Decision Making*.

Gaudeul, Alexia ⓡ Paolo Crosetto. 2019. “Fast then slow: A choice process explanation for the attraction effect.” Grenoble Applied Economics Laboratory Working Paper 2019-06.

Giné, Xavier, Jessica Goldberg, Dan Silverman, and Dean Yang. 2018. “Revising Commitments: Field Evidence on the Adjustment of Prior Choices.” *Economic Journal* 128(608): 159–88.

Goldin, Jacob. 2015. “Which Way to Nudge: Uncovering Preferences in the Behavioral Age.” *Yale Law Journal* 125: 226.

Goldin, Jacob, and Daniel Reck. Forthcoming. “Revealed Preference Analysis with Framing Effects.” *Journal of Political Economy*.

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): 504–528.

Halevy, Yoram. 2007. “Ellsberg Revisited: An Experimental Study.” *Econometrica* 75(2): 503–536.

Hausman, Daniel M. 2011. *Preference, Value, Choice, and Welfare*. Cambridge: Cambridge University Press.

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): 263–291.

Karni, Edi, and David Schmeidler. 1991. Atemporal dynamic consistency and expected utility theory. *Journal of Economic Theory* 54: 401–408.

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): 1028–1038.

Kőszegi, Botond, and Matthew Rabin. 2006. “A Model of Reference-Dependent Preferences.” *Quarterly Journal of Economics* 121(4): 1133–1166.

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

Loewenstein, George, and Nachum Sicherman. 1991. "Do Workers Prefer Increasing Wage Profiles?" *Journal of Labor Economics* 9(1): 67–84.

Loomes, Graham, and Robert Sugden. 1982. "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty." *The Economic Journal* 92(368): 805–824.

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. Springer Netherlands.

McArdle, John, Willard Rodgers, and Robert Willis. 2015. "Cognition and Aging in the USA (CogUSA) 2007-2009." ICPSR36053-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2015-04-16 (accessed 2020-10-01).

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, 562–568. Cambridge: Cambridge University Press.

Morgenstern, Oskar. 1979. "Some Reflections on Utility." In *Expected Utility Hypotheses and the Allais Paradox*, 175–183. Dordrecht: Springer.

Moskowitz, Herbert. 1974. "Effects of problem representation and feedback on rational behavior in Allais and Morlat-type problems." *Decision Sciences* 5(2): 225-242.

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. Accessed on 2020-10-01.

Raiffa, Howard. 1961. "Risk, Ambiguity, and the Savage Axioms: Comment." *The Quarterly Journal of Economics* 75 (4): 690–694.

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

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

Rawls, John. 1971. *A Theory of Justice*. Cambridge, Mass: Belknap Press of Harvard University Press.

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

Shechter, Steven. 2020. "Deconstructing the Allais Paradox: The Reduction of Compound Lotteries vs. the Independence Axiom." Available at SSRN: https://ssrn.com/abstract=3521572. January 17.

Segal, Uzi. 1990. "Two-Stage Lotteries without the Reduction Axiom." *Econometrica* 58: 349–377.

Slovic, Paul, and Amos Tversky. 1974. "Who accepts Savage's axiom?" *Behavioral Science* 19(6): 368–373.

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.

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

van de Kuilen, Gijs, and Peter P. Wakker. 2006. "Learning in the Allais paradox." *Journal of Risk and Uncertainty* 33(3): 155–164.

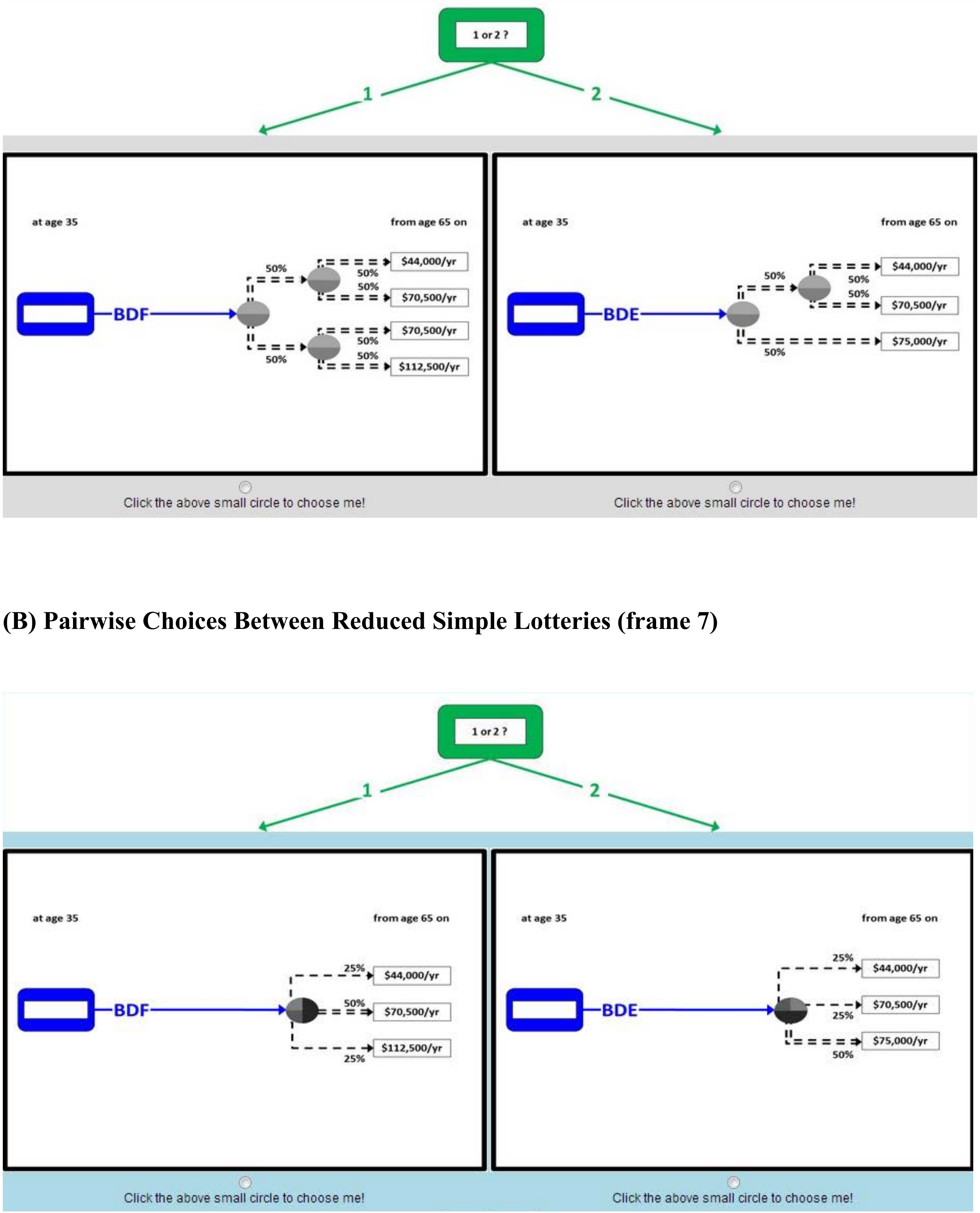
Volij, Oscar. 1994. "Dynamic consistency, consequentialism and reduction of compound lotteries." *Economic Letters* 46: 121–129.

Yu, Chi Wai, Y. Jane Zhang, and Sharon X. Zuo. Forthcoming. "Multiple switching and data quality in the multiple price list." *Review of Economics and Statistics*.

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

Zizzo, Daniel J. 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)

Figure 2: Master decision tree, which is also the frame Complete Contingent Action Plan (frame 4)

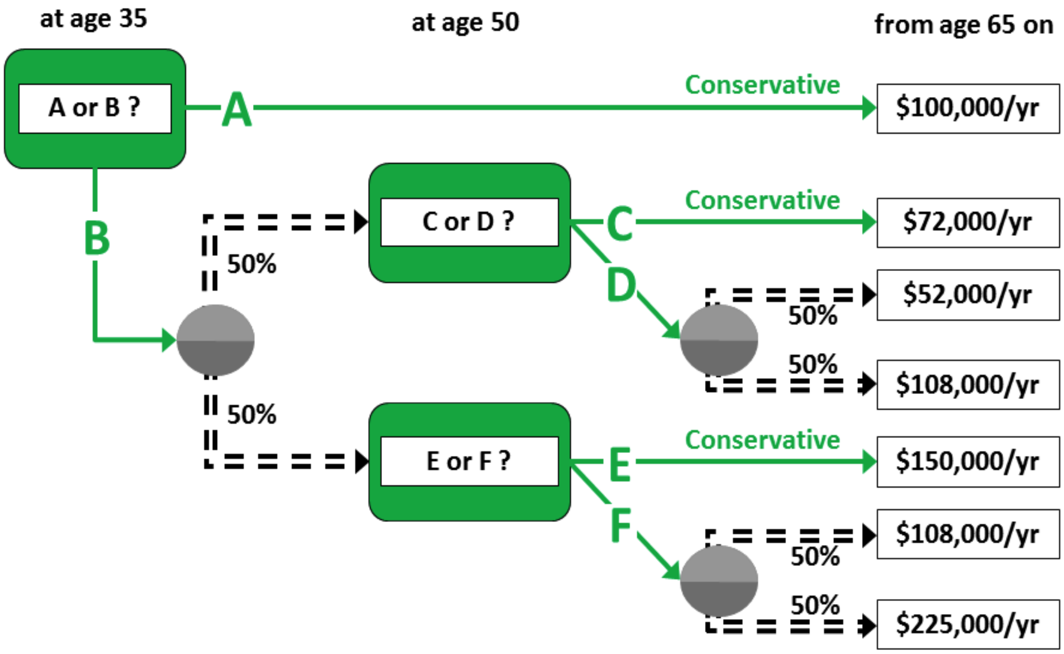
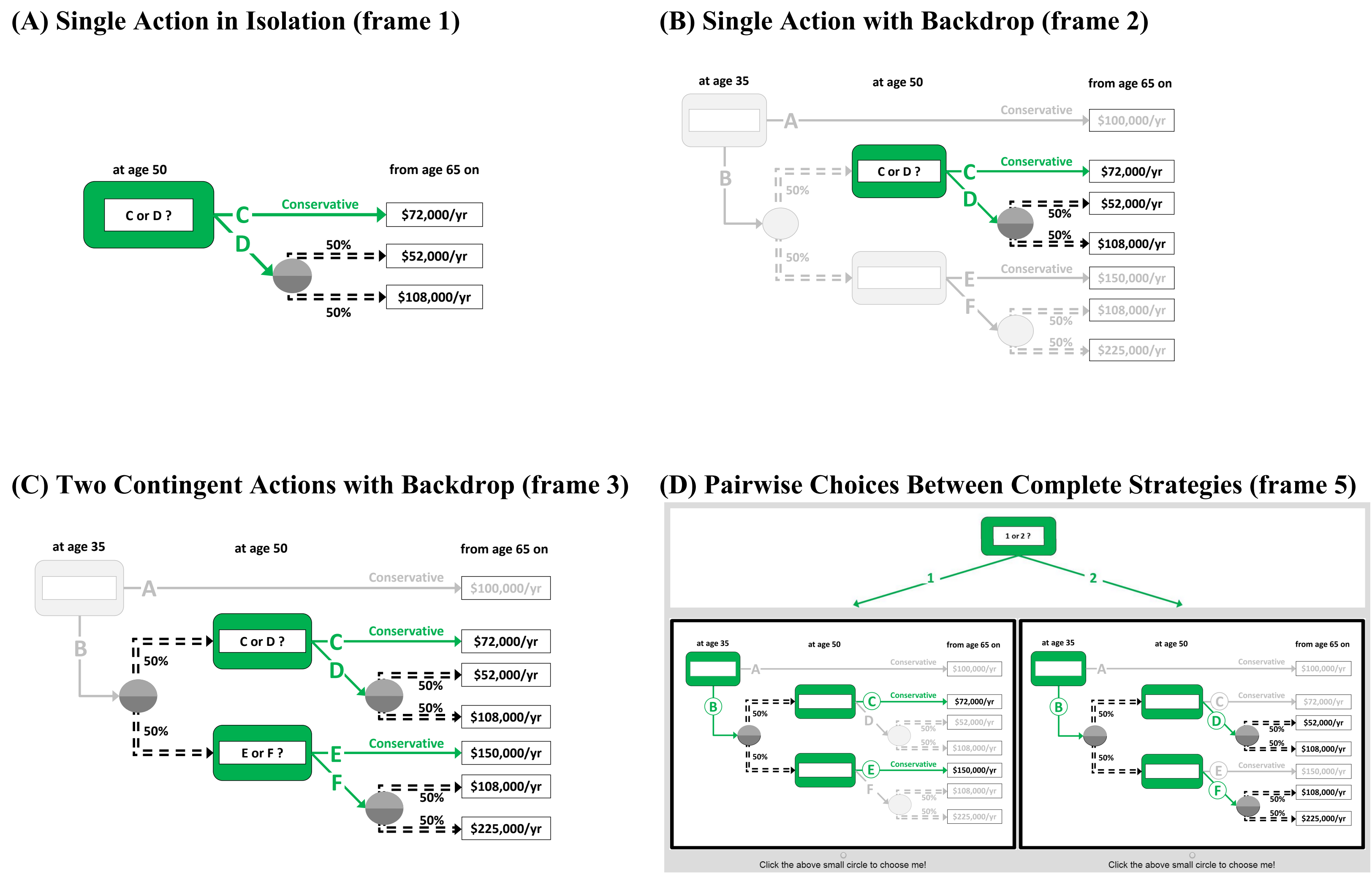


Figure 3: Examples of the remaining frames



(A) Single Action in Isolation (frame 1)

(B) Single Action with Backdrop (frame 2)

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

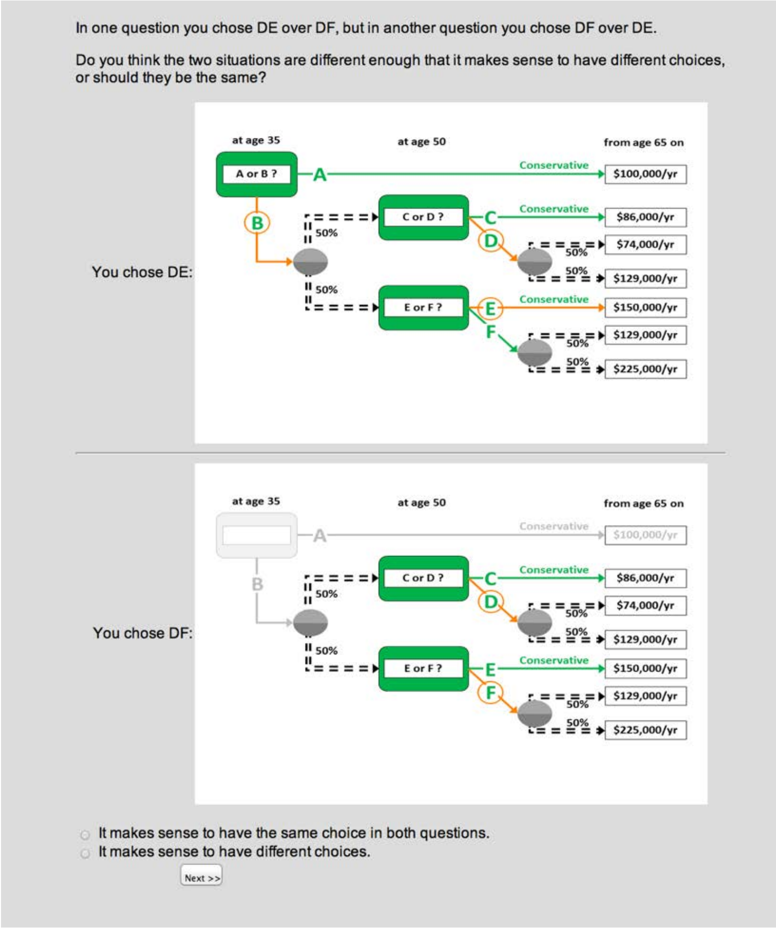
(D) Pairwise Choices Between Complete Strategies (frame 5)

○ Click the above small circle to choose me!

○ Click the above small circle to choose me!

Figure 4: Screenshots of inconsistency and intransitivity reconsiderations

(A) Inconsistency reconsideration



In one question you chose DE over DF, but in another question you chose DF over DE.

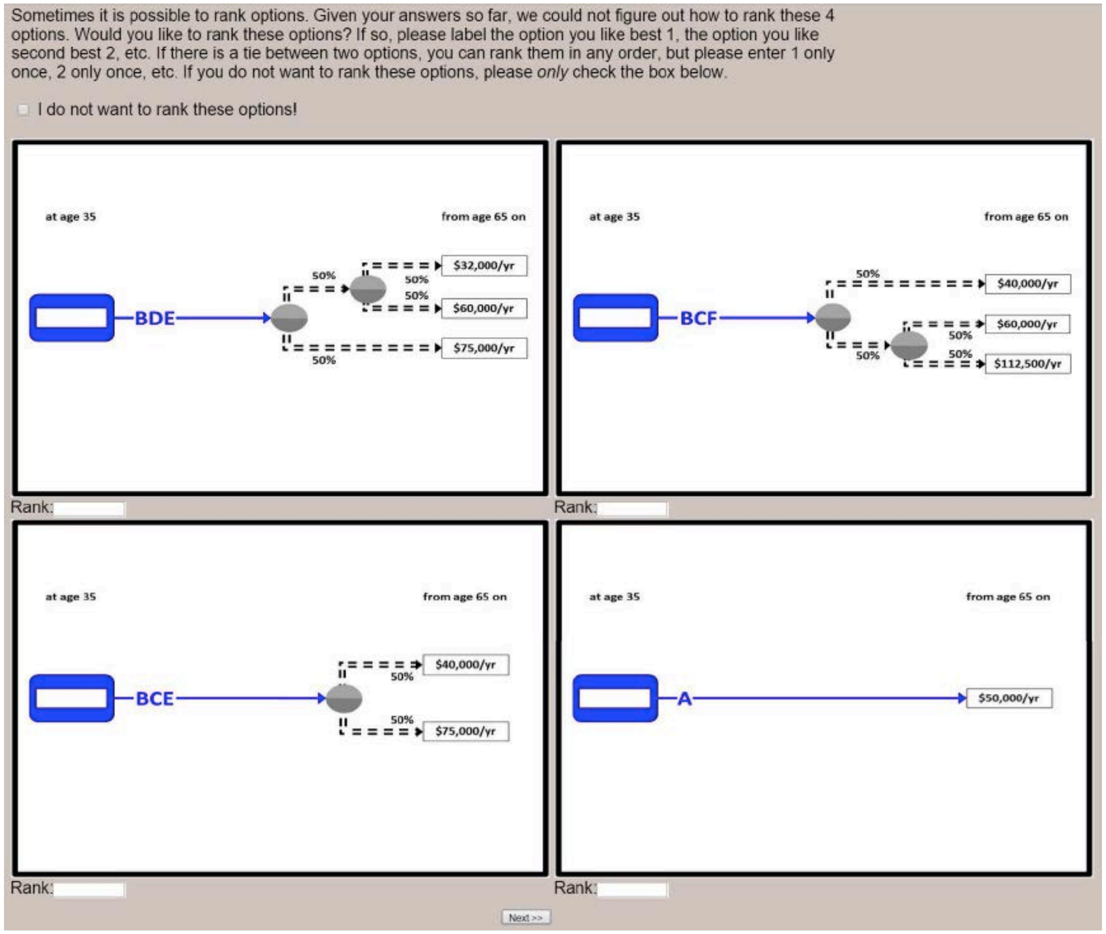
Do you think the two situations are different enough that it makes sense to have different choices, or should they be the same?

You chose DE:

You chose DF:

○ It makes sense to have the same choice in both questions.
○ It makes sense to have different choices.

(B) Intransitivity reconsideration



Sometimes it is possible to rank options. Given your answers so far, we could not figure out how to rank these 4 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.

□ I do not want to rank these options!

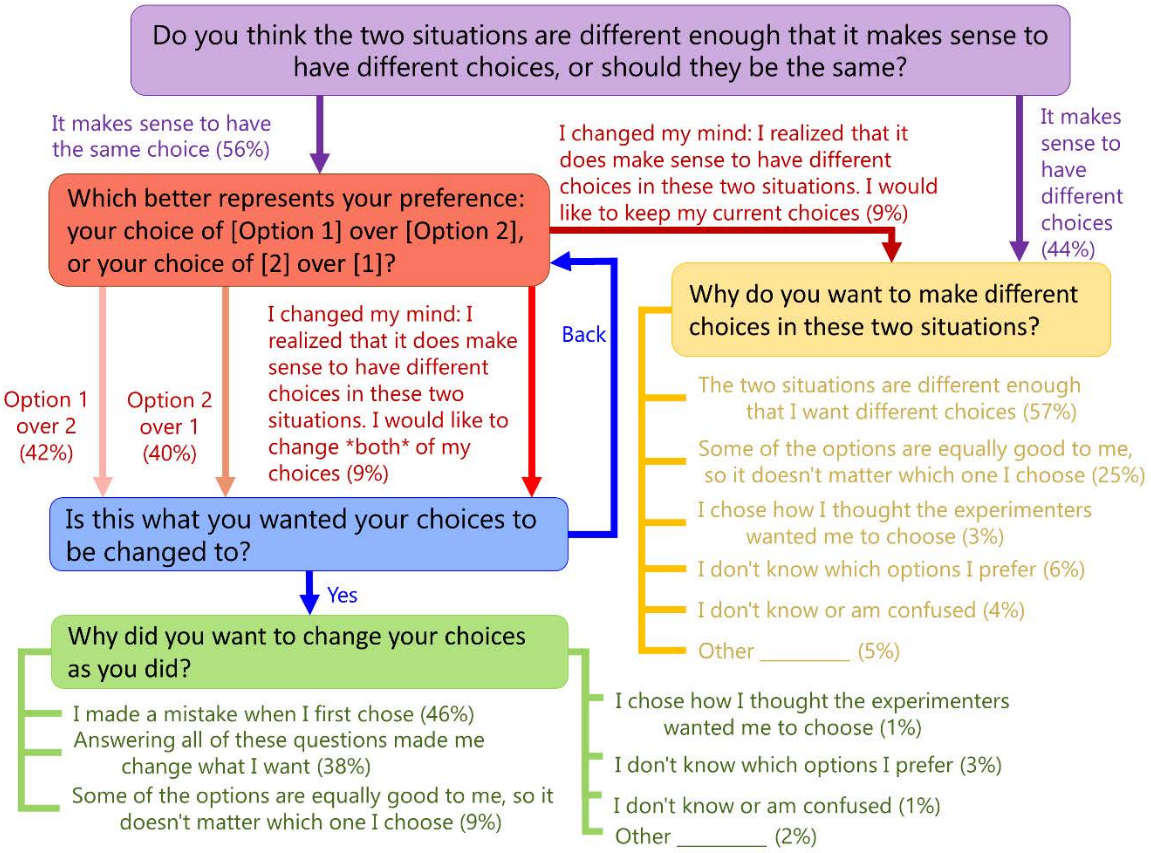
Rank:

Rank:

Rank:

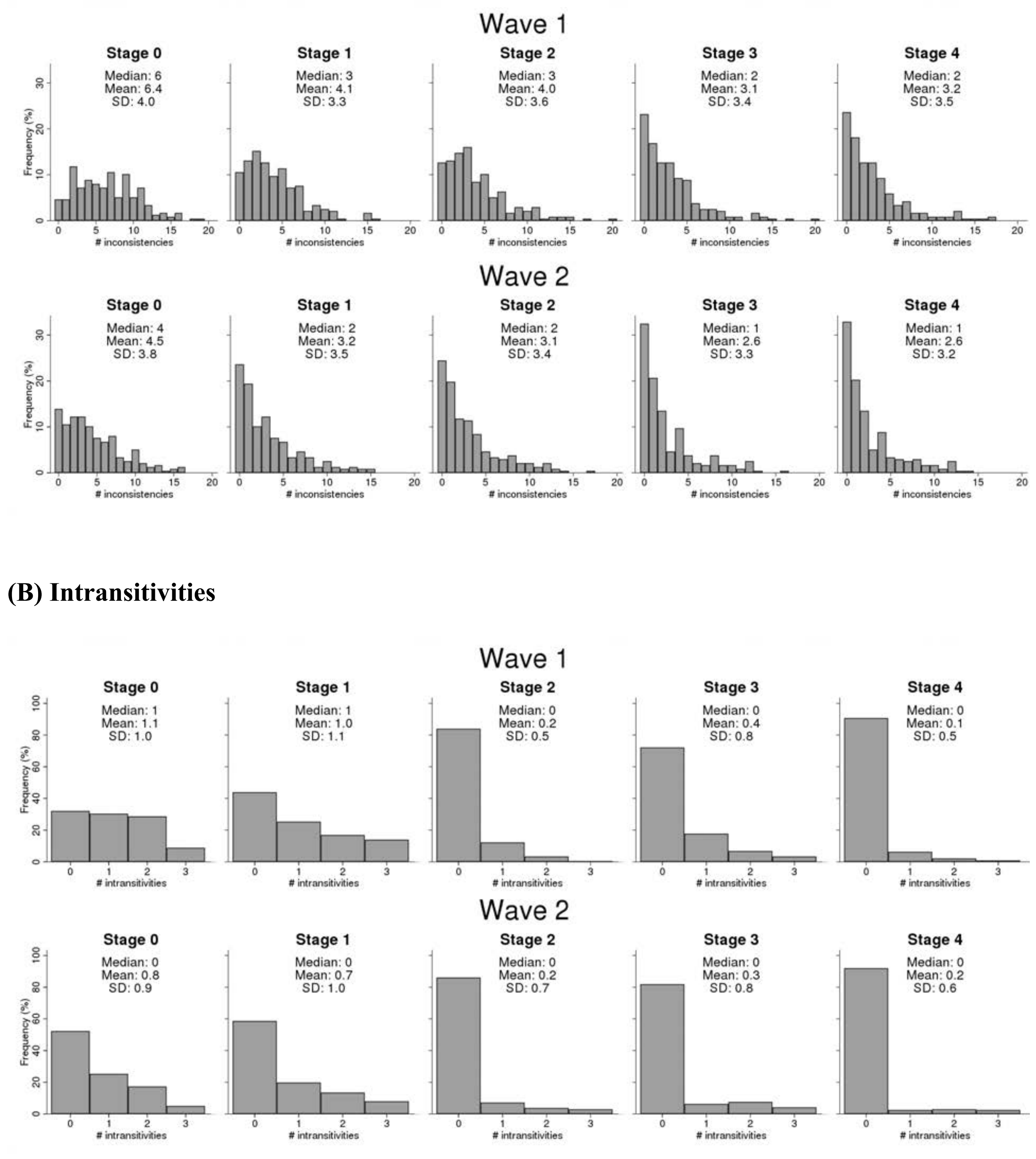
Rank:

Figure 5: Flow chart for the inconsistency reconsideration procedure


Do you think the two situations are different enough that it makes sense to have different choices, or should they be the same?
It makes sense to have the same choice (56%)
It makes sense to have different choices (44%)
Which better represents your preference: your choice of [Option 1] over [Option 2], or your choice of [2] over [1]?
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 (9%)
Option 1 over 2 (42%)
Option 2 over 1 (40%)
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 (9%)
Back
Is this what you wanted your choices to be changed to?
Yes
Why did you want to change your choices as you did?
I made a mistake when I first chose (46%)
Answering all of these questions made me change what I want (38%)
Some of the options are equally good to me, so it doesn't matter which one I choose (9%)
I chose how I thought the experimenters wanted me to choose (1%)
I don't know which options I prefer (3%)
I don't know or am confused (1%)
Other ________ (2%)
Why do you want to make different choices in these two situations?
The two situations are different enough that I want different choices (57%)
Some of the options are equally good to me, so it doesn't matter which one I choose (25%)
I chose how I thought the experimenters wanted me to choose (3%)
I don't know which options I prefer (6%)
I don't know or am confused (4%)
Other ________ (5%)

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 6: Histograms of number of inconsistencies and intransitivities

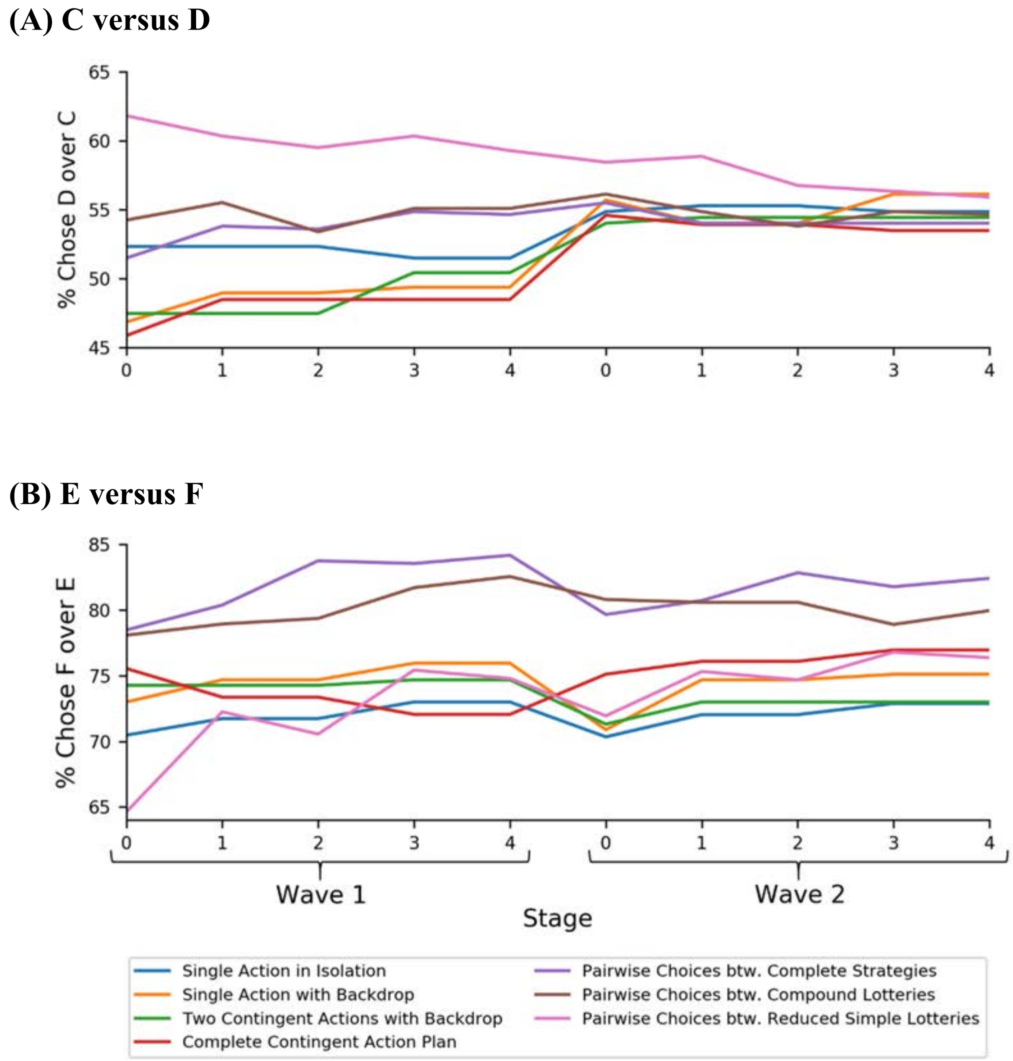


(A) Inconsistencies

(B) Intransitivities

Note: Wave 1+2 sample.

Figure 7: Percentage of participants who make the risky choice in simple lotteries



(A) C versus D

(B) E versus F

Note: Wave 1+2 sample. 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", "I should be able to rank the options, but it's extremely hard", 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: Average consistency rates by frame across stages and waves

	(1)	(2)		(3)	(4)	(5)	
Frame	**# Pot. Incons.**	**Wave 1 vs. 2 Consistency Rate**		**Random Choice Consistency Rate**	***p*-value: Stage-0 Consistency Rate = Stage-4 Consistency Rate**	***p*-value: Stage-0 Consistency Rate = Random Consistency 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. Consistency rate = (total number of consistencies)/(number of potential inconsistencies). *P*-values are from two-sided tests for differences in proportions. Percentages are rounded to the nearest 0.1%.

Table 5: Percentage of participants whose reconsidered choices can be rationalized by a simple heuristic

Heuristic (choose for all 68 questions)	Frequency
Maximize expected value	6.4%
Minimize expected value	0.0%
Choose top option	1.8%
Choose bottom option	4.6%
None of the above	93.6%
Heuristic (choose for at least 60 of 68 questions)	**Frequency**
Maximize expected value	19.3%
Minimize expected value	0.0%
Choose top option	4.6%
Choose bottom option	14.7%
None of the above	80.7%

Note: Wave 1+2 sample, restricted to the 109 participants randomized to the rightside-up orientation (i.e., with option A at the top and E at the bottom, as in the master decision tree) in one wave and the upside-down orientation (i.e., with E at the top and A at the bottom) in another. The table shows the frequencies among reconsidered (stage-4) choices, averaged across waves 1 and 2. In the rightside-up orientation, the top option always minimizes expected value (because it is the safe option), and the bottom choice always maximizes expected value (because it is the risky option); vice-versa in the upside-down orientation. Therefore, in each panel, the sum of "Maximize expected value" and "Minimize expected value" equals the sum of "Choose top option" and choose "Choose bottom option." Percentages are rounded to the nearest 0.1%.

Table 6: Direction of revising inconsistencies

Axiom	**Choose**	**Choose**	***p*-value**	**No**	**Swap**	**#Obs**
	Frame *j*	**Frame *j+1***	***j* = *j*+1**	**Update**		
Choice in frame *j* was riskier						
Irrelevance of Background Counterfactuals	**33.3%**	20.5%	0.1236	43.6%	2.6%	78
Simple Actions = State-Contingent Actions	**25.6**	6.1	0.0013	64.6	3.7	82
Irrelevance of Counterfactual Choices	25.8	**26.8**	0.8895	41.2	6.2	97
Fusion + Shift from Nodewise to Pairwise	**23.2**	17.6	0.2951	56.3	2.8	142
Complete Strategies = Implied Lotteries	**26.0**	21.5	0.1219	48.5	4.0	550
Reduction of Compound Lotteries	**33.2**	16.0	<0.0001	45.2	5.6	832
Overall	**29.4**	18.1	<0.0001	47.7	4.7	1781
Choice in frame *j+1* was riskier						
Irrelevance of Background Counterfactuals	14.4	**32.5**	0.0154	50.6	2.4	83
Simple Actions = State-Contingent Actions	12.3	**16.0**	0. 5349	66.7	4.9	81
Irrelevance of Counterfactual Choices	**32.7**	13.9	0.0050	48.5	5.0	101
Fusion + Shift from Nodewise to Pairwise	**18.5**	17.5	0.8192	57.8	6.2	211
Complete Strategies = Implied Lotteries	17.3	**31.0**	<0.0001	46.6	5.2	562
Reduction of Compound Lotteries	20.2	**25.4**	0.0522	49.0	5.5	635
Overall	19.1	**25.5**	0.0001	50.2	5.3	1673
Choices in frames *j* and *j+1* are not risk-ranked						
Irrelevance of Background Counterfactuals	-	**-**	-	-	-	0
Simple Actions = State-Contingent Actions	-	**-**	-	-	-	0
Irrelevance of Counterfactual Choices	14.3	**28.6**	0.2554	46.4	10.7	28
Fusion + Shift from Nodewise to Pairwise	**23.8**	11.9	0.2003	57.1	7.1	42
Complete Strategies = Implied Lotteries	20.0	**23.8**	0.4382	52.4	3.8	185
Reduction of Compound Lotteries	**28.0**	18.3	0.0243	48.4	5.3	246
Overall	**24.0**	20.4	0.2274	50.5	5.2	501

Note: Wave 1+2 sample. *P*-values are from two-sided tests for differences in proportions. To facilitate reading the table, we have bolded whichever frame-*j* or frame-(*j*+1) number is larger in each row. Percentages are rounded to the nearest 0.1% and therefore row percentages may not add up to 100%.

Table 7: Average inconsistency rates by axiom

Axiom	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	#Obs
	Inconsistency Rate Wave 1		**Inconsistency Rate Wave 2**		***p*-value (1)-(2)**	***p*-value (1)-(3)**	***p*-value (3)-(4)**	***p*-value (1)-(4)**	
	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 tests for differences in proportions.

Table 8: Results from structural estimation

A. Pairwise Choices Between Complete Strategies

	(1)	(2)	(3)	(4)	(5)
	ln(CRRA)	ln(SD(error))	σ_{ν}^2	$\sigma_{\eta_1}^2$	$\sigma_{\eta_2}^2$
wave2	0.803	-0.735			
	(1.192)	(0.550)			
stage	-0.0117	-0.0953			
	(0.0363)	(0.0151)			
wave2*stage	0.0172	0.0589			
	(0.0414)	(0.0208)			
constant	-1.312	1.300	4.988	7.405	1.249
	(1.103)	(0.455)	(2.937)	(9.220)	(3.368)

B. Pairwise Choices Between Compound Lotteries

	(1)	(2)	(3)	(4)	(5)
	ln(CRRA)	ln(SD(error))	σ_{ν}^2	$\sigma_{\eta_1}^2$	$\sigma_{\eta_2}^2$
wave2	0.216	-0.477			
	(0.871)	(0.459)			
stage	0.0303	-0.102			
	(0.0308)	(0.0148)			
wave2*stage	-0.00574	0.062			
	(0.0381)	(0.0208)			
constant	-0.919	1.101	4.599	2.814	2.460
	(0.681)	(0.333)	(2.251)	(3.727)	(3.478)

C. Pairwise Choices Between Reduced Simple Lotteries

	(1)	(2)	(3)	(4)	(5)
	ln(CRRA)	ln(SD(error))	σ_{ν}^2	$\sigma_{\eta_1}^2$	$\sigma_{\eta_2}^2$
wave2	0.0489	-0.390			
	(0.707)	(0.431)			
stage	-0.0205	-0.111			
	(0.0299)	(0.0151)			
wave2*stage	0.00834	0.054			
	(0.0376)	(0.0205)			
constant	-0.575	1.144	3.886	1.930	2.933
	(0.515)	(0.300)	(1.816)	(2.615)	(3.344)

Note: Wave 1+2 sample. #Obs is 21330 choices. Standard errors in parentheses. In panels A, B, and C, respectively, the difference between $\sigma_{\eta_1}^2$ (columns 4) and $\sigma_{\eta_2}^2$ (columns 5) is -6.155 (SE = 12.053), -0.354 (SE = 6.738), and 1.003 (SE = 5.545).