American Economic Review 2022, 112(11): 3584–3626
https://doi.org/10.1257/aer.20210290

Evaluating Deliberative Competence: A Simple Method with an Application to Financial Choice†

By Sandro Ambuehl, B. Douglas Bernheim, and Annamaria Lusardi*

We examine methods for evaluating interventions designed to improve decision-making quality when people misunderstand the consequences of their choices. In an experiment involving financial education, conventional outcome metrics (financial literacy and directional behavioral responses) imply that two interventions are equally beneficial even though only one reduces the average severity of errors. We trace these failures to violations of the assumptions embedded in the conventional metrics. We propose a simple, intuitive, and broadly applicable outcome metric that properly differentiates between the interventions, and is robustly interpretable as a measure of welfare loss from misunderstanding consequences even when additional biases distort choices. (*JEL* D83, D91, G51, G53)

Mounting evidence documents low-quality decision-making in a wide range of policy-relevant contexts, such as household finance (Beshears et al. 2018), health insurance (Loewenstein et al. 2013), and the consumption of durable goods (Grubb and Osborne 2015; Allcott and Taubinsky 2015). An important class of corrective policies seeks to address the resulting inefficiencies by influencing consumers' decisions without changing either their choice sets or the factual information to which they have access (*opportunity-neutral* interventions). Examples include (i) programs that seek, through education and training, to equip decision-makers with the knowledge and skills necessary to process and interpret naturally occurring information; (ii) efforts to "nudge" people toward better decisions through salient or graphic labels and other measures that attempt to change the "framing" of decision tasks (also known as *choice architecture*; Thaler and Sunstein 2009); and (iii) measures that influence decision-making by altering the nature or frequency of personal interactions with professional advisers, family members, or peers. Evaluating the welfare effects of such policies presents challenges. For example, if we proceed

*Ambuehl: University of Zurich, Department of Economics (email: sandro.ambuehl@econ.uzh.ch); Bernheim: Department of Economics (email: bernheim@stanford.edu); Lusardi: George Washington University School of Business (email: alusardi@email.gwu.edu). Stefano DellaVigna was the coeditor for this article. This paper incorporates some material contained in previous working papers entitled "Financial Education, Financial Competence, and Consumer Welfare" and "A Method for Evaluating the Quality of Financial Decision Making, with an Application to Financial Education." We thank Charles Sprenger, Steven Sheffrin, Glen Weyl, our referees, and participants at various conferences and seminars for helpful comments and suggestions. Fulya Ersoy and David Zuckerman provided excellent research assistance. We appreciate funding from the TIAA Institute, the Alfred P. Sloan Foundation, and the Department of Economics at Stanford University. Experiment A was approved in Stanford IRB protocol 29615, experiment B in University of Toronto REB protocol 34511.

†Go to https://doi.org/10.1257/aer.20210290 to visit the article page for additional materials and author disclosure statements.

from the premise that financial decisions are often misguided, then we cannot evaluate financial well-being by deferring indiscriminately to choices, as standard welfare analysis prescribes. Instead, we require measures of decision quality. Moreover, suitable methods must account for the various ways in which different biases (and the policies that seek to address them) counteract or reinforce each other. The purpose of this paper is to devise a simple method of evaluation that overcomes these challenges.

While the approach we develop targets a variety of potential applications, for the sake of concreteness we focus on the domain of personal financial decisions. Three features of this setting make it ideal for our purposes. First, because financial illiteracy is pervasive (see, e.g., Hastings, Madrian, and Skimmyhorn 2013; Lusardi and Mitchell 2014), poor decision-making is likely widespread. Second, it is easy to imagine opportunity-neutral interventions that could potentially improve the quality of financial decision-making. Third, financial education is already the subject of a vast academic literature (see Hastings, Madrian, and Skimmyhorn 2013; Lusardi and Mitchell 2014; Fernandes, Lynch, and Netemeyer 2014; Beshears et al. 2018; Kaiser et al. 2022 for reviews), and has become a ubiquitous public policy tool.[1]

Two methods for evaluating financial education interventions dominate the literature (see the reviews cited above). The first investigates whether an intervention changes behavior in a direction that counteracts a suspected bias. For example, on the assumption that people save too little, some studies ask whether financial education leads them to save more (e.g., Bernheim, Garrett, and Maki 2001; Bernheim and Garrett 2003; Cole and Shastry 2010; Bruhn, Ibarra, and McKenzie 2014; Bruhn et al. 2016).[2] A second approach uses exam-style questions to assess comprehension of financial principles (e.g., Mandell 2009; Mandell and Klein 2009; Walstad, Rebeck, and MacDonald 2010; Carpena et al. 2011; Collins 2013; Heinberg et al. 2014; Bruhn, Ibarra, and McKenzie 2014; Bruhn et al. 2016; Lusardi et al. 2017). Both methods are easy to implement, but neither measures the quality of decision-making directly. Still, if education not only improves financial literacy but also shifts behavior in a direction that counters a presumed bias (as the recent meta-analysis of Kaiser et al. 2022 suggests), there is a tendency to infer that it achieves the right changes for the right reasons and consequently makes consumers better off.

The first part of this paper (Section I) explores these evaluative metrics in the context of a laboratory experiment, and then develops our alternative. Experimental methods are widely used in research on the quality of decision-making both in the lab and in the field because they provide the ability to control the environment tightly and to collect the type of detailed measurements required for rigorous inference about welfare (see, e.g., Bernheim and Taubinsky 2018; Beshears et al. 2018 for

[1] Most member countries of the Organisation for Co-operation and Economic Development (OECD), as well as India and China, have implemented policies to enhance financial education. Altogether, these initiatives seek to reach more than five billion people in 60 countries (OECD 2015). The European Commission recently identified financial literacy as a key objective in its *capital markets union* 2020 *action plan* (European Commission 2020) and, in cooperation with the OECD, has launched the development of a financial competence framework for its 27 member states (OECD 2021). Additionally, in 2013, private nonprofit organizations spent nearly half a billion dollars on the provision of financial education in the United States alone (CFPB 2013).

[2] Other directional biases include the tendency to choose low-deductible health insurance plans (Sydnor 2010; Loewenstein et al. 2013), naïve diversification (Benartzi and Thaler 2001), and investors' tendency to exit the market in downturns (Bennyhoff and Kinniry 2011).

reviews).[3] We focus on a fundamental concept in personal finance: compound interest. Some of our subjects receive financial education pertaining to compound interest, while others do not. We use two educational interventions that closely follow a popular text on investing and differ from each other only in that one provides practice with individualized feedback, while the other does not. We elicit subjects' reservation prices for interest-bearing investments (e.g., "We will invest \$*a* in an account with *X%* interest per day. Interest is compounded daily. We will pay you the proceeds in *t* days."), and for each corresponding future payment stated transparently (e.g., "We will pay you \$*r* in *t* days."). We label the former *complex framing* and the latter *simple framing*. Given the well-documented tendency for people to underestimate exponential growth (*exponential growth bias*; Wagenaar and Sagaria 1975; Eisenstein and Hoch 2007; Stango and Zinman 2009; Almenberg and Gerdes 2012; Levy and Tasoff 2016), there is a strong presumption that effective education would increase the demand for interest-bearing assets. Subjects also take an incentivized test on compound interest that requires them to apply, rather than merely to recite, the principles they learned.

According to both conventional metrics, the two interventions are effective: they improve performance on the exam-style questions, and they increase reservation prices for interest-bearing investments. Moreover, in each case, their effects are nearly identical. Such analyses appear to imply that the basic intervention significantly improves the quality of decision-making, and that the addition of practice and feedback neither amplifies nor diminishes these benefits.

Next, we demonstrate that the inferences one would draw from the conventional metrics are, in fact, misleading. Because subjects performed equivalent valuation tasks with both complex and simple framing, we can assess the magnitudes of the biases resulting from misapprehension of compound interest. A subject who understands interest compounding and who applies that knowledge when making decisions will exhibit the same valuation for an interest-bearing investment as for the future payment it yields. For subjects with poor comprehension, valuations will differ across the frames. Using these data, we show that the two interventions have profoundly different effects. For the ideal opportunity-neutral intervention, the treatment effect would be positively correlated with the initial bias: it would be larger for those who undervalue interest-bearing assets more severely. And yet, without practice and feedback, the intervention increases subjective valuations of interest-bearing assets across the board, even though subjects' initial biases are highly heterogeneous. It therefore leads some subjects to overcorrect for their initial biases, and it yields greater bias among those whose initial estimates of exponential growth are roughly

[3] A reasonable question is whether, with larger real-life stakes, people might deliberate more carefully and at greater length, and might be more likely to employ analytic tools or seek advice. In practice, most people make financial decisions without qualified assistance (Bernheim 1998; Benartzi and Thaler 1999; Lusardi and Mitchell 2011), spend less than an hour making retirement saving decisions (Benartzi and Thaler 1999), fail to develop explicit financial or retirement plans (Lusardi and Mitchell 2011), rely on simple rules of thumb (Bernheim 1994), and respond to nonsubstantive nudges (Madrian and Shea 2001; Bernheim, Popov, and Fradkin 2015). Notably, the fraction of subjects reporting that they did not seek advice in our experiment (three quarters) matches experience in the field (Lusardi and Mitchell 2011). See also Enke et al. (forthcoming) for general evidence that large stakes do not remove cognitive biases.

correct or excessive.[4] In contrast, with practice and feedback, initial bias and treatment effects are indeed positively correlated: the intervention reduces the degree to which subjects undervalue complexly framed opportunities without increasing the prevalence or degree of overvaluation. The benefits of education are therefore unambiguous with practice and feedback, but not without it. Figure 3 of Section ID encapsulates these findings.

Why do the conventional metrics lead us astray? When an improvement in knowledge accompanies a directionally appropriate shift in behavior, we tend to assume that the relationship is causal—i.e., that behavior changes *because* knowledge improves. Further analysis involving two additional treatments falsifies that assumption. For one treatment, we remove most of the substantive material from the intervention and maintain only its rhetorical elements (e.g., "Albert Einstein is said to have called compound interest the most powerful force in the universe"). The resulting intervention largely reproduces the behavioral effect of the full intervention, but has almost no effect on measured comprehension. For the second treatment, we retain the substantive material and remove the rhetoric. That treatment roughly reproduces the effect of the original intervention on measured comprehension, but largely fails to alter behavior.[5] Thus, without practice and feedback, the effects of the educational intervention on test performance arise mainly from the substantive elements of instruction, but its behavioral effects arise mainly from the rhetorical elements. These findings explain why the intervention can enhance test performance and shift behavior in the "desired" direction without improving the quality of decision-making overall.

Our critique of the conventional evaluative metrics points to a simple alternative: compute the average distance between simply and complexly framed valuations in equivalent tasks (our proposed measure of *deliberative competence*). This metric is the absolute value (or square) of the *price-metric bias* studied in behavioral public economics (see Bernheim and Taubinsky 2018). In the absence of biases other than consumers' misunderstanding of the consequences of their choices, the price-metric bias is interpretable as the ideal price correction (tax or subsidy). As we explain, it has another interpretation that renders it useful for welfare analyses of opportunity-neutral policies: intuitively, the absolute size of the bias gauges the dollar denominated harm the consumer might suffer due to errors resulting from misunderstanding her options. Additional analysis of the data from our experiment reveals that our measure of deliberative competence captures the essential patterns that conventional evaluative metrics miss: it implies that the intervention is effective if it includes practice and feedback, but is not effective if these elements are absent.

[4] Goda et al. (2019) and Levy and Tasoff (2017) also document considerable heterogeneity with respect to the perceived benefits of compounding. Relatedly, Harrison, Morsink, and Schneider (2020) finds that increased take-up of seemingly beneficial products (index insurance) can lead to welfare losses due to subject heterogeneity. See also Oster (2020), who shows that health recommendations have the greatest impact on those who are already attuned to health issues, which raises the analogous possibility that treatment effects are negatively correlated with biases in that context.

[5] Related dissociations have been observed in other contexts. Enke and Zimmermann (2019) shows that many people tend to neglect correlations in decision-making, despite knowing how to account for them when prompted explicitly. Taubinsky and Rees-Jones (2018) find that many consumers underreact to excise taxes, even though they can properly compute tax-inclusive prices when prompted to do so.

The second part of this paper (Section II) supplements our intuitive welfare arguments with formal analysis. Our first objective is to establish that deliberative competence, as we define it, is interpretable as a measure of the welfare loss resulting from poor comprehension of the consequences that follow from choice options (rather than merely as a measure of bias). Our second objective is to study the robustness of this measure with respect to the possibility that other biases may counteract or reinforce the effects of poor comprehension.

We begin by providing theoretical foundations for interpreting deliberative competence as a welfare measure under the restrictive assumption that the consumer's decision-making apparatus involves no defects other than the misunderstandings that the pertinent policy interventions target. Loosely, this assumption allows us to treat choices in the simple frame as revealing "true" preferences. Methodologically, our analysis of this special case parallels other recent work in behavioral public economics, in that we evaluate naturalistic choices based on the decisions consumers would make if they properly understood the relationships between options and consequences.[6] In the case of our experiment, the complexly framed tasks are naturalistic, and the simply framed tasks avoid misunderstandings associated with compound interest.

An obvious limitation of the analysis summarized in the last paragraph is that the consumer's decision-making apparatus may well involve additional defects. If other biases contaminate simply framed choices, those choices may not provide valid normative benchmarks. Such possibilities raise *second-best* issues in the sense of Lipsey and Lancaster (1956). Concretely, in our experiment, considerations such as present bias and excessive suspicion concerning the likelihood the experimenter will renege on promised future payments may suppress simply framed valuations. Arguably, a policy that overcorrects for exponential growth bias is beneficial because it counteracts these additional biases.[7] We reject this argument on the grounds that it implicitly addresses the wrong question. Specifically, it introduces additional biases without considering the interventions policymakers might use to address them (e.g., for present bias, commitment options and corrective taxation).

Ideally, one would account for second-best concerns by analyzing all biases and all associated policy options comprehensively. Unfortunately, we lack the requisite comprehensive understanding of human behavior to proceed in this manner. Practical considerations necessitate a compartmentalized approach, which is precisely what one finds in the existing literature on behavioral public economics.

The literature does not, however, offer a fully satisfactory theoretical framework for conducting compartmentalized welfare analysis. The typical paper focuses on one or two biases at a time and assumes (usually implicitly) that the process generating the observed data is otherwise free of defects—in effect, that any other biases are either orthogonal or have been resolved. In an analogous framework involving

[6] For foundations, see Bernheim and Rangel (2009); Bernheim (2009); Bernheim (2016); Bernheim and Taubinsky (2018); Bernheim (2021). Bernheim and Taubinsky (2018) also survey applications.

[7] Similar issues arise elsewhere in behavioral public economics. For example, as noted by Blumenthal and Volpp (2010), accurate information about the caloric content of food can exacerbate excessive consumption if people tend to overestimate calories. Likewise, according to Downs, Loewenstein, and Wisdom (2009, p. 159), "dietary information is likely to improve self-protective behavior only if existing biases encourage unhealthy eating, but the reverse is equally likely. When it comes to smoking, for example, there is evidence that smokers tend to overestimate the health risks, in which case providing risk information could undermine their motivation to quit."

multiple distortions, Lipsey and Lancaster (1956) demonstrated that this approach to welfare analysis is conceptually fraught. Pasting together a collection of such solutions does not generally yield an overall optimum. We fill this gap in the literature by proposing the notion of *idealized welfare analysis*, which acknowledges that other biases infect the available data, but assumes that remedies for them will be forthcoming. In contrast to the usual compartmentalized approach, this perspective allows the analyst to arrive at an overall solution to multiple interacting problems by solving them one at a time.

We then demonstrate that deliberative competence is interpretable as a valid measure of the idealized welfare loss despite the presence of biases outside the scope of the analysis, even when relatively little is known about their nature and magnitudes. This result requires a restrictive separability assumption, but it encompasses many of the types of biases that might infect a study such as ours. We also address other concerns that could in principle limit the applicability of our approach, showing for example that a simple adjustment removes a class of potential confounds.

This paper contributes to several literatures. It adds to the growing literatures in behavioral public economics and behavioral welfare economics (cited above) by showing that, under specified conditions, simple transformations of the price-metric bias are robustly interpretable as measures of the welfare loss from poor decision-making, suitable for evaluating opportunity-neutral interventions. With respect to the literature on financial education (also cited above), it documents important limitations of conventional evaluative metrics and offers an easily implemented alternative. It also contributes to a smaller methodological literature on measures of decision-making quality. Alternative measures include the frequencies of (i) either dominated or dominant choices (Ernst, Farris, and King 2004; Calvet, Campbell, and Sodini 2007, 2009; Agarwal et al. 2009; Baltussen and Post 2011; Choi, Laibson, and Madrian 2011; Aufenanger, Richter, and Wrede 2016; Bhargava, Loewenstein, and Sydnor 2017), or (ii) violations of the weak or the generalized axiom of revealed preference (Afriat 1972; Echenique, Lee, and Shum 2011; Choi et al. 2014). Those methods are not easily adapted to the types of decisions that provide the focus for applications such as ours. Consistency-based methods are unlikely to detect problems arising from a consistent misunderstanding (e.g., of compound interest), and dominance-based methods are inapplicable when the best choice depends on preferences.

I. Experimental Evaluations of Financial Education Interventions

For our application, we investigate the efficacy of two educational interventions aimed at improving intertemporal decision-making by enhancing comprehension of compound interest. Interest compounding is one of the most fundamental concepts in finance, and a core topic in the vast majority of courses and books on the principles of financial decision-making. We begin by describing the experiment (Section IA) and its implementation (Section IB). After evaluating the educational interventions based on conventional metrics (Section IC), we demonstrate that their implications are untenable and their underlying assumptions invalid (Section ID). We then propose, intuitively justify, and implement a simple alternative outcome metric and show that it yields sensible conclusions. Section II provides formal theoretical foundations for this alternative approach.

A. *Design*

Our investigation involves two experiments. Both consist of three stages. First, subjects participate in a randomly assigned financial education intervention. Second, subjects complete paired valuation tasks. Third, subjects answer a battery of exam-style questions on compound interest. Table 1 presents a schematic overview. Additional detail concerning each stage follows.

Stage 1. Education Intervention: Using material from a popular investment guide, *The Elements of Investing: Easy Lessons for Every Investor* by Malkiel and Ellis (2013), we produced an instructional video on compound interest that covers the topic through a narrated slide presentation.[8] The narration is verbatim from the text (with a few minor adjustments), while the slides summarize key points. Stylistically, the video resembles those offered through the popular educational internet platform, Khan Academy.[9]

For experiment A, the video constitutes the entire treatment. For experiment B, we split the video into three parts, each of which is followed by practice problems with automated, individualized feedback. If a subject provides an answer that is mistaken but consistent with the calculation of simple interest, the intervention mentions the likely source of the error, explains why the answer is wrong, and suggests how to get started with a calculation of the correct response. There are six practice questions in total. For the first four, subjects who do not answer correctly in three attempts move on to the next part of the treatment. For the last two questions, subjects still receive feedback, but they must select the unique correct answer from 13 options before they can continue.[10] We reproduce the complete practice stage in online Appendix F.

Subjects in the control conditions for each experiment participate in interventions that are stylistically similar and that require comparable amounts of cognitive effort as the respective treatments. In experiment A, subjects watch a video based on a section about index funds from the same investment guide. Experiment B employs a control condition concerning portfolio diversification that also contains practice problems with personalized feedback, based on Malkiel (1985), supplemented with material from Malkiel and Ellis (2013). Neither control intervention mentions compound interest or the time value of money.

In order to isolate the features of the intervention that affect test scores and behavior, experiment A employs two additional treatments that dissect the main treatment

[8] We chose this approach because existing research indicates that financial education videos are generally more effective than written text (Lusardi et al. 2017).

[9] While the intervention is brief, it is important to bear in mind that financial education in the workplace is also brief. A meta-analysis by Fernandes, Lynch, and Netemeyer (2014) finds that the average financial education program involves only 9.7 hours of instruction. That time is divided among a long list of complex topics. For example, Skimmyhorn (2016) reports that a financial education program used by the US military covers compound interest, the focus of our current study, along with a collection of several more complex topics—retirement concepts, the Thrift Savings Plan, military retirement programs, and investments—all within a single two-hour session.

[10] A single subject dropped out of the study when attempting to answer these questions.

Table 1—Experiments Overview

	Experiment A				Experiment B	
	Treatment	Control	Rhetoric only	Subst. only	Treatment	Control
Stage 1: Education intervention						
Topic	Compound interest	Index funds	Compound interest	Compound interest	Compound interest	Portfolio diversification
Rhetoric	✓	✓	✓		✓	✓
Substance	✓	✓		✓	✓	✓
Practice and feedback					✓	✓
Stage 2: Valuation tasks	✓	✓	✓	✓	✓	✓
Stage 3: Exam-style questions						
Interest compounding	✓	✓	✓	✓	✓	✓
Index funds	✓	✓	✓	✓		
Portfolio diversification					✓	✓

into its constituent parts. The complete educational module begins with a simple explanation of compound interest illustrated through a standard iterative calculation. The remainder of the text consists of two components:

(i) Substantive content. An explanation of a simple, memorable, and potentially valuable heuristic, the rule of 72, along with five illustrative applications.[11] The rule of 72 is a method for approximating an investment's doubling period; one can also use it to approximate the growth in an investment's value over a fixed holding period. It states that the percentage interest rate on an investment multiplied by the number of periods required for its value to double equals 72 (approximately).

(ii) Rhetorical material. The section opens with the observation that "Albert Einstein is said to have described compound interest as the most powerful force in the universe." It provides various anecdotes concerning small investments that grew to impressive sums (in some cases millions of dollars) over long time periods. These anecdotes do not include any computations, and hence are not helpful for understanding the mechanics of compound interest. The text also explicitly exhorts readers to behave frugally, and characterizes compounding as a "miracle."

Subjects in the *substance-only* treatment view a video covering all of the substantive material, but omitting exhortations and atmospheric quotes.[12] In contrast, subjects in the *rhetoric-only* treatment view a video containing all of the rhetorical material, as well as the introductory explanation of compound interest, but omitting all material on the rule of 72.

[11] We used this particular investment guide in part because it teaches a useful quantitative heuristic. Some investment guides and educational interventions cover this topic without offering useful quantitative tools.

[12] In cases where it was impossible to remove sentences containing rhetorical material, we substituted neutral language. For instance, the first example of compounding presented in the original text is preceded by the transitional question, "Why is compounding so powerful?" In the substance-only treatment, we substituted the question, "How does compounding work?"

	You will get the specified dollar amount within two days from today	We will invest $a in an account with X percent interest per day. Interest is compounded daily. We will pay you the proceeds in t days.
$20	○	○
$18	○	○
$16	○	○
…	…	…
$2	○	○
$0	○	○

Figure 1. Elicitation of Valuations

Notes: The figure displays the first decision list of a round with complex framing. Payment amounts on the left include all even dollar amounts between and including $0 and $20. For simply framed choices, the text on the upper right is replaced by "We will pay you *R* in *t* days."

Stage 2. Valuation Tasks: Subjects perform ten pairs of valuation tasks. Each task elicits an equivalent current dollar value for a reward *r* to be received in either 36 or 72 days. With *simple framing*, we describe the reward as follows: "We will pay you $*r* in *t* days." With *complex framing*, we describe the same reward in terms of a return on an initial investment, as follows: "We will invest $*a* at an interest rate of *X*% per day. Interest is compounded daily. We will pay you the proceeds in *t* days." Subjects make two sets of choices pertaining to each future reward, one with simple framing, the other with complex framing. For each frame *f*, we elicit each subject *j*'s immediate dollar equivalent of a payment *R* received in *t* days, using the iterated multiple price list method with a resolution of $0.20 (Andersen et al. 2006). Figure 1 presents an example of the first decision list of a round for a complexly framed prospect.

We randomize the order of the valuation tasks at the subject level. Subjects are not told that some of the tasks are substantively equivalent, and they typically do *not* perform equivalent simply and complexly framed tasks consecutively.

Table 2 lists the parameters *t*, *X*, *a*, and *R* used for the paired valuation tasks. We chose time horizons of 36 and 72 days to simplify applications of the rule of 72.[13] This feature increases the likelihood our study will find beneficial behavioral effects of the treatment interventions.

Stage 3. Exam-Style Questions: The final stage of each experiment is an incentivized test consisting of the five questions about compound interest listed in Table 3, as well as five questions about the material covered in the video shown to the control group.[14] The exam-style questions have right and wrong answers which we specifically ask subjects to determine. These features distinguish the exam-style questions from the valuation tasks: for the latter, best choices depend on preferences, and we

[13] We use two different time frames so subjects face a greater variety of decision problems, and hence are less likely to consider successive problems highly similar.

[14] The test questions for the material in the control interventions are in online Appendix Table A2. We randomize the order of all ten test questions at the subject level.

Table 2—Decision Problems

Future reward R	Investment amount a	Daily interest rate X	Number of doublings
		Duration: 72 days	
\$20	\$10	0.01	1
\$18	\$4.5	0.02	2
\$16	\$2	0.03	3
\$14	\$0.9	0.04	4
\$12	\$2	0.025	2.5
		Duration: 36 days	
\$20	\$10	0.02	1
\$18	\$4.5	0.04	2
\$16	\$2	0.06	3
\$14	\$0.9	0.08	4
\$12	\$2	0.05	2.5

Notes: Number of doublings is the number of times the initial investment doubles over the investment horizon according to the rule of 72. Final amounts are calculated using the rule of 72. Exact final amounts differ by no more than \$0.80, except for the 4 percent interest rate over 72 days, where the rule understates the future value by \$1.16. Our analysis controls for these differences.

Table 3—Exam-Style Questions

Q1. If the interest rate is 10% per year (interest is compounded yearly), how many years does it take until an investment doubles?

7 years, *7.2 years*, *7.4 years*, *7.8 years*, *8 years*

Q2. If somebody tells you an investment should double in four years, what rate of return (per year) is he promising?

15%, *16%*, *17%*, *18%*, *19%*, *20%*

Q3. If the interest rate is 7% per year (interest is compounded yearly), about how long does it take until an investment has grown by a factor of four (i.e., is four times as large as it was originally)?

About 5 *years* to *about* 40 *years*, in steps of 5 years.

Q4. Paul had invested his money into an account which paid 9% interest per year (interest is compounded yearly). After 8 years, he had \$500. How big was the investment that Paul had made 8 years ago?

\$200 to *\$400* in steps of \$10

Q5. If an investment grows at 8 percent per year (interest is compounded yearly), by how much has it grown after 4 years?

By 30%, to *by 40%* in steps of one percentage point.

Note: Questions were presented in random order and intermingled with the five questions concerning material covered in the respective control interventions.

do not explicitly ask subjects to compute future values. Subjects are aware that test performance is incentivized prior to participating in the education intervention.

Payment: All subjects know from the outset that they will be paid according to one randomly selected decision from the entire experiment with 75 percent probability, and according to their performance on the test with 25 percent probability, in which case they earn \$1 for each of the ten questions answered correctly. We guarantee subjects that they will receive payment within at most two days of the promised date. The instructions emphasize that subjects "should make every decision as if it is the one that counts, because it might be!" In addition to the incentive payment, each subject receives a completion payment of $10.

B. *Implementation and Preliminary Analysis*

We conducted both experiments through Amazon Mechanical Turk (AMT), an online labor market. Our primary reason for employing this subject pool is that the typical member has a poor understanding of compound interest.[15] Also, this group resembles the target populations for many financial education programs in terms of age and income. We ran experiment A in eight sessions with a total of 504 subjects during April and May 2014 and experiment B in two sessions with a total of 401 subjects in October 2018, all on weekday mornings. Each subject participated in exactly one treatment of exactly one experiment. We restricted participation to subjects who resided in the United States and were at least 18 years of age. We took several precautionary measures to ensure that subjects were able to view the videos and that they would pay attention. We describe these measures and additional implementation details in online Appendix C.[16]

Each experiment begins with a two-and-a-half minute video recording of one of the authors (Bernheim) vouching that we will pay subjects exactly the amount we promise within at most two days of the promised date,[17] and ends with a number of nonincentivized questions about subjects' decision-making.[18] We intentionally place no restriction on the use of resources such as calculators, the internet, or personal advice when making decisions, as subjects always have those options when making real-world decisions. Roughly a quarter of our subjects in experiment A and less than eight percent in experiment B report using such resources when completing the incentivized test, a fraction that does not vary meaningfully across treatments. Subjects could complete all valuation tasks at their own pace. Subjects in experiment A took 62 minutes on average (standard deviation = 22 minutes) to complete the study, while those in experiment B took 75 minutes (standard deviation = 25 minutes). The difference in completion times is attributable to the longer interventions in experiment B.

On average, subjects earned $22.86 in experiment A and $23.14 in experiment B. In both experiments, earnings include a fixed $10 participation fee and range from a low of $10 to a high of $30.47. In comparison, AMT participants typically earn about $5 per hour (Mason and Suri 2012; Hara et al. 2018).

Multiple Switching: Any subject with coherent preferences will switch her choice from the immediate payment to the future reward at most once within a

[15] We opted for this subject pool after pilot experiments at Stanford and at The Ohio State University indicated that the vast majority of student subjects at these universities correctly apply compound interest calculations.

[16] A replication package is available in Ambuehl, Bernheim, and Lusardi (2022).

[17] The video invites subjects to click a link to the authors' homepages so they can verify the authenticity of the video. Before participating in the main stages of the experiment, subjects also complete an unincentivized questionnaire concerning demographics, as well as a standard battery of five questions designed to assess financial literacy (Lusardi and Mitchell 2017). We reproduce the five questions in online Appendix Table A1.

[18] Online Appendix Table A4 details the questions and responses by experiment and treatment.

single price list. Hence, we informed subjects that "most people begin a decision list by preferring the option on the left and then switch to the option on the right." In practice, 9.7 percent of subjects (49 of 504) in experiment A and 13.2 percent of subjects (53 of 401) in experiment B switched two or more times in at least one price list. This number does not differ significantly across treatments ($p = 0.85$ and $p = 0.18$ for experiments A and B, respectively). In laboratory studies of risky choices by undergraduate subjects (such as Holt and Laury 2002), the corresponding figure typically falls in the range of 10 to 15 percent. Following the usual convention (see, for example, Harrison et al. 2005), we focus attention on the 803 subjects who respected monotonicity.

Demographics: While our conclusions about the validity of methods for assessing decision quality do not depend on the demographic composition of our sample, we note that our participants are slightly more financially literate than other pools of US subjects (see Lusardi and Mitchell 2017 and Lusardi 2011). There are also differences between the samples used for experiments A and B, possibly due to changes in the composition of the AMT subject pool.[19] Subjects in experiment A are on average five years younger, the fraction of women is 6.6 percentage points lower, and their median income is 30 percent lower. Online Appendix D.1 lists all demographic details. Due to the differences between the samples, we analyze each experiment separately, without making direct statistical comparisons. Online Appendix D.2 complements our main analysis by combining data across experiments. It shows that the relative performance of the two compound-interest interventions are driven by their differences rather than by divergences between the subject pools.

Randomization and Attrition: Randomization into treatments was successful. Of the 68 F-tests we perform to assess the differences in demographic characteristics across treatments and experiments (one for each characteristic and each experiment), two are significant at the 5 percent level, and five more are significant at the 10 percent level (see online Appendix D.1). These figures are well within the expected range.

Because we conducted the experiment over the internet, attrition is a possibility. In experiment A, we find it to be negligible and unrelated to the treatments. Only four subjects who reached the stage at which they may have viewed a treatment video failed to complete the study (compared to 504 who completed it).[20] In experiment B, two subjects in the control condition and nine subjects in the treatment condition began but did not finish the intervention. While these magnitudes are small compared to the sample size of 401 subjects, they differ statistically at the 5 percent level.

[19] In addition, for experiment B, we required subjects to have at least 500 completed human intelligence tasks (HITs) and at least a 98 percent approval rating.

[20] A larger number of subjects quit before reaching that stage, but that type of attrition is necessarily independent of the treatment, and hence largely innocuous.

C. *Assessment Based on Conventional Outcome Measures*

In this section, we analyze the effects of the treatment interventions in each experiment using the two conventional metrics, exam-style questions that test comprehension, and directional effects on behavior. We defer the analysis of the substance-only and rhetoric-only treatments of experiment A to Section ID.[21]

Effects on Tested Knowledge: Panel A of Figure 2 shows how the treatments affect subjects' average scores on the five test questions pertaining to compound interest. In the control condition of experiment A, the average subject answers just under two of five, or 39 percent, of the questions correctly. The treatment intervention increases the average score by roughly 29 percentage points (1.4 additional correct answers), to 68 percent.[22] The numbers in experiment B are remarkably similar, with 37 percent correct responses in the control condition and 69 percent in the treatment condition.

Columns 1 and 2 of Table 4 summarize these results and present tests of equality between the treatment and control conditions using ordinary least squares (OLS) regressions. The treatment effects in both experiments are highly statistically significant ($p < 0.01$). Columns 3 and 4 establish that the improvements in test performance are not due to effects of the treatment on general motivation or attention. If the effects were spurious, we would find comparable effects for subjects' scores on the five test questions concerning the control interventions. On the contrary, in each experiment, subjects in the treatment condition perform substantially worse on questions about the respective control intervention than subjects in the control condition. The differences of 1.06 and 0.37 questions in experiments A and B, respectively, are again highly statistically significant ($p < 0.01$). The variation in effect size across experiments is attributable to the fact that the two experiments employ different control conditions and different test questions about those conditions.

Directional Behavioral Effects: As we have noted, it is well established that people, on average, tend to underestimate the power of compound interest. Consequently, following the approach sometimes adopted in the literature, one would deem an intervention potentially welfare-improving if it causes subjects to state higher valuations for investments that involve compound interest. To make these comparisons, we express each valuation as a percentage of the associated future reward. Formally, letting $\tilde{r}_f^{i,R,t}$ denote subject i's valuation for the future reward R received with delay t when presented in frame $f \in \{s,c\}$, we define the normalized valuation as $r_f^{i,R,t} = \tilde{r}_f^{i,R,t}/R$.[23]

Panel B of Figure 2 displays normalized valuations for complexly framed prospects by experiment and condition. Averaged across reward amounts and time frames, subjects in the control condition in both experiments regard the interest-bearing investment as equivalent to an immediate payment of about 60 percent of the future

[21] All results reported in this section are robust with respect to various statistical controls and alternative specifications. For details, see online Appendix D.

[22] See online Appendix D.3 for the effects on individual test questions.

[23] Because we elicit valuations in multiple price lists, they are interval coded. Throughout, we use the midpoint of the pertinent interval for analysis. For further details on the iterated multiple price lists, see online Appendix C.

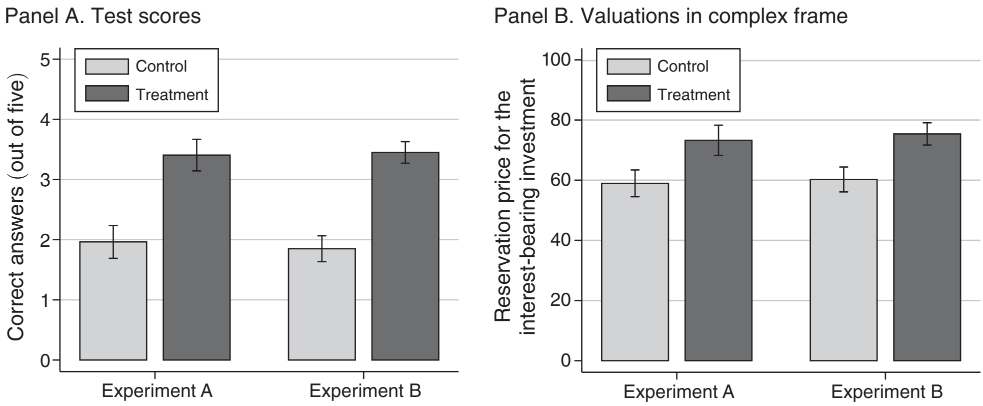

Figure 2. Assessment by Conventional Methods

Notes: Panel A: fraction of correct answers on the five exam-style questions about compound interest. Panel B: mean valuation of complexly framed future payments, $r_c^{i,R,t}$, rescaled to range from 0 to 100 (percentage points). Whiskers indicate 95 percent confidence intervals. Panel B includes multiple observations per subject; standard errors are clustered at the subject level.

reward amount. Subjects in the respective treatment conditions state valuations that are about 15 percentage points higher on average. This pattern is precisely what one would expect if the treatment in each experiment successfully counteracts exponential growth bias. Given the magnitude of the bias documented in the existing literature (Stango and Zinman 2009), the size of the average treatment effect does not appear to raise concerns about systematic overcorrection. Following the standard approach in the literature, we would thus conclude that the treatment interventions substantially increase the quality of financial decision-making, and do so similarly across the two experiments.[24]

We formalize these findings by regressing normalized valuations in the complex frame, $r_c^{i,R,t}$, on treatment indicators, clustering standard errors by subject. Column 1 of Table 5 shows the results for experiment A. In the control condition, subjects value an investment that compounds to one future dollar at 59 cents on average. The treatment condition raises average valuations by a substantial and statistically significant increment, to 73 cents $(p < 0.01)$. The estimates for experiment B, displayed in column 2, are remarkably similar. Subjects in the control condition value a future dollar at 60 cents on average, and the treatment condition increases this valuation by a substantial and statistically significant increment, to 75 cents $(p < 0.01)$. Columns 3–6 of Table 5 replicate this analysis separately for decisions with 36-day delays and those with 72-day delays, respectively. The same conclusions follow in each instance.

[24] Some prominent studies do not directly observe behavior, but rely on self-reported data about behavior (e.g., Bruhn et al. 2016). Our experiments also elicit such data. At the end of the study, subjects report the number of decision problems for which they explicitly calculated future values, whether they have used the rule of 72 in complexly framed problems, whether they have used it in simply framed problems, and whether they have used external help to make their decisions. As we find in our analysis on directional behavioral effects, our analysis of self-reports suggests that both interventions improve decisions and that they are similarly effective. See online Appendix D.4 for details.

Table 4—Performance in Exam-Style Test

	Test score compounding (out of 5)		Test score control module (out of 5)	
Experiment:	A	B	A	B
	(1)	(2)	(3)	(4)
Levels				
Control	1.963	1.849	3.284	3.078
	(0.140)	(0.110)	(0.114)	(0.101)
Treatment	3.406	3.450	2.226	2.704
	(0.135)	(0.092)	(0.092)	(0.098)
Difference	1.442	1.601	−1.058	−0.374
	(0.194)	(0.143)	(0.146)	(0.141)
Observations	215	348	215	348

Notes: Each column displays the coefficients of a separate OLS regression. Standard errors in parentheses.

Table 5—Valuations in Complexly Framed Tasks

Delay in days:	72 and 36	72 and 36	72	72	36	36
Experiment:	A	B	A	B	A	B
	(1)	(2)	(3)	(4)	(5)	(6)
Levels						
Control	58.949	60.244	56.401	58.181	61.496	62.308
	(2.275)	(2.115)	(2.336)	(2.134)	(2.374)	(2.169)
Treatment	73.261	75.378	70.629	71.747	75.893	79.010
	(2.569)	(1.886)	(2.718)	(1.904)	(2.621)	(2.001)
Difference	14.312	15.134	14.227	13.566	14.398	16.702
	(3.431)	(2.834)	(3.583)	(2.860)	(3.536)	(2.951)
Observations	2,150	3,480	1,075	1,740	1,075	1,740
Subjects	215	348	215	348	215	348

Notes: Valuations in the complex frame, $r_c^{i,R,t}$, in percentage points. Each column displays the coefficients of a separate OLS regression. Standard errors in parentheses, clustered by subject.

Taken at face value, results based on these conventional metrics seem to imply that both versions of the educational intervention are highly beneficial: they shift behavior in the desired direction, apparently for the right reasons. Moreover, given the similarity of their effects, it is tempting to conclude that the provision of opportunities for practice and individualized feedback yield no incremental benefits.

D. *Problems with Assessments Based on Conventional Outcome Measures*

Further analysis reveals that assessments based on conventional outcome metrics are misleading in this setting. Moreover, their failures here are traceable to problems that could plausibly arise in other applications.

Financial education interventions seek to ensure that people understand the consequences of their options when they make choices. In our experiment, simple framing simulates perfect education by making those consequences transparent. The degree

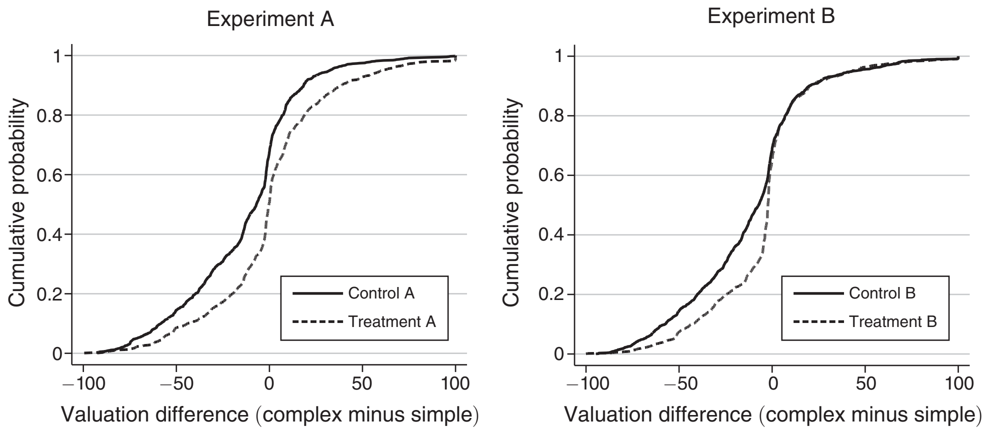

FIGURE 3. CDFS OF VALUATION DIFFERENCES

Notes: CDFs of valuation differences $r_c^{i,R,t} - r_s^{i,R,t}$ across framings for equivalent instruments in the treatment and control conditions of experiments A and B, respectively. Mass to the left of zero indicates underestimation of compound interest (exponential growth bias). Mass to the right of zero indicates overestimation.

to which financial education closes the gap between complexly and simply framed valuations, $r_c^{i,R,t}$ and $r_s^{i,R,t}$, therefore provides a gauge of the intervention's success.

Figure 3 displays the empirical cumulative distribution function of the valuation difference $r_c^{i,R,t} - r_s^{i,R,t}$ for the treatment and control conditions in each experiment. In each control condition, roughly 60 percent of subjects are afflicted by exponential growth bias $(r_c^{i,R,t} < r_s^{i,R,t})$, while a nontrivial fraction of subjects are well calibrated, or overestimate compound interest $(r_c^{i,R,t} \geq r_s^{i,R,t})$.[25]

An effective intervention would increase valuations in complexly framed tasks for subjects who underestimate compound interest, and would *decrease* them for subjects who overestimate compound interest. As a result, the cumulative distribution function (CDF) of the valuation difference $r_c^{i,R,t} - r_s^{i,R,t}$ would become steeper around zero. We observe a much different effect in experiment A: the treatment shifts the entire CDF to the right. It increases valuations for complexly framed tasks across the board, irrespective of whether a subject initially underestimates compound interest, estimates it correctly, or overestimates it. This indiscriminate effect helps some subjects but hurts others. In sharp contrast, the effects of the treatment in experiment B are sensitive to subjects' initial bias: it increases valuations of complexly framed tasks only for subjects who would have underestimated compound interest, and not for those who would have estimated compound interest correctly, or who would have overestimated it.[26] Unlike the experiment A treatment, the experiment B treatment is therefore unambiguously successful. Yet neither of the conventional outcome metrics detects this difference.

[25] Online Appendix D.7 tests and refutes the hypothesis that the overvaluation of complexly framed opportunities is solely attributable to noisy choice. A fraction of subjects in Goda et al. (2019) and in Levy and Tasoff (2016) also overestimate compound interest.

[26] In principle, a change in the valuation difference $r_c^{i,R,t} - r_s^{i,R,t}$ may reflect changes in either $r_c^{i,R,t}$ or $r_s^{i,R,t}$. We perform this decomposition in Section IIE.

Why do the conventional metrics fail to reveal the problems with the experiment A treatment? With respect to directional effects on behavior, the issues are obvious in light of Figure 3: asking whether the average change in complexly framed valuations counters the typical initial bias ignores heterogeneity in both. Despite the fact that the average complexly framed valuation moves toward the average simply framed valuation, many subjects overshoot, and some apparently double down on their preexisting biases.

The problems with knowledge assessments are more subtle. Improved performance on batteries of exam-style questions implies better financial decisions if (i) subjects employ that knowledge when making decisions, and (ii) changes in financial choices are only due to the additional knowledge, and do not involve mechanisms such as propaganda or browbeating.[27] Further investigation falsifies these assumptions.

We evaluate assumptions (i) and (ii) directly by studying the effects of the rhetoric-only and substance-only interventions in experiment A. First, we examine their effects on assessed comprehension by regressing test scores on each of the four treatments in experiment A, clustering standard errors at the subject level. Column 1 of Table 6 displays the results. The substance-only intervention increases performance on test questions concerning compound interest by a similar amount as the full intervention, while the rhetoric-only intervention has a much smaller effect.[28] To verify that these results are not due to an effect of the substance-only intervention on general motivation, Column 2 uses performance on test questions about the control intervention as the dependent variable. Indeed, subjects in the control intervention perform significantly better on these questions than subjects in any other treatment. Accordingly, increases in performance on our assessment of comprehension arise through the expected mechanism (substantive instruction).

Next, we probe the connection between effects on comprehension and effects on behavior. Assumption (i) maintains that behavior responds to knowledge. Because we have just demonstrated that the substance-only treatment improves knowledge, this assumption implies that it ought to change valuations in complexly framed decision problems. Assumption (ii) maintains that behavior only responds to features of the intervention that enhance comprehension. Because we have just demonstrated that the rhetoric-only treatment has only a modest effect on knowledge, this assumption implies that it ought to change valuations in complexly framed decisions to a much smaller degree than the full intervention.

In fact, we find precisely the opposite. Column 3 displays the coefficients of a regression of complexly framed valuations on treatment indicators. The substance-only intervention does not have a statistically discernible effect on these valuations. In contrast, the rhetoric-only intervention yields effects comparable to those of the full intervention. Column 4 confirms that the effects of the interventions

[27] This approach also assumes that the effects of knowledge on the quality of decision-making are monotonic, which may not be correct if "a little knowledge is a dangerous thing."

[28] The fact that the effect of that intervention is positive may be attributable to the inclusion of an example that illustrates the calculation of compound interest.

TABLE 6—SEPARATE EFFECTS OF RHETORIC AND SUBSTANCE IN EXPERIMENT A

	Test scores on questions about		Valuations in frame	
	Treatment	Control	Complex	Simple
	(1)	(2)	(3)	(4)
Levels				
Substance only	3.234	1.945	62.969	72.273
	(0.123)	(0.099)	(2.373)	(2.030)
Rhetoric only	2.455	2.205	77.538	77.623
	(0.146)	(0.095)	(2.785)	(2.119)
Treatment A	3.406	2.226	73.261	72.657
	(0.135)	(0.092)	(2.566)	(2.139)
Control A	1.963	3.284	58.949	72.255
	(0.140)	(0.114)	(2.272)	(2.089)
p-value of difference to control				
Treatment A	0.000	0.000	0.000	0.893
Substance only	0.000	0.000	0.222	0.995
Rhetoric only	0.015	0.000	0.000	0.072
Observations	455	455	4,550	4,550
Subjects	455	455	455	455

Notes: Each column displays the coefficients of a separate OLS regression. Standard errors in parentheses, clustered by subject.

are largely confined to the complexly framed valuations; simply framed valuations are largely unaffected.[29]

We conclude that the financial education intervention largely impacts knowledge through its substantive component, but largely impacts behavior through its rhetorical component. In other words, behavioral effects are wholly disconnected from effects on comprehension. Thus, neither assumption (i) nor assumption (ii) holds in experiment A. These findings invalidate inferences about the quality of decision-making based on assessments of comprehension in experiment A, and raise more general questions about the reliability of programmatic evaluations based on measures of knowledge and directional behavioral effects.

E. *Assessments Based on Deliberative Competence*

In this subsection, we propose a simple alternative outcome metric, justify it intuitively, and use it to evaluate the various interventions studied in experiments A and B. We demonstrate formally that this metric robustly captures welfare losses in Section II.

Divergences between valuations in equivalent simply and complexly framed tasks are of concern because they imply that people may commit errors when deciding whether to purchase instruments. The greater the divergence, the greater the potential harm these errors cause. For a given valuation pair involving future reward R and delay t, we therefore gauge subject i's *deliberative competence* by computing

[29] Simply framed valuations are a bit higher for the rhetoric-only treatment than for the control ($p < 0.10$). For the other three treatments, these valuations are essentially identical.

the absolute value of the difference between the normalized complexly and simply framed valuations:

$$d_M^{i,R,t} = -\left|r_c^{i,R,t} - r_s^{i,R,t}\right|.$$

In the language of behavioral public economics, $d_M^{i,R,t}$ is the absolute value of subject i's *price-metric bias* (Bernheim and Taubinsky 2018). The sign convention used here aids interpretation because it associates larger numerical values with higher quality decisions.[30] We average over tasks to arrive at the typical magnitude of decision error for subject i, and then average over subjects to obtain an overall measure of deliberative competence.

By definition, the absolute difference between $r_c^{i,R,t}$ and $r_s^{i,R,t}$ measures the consumer's valuation error. However, we are interested in measuring something else: the harm (or welfare loss) that such errors inflict on consumers in settings where they must decide whether to purchase instruments. Our theoretical analysis shows that the same absolute differential also measures the potential harm.

To understand this point, note that harm can befall the consumer either because she purchases an instrument when she should not, or because she fails to purchase it when she should. Suppose first that a complexly framed security paying R at time t is available for a price $p < \tilde{r}_s^{i,R,t}$, and that subject i chooses not to buy it because $\tilde{r}_c^{i,R,t} < p$. In that case, she forgoes a net gain of $\tilde{r}_s^{i,R,t} - p$. Subject to the constraint that the price remains between the two valuations (so the decisions differ), this expression reaches a maximum of $\tilde{r}_s^{i,R,t} - \tilde{r}_c^{i,R,t}$ at $p = \tilde{r}_c^{i,R,t}$. Next suppose that the same security is available for a price $p > \tilde{r}_s^{i,R,t}$, and that subject i chooses to buy it because $\tilde{r}_c^{i,R,t} > p$. In that case, she overpays by $p - \tilde{r}_s^{i,R,t}$. Subject to the constraint that the price remains between the two valuations (so the decisions differ), this expression reaches a maximum of $\tilde{r}_c^{i,R,t} - \tilde{r}_s^{i,R,t}$ at $p = \tilde{r}_c^{i,R,t}$. Thus, in both cases, $d_M^{i,R,t}$ represents the maximum loss to which the subject is exposed, normalized by R. Section II formalizes this logic and explores its robustness.

Figure 4 shows the levels of deliberative competence across control and treatment groups in experiments A and B, respectively, averaged over all valuation pairs and subjects. In each experiment, deliberative competence in the control condition is approximately −25 percent (−24.4 percent in experiment A and −25.7 percent in experiment B). A key finding is that treatment effects differ dramatically across experiments A and B. For experiment A, the treatment has no discernible effect on deliberative competence. In stark contrast, for experiment B, the treatment increases deliberative competence by a substantial and statistically significant increment, from −25.7 percent to −18.6 percent. Anticipating the results of Section II, we can say that the treatment in experiment B eliminates 27.6 percent of the welfare loss associated with poor comprehension of consequences in the complex frame $(1 - 0.186/0.257 = 0.276)$.[31] As a practical matter, this finding suggests that incorporating practice and feedback can have a dramatic impact on the efficacy of

[30] As discussed in Section II, the distance metric $-\frac{1}{2}\left(r_c^{i,R,t} - r_s^{i,R,t}\right)^2$ also has a valid welfare interpretation. Online Appendix D.5 reproduces our analysis using this metric. Results are generally similar.

[31] These estimates are based on equal weighting of observed deliberative competence across subjects and decision problems. We discuss the implications of equal weighting of subjects in Section IIF.

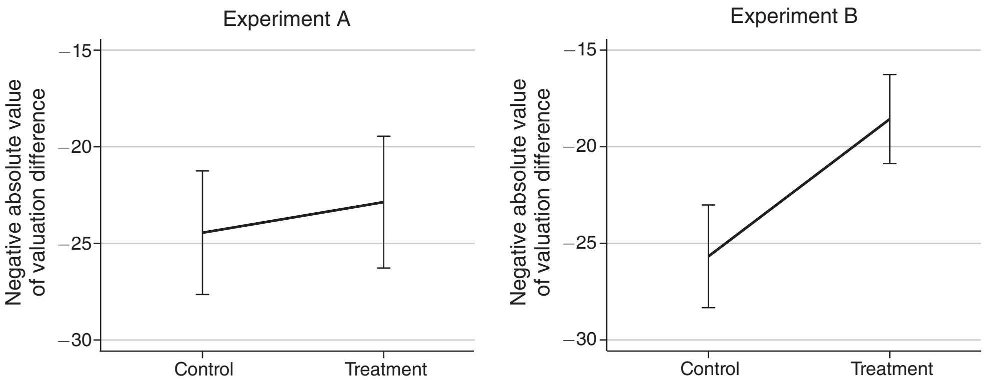

Figure 4. Assessment Based on Deliberative Competence

Notes: Mean deliberative competence, $d_M^{i,R,t} = -|r_c^{i,R,t} - r_s^{i,R,t}|$, with $r_f^{i,R,t}$ rescaled to range from 0 to 100 (percentage points). Whiskers represent 95 percent confidence intervals with standard errors clustered at the subject level.

financial education, even when the feedback is fully automated and relatively simple. Thus, using deliberative competence rather than conventional outcome metrics reverses our conclusions concerning the role of practice and feedback.

To evaluate treatment effects econometrically, we regress $d_M^{i,R,t}$ on a treatment indicator and cluster standard errors by subject, separately for each experiment. Table 7 displays the results. Column 1 shows that the treatment intervention in experiment A increases deliberative competence by a small and statistically insignificant 1.5 percentage points. Column 2 exhibits the corresponding regression for experiment B. It shows that while control subjects display a level of deliberative competence similar to those in experiment A, the treatment intervention substantially raises deliberative competence. The increase of 7.1 percentage points is highly statistically significant $(p < 0.01)$. Columns 3 and 4 replicate this analysis separately using the set of investments with 36-day delays, whereas Columns 5 and 6 do so for investments with 72-day delays. The same conclusions follow in both cases.

Column 1 also shows that the substance-only intervention leaves deliberative competence unchanged, despite its positive effect on knowledge. While the rhetoric-only intervention produces a small improvement in deliberative competence, we know from our earlier analysis that this effect does not flow from improved comprehension. Rather, the motivational rhetoric provides a nudge that proves fortuitously beneficial to a small degree. Columns 3 and 5 show that we obtain similar results for tasks with 36-day and 72-day delays.

II. Theoretical Framework and Supplemental Empirical Analyses

In Section IE, we provided an intuitive justification for using differentials between simply and complexly framed valuations of the same future payment to evaluate policy interventions that seek to improve comprehension of the consequences that follow from choices. According to our empirical analysis, such metrics capture

Table 7—Deliberative Competence

Delay in days:	72 and 36	72 and 36	72	72	36	36
Experiment:	A	B	A	B	A	B
	(1)	(2)	(3)	(4)	(5)	(6)
Levels						
Control	−24.448	−25.673	−24.076	−25.945	−24.820	−25.401
	(1.633)	(1.357)	(1.744)	(1.428)	(1.682)	(1.393)
Treatment	−22.864	−18.569	−22.959	−18.441	−22.769	−18.697
	(1.741)	(1.177)	(1.799)	(1.208)	(1.886)	(1.253)
Substance only	−22.012		−21.257		−22.767	
	(1.433)		(1.422)		(1.620)	
Rhetoric only	−19.797		−20.191		−19.403	
	(1.406)		(1.458)		(1.506)	
p-value of difference to control						
Treatment	0.507	0.000	0.656	0.000	0.418	0.000
Substance only	0.263		0.211		0.380	
Rhetoric only	0.031		0.088		0.017	
Observations	4,550	3,480	2,275	1,740	2,275	1,740
Subjects	455	348	455	348	455	348

Notes: Each column displays the coefficients of a separate OLS regression of deliberative competence on treatment indicators. Standard errors in parentheses, clustered by subject. Deliberative competence measured as $d_M^{i,R,t} = -|r_c^{i,R,t} - r_s^{i,R,t}|$, with $r_c^{i,R,t}$ rescaled to range from 0 to 100 (percentage points).

important aspects of policy effects that conventional outcome measures overlook. The objectives of this section are to demonstrate that deliberative competence, as we define it, is robustly interpretable as a measure of the welfare loss resulting from poor decision-making, even in contexts that raise second-best issues, and to clarify the types of settings to which these conclusions apply.

After describing the setting (Section IIA), we articulate our welfare criteria (Section IIB). The most novel aspect of this discussion concerns our treatment of decision-making defects outside the scope of analysis. We proceed from the premise that it is impractical to analyze all biases and remedial policies at once (*comprehensive* welfare analysis). Instead, studies in behavioral public economics invariably focus on one or two biases and policies at a time (*compartmentalized* welfare analysis). Yet the literature offers no framework for conducting such analyses in a manner that avoids the conceptual "second-best" problems identified by Lipsey and Lancaster (1956). To fill this gap, we propose and justify the concept of *idealized welfare analysis*.

We formalize the welfare interpretation of our deliberative competence measures under the restrictive assumption that the consumer's decision-making apparatus involves no defects other than the misunderstandings that the pertinent policy interventions target (Section IIC), as well as under less restrictive assumptions that permit us to conduct idealized welfare economics while acknowledging the existence of other biases (Section IID). We then discuss a method for detecting and adjusting for a class of potential confounds (Sections IIE), as well as a strategy for addressing issues involving interpersonal aggregation (Section IIF), both of which we implement using our experimental data.

A. *Setting*

Following the literature on behavioral public economics (reviewed in Bernheim and Taubinsky 2018), we distinguish between choice options and the preference-relevant outcomes they induce. For example, people make choices over financial portfolios but care about the patterns of returns those portfolios generate, rather than the portfolios themselves. Likewise, people make choices among over-the-counter medications but care about the health consequences those products yield, rather than the medications themselves. We refer to such choice options as *instruments* because they are a means to an end; they do not matter intrinsically. The distinction between an instrument and its intrinsically valued consequences is important because it allows for the possibility that people may not understand the implications of their choices, a phenomenon known in behavioral public economics as *characterization failure* (Bernheim and Taubinsky 2018).

Our method of measuring deliberative competence involves simple binary decisions: whether to purchase an instrument $I \in \mathcal{I}$ at a price p. We assume the consumer cares about current cash, m, and a vector y of future consequences (e.g., specifying available cash at each point in time and state of nature). She starts with current cash m_0; without loss of generality, we simplify notation by setting $m_0 = 0$. If she forgoes the instrument, she anticipates the consequences y_0. In frame f, she interprets the instrument I as offering consequences y_f. In other words, she thinks the incremental effect of the instrument will be $y_f - y_0$. We assume there is a simple way to frame the instrument, $f = s$, that presents its consequences transparently, thereby ensuring the consumer interprets them correctly. Thus, y_s represents the instrument's actual consequences. With naturalistic (complex, nontransparent) framing, $f = c$, she misinterprets the instrument as offering the outcome $y_c \neq y_s$.[32] In other words, she experiences a form of characterization failure. Our object is to evaluate policies designed to mitigate those characterization failures. Where we distinguish between policies explicitly, we use $y_{c\theta}$ to denote the consumer's beliefs in the complex frame about the instrument's consequences when policy θ is in place. In our experiment, the instrument is a claim on a future payment, simply framed tasks transparently specify the future payment associated with the available instrument, complexly framed tasks describe the future payment in terms of compound interest, and the policies of interest target comprehension of compounding.

We posit the existence of a utility function, $V(m, y_f)$, that rationalizes observed choices. By assumption, the characteristics of the instrument and the frame only impact the consumer's choices insofar as they determine her perceptions of consequences. Some of our analysis concerns cases in which the decision-maker suffers

[32] As noted in the text, this framework accommodates uncertainty, in the sense that we can interpret y as a vector of state-contingent consequences. Under that interpretation, it would appear that false beliefs must pertain to those consequences and not to their likelihoods. However, if baseline consequences (y_0) are either constant over states or independent of the instrument's returns, we can reformulate a setting with false beliefs about probabilities as one with correct beliefs about probabilities and false beliefs about state-contingent payments, which our framework subsumes. To take a simple example, imagine that a lottery pays x with probability $\frac{1}{3}$ and z with probability $\frac{2}{3}$, but the consumer falsely believes the probabilities are reversed. We can represent this setting as one in which the lottery yields a payment y_s^ω in state of nature $\omega \in \{1,2,3\}$, where $y_s^1 = x$, $y_s^2 = y_s^3 = z$, and the states are equally probable; furthermore, the consumer has correct beliefs about the probabilities of the underlying states but incorrectly believes the lottery pays $y_c^1 = y_c^2 = x$ and $y_c^3 = z$.

from additional and possibly unknown valuation biases that the aforementioned simple framing does not remove. We will assume that, with these ancillary biases removed through other means, the utility function $U(m, y_f)$ would rationalize choices in the simple and complex frames. Under these assumptions, $U(m, y_s)$ plays the role of "true preferences."[33] We interpret V and U as indirect utility functions, in the sense that m and y could represent income rather than consumption. We assume throughout that these functions are unbounded above and below in m, that they have bounded first and second derivatives, and that their first derivatives with respect to m are strictly positive.

We define the consumers' reservation price (alternatively, *valuation*) given her interpretation of the instrument's consequences, y_f, assessed according to utility function $u \in \{U, V\}$, as the highest price at which she purchases the instrument (denoted $r^u(y_f)$). This valuation is given by the following condition:[34]

$$u\big(-r^u(y_f), y_f\big) = u(0, y_0). \tag{1}$$

B. *Framework for Welfare Analysis*

We consider a social planner who seeks to quantify the utility loss a consumer sustains from characterization failures in the domain targeted by the policies we aim to evaluate. The planner uses a utility function u (the *evaluation standard*) for this purpose. For reasons that will become apparent in Section IID, we allow for the possibility that u is either U or V.

1. Welfare Criteria.—We evaluate welfare treating the price p of an instrument I as a random realization. This perspective is appropriate in the context of our experiment because p is literally random. Of course, measuring welfare is important in an experiment such as ours not because we care about the experimental tasks per se, but because those tasks arguably distill the types of trade-offs people may confront in practice. When interpreting our welfare measures, one can think of the randomness in p as representing uncertainty about the particular trade-offs that will prove relevant to real-world decisions.

We adopt two approaches to aggregating welfare over possible realizations. The first envisions a welfarist planner concerned with the consumer's expected utility. We assume the future price p for the instrument I is distributed according to some atomless CDF H with density h. The loss, in terms of expected utility, associated with employing valuation r rather than the utility-maximizing valuation $r^u(y_s)$ is then

$$l_W^u(r, y_s, H) = \begin{cases} \int_r^{r^u(y_s)} \big[u(-p, y_s) - u(0, y_0)\big] h(p)\,dp, & \text{if } r \le r^u(y_s); \\ \int_{r^u(y_s)}^r \big[u(0, y_0) - u(-p, y_s)\big] h(p)\,dp, & \text{if } r \ge r^u(y_s). \end{cases} \tag{2}$$

[33] To be clear, our approach does not require the existence of a single "true preference" relation. As we discuss in Section IID, it extends directly to cases in which welfare-relevant choices are context dependent, and consequently admit multiple frame-dependent utility representations, as permitted in the welfare framework of Bernheim and Rangel (2009).

[34] Here, we assume that in case of indifference, the consumer purchases the instrument. Existence of r^u follows from continuity, monotonicity, and unboundedness.

The subscript W reminds us that this measure reflects a welfarist perspective.

The welfarist approach assumes the probability distribution of the price p is known. While it is known for any given experiment, it is hard to say which prices give rise to trade-offs that are likely to emerge in practice. Treating the probability distribution as unknown, we can employ an alternative approach from computer science that has recently found its way into economics: evaluate an option based on the maximum possible loss it could induce.[35] The maximal loss associated with using the valuation r rather than the utility-maximizing valuation $r^u(y_s)$ is

$$(3) \qquad l_M^u(r, y_s) = \begin{cases} \min\limits_{p \in [r, r^u(y_s)]} \{u(-p, y_s) - u(0, y_0)\}, & \text{if } r \leq r^u(y_s); \\ \min\limits_{p \in [r^u(y_s), r]} \{u(0, y_0) - u(-p, y_s)\}, & \text{if } r \geq r^u(y_s). \end{cases}$$

The subscript M reminds us that this function measures the maximal loss.

Because the scale of u is arbitrary, in both cases we convert to money-metric utility. Specifically, we divide the welfare loss by the perceived marginal utility of income according to u when the individual does not purchase the instrument:[36]

$$L_W^u(r, y_s, H) = \frac{1}{u_m(0, y_0)} l_W^u(r, y_s, H),$$

$$L_M^u(r, y_s) = \frac{1}{u_m(0, y_0)} l_M^u(r, y_s).$$

2. *Accounting for Additional Biases.*—In many applications, it is reasonable to think that consumers suffer from biases other than misapprehension of the consequences y_s. The existence of such biases may cast doubt on our assumption that simply framed choices provide a valid normative benchmark. Loosely speaking, if we suspect that biases infect simply framed choices, we must accept the possibility that they may not reveal "true preferences." Formally, the issue is that such biases cause the utility function V to diverge from U (other than in the sense of a monotonic transformation), affecting choices in both frames.

As an illustration of the problem, in our experiment, simple framing eliminates the need to understand compound interest, but it does not address considerations such as present focus or excessive suspicion that the experimenter will renege on promised payments.[37] If these considerations bias subjects' choices toward lower valuations of future prospects in *both* frames, then one might be tempted to conclude that the "overshooting" documented in Section ID is actually beneficial.

Lipsey and Lancaster (1956) reasoned that the right way to handle these types of issues is to analyze all distortions (here, biases affecting the valuation $r^V(r_c)$) and all

[35] See, for example, Carroll (2019), who adopts this approach to study optimal mechanism design in settings where the distribution of agent characteristics is unknown. Other applications include decision theory (Machina and Siniscalchi 2014) and public policy (Sunstein 2020).

[36] This adjustment approximates the reduction in income without the instrument that would produce the same utility decrement as the consumer's decision error. Alternatively, one could define L_e^u (for $e \in \{W, M\}$) as the *exact* reduction in income sufficient to produce that utility decrement: $u(0, y_0) - u(-L_e^u, y_0) = l_e^u$. Because we ultimately take limiting approximations, these two scaling methods yield the same results.

[37] Present focus may not be an issue in our experiment because the earliest payments involve a two-day delay.

remedies for them simultaneously. We call this approach *comprehensive second-best analysis*. While conceptually attractive, it is infeasible in practice.[38] Indeed, the field of behavioral economics has not yet arrived at the comprehensive structural understanding of imperfections in human decision-making that this approach requires.

Because of these practical challenges, the overwhelming preference of economists, as revealed by research in this area, is to analyze policies one (or two) at a time. The idea is to focus on solving individual pieces of the overall welfare puzzle, thereby progressively building up a complete solution. But the compartmentalization of policy analysis raises challenging conceptual issues. Indeed, any approach other than comprehensive second-best analysis necessarily involves conceptual compromises. Still, to the extent the requirement of tractability necessitates a compromise, we ought to make the best one possible. Table 8 lists three frameworks for compartmentalized welfare analysis. Two are familiar; the third is our proposal.

The first framework, which we label *narrow second-best welfare analysis*, involves a perspective that Lipsey and Lancaster (1956) attribute to Meade (1955). We imagine that there are many biases affecting the valuation $r^V(r_c)$ and corresponding policy options that target them. We evaluate one policy targeting one bias (or a small collection of policies targeting a small number of biases) accounting for the existence of the other biases, but holding the rest of the policies fixed, even if they are suboptimal. Within our setting, this approach quantifies the welfare cost of choosing $r^V(y_{c\theta})$ (which reflects all biases, as they exist under prevailing policies) rather than $r^U(y_s)$ (which reflects the absence of all biases), using U as the standard of evaluation. In other words, as shown in Table 8, the pertinent measures of the welfare loss for policy θ are $L_M^U\big(r^V(y_{c\theta}), y_s\big)$ and $L_W^U\big(r^V(y_{c\theta}), y_s, H\big)$.

Setting aside the enormous practical difficulty of identifying and properly accounting for all pertinent biases, this approach suffers from a critical conceptual flaw: if the policy of interest is merely one component of a broader effort to design interventions that improve the quality of decision-making, then treating other policies as fixed constraints can yield incorrect conclusions. To illustrate, consider the subjects in our experiment who overstate the returns to investments paying compound interest. If they suffer from present bias, then, as suggested at the outset of this section, one might be tempted to conclude that any improvement in their understanding of compounding would cause them harm. And yet, they would clearly derive the greatest benefit from a combination of policies that address both biases, such as education along with commitment devices. Thus, insisting that a policy ought to compensate for all biases that may infect choice, and not merely for the one it is designed to address, can yield misleading conclusions concerning the policy of interest.

The second framework in Table 8, *myopic welfare analysis*, focuses on a small number (usually just one) of biases and associated policy instruments while assuming, at least implicitly, that the consumer's decision-making apparatus is otherwise faultless, perhaps because effective remedies for other valuation biases are already in place. Within our setting, this approach calls for evaluating the loss from choosing $r^V(y_{c\theta})$, the naturalistic selection, rather than $r^V(y_s)$, according to the utility function

[38] For example, while the general theoretical framework of Farhi and Gabaix (2020) encompasses such analyses, they acknowledge that empirical implementation would be a "momentous task" (p. 311).

Table 8—Frameworks for Compartmentalized Welfare Analysis

Framework	Evaluated choice	Evaluation standard	Welfare criterion
Narrow second-best welfare analysis	$r^V(y_{c\theta})$	U	$L_M^U(r^V(y_{c\theta}), y_s)$ or $L_W^U(r^V(y_{c\theta}), y_s, H)$
Myopic welfare analysis	$r^V(y_{c\theta})$	V	$L_M^V(r^V(y_{c\theta}), y_s)$ or $L_W^V(r^V(y_{c\theta}), y_s, H)$
Idealized welfare analysis	$r^U(y_{c\theta})$	U	$L_M^U(r^U(y_{c\theta}), y_s)$ or $L_W^U(r^U(y_{c\theta}), y_s, H)$

V. In other words, as shown in Table 8, the pertinent measures of the welfare loss for policy θ are $L_M^V(r^V(y_{c\theta}), y_s)$ and $L_W^V(r^V(y_{c\theta}), y_s, H)$. This form of compartmentalized policy analysis dominates the literature. Examples of studies that focus on a single bias include Gruber and Kőszegi (2001, 2004); DellaVigna and Malmendier (2004); and O'Donoghue and Rabin (2006). A few studies consider two biases (and associated policies) at a time, for instance Allcott, and Mullainathan, and Taubinsky (2014). While such analyses highlight some important second-best considerations, they inevitably ignore others.

The problem with myopic welfare analysis is that the data are not in fact generated by an otherwise faultless decision process. As Lipsey and Lancaster (1956) pointed out in an analogous setting involving multiple distortions, there is no reason to think that solutions derived in this manner (which they called "piecemeal") are in fact desirable. Nor can one coherently paste together solutions for different biases derived in this way, because each solution implicitly assumes that the other solutions are unnecessary. In the context of our experiment, for instance, pasting together a solution for poor comprehension, devised under the assumption of no remedy for present bias, with a solution for present bias, devised under the assumption of no remedy for poor comprehension, does not generally yield policies that complement each other well and produce a desirable overall outcome.

The third approach listed in Table 8, *idealized welfare analysis*, reformulates piecemeal policy analysis in a manner that avoids Lipsey and Lancaster's criticism. Viewing the analysis as part of a broader effort to diagnose and correct flaws in decision-making, one evaluates policies designed to address a single bias (or a small collection of biases) under the assumption that effective remedies for other biases *will be* forthcoming. In contrast to myopic welfare analysis, which assumes that other biases have *already been* addressed, one must therefore account for the possibility that the observed data are infected by biases other than those the policy of interest seeks to target. Approaching each component problem in this manner, one can potentially arrive at a medley of compartmentalized solutions that complement each other and together achieve an overall solution.[39] For our setting, idealized welfare analysis gauges the welfare cost of choosing $r^U(y_{c\theta})$ rather than $r^U(y_s)$ (unobserved valuations purged of other biases), which potentially differ from $r^V(y_{c\theta})$ and $r^V(y_s)$ (observed valuations infected with other biases), using U as the evaluation standard. Thus, unlike myopic welfare analysis, it does not take observed

[39] By way of analogy, consider the problem of maximizing a function $f(x_1, \ldots, x_N)$. Let x^* denote the solution. Idealized welfare analysis is akin to maximizing $f(x_n, x_{-n}^*)$ for each n.

simply framed valuations at face value as valid normative benchmarks. In the context of our experiment, the approach calls for assessing the loss associated with the complexly framed valuations we would elicit in an idealized setting that removes all valuation biases other than a misunderstanding of compound interest, based on simply framed valuations elicited in that same scenario. As shown in Table 8, the pertinent measures of the welfare loss for policy θ are $L_M^U\big(r^U(y_{c\theta}), y_s\big)$ and $L_W^U\big(r^U(y_{c\theta}), y_s, H\big)$.

A natural first reaction to this proposed framework is that, if practical limitations render comprehensive second-best welfare analysis infeasible, then they must also preclude idealized welfare analysis. After all, idealized welfare analysis, too, requires knowledge of U, which the data do not reveal. And yet, as we prove below, there are reasonably general conditions under which idealized and myopic welfare measures coincide (up to a multiplicative factor), at least within the specific context of analyses involving comparisons of valuations between simple and complex frames.

Anything short of comprehensive welfare analysis inevitably involves conceptual compromises, and idealized welfare analysis is no exception. One must assume that other biases will be fully resolved, rather than either partially resolved or left untreated, and that the policies targeting them will not directly impact the bias of interest—for instance, that a policy targeting present bias or false beliefs concerning experimenter reliability will not meaningfully improve comprehension of compound interest.

The fact that one can perform idealized welfare analysis without recovering the idealized utility function U—potentially without understanding or even knowing about other biases—is counterintuitive. After providing formal results, we therefore offer a simple illustrative example, which also shows that the local approximations used in our main results can perform well for large distortions.

3. *Approximations.*—In the following subsections, we provide formal justifications for our metrics of deliberative competence. Our approach is, in effect, to employ first- and second-order Taylor series approximations for welfare losses. Within public economics, this is a standard strategy for quantifying inefficiencies resulting from market distortions. For example, it provides the foundation for the famous Harberger triangle (Harberger 1964).

Formally, we define

$$y_f^\alpha \equiv \alpha y_f + (1 - \alpha) y_0.$$

This formulation allows us to rescale the consequences of the instrument as perceived in the simple and complex frames up and down by the factor α. As the scale of the instrument shrinks, the pertinent distribution of prices (which we write as H^α to reflect the dependence on α) presumably becomes more concentrated near zero. Absent information about the distribution of prices encountered in real-world analogs to our experimental tasks, we assume for the purpose of the welfarist metric that H^α is uniform in the pertinent region, with density $\frac{h}{\alpha}$.

Defining distance metrics $d_M(r,r') \equiv |r - r'|$ and $d_W(r,r') \equiv \frac{1}{2}(r - r')^2$, we show that $d_M(r^u(y_c), r^u(y_s))$ provides a first-order approximation for $L_M^u(r^u(y_c), y_s)$, and that $d_W(r^u(y_c), r^u(y_s))$ provides a second-order approximation for $L_W^u(r^u(y_c), y_s, H)$ (hence our subscript convention).[40]

In taking limits, we hold the consumer's misunderstanding of the instrument fixed and let the scale of its consequences shrink. Another possibility is to hold the scale of the instrument's consequences fixed and let the magnitude of the misunderstanding shrink. Potentially, the latter approach may provide better approximations for large instruments, but worse approximations for large misunderstandings. It may also be more suitable for settings where risk plays a central role. As shown in online Appendix B, this alternative approach yields similar approximation results.

C. *Welfare Analysis When Simple Framing Involves No Biases*

We begin by assuming that the decision-maker suffers from no biases other than misapprehension of an instruments' consequences, y_s. In this case, $V = U$. Our first result shows that one can approximate welfare losses by measuring distances between paired valuations.

PROPOSITION 1: *For all y_c and y_s satisfying $\left.\frac{dr^V(y_c^\alpha)}{d\alpha}\right|_{\alpha=0} \neq \left.\frac{dr^V(y_s^\alpha)}{d\alpha}\right|_{\alpha=0}$, we have*

$$\lim_{\alpha \to 0}\left(\frac{L_M^V(r^V(y_c^\alpha), y_s^\alpha)}{d_M(r^V(y_c^\alpha), r^V(y_s^\alpha))}\right) = 1 \tag{4}$$

and

$$\lim_{\alpha \to 0}\left(\frac{L_W^V(r^V(y_c^\alpha), y_s^\alpha, H^\alpha)}{d_W(r^V(y_c^\alpha), r^V(y_s^\alpha))}\right) = h. \tag{5}$$

Equation (4) provides a formal foundation for treating $d_M(r^V(y_c), r^V(y_s))$ as an approximate measure of the maximal money-metric welfare loss, $L_M^V(r^V(y_c), y_s)$. Suppose our objective is to evaluate two corrective policies, θ and θ', that lead the consumer to interpret the instrument as offering $y_{c\theta}$ and $y_{c\theta'}$, respectively. According to Proposition 1, $d_M(r^V(y_{c\theta}), r^V(y_s)) - d_M(r^V(y_{c\theta'}), r^V(y_s))$ approximates the difference in the maximal money-metric welfare losses, $L_M^V(r^V(y_{c\theta}), y_s)) - L_M^V(r^V(y_{c\theta'}), y_s)$. Similarly, for the welfarist criterion, equation (5) implies that $L_W^V(r^V(y_c), y_s, H)$ is approximately proportional to $d_W(r^V(y_c), r^V(y_s))$, where the constant of proportionality is independent of the instrument's perceived consequences, and hence of the corrective policy. While this measure does not tell us the absolute level of the welfare loss associated with any policy, it does allow us to rank any two

[40] We use a second-order approximation for L_W^u rather than a first-order approximation because the first-order term is identically equal to zero.

policies correctly, and to approximate the ratio of their associated money-metric welfare losses, $\frac{L_W^V\left(r^V(y_{c\theta}), y_s, H\right)}{L_W^V\left(r^V(y_{c\theta'}), y_s, H\right)}$, with $\frac{d_W\left(r^V(y_{c\theta}), r^V(y_s)\right)}{d_W\left(r^V(y_{c\theta'}), r^V(y_s)\right)}$.

Despite the fact that the expected welfare loss L_W^V and the maximal welfare loss L_M^V reflect different approaches to conceptualizing the loss function, they admit similar welfare approximations, inasmuch as $\left(r^V(y_c) - r^V(y_s)\right)^2$ is simply the square of $\left|r^V(y_c) - r^V(y_s)\right|$. While the expected loss reflects a more conventional approach, there are two reasons to prefer the maximal loss. A practical advantage is that $\left|r^V(y_c) - r^V(y_s)\right|$ is less sensitive to outliers than $\left(r^V(y_c) - r^V(y_s)\right)^2$. A conceptual advantage is that the maximal loss is directly interpretable as a dollar-value measure, while the expected loss is a rescaling of a dollar-value measure by the unknown density h. The latter remains useful, however, in that it correctly ranks policies and accurately measures the ratio of their benefits.

In Section IE, we averaged d_M over a collection of instruments. Proposition 1 also allows us to interpret such averages. Suppose our objective is to determine $L_M^V(\mathcal{I}) \equiv \sum_{I\in\Im}\lambda_I L_M^V\left(r^V(y_c^I), y_s^I\right)$, where $\mathcal{I}$ includes the instruments of interest, and y_f^I denotes the perceived outcome for instrument I in frame f. The weight λ_I could reflect judgments of each instrument's relevance for real-world choices, or about the relative importance of an incremental dollar in the types of circumstances that are associated with the instrument's availability. Proposition 1 tells us that we can approximate $L_M^V(\mathcal{I})$ by computing $\sum_{I\in\Im}\lambda_I d_M\left(r^V(y_c^I), r^V(y_s^I)\right)$. Likewise, the proposition tells us that we can approximate $L_W^V(\mathcal{I})$ (similarly defined) up to a constant of proportionality (which allows us to rank policies and measure their relative benefits) by computing $\sum_{I\in\Im}h_I\lambda_I d_W\left(r^V(y_c^I), r^V(y_s^I)\right)$, where h_I is the density of the price distribution for instrument I. Because the pertinent price distributions are unknown, it is reasonable to treat h_I as either a constant, or at least uncorrelated with the welfare loss. In either case, computing $\sum_{I\in\Im}\lambda_I d_W\left(r^V(y_c^I), r^V(y_s^I)\right)$ provides the desired approximation. Aggregation over consumers raises related issues; see Section IIF.

A limitation of Proposition 1 and other results to follow is that they pertain to binary choices rather than to selections from arbitrary menus. Because binary comparisons lie at the core of more complex choices, this feature is less of a limitation than one might think. Suppose the consumer must purchase an instrument from the set $M \cup \{0\}$, where 0 represents the status quo. As an example, different elements of this set might represent different quantities of the same instrument. Suppose further that we have assessed simply and complexly framed valuations for all members of M. Letting M_L (M_H) represent the subset of instruments for which the consumer's valuation is too low (too high), the greatest possible welfare loss is then proportional to $\max_{I\in M_H} d_M\left(r^V(I,c), r^V(I,s)\right) + \max_{I\in M_L} d_M\left(r^V(I,c), r^V(I,s)\right)$ (where $r^u(I,f)$ is shorthand for the valuation of instrument I in frame f and, by convention, $\max_{I\in\emptyset} L_M(I) \equiv 0$)—i.e., it is the loss incurred when the consumer purchases the most overvalued instrument at the worst possible price and forgoes the most undervalued instrument at the best price consistent with purchasing the first instrument.[41]

[41] Suppose the consumer purchases I at price p when she should have purchased I' at price p'. We then have $r^u(I,c) - p \geq \max\{0, r^u(I',c) - p'\}$ and $r^u(I',s) - p' \geq \max\{0, r^u(I,s) - p\}$. For any given p, the loss is maximized by setting p' as low as possible, consistent with these inequalities. Ignoring the nonnegativity constraints for

Welfarist extensions appear more challenging due to the need for assumptions concerning the joint distribution of prices, but may be possible in cases where such assumptions are natural (for example, when an instrument is available in various quantities at a fixed unit price).

D. *Welfare Analysis When Simple Framing Involves Additional Biases*

1. The General Result.—We now relax the assumption that simple framing, by itself, removes all pertinent biases. Formally, we extend our analysis to allow for arbitrary differences between U and V. The main result of this section shows that our measure of deliberative competence, $d_e\left(r^V\left(y_c^\alpha\right), r^V\left(y_s\right)\right)$, approximates the corresponding idealized welfare measure up to a multiplicative constant for each welfare perspective $e \in \{M, W\}$ if a certain separability assumption holds. The assumption requires that we can write the utility functions as $V\left(m, \varphi(y)\right)$ and $U\left(m, \varphi(y)\right)$ for some subutility function φ.

When y is a scalar (as in our experiment), the separability assumption involves no loss of generality. When y is multidimensional, it is restrictive, but it nevertheless subsumes important possibilities, including the following.

(i) Instruments are investments that yield financial returns in future periods and states of nature, which the decision-maker misunderstands. Our objective is to evaluate policies that target those misunderstandings. A complicating factor is that the consumer may also be present biased to an unknown degree. In that case, the utility functions may take the form $V(m,y) = \nu(m) + \beta\varphi(y)$ and $U(m,y) = \nu(m) + \varphi(y)$, where $\varphi(y)$ is δ-discounted utility and β is the unknown present-bias parameter.

(ii) Instruments are insurance contracts that yield future payments contingent on the occurrence of a particular event. The consumer misunderstands the formulas that determine those payments, and our objective is to evaluate policies that target those misunderstandings. A complicating factor is that the consumer may misjudge the probability of the event to an unknown degree. In that case, the utility functions may take the form $V(m,y) = \nu(m) + \tilde{\pi}\varphi(y) + (1 - \tilde{\pi})u_0$ and $U(m,y) = \nu(m) + \pi\varphi(y) + (1 - \pi)u_0$, where $\tilde{\pi}$ is the unknown perceived probability, π is the true probability, $\varphi(y)$ is expected utility conditional on the event, and u_0 is expected utility conditional on the event's absence.

(iii) Instruments are medications that impact various aspects of future health in ways consumers misunderstand. Our objective is to evaluate policies that target those misunderstandings. A complicating factor is that the consumer may

the moment, we have $p' = r^u(I',c) - r^u(I,c) + p$. Setting $p = r^u(I,c)$ and $p' = r^u(I',c)$ satisfies nonnegativity. The total loss, which we approximate as $\left[r^u(I',s) - p'\right] + \left[p - r^u(I,s)\right]$, is then $\left[r^u(I',s) - r^u(I',c)\right] + \left[r^u(I,c) - r^u(I,s)\right]$. Choosing I and I' to be the instruments with, respectively, the greatest overvaluation and the greatest undervaluation maximizes this expression. Note that this argument subsumes the case in which either I or I' is 0.

also be imperfectly attentive to health consequences. In that case, the pertinent utility functions might take the assumed forms, with $\varphi(y)$ representing a composite health good, and with the discrepancy between V and U reflecting inattention.

(iv) An additional complicating factor in any of the preceding examples is that the method used for eliciting valuations may activate biases. For example, multiple price lists can induce "end-point" effects (Andersen et al. 2006) and, more generally, framing can "anchor" subjects' assessments (e.g., Ariely, Loewenstein, and Prelec 2003). In such cases, we might have $U(m,y) = \nu(m) + \varphi(y)$ and $V(m,y) = \mu(m) + \xi(\varphi(y))$ for functions μ and ξ.

While this list is not exhaustive, it is important to understand that the separability assumption limits potential applications by ruling out biases that impact trade-offs among the elements of the vector y. For example, treating departures from exponential discounting as biases, our setting accommodates quasi-hyperbolic discounting but excludes hyperbolic discounting.[42]

The following result shows, in effect, that the idealized welfare loss, $L_M^U(r^U(y_c), y_s)$ or $L_W^U(r^U(y_c), y_s, H)$, and the corresponding measure of deliberative competence, $d_e(r^V(y_c), r^V(y_s))$ for $e \in \{W, M\}$, are approximately proportional (in the sense that they become exactly proportional in the limit as $\alpha \to 0$), where the factor of proportionality does not depend on the decision-maker's misperceptions of consequences.

PROPOSITION 2: *For each $e \in \{W, M\}$, there exists a strictly positive constant, K_e, such that for all y_c and y_s satisfying $\left.\frac{dr^u(y_c^\alpha)}{d\alpha}\right|_{\alpha=0} \neq \left.\frac{dr^u(y_s^\alpha)}{d\alpha}\right|_{\alpha=0}$, we have*

$$\lim_{\alpha \to 0}\left(\frac{L_M^U(r^U(y_c^\alpha), y_s^\alpha)}{d_e(r^V(y_c^\alpha), r^V(y_s^\alpha))}\right) = K_M, \tag{6}$$

$$\lim_{\alpha \to 0}\left(\frac{L_W^U(r^U(y_c^\alpha), y_s^\alpha, H^\alpha)}{d_e(r^V(y_c^\alpha), r^V(y_s^\alpha))}\right) = K_W. \tag{7}$$

Proposition 2 (for idealized welfare analysis) is weaker than Proposition 1 (for myopic welfare analysis) in only one respect: it may be the case that $K_M \neq 1$. Even with that qualification, several important conclusions follow.

First, we can conduct useful welfare analysis in a compartmentalized fashion *even when we do not have estimates of the parameters governing other biases*

[42] In the case of dynamic choices, biases can affect a given decision either directly or indirectly through future choices that shape the consequences of the current choice. Our method is suitable for measuring the direct effects. Specifically, it allows us to gauge the magnitudes of mistakes people make in one-shot decisions, with future consequences fixed experimentally. Such decisions are the building blocks of dynamic decisions, in the sense that dynamic problems require the consumer to solve one-shot problems where the future consequences are fixed at the continuation solutions.

outside the scope of the analysis.[43] Specifically, analogously to Proposition 1 (which concerned the myopic welfare losses $L_M^V\big(r^V(y_c), y_s\big)$ and $L_W^V\big(r^V(y_c), y_s, H\big)\big)$, even though the bias-free utility function U is potentially unknown and unobservable, we can infer the ranking of corrective policies according to the idealized welfare loss, $L_M^U\big(r^U(y_c), y_s\big)$ or $L_W^U\big(r^U(y_c), y_s, H\big)$, by ranking them according to the corresponding index of deliberative competence, $d_e\big(r^V(y_c), r^V(y_s)\big)$, which is measurable. We can even gauge the percentage differences between the dollar equivalents of the benefits flowing from these policies. We can also aggregate over instruments, subject to the same qualifications as in Section IIC. Thus, for our experiment, columns 1 and 2 of Table 7 show that the intervention in experiment A reduces the idealized welfare loss associated with poor comprehension of compound interest by only 6.5 percent (that is, $1 - 22.864/24.448 = 0.065$, standard error $= 0.095$), while the intervention in experiment B reduces it by 27.7 percent (that is, $1 - 18.569/25.673 = 0.277$, standard error $= 0.060$).

Second, $d_e\big(r^V(y_c), r^V(y_s)\big)$ remains a valid approximation of idealized welfare (up to an unknown multiplicative scalar) even when the analyst is unaware that certain types of other biases exist. This robustness property is reassuring.

Third, to the extent our method is robust to certain known and unknown biases (up to a multiplicative scalar), it is also robust to normative ambiguity about such biases, i.e., the possibility that two different utility functions and may both reflect legitimate normative perspectives, as permitted in the welfare framework of Bernheim and Rangel (2009). Hence, we can assess the effect of a policy θ on the welfare loss from mischaracterization within the framework of idealized welfare economics without needing to take a stance on thorny issues such as the normative interpretation of present focus (that is, whether and in what sense it constitutes a bias).

2. *An Example*.—Proposition 2 may initially strike the reader as counterintuitive, because it allows us to rank corrective policies according to an unknown function. The following simple example makes the logic more transparent.

For the purpose of this example, we interpret m as current consumption, y as a vector describing state-contingent consumption in future periods, and I as a financial instrument that enables the consumer to convert the former into the latter. The consumer's beliefs about the instrument are mistaken in two ways. First, due to complex framing, she thinks the instrument promises to deliver $y_c - y_0$ rather than $y_s - y_0$. Second, she entertains the objectively groundless suspicion that the instrument is a scam that will pay nothing, leaving her future outcomes unchanged with probability $1 - \pi$. We also assume that immediate and future consequences, m and y, enter decision utility additively through separate functions, $\nu(\cdot)$ and $\varphi(\cdot)$, respectively, where $\nu(m) = m$. Finally, we assume the consumer discounts future expected utility at the rate $\beta\delta$, where β measures the extent of present bias: $V(m, y) = m + \beta\delta\varphi(y)$, and $U(m, y) = m + \delta\varphi(y)$. Our objective is to evaluate the idealized welfare effects of policies that seek to improve the consumer's understanding of promised

[43] Naturally, if we can measure the parameters governing the other biases, we can compute the factors of proportionality, $K_e(y_s)$, and then estimate the absolute welfare effects of various corrective policies by making simple multiplicative adjustments. Such exercises would still fall within the scope of idealized welfare analysis rather than comprehensive second-best analysis, in that they would use estimates of other biases to "correct" the data before evaluating the policy of interest, without explicitly modeling the ancillary remedies that would achieve those corrections.

consequences (i.e., to reduce the discrepancy between y_s and y_c), treating excessive skepticism and present focus as ancillary biases, of which we may or may not be aware.

In frame f (with the ancillary biases present), the consumer decides whether to purchase the instrument by comparing the perceived utility derived from purchase, $V(m_0 - p, y_f) = m_0 - p + \pi\beta\delta\varphi(y_f) + (1 - \pi)\beta\delta\varphi(y_0)$, with the perceived utility derived from no purchase, $V(m_0, y_0) = m_0 + \beta\delta\varphi(y_0)$. The consumer's reservation price, $r^V(y_f)$, is the value of p that equates these expressions: $r^V(y_f) = \pi\beta\delta\big(\varphi(y_f) - \varphi(y_0)\big)$. Similarly, $r^U(y_f) = \delta\big(\varphi(y_f) - \varphi(y_0)\big)$. In the case where $\varphi(y_s) > \varphi(y_c)$, we then have

$$\begin{aligned} L_M^U\big(r^U(y_c), r^U(y_s)\big) &= U\big(m_0 - r^U(y_c), \varphi(y_s)\big) - U\big(m_0, \varphi(y_0)\big) \\ &= \big[m_0 - r^U(y_c) + \delta\varphi(y_s)\big] - \big[m_0 + \delta\varphi(y_0)\big] \\ &= \delta\big[\varphi(y_s) - \varphi(y_c)\big]. \end{aligned}$$

We also have

$$d_M\big(r^V(y_c), r^V(y_s)\big) = r^V(y_s) - r^V(y_c) = \pi\beta\delta\big[\varphi(y_s) - \varphi(y_c)\big].$$

Putting these observations together yields

$$L_M^U\big(r^U(y_c), y_s\big) = \frac{1}{\pi\beta} d_M\big(r^V(y_c), r^V(y_s)\big).$$

A similar argument applies when $\varphi(y_s) < \varphi(y_c)$. Thus, $K_M(y_s) = \frac{1}{\pi\beta}$. If H is uniform within the relevant interval, we can likewise show that $L_W^U\big(r^U(y_c), y_s, H\big)$ is exactly proportional to $d_W\big(r^V(y_c), r^V(y_s)\big)$.

The implications of this analysis merit emphasis. Suppose an analyst evaluates educational interventions based on d_M or d_W, but is unaware that consumers suffer from either present bias or false beliefs about experimenter reliability. The analysis would still yield valid measures of idealized welfare effects, up to an unknown constant of proportionality. Comparisons of relative benefits across policies would remain valid.

A notable feature of this example is that the relationship between the idealized welfare loss and our measure of deliberative competence is exact. Accordingly, the example shows that our approximations can yield accurate answers even when consequences and misunderstandings are both large.

Within the idealized welfare framework, these interpretations follow under the assumption that issues of present bias and excessive suspicion about experimenter reliability will be fully resolved through other corrective policies. In settings where this assumption is unacceptable, and in which there are no biases other than these two, narrow second-best welfare analysis is more appropriate. According to the logic of that framework, it is typically optimal for the consumer to entertain misunderstandings that counteract other biases. In the current example, the consumer incurs a welfare loss of $L_M^U\big(r^V(y_s), y_s\big) = \big|(\beta\pi - 1)\delta\big(\varphi(y_s) - \varphi(y_0)\big)\big|$ when she

TABLE 9—VALUATIONS IN SIMPLY FRAMED TASKS

Delay in days:	72		36	
Experiment:	A	B	A	B
	(1)	(2)	(3)	(4)
Levels				
Control	68.637	69.698	75.874	75.343
	(2.246)	(1.766)	(2.015)	(1.628)
Treatment	69.513	77.908	75.801	82.791
	(2.265)	(1.707)	(2.076)	(1.451)
Difference	0.876	8.210	−0.073	7.448
	(3.190)	(2.456)	(2.893)	(2.181)
Observations	1,075	1,740	1,075	1,740
Subjects	215	348	215	348

Notes: Valuations in simply framed tasks, $r_s^{i,R,t}$, rescaled to range from 0 to 100 (percentage points). Each column displays the coefficients of a separate OLS regression. Standard errors in parentheses, clustered by subject.

understands the instrument's consequences perfectly. As $\beta\pi$ shrinks, the potential gains from counteracting the associated bias grow, and consequently correcting the consumer's misunderstanding of y_s becomes less beneficial.

E. *Policy-Induced Confounds*

So far, we have assumed that neither V nor U depends directly on the policies we seek to compare (e.g., θ and θ')—in other words, that there are no policy-induced confounds. This assumption is responsible for the invariance of the multiplicative constant, K_e, across policies, which these comparisons exploit. Yet, it is possible that an intervention changes V, causing the consumer's simply framed valuations to vary with the policy. For example, if an educational intervention targeting the mechanics of interest compounding exhorts people to be patient, simply framed valuations for delayed payments might increase.

This concern arises in our experiment. Table 9 presents regressions describing valuations in the simple frame, separately for each time frame. Columns 1 and 3 show that, in experiment A, there are no treatment effects on these valuations. In each condition, subjects value a dollar received with a 72-day delay at roughly 70 cents, and a dollar received with a 32-day delay at roughly 75 cents.[44] In contrast, a treatment effect is present in experiment B. The valuations of subjects in the control condition are similar to those elicited from subjects in experiment A, as columns 2 and 4 show. Average simply framed valuations in the treatment condition, however, are higher by 8.2 and 7.4 cents for the 72-day and 36-day time frames, respectively.

How should an analyst proceed if the data reveal these types of confounds? As we show next, a simple adjustment to our measure of deliberative competence preserves all the implications of Proposition 2.

[44]These magnitudes are typical for studies that elicit time preferences over short horizons (Frederick, Loewenstein, and O'Donoghue 2002). Not only do subjects discount the future heavily, but the discount function decreases much less steeply for longer delays than for short delay. Part of the explanation may involve perceptions of experimenter reliability.

Formally, we allow the indirect utility function V to depend on policies θ directly rather than only through the subjects' misperception of the instrument's consequences, y. That is, we write $V(m, \varphi(y), \theta)$. Because θ influences behavior through valuation biases, we assume it does not directly affect the bias-free utility function, $U(m, \varphi(y))$. Notationally, we include θ as an argument of r^V to accommodate the confounding framing effects.

Our next proposition identifies the adjustment to measured deliberative competence required to restore comparability across policies.

PROPOSITION 3: *There exists a strictly positive function,* $\kappa(y_s)$, *such that for all values of* $y_{c\theta}$ *and* y_s *satisfying* $\left.\frac{dr^u(y_{c\theta}^\alpha)}{d\alpha}\right|_{\alpha=0} \neq \left.\frac{dr^u(y_s^\alpha)}{d\alpha}\right|_{\alpha=0}$, *and all policies* θ, *we have*

$$\lim_{\alpha\to 0}\left[\left(\frac{L_M^U\left(r^U(y_{c\theta}^\alpha), y_s^\alpha\right)}{d_M\left(r^V(y_{c\theta}^\alpha, \theta), r^V(y_s^\alpha, \theta)\right)}\right) - \frac{\alpha\kappa(y_s)}{r^V(y_s^\alpha, \theta)}\right] = 0 \tag{8}$$

and

$$\lim_{\alpha\to 0}\left[\left(\frac{L_W^U\left(r^U(y_{c\theta}^\alpha), y_s^\alpha, H^\alpha\right)}{d_W\left(r^V(y_{c\theta}^\alpha, \theta), r^V(y_s^\alpha, \theta)\right)}\right) - h\left(\frac{\alpha\kappa(y_s)}{r^V(y_s^\alpha, \theta)}\right)^2\right] = 0. \tag{9}$$

Furthermore, when y *is a scalar,* $\kappa(y_s) = (y_s - y_0)\kappa^*$ *for some strictly positive constant* κ^*.

According to this proposition, we can address the confound simply by rescaling our measures of deliberative competence. Specifically, using the proposition to generate approximations for the case of $\alpha = 1$, we conclude that $L_M^U\left(r^U(y_{c\theta}), y_s\right)$ is proportional to $\frac{d_M\left(r^V(y_{c\theta}, \theta), r^V(y_s, \theta)\right)}{r^V(y_s, \theta)}$, and that $L_W^U\left(r^U(y_{c\theta}), y_s, H\right)$ is proportional to $\frac{d_W\left(r^V(y_{c\theta}, \theta), r^V(y_s, \theta)\right)}{\left(r^V(y_s, \theta)\right)^2}$, where the constants of proportionality are independent of y_c, and consequently of the policy θ.

The fact that the constants of proportionality depend on y_s (through $\kappa(y_s)$) introduces a qualification when aggregating over instruments. However, for the case in which y_s is a scalar, the proposition implies that $L_M^U\left(r^U(y_{c\theta}), y_s\right)$ is approximately proportional to $\frac{d_M\left(r^V(y_{c\theta}, \theta), r^V(y_s, \theta)\right)}{r^V(y_s, \theta)/(y_s - y_0)}$, and that $L_W^U\left(r^U(y_{c\theta}), y_s, H\right)$ is approximately proportional to $\frac{d_W\left(r^V(y_{c\theta}, \theta), r^V(y_s, \theta)\right)}{\left(r^V(y_s, \theta)/(y_s - y_0)\right)^2}$, where the constants of proportionality are also independent of y_s (apart from the possibility that y_s is related to h). In other words, to restore all the implications of Proposition 2, we simply rescale the distance metrics by $r^V(y_s, \theta)/(y_s - y_0)$ and $\left(r^V(y_s, \theta)/(y_s - y_0)\right)^2$, respectively.

We use this proposition to re-evaluate the data from our experiment.[45] The correction effectively changes the units of measurement: the differences in Table 7 reflect

[45] Recall that our empirical analysis involves normalized valuations, which equal the raw valuations divided by the future payments. To mitigate noise in the simply framed elicitations, we average over each subject's simply

Table 10—Deliberative Competence Corrected for Policy-Induced Framing Effects

Delay in days:	72 and 36	72 and 36	72	72	36	36
Experiment:	A (1)	B (2)	A (3)	B (4)	A (5)	B (6)
Levels						
Control	−34.639 (2.115)	−37.541 (1.906)	−36.246 (2.535)	−39.326 (2.127)	−33.032 (2.053)	−35.756 (2.001)
Treatment	−37.204 (3.787)	−25.491 (1.775)	−38.363 (3.882)	−26.293 (1.865)	−36.044 (4.202)	−24.689 (1.882)
Substance only	−37.606 (4.710)		−40.610 (6.550)		−34.602 (3.210)	
Rhetoric only	−29.728 (2.427)		−32.242 (2.808)		−27.214 (2.337)	
p-value of difference to control						
Treatment	0.555	0.000	0.648	0.000	0.520	0.000
Substance only	0.566		0.535		0.680	
Rhetoric only	0.128		0.290		0.062	
Observations	4,550	3,470	2,275	1,735	2,275	1,735
Subjects	455	347	455	347	455	347

Notes: Each column displays the coefficients of a separate OLS regression of deliberative competence $d_M^{i,R,t} = -|r_c^{i,R,t} - r_s^{i,R,t}|$, corrected for the change in simply framed valuations, on treatment indicators. One subject in experiment B revealed valuations of zero in each simply framed task. Because the correction entails dividing by the within-subject mean of simply framed valuations, we exclude this subject from the analysis.

subjects' valuations of future prospects in terms of immediate payments, while division by simply framed valuations reflects subjects valuations of future prospects in terms of future payments. Table 10 reproduces the results of Table 7 with the corrected measure of financial competence. Columns 1 and 2 average across both time frames; the remaining columns display estimates separately within each time frame. We continue to observe large and highly statistically significant improvements in deliberative competence in experiment B; the estimated relative improvement is now $1 - 25.491/37.541 = 0.321$ (standard error $= 0.058$). Treatment effects in experiment A remain insubstantial and statistically distinguishable from zero.

F. *Interpersonal Aggregation*

So far, we have focused on evaluating the welfare effects of opportunity-neutral interventions on a single individual. How can an analyst assess the aggregate welfare effect of such interventions on a heterogeneous population? Equipped with a set of welfare weights that reflect the relative social value of an extra dollar received by each individual, one can meaningfully aggregate money-metric welfare losses over members of a population. That standard procedure becomes more challenging in the current context because, unless we adopt the maximal loss criterion and assume away biases outside the scope of the analysis, we measure dollar-equivalent welfare

framed valuations for each time frame and use the resulting mean to correct complexly framed choices by that subject in that time frame. Note that the normalized future payoff, $y_s(I) - y_0$, is 1. Consequently, when we rescale the distance metrics by $r^V(y_s,\theta)$ and $\left(r^V(y_s,\theta)\right)^2$, respectively, we are also rescaling by $r^V(y_s,\theta)/(y_s - y_0)$ and $\left(r^V(y_s,\theta)/(y_s - y_0)\right)^2$, which gives us comparability across instruments (as noted above).

losses up to a constant of proportionality that can potentially vary from one person to the next. To focus on that issue, we will assume for simplicity that the social value of an extra dollar is the same for everyone.

Proposition 2 tells us that, as an approximation, the welfare loss for individual j given perspective $e \in \{W, M\}$ (for some specified instrument and policy), abbreviated $L_e^U(j)$, is approximately equal to measured deliberative competence for that individual, abbreviated $d_e^V(j)$, times a constant $K_e(j)$. It follows that the aggregate (average) welfare loss is $\overline{L}_e^U = \overline{d}_e \overline{K}_e + \text{cov}(d_e, K_e)$, where we use bars to denote means across individuals for each variable.

Our objective is to compare policies according to their average idealized welfare losses, $\overline{L}_e^U$. Assume for the moment that decision-making defects within the scope of the analysis are uncorrelated with defects outside the scope of the analysis—in other words, $\text{cov}(d_e^V, K_e) = 0$. In that case, the invariance of each $K_e(j)$, and hence of $\overline{K}_e$, across corrective policies implies that the change in the aggregate welfare loss ($\overline{L}_e^U$) is a fixed multiple of the change in average deliberative competence ($\overline{d}_e$), which we can measure.

Unfortunately, biases tend to be correlated across domains (Dean and Ortoleva 2019; Stango and Zinman forthcoming; Chapman et al. 2018).[46] The correlation is also unlikely to be stable across policies or to vary in proportion to $\overline{d}_e$. To illustrate, suppose present bias is associated with greater underestimation (less overestimation) of compound interest. A policy that shifts the entire population from underestimation to overestimation will reverse the correlation between $d_e^V(j)$ and the ancillary bias.

One strategy for addressing potentially problematic correlations between the biases of interest and other decision-making defects is to disaggregate into more homogeneous subsamples, and to evaluate whether results are robust with respect to subsample weights. While we obviously cannot disaggregate based on the magnitudes of unobserved biases, we can exploit the fact that such biases tend to impact simply framed valuations. For example, in our experiment, considerations such as present bias and false beliefs about experimenter reliability would suppress valuations of future rewards.

Accordingly, we sort subjects into quartiles based on their mean valuations of simply framed prospects (separately for each treatment group to account for the fact that the treatment intervention in experiment B affected those valuations). For each experiment, we then regress measured deliberative competence on a treatment dummy, an indicator for each quartile of simply framed valuations, and an interaction between those indicators and the treatment dummy. We cluster standard errors at the subject level.

Table 11 displays the results. Within each quartile, the treatment effect is larger in experiment B than in experiment A by several percentage points. For experiment B, the benefit of the intervention is roughly zero for subjects in the bottom quartile, positive for all other quartiles, and statistically significant for the top two quartiles. For experiment A, the benefit is negative and substantial (though not statistically

[46] In the case of $e = W$, the constant of proportionality $K_W(j)$ incorporates the density parameter $h(r^V(y_s))$ for individual j. There is no particular reason to think this term is correlated with either $d_W^V(j)$ or with the other components of $K_W(j)$. To the extent these correlations are zero, heterogeneity in the density term does not impact $\bar{L}_W^U$.

Table 11—Deliberative Competence by Quartiles of Simply Framed Valuations

	Deliberative competence			
	Quartile simply framed valuation			
	Bottom	2nd from bottom	2nd from top	Top
Panel A. Experiment A				
Levels				
Control	−17.007	−21.693	−26.589	−32.216
	(2.037)	(2.596)	(3.410)	(3.893)
Treatment	−24.967	−22.345	−22.855	−21.367
	(4.484)	(2.779)	(2.559)	(3.772)
Effect	−7.961	−0.652	3.734	10.849
	(4.925)	(3.803)	(4.263)	(5.421)
Observations		2,150		
Subjects		215		
Panel B. Experiment B				
Levels				
Control	−17.734	−25.920	−25.706	−33.156
	(1.893)	(2.073)	(2.414)	(3.605)
Treatment	−17.818	−21.239	−17.215	−18.018
	(2.155)	(2.392)	(2.254)	(2.547)
Effect	−0.085	4.681	8.492	15.138
	(2.869)	(3.165)	(3.302)	(4.414)
Observations		3,480		
Subjects		348		

Notes: Effect of treatments on deliberative competence, $d_M^{i,R,t} = -|r_c^{i,R,t} - r_s^{i,R,t}|$, by quartiles of simply framed valuations, rescaled to range from 0 to 100 (percentage points). Each panel presents the output of a single OLS regression. Standard errors in parentheses, clustered by subject.

significant at conventional levels) for those in the lowest quartile, slightly negative for those in the second quartile, and positive for the top two quartiles, but statistically significant only for the top quartile. We conclude that the treatment intervention in experiment B has a (weakly) positive effect on deliberative competence regardless of the welfare weights assigned to different subjects based on a proxy for biases outside the scope of the analysis, whereas the treatment intervention in experiment A has mixed effects, clearly benefiting only those with the highest simply framed valuations. Because the treatment intervention in experiment B has a larger positive effect within each quartile than the intervention in experiment A, it is unlikely that our conclusions concerning the benefits of practice and individualized feedback are driven by a spurious relationship between welfare weights and extraneous biases.

III. Conclusion

We have developed a simple method for evaluating opportunity-neutral policies that seek to improve the quality of decision-making. It involves simple transformations of the price-metric bias, which are interpretable robustly as measures of the welfare loss resulting from poor comprehension of consequences. We illustrated its advantages over conventional evaluative metrics in an application involving financial education.

The method lends itself to many other applications. Maintaining a focus on financial education, one could use it to evaluate training pertaining to loans and leases, rent-to-own agreements, portfolio diversification, contracts with multipart tariffs, and insurance products. One could also evaluate alternatives to education, such as peer advice (as we show in Ambuehl et al. forthcoming), professional guidance, disclosure rules, and informational labels.

Outside the financial domain, our method offers a potential avenue for evaluating a variety of opportunity-neutral policies that target specific biases. As in other recent work in behavioral public economics (reviewed in Bernheim and Taubinsky 2018), the main requirement is that one can contrive versions of valuation tasks that plausibly remove the bias of interest. This requirement is most easily satisfied when the bias pertains to a specific misconception about the consequences of particular actions, and where one can describe the options in terms of their consequences without triggering the misconception. For example, to evaluate the benefits of "traffic light" labels concerning energy efficiency for household appliances, "simple framing" might entail full and correct information concerning energy costs (as in Allcott and Taubinsky 2015). One can imagine similar applications involving nutritional labels and other health advisories. Similarly, using the type of data gathered in Akesson et al. (2022), one could measure the welfare benefits of interventions that aim to help consumers identify fake reviews on online platforms.

Naturally, the approach has limitations, particularly regarding robustness with respect to biases outside the scope of the analysis, which we highlighted in our theoretical discussion. Idealized welfare analysis overcomes the Lipsey-Lancaster critique of "piecemeal" welfare analysis by making a couple of conceptual compromises, which we have stated explicitly. It also involves a separability assumption that may or may not be plausible in the context of any given application. A practical limitation is that implementation requires the analyst to observe an individual's simply and complexly framed valuations for the same instrument. Because naturally occurring processes do not generally produce such data, potential applications are limited to experiments. While the current investigation involves a laboratory experiment with stylized tasks, applications involving field experiments with more naturalistic tasks are also feasible.

REFERENCES

Afriat, Sidney N. 1972. "Efficiency Estimation of Production Functions." *International Economic Review* 13 (3): 568–98.

Agarwal, Sumit, John C. Driscoll, Xavier Gabaix, and David Laibson. 2009. "The Age of Reason: Financial Decisions over the Life Cycle and Implications for Regulation." *Brookings Papers on Economic Activity* 39 (2): 51–101.

Akesson, Jesper, Robert W. Hahn, Robert D. Metcalfe, and Manuel Monti-Nussbaum. 2022. "The Impact of Fake Reviews on Demand and Welfare." Unpublished.

Allcott, Hunt, Sendhil Mullainathan, and Dmitry Taubinsky. 2014. "Energy Policy with Externalities and Internalities." *Journal of Public Economics* 112: 72–88.

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105 (8): 2501–38.

Almenberg, Johan, and Christer Gerdes. 2012. "Exponential Growth Bias and Financial Literacy." *Applied Economics Letters* 19 (16–18): 1693–96.

Ambuehl, Sandro, B. Douglas Bernheim, Fulya Ersoy, and Donna Harris. Forthcoming. “Peer Advice on Financial Decisions: A Case of the Blind Leading the Blind?” *Review of Economics and Statistics*.

Ambuehl, Sandro, B. Douglas Bernheim, and Annamaria Lusardi. 2022. “Replication Data for: Evaluating Deliberative Competence: A Simple Method with an Application to Financial Choice.” American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E171681V1.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutstrom. 2006. “Elicitation using Multiple Price List Formats.” *Experimental Economics* 9 (4): 383–405.

Ariely, Dan, George Loewenstein, and Drazen Prelec. 2003. “‘Coherent Arbitrariness’: Stable Demand Curves without Stable Preferences.” *Quarterly Journal of Economics* 118 (1): 73–106.

Aufenanger, Tobias, Friedemann Richter, and Matthias Wrede. 2016. “Measuring Decision-Making Ability in the Evaluation of Financial Literacy Education Programs.” Unpublished.

Baltussen, Guido, and Gerrit T. Post. 2011. “Irrational Diversification: An Examination of Individual Portfolio Choice.” *Journal of Financial and Quantitative Analysis* 46 (5): 1463–91.

Benartzi, Shlomo, and Richard H. Thaler. 1999. “Risk Aversion or Myopia? Choices in Repeated Gambles and Retirement Investments.” *Management Science* 45 (3): 364–81.

Benartzi, Shlomo, and Richard H. Thaler. 2001. “Naive Diversification Strategies in Defined Contribution Saving Plans.” *American Economic Review* 91 (1): 79–98.

Bennyhoff, Donald G., and Francis M. Kinniry Jr. 2011. “Advisor’s Alpha.” Vanguard. https://web.archive.org/web/20141023045402/https://advisors.vanguard.com/iwe/pdf/ICRAA.pdf.

Bernheim, B. Douglas. 1994. “Personal Saving, Information, and Economic Literacy: New Directions for Public Policy.” In *Tax Policy for Economic Growth in the 1990s*, 53–78. Washington, DC: American Council for Capital Formation.

Bernheim, B. Douglas. 1998. “Financial Illiteracy, Education, and Retirement Saving.” In *Living with Defined Contribution Pensions. Remaking Responsibility for Retirement*, edited by Olivia S. Mitchell and Sylvester J. Schieber, 38–68. Philadelphia: University of Pennsylvania Press.

Bernheim, B. Douglas. 2009. “Behavioral Welfare Economics.” *Journal of the European Economic Association* 7 (2-3): 267–319.

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. 2021. “In Defense of Behavioral Welfare Economics.” *Journal of Economic Methodology* 28 (4): 385–400.

Bernheim, B. Douglas, and Daniel M. Garrett. 2003. “The Effects of Financial Education in the Workplace: Evidence from a Survey of Households.” *Journal of Public Economics* 87 (7–8): 1487–519.

Bernheim, B. Douglas, Daniel M. Garrett, and Dean M. Maki. 2001. “Education and Saving: The Long-Term Effects of High School Financial Curriculum Mandates.” *Journal of Public Economics* 80 (3): 435–65.

Bernheim, B. Douglas, Igor Popov, and Andrey Fradkin. 2015. “The Welfare Economics of Default Options in 401(k) Plans.” *American Economic Review* 105 (9): 2798–2837.

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

Bernheim, B. Douglas, and Dmitry Taubinsky. 2018. “Behavioral Public Economics.” In *Handbook of Behavioral Economics: Applications and Foundations 1*. Vol. 1, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, 381–516. Amsterdam: Elsevier.

Beshears, John, James J. Choi, David Laibson, and Brigitte C. Madrian. 2018. “Behavioral Household Finance.” In *Handbook of Behavioral Economics: Applications and Foundations 1*. Vol. 1, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laison, 177–276. Amsterdam: Elsevier.

Bhargava, Saurabh, George Loewenstein, and Justin Sydnor. 2017. “Choose to Lose: Health Plan Choices from a Menu with Dominated Option.” *Quarterly Journal of Economics* 132 (3): 1319–72.

Blumenthal, Karen, and Kevin G. Volpp. 2010. “Enhancing the Effectiveness of Food Labeling in Restaurants.” *JAMA* 303 (6): 553–54.

Bruhn, Miriam, Luciana de Souza Leão, Arianna Legovini, Rogelio Marchetti, and Bilal Zia. 2016. “The Impact of High School Financial Education: Evidence from a Large-Scale Evaluation in Brazil.” *American Economic Journal: Applied Economics* 8 (4): 256–95.

Bruhn, Miriam, Gabriel Lara Ibarra, and David McKenzie. 2014. “The Minimal Impact of a Large-Scale Financial Education Program in Mexico City.” *Journal of Development Economics* 108: 184–89.

Calvet, Laurent E., John Y. Campbell, and Paolo Sodini. 2007. “Down or Out: Assessing the Welfare Costs of Household Investment Mistakes.” *Journal of Political Economy* 115 (5): 707–47.

Calvet, Laurent E., John Y. Campbell, and Paolo Sodini. 2009. "Measuring the Financial Sophistication of Households." *American Economic Review* 99 (2): 393–98.

Carpena, Fenella, Shawn Cole, Jeremy Shapiro, and Bilal Zia. 2011. "Unpacking the Causal Chain of Financial Literacy." World Bank Policy Research Working Paper 5798.

Carroll, Gabriel. 2019. "Robustness in Mechanism Design and Contracting." *Annual Review of Economics* 11: 139–66.

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer. 2018. "Econographics." NBER Working Paper 24931.

Choi, James J., David Laibson, and Brigitte C. Madrian. 2011. "$100 Bills on the Sidewalk: Suboptimal Investment in 401(k) Plans." *Review of Economics and Statistics* 93 (3): 748–63.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman. 2014. "Who Is (More) Rational?" *American Economic Review* 104 (6): 1518–50.

Consumer Financial Protection Bureau (CFPB). 2013. *Navigating the Market: A Comparison of Spending on Financial Education and Financial Marketing*. Washington, DC: Consumer Financial Protection Bureau.

Cole, Shawn, and Gauri Kartini Shastry. 2010. "Is High School the Right Time to Teach Self-Control? The Effect of Financial Education and Mathematics Courses on Savings Behavior." Unpublished.

Collins, J. Michael. 2013. "The Impacts of Mandatory Financial Education: Evidence from a Randomized Field Study." *Journal of Economic Behavior and Organization* 95: 146–58.

Dean, Mark, and Pietro Ortoleva. 2019. "The Empirical Relationship between Nonstandard Economic Behaviors." *Proceedings of the National Academy of Sciences* 116 (33): 16262–67.

DellaVigna, Stefano, and Ulrike Malmendier. 2004. "Contract Design and Self-Control: Theory and Evidence." *Quarterly Journal of Economics* 119 (2): 353–402.

Downs, Julie S., George Loewenstein, and Jessica Wisdom. 2009. "Strategies for Promoting Healthier Food Choices." *American Economic Review* 99 (2): 159–64.

Echenique, Federico, Sangmok Lee, and Matthew Shum. 2011. "The Money Pump as a Measure of Revealed Preference Violations." *Journal of Political Economy* 119 (6): 1201–23.

Eisenstein, Eric M., and Stephen J. Hoch. 2007. "Intuitive Compounding: Framing, Temporal Perspective, and Expertise." Unpublished.

Enke, Benjamin, Uri Gneezy, Brian Hall, David Martin, Vadim Nelidov, Theo Offerman, and Jeroen van de Ven. Forthcoming. "Cognitive Biases: Mistakes or Missing Stakes?" *Review of Economics and Statistics*.

Enke, Benjamin, and Florian Zimmermann. 2019. "Correlation Neglect in Belief Formation." *Review of Economic Studies* 86 (1): 313–32.

Ernst, Keith, John Farris, and Uriah King. 2004. *Quantifying the Economic Cost of Predatory Payday Lending*. Washington, DC: Center for Responsible Lending.

European Commission. 2020. *A Capital Markets Union for People and Businesses-New Action Plan*. Brussels: European Commission.

Farhi, Emmanuel, and Xavier Gabaix. 2020. "Optimal Taxation with Behavioral Agents." *American Economic Review* 110 (1): 298–336.

Fernandes, Daniel, John G. Lynch Jr., and Richard G. Netemeyer. 2014. "Financial Literacy, Financial Education, and Downstream Financial Behaviors." *Management Science* 60 (8): 1861–83.

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.

Goda, Gopi Shah, Matthew Levy, Colleen Flaherty Manchester, Aaron Sojourner, and Joshua Tasoff. 2019. "Predicting Retirement Savings using Survey Measures of Exponential-Growth Bias and Present Bias." *Economic Inquiry* 57 (3): 1636–58.

Grubb, Michael D., and Matthew Osborne. 2015. "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock." *American Economic Review* 105 (1): 234–71.

Gruber, Jonathan, and Botond Kőszegi. 2001. "Is Addiction 'Rational'? Theory and Evidence." *Quarterly Journal of Economics* 116 (4): 1261–1303.

Gruber, Jonathan, and Botond Kőszegi. 2004. "Tax Incidence When Individuals Are Time-Inconsistent: The Case of Cigarette Excise Taxes." *Journal of Public Economics* 88 (9–10): 1959–87.

Hara, Kotaro, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. "A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk." In *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. New York: Association for Computing Machinery.

Harberger, Arnold C. 1964. "The Measurement of Waste." *American Economic Review* 54 (3): 58–76.

Harrison, Glenn W., Morten Igel Lau, E. Elisabet Rutström, and Melonie B. Sullivan. 2005. "Eliciting Risk and Time Preferences Using Field Experiments: Some Methodological Issues." *Field Experiments in Economics* 10: 125–218.

Harrison, Glenn, Karlijn Morsink, and Mark Schneider. 2020. "Do No Harm? The Welfare Consequences of Behavioral Interventions." Unpublished.

Hastings, Justine S., Brigitte C. Madrian, and William L. Skimmyhorn. 2013. "Financial Literacy, Financial Education, and Economic Outcomes." *Annual Review of Economics* 5 (1): 347–73.

Heinberg, Aileen, Angela Hung, Arie Kapteyn, Annamaria Lusardi, Anya Savikhin Samek, and Joanne Yoong. 2014. "Five Steps to Planning Success: Experimental Evidence from US Households." *Oxford Review of Economic Policy* 30 (4): 697–724.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92 (5): 1644–55.

Kaiser, Tim, Annamaria Lusardi, Lukas Menkhoff, and Carly Urban. 2022. "Financial Education Affects Financial Knowledge and Downstream Behaviors." *Journal of Financial Economics* 145 (2): 255–72.

Levy, Matthew, and Joshua Tasoff. 2016. "Exponential Growth Bias and Lifecycle Consumption." *Journal of the European Economic Association* 14 (3): 545–83.

Levy, Matthew R., and Joshua Tasoff. 2017. "Exponential-Growth Bias and Overconfidence." *Journal of Economic Psychology* 58: 1–14.

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

Loewenstein, George, Joelle Y. Friedman, Barbara McGill, Sarah Ahmad, Suzanne Linck, Stacey Sinkula, John Beshears, et al. 2013. "Consumers' Misunderstanding of Health Insurance." *Journal of Health Economics* 32 (5): 850–62.

Lusardi, Annamaria. 2011. "Americans' Financial Capability." NBER Working Paper 17103.

Lusardi, Annamaria, and Olivia S. Mitchell. 2011. "Financial Literacy and Planning: Implications for Retirement Well-being." In *Financial Literacy. Implications for Retirement Security and the Financial Marketplace*, edited by Annamaria Lusardi and Olivia S. Mitchell, 17–39. Oxford: Oxford University Press.

Lusardi, Annamaria, and Olivia S. Mitchell. 2014. "The Economic Importance of Financial Literacy: Theory and Evidence." *Journal of Economic Literature* 52 (1): 5–44.

Lusardi, Annamaria, and Olivia S. Mitchell. 2017. "How Ordinary Consumers Make Complex Economic Decisions: Financial Literacy and Retirement Readiness." *Quarterly Journal of Finance* 7 (03): 1–31.

Lusardi, Annamaria, Anya Samek, Arie Kapteyn, Lewis Glinert, Angela Hung, and Aileen Heinberg. 2017. "Visual Tools and Narratives: New Ways to Improve Financial Literacy." *Journal of Pension Economics and Finance* 16 (3): 297–323.

Machina, Mark J., and Marciano Siniscalchi. 2014. "Ambiguity and Ambiguity Aversion." In *Handbook of the Economics of Risk and Uncertainty*. Vol. 1, edited by Mark J. Machina and W. Kip Viscusi, 729–807. Amsterdam: Elsevier.

Madrian, Brigitte C., and Dennis F. Shea. 2001. "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior." *Quarterly Journal of Economics* 116 (4): 1149–87.

Malkiel, Burton Gordon. 1985. *A Random Walk Down Wall Street: Including a Life-Cycle Guide to Personal Investing.* New York: W.W. Norton and Company.

Malkiel, Burton G., and Charles D. Ellis. 2013. *The Elements of Investing: Easy Lessons for Every Investor.* New Jersey: Wiley.

Mandell, Lewis. 2009. *The Financial Literacy of Young American Adults: Results of the 2008 National Jump$tart Coalition Survey of High School Seniors and College Students*. Washington, DC: Jump$tart Coalition.

Mandell, Lewis, and Linda Schmid Klein. 2009. "The Impact of Financial Literacy Education on Subsequent Financial Behavior." *Journal of Financial Counseling and Planning* 20 (1): 15–24.

Mason, Winter, and Siddarth Suri. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods* 44 (1): 1–23.

Meade, James Edward. 1955. *The Theory of Customs Unions.* Amsterdam: North-Holland Publishing.

O'Donoghue, Ted, and Matthew Rabin. 2006. "Optimal Sin Taxes." *Journal of Public Economics* 90 (10–11): 1825–49.

Organisation for Economic Co-operation and Development (OECD). 2015. *National Strategies for Financial Education*. Paris: Organisation for Economic Co-operation and Development.

Organisation for Economic Co-operation and Development (OECD). 2021. "Launch of the EC/OECD-INFE Project to Develop a Financial Competence Framework for the EU." Panel, 26 April 2021. https://web.archive.org/web/20220128161019/https://www.oecd.org/financial/education/launch-eu-financial-competence-framework.htm.

Oster, Emily. 2020. "Health Recommendations and Selection in Health Behaviors." *American Economic Review: Insights* 2 (2): 143–60.

Skimmyhorn, William. 2016. "Assessing Financial Education: Evidence from Boot Camp." *American Economic Journal: Economic Policy* 8 (2): 322–43.

Stango, Victor, and Jonathan Zinman. 2009. "Exponential Growth Bias and Household Finance." *Journal of Finance* 64 (6): 2807–49.

Stango, Victor, and Jonathan Zinman. Forthcoming. "We Are All Behavioral, More or Less: Measuring and Using Consumer-Level Behavioral Sufficient Statistics." *Review of Economic Studies*.

Sunstein, Cass R. 2020. "Maximin." *Yale Journal on Regulation* 37: 940

Sydnor, Justin. 2010. "(Over) Insuring Modest Risks." *American Economic Journal: Applied Economics* 2 (4): 177–99.

Taubinsky, Dmitry, and Alex Rees-Jones. 2018. "Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment." *Review of Economic Studies* 85 (4): 2462–96.

Thaler, Richard H., and Cass R. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* New Haven: Yale University Press.

van Rooij, Maarten, Annamaria Lusardi, and Rob Alessie. 2011. "Financial Literacy and Stock Market Participation." *Journal of Financial Economics* 101 (2): 449–72.

Wagenaar, William M., and Sabato D. Sagaria. 1975. "Misperception of Exponential Growth." *Perception and Psychophysics* 18 (6): 416–22.

Walstad, William B., Ken Rebeck, and Richard A. MacDonald. 2010. "The Effects of Financial Education on the Financial Knowledge of High School Students." *Journal of Consumer Affairs* 44 (2): 336–57.