Journal of Banking and Finance 152 (2023) 106851

Contents lists available at ScienceDirect

Journal of Banking and Finance

journal homepage: www.elsevier.com/locate/jbf

Experimental Research on Retirement Decision-Making: Evidence from Replications[☆]

Kremena Bachmann[a,*], Andre Lot[b], Xiaogeng Xu[c], Thorsten Hens[d]

[a] University of Zurich and Zurich University of Applied Sciences, Switzerland
[b] Norwegian School of Economics, Norway
[c] Hanken School of Economics and Helsinki GSE, Finland
[d] University of Zurich, University of Lucerne, Switzerland, and Norwegian School of Economics, Norway

ARTICLE INFO

Article history:
Received 9 September 2022
Accepted 5 April 2023
Available online 17 April 2023

JEL classification:
C91
D15
G51
J26

Keywords:
Household finance
Retirement decision
Savings
Annuities
Life-cycle optimization
Income smoothing
Experiments
Replications

ABSTRACT

We adapt the design of four experimental studies on retirement decision-making and conduct replications with a larger online sample from the broader population. We replicate most of the main effects of the original studies. In particular, we confirm that consumption decisions are less efficient when subjects need to borrow from the future than when they need to save from the present. When subjects collect retirement benefits as lump sum instead of annuities, they choose to retire later, as suggested by the original study. We also confirm that savings are higher when they are incentivized with matching contributions than when incentivized with tax rebates. However, when faced with varying survival risks, subjects in our replication make only partial adjustments to spending paths when ambiguity is reduced. We also propose a further experimental research agenda in related topics and discuss practical issues on subject recruitment, attrition, and redesign of complex tasks.

1. Introduction

Retirement financial decisions over the life cycle exhibit puzzling patterns in the field, such as subjects not converting savings into life annuities, saving too little before retirement, or spending their savings too slowly after retirement (Lugilde et al., 2019; Peijnenburg et al., 2016; Feigenbaum et al., 2013; Heimer et al., 2019). Some of these patterns may be related to the nature of the decision problem. Financial decisions over the life cycle are complex and require high cognitive skills and financial knowledge. The long spans between decisions and observable outcomes, as well as a low decision frequency, limit the ability to learn from the own experience. Normative institutional settings and strong social norms around these decisions impose further challenges for researchers seeking to identify the underlying drivers of observed behaviour.

Experimental studies on retirement decisions have addressed the empirical challenges associated with these decisions in the field. However, many of these studies have relied on student samples that do not vary with respect to characteristics that can be related to the studied treatment effects. In general, such homogeneity with respect to those characteristics may hinder conclusions on whether the observed causal relationship has external validity.

In the context of retirement decision-making, using student samples can be problematic since students are more likely than individuals from the general population to use hyperbolic discounting (Carbone, 2006), which can cause differences in the behaviour when dealing with life cycle optimization problems. Higher cognitive abilities within student samples could also conceal the limitations faced by the representative agent in the population making retirement decisions motivated by myopic planning (Ballinger et al., 2011). Students' lack of experience with long-term

[☆] We are grateful for the comments and suggestions of three anonymous referees, the special issue editors (Michael Kirchler and Utz Weitzel), Christian König gen. Kersting, Thomas de Haan, and Alexander Ziegler; and the audiences at Financial Economics Colloquium at University of Zurich (2021), Experimental Finance conference in Bonn (2022), and FAIR Midway conference in Tromsø (2022). Financial support from the Swiss National Science Foundation [grant number 100018_189107] is greatly acknowledged.

[*] Corresponding author.
E-mail address: kremena.bachmann@zhaw.ch (K. Bachmann).

https://doi.org/10.1016/j.jbankfin.2023.106851

debt management could also plausibly explain certain suboptimal life-cycle optimization results observed in student samples, such as those in Meissner (2016). Such individual-level characteristics of different samples can potentially moderate the treatment effects in studies on retirement decision-making, which calls the generalisability of the reported causal effects into question.

This paper aims to evaluate the external validity of some main findings in the experimental literature on retirement decision-making by using online samples from the general population. We selected four experimental studies addressing different aspects of the retirement decision-making problem, in which the observed effects can potentially depend on individual characteristics such as experience with specific decisions and general ability to deal with complex decision problems. By using samples from the general population that differ more with respect to these characteristics than the samples used in the original studies, we evaluate which findings of the original studies can be replicated. All our replication studies have been preregistered.

We successfully replicated most of the main effects of the selected experimental studies. In particular, we found that although subjects behaved less optimally than the subjects in the original study, their decisions were still better when they needed to save for the future than when they needed to borrow from the future (as in Meissner, 2016). In addition, we find that the impact of this debt aversion remained after considering individual differences in patience and risk preferences. In the face of survival risk, the subjects in our replication were also more likely to delay the timing of retirement when collecting benefits as lump sum than as annuities (like those in Fatas et al., 2007). In addition, we find that these timing decisions were affected by the survival risk that subjects experienced in previous rounds of the experiment. When incentives to save were offered as matching contributions rather than tax rebates, the effective savings rates were higher (in line with the observations of Blaufus and Milde, 2021). Finally, when facing varying survival risk, subjects in our sample adjusted their spending (as observed by Anderhub et al., 2000). However, in our sample, the response of spending to changes in the resolution of ambiguity of varying survival risk was insufficient and weaker than in the original study.

In addition to testing the replicability of the original studies, we document evidence of substantial suboptimal decision-making behaviour. Subjects consistently under-consumed their lifetime income, or consistently did not save enough, going bankrupt when needing to fund mandatory expenses. Such inefficiencies remain hidden in experiments with enforced lifetime budgets.

At last, we present and discuss some important methodological challenges and practical issues concerning the modification of original tasks, the implementation of such experiments with online panels from the general population, and the efficiency of decision-making within the tasks. We then propose a further experimental research agenda on relevant topics and themes to address lingering questions arising from the current state of the empirical field and experimental literature.

By replicating the main effects of several experimental studies on retirement decision-making using larger and more heterogeneous samples than the original studies, our paper mainly contributes to the discussion of whether the experimental findings on this topic are externally valid. Although student samples can be generally criticized as they are on average more homogeneous than non-student samples (Peterson, 2001) and show different personal and attitudinal characteristics (Hanel and Vione, 2016), research has been sensitive enough to note that the usefulness of a sample should be judged upon having variance on relevant moderators (Druckman and Kam, 2011). The usefulness of student samples has been studied in various areas of research. In political science research, Krupnikov and Levine (2014) found that both student and diverse national adult samples behave consistently and in line with theoretical predictions once relevant moderators are taken into account. In economics, Horton et al. (2011) found that the main effects of common experiments in economics (such as prisoner's dilemma, priming, and framing effect in risk-taking) also hold true among Amazon Mechanical Turk workers. In retirement decision-making, Carbone (2005) found that differences in demographic characteristics do not affect the strategies used to solve the life-cycle optimization problem. However, Carbone (2006) found that people from the general population have a shorter planning horizon than students, and students are more likely to discount hyperbolically. Our study contributes to this discussion by showing that such differences between different samples have only a limited impact on the main effects of experimental research studying the behaviour in the context of financial retirement decision-making.

Our results also support the point of view that the complexity of financial retirement decisions per se could be an obstacle to efficient decision-making. Previous research has shown that the complexity of decisions can motivate myopic planning (Ballinger et al., 2011) or the use of heuristics, which could potentially lead to suboptimal decisions. With respect to the implications of heuristics, Winter et al. (2012) found that the outcome of such heuristics does not need to be different from the outcomes of the underlying life-cycle dynamic optimization problems. Our research contributes to this discussion by showing that the complexity of the decision problem may lead to suboptimal behaviour, as it can potentially motivate decisions that are not sensitive enough to changes in the characteristics of the decision problem.

Finally, our findings have implications for policymakers who consider pension reforms that allow more discretion in retirement decisions or relax compulsory mandates. In this context, policymakers often assume that individuals would make retirement financial decisions in line with their individual preferences and economic constraints. Our results suggest that the financial retirement planning might be too complex for individuals to respond optimally to changes in the decision environment, and the suboptimal decision behaviour may impose a restriction on the efficacy of policy reforms.

In Section 2, we present an overview of the relevant experimental literature. Then, in Section 3, we introduce the original studies and present the results of our replications. We discuss the implications of our results and propose a future research agenda on this topic in Section 4, and conclude in Section 5. Additional experimental materials, original data and analysis code are available in the Online Repository.

2. Experimental literature on retirement decision-making

Experiments on individual retirement decision-making have investigated the importance of its various driving factors by employing different task designs. In the first subsection, we present an overview of the literature, along with the factors that previous studies considered as potential drivers for the observed decision-making behaviour. In the second subsection, we then discuss in more detail the most common experimental task features that distinguish experiments in this domain. Table 1 summarizes the studies in terms of their main findings and distinguishing features with respect to the experimental design.

2.1. Drivers of retirement decision-making behaviour

One strand of experimental studies investigates how specific features of the decision problem affect people's decision behaviour. Carbone and Hey (2004) investigated how people adjust their consumption behaviour to the possibility of unemployment, and

Table 1
An Overview of Experimental Studies on Retirement Decision-Making.

Study	Focus	Main Dependant Variable(s)	Main Finding(s)	N	MP	LR	InU	ME	IoS	ELB	SWRP	IU	Sample
Agnew et al. (2015)	behavioural biases	demand for annuity	Past market performance influences the demand for annuities.	1093	X	X		X					general population (lab)
Anderhub et al. (2000)	**decision problem feature**	**consumption**	**Observed consumption paths are qualitatively correct with respect to optimal ones.**	**100**	**X**	**X**						**X**	**students (lab)**
Ballinger et al. (2011)	heterogeneity	decision performance	Cognitive abilities predict performance.	192	X		X					X	students (lab)
Ballinger et al. (2003)	learning	consumption	Later generations perform better than earlier generations.	36	X		X					X	students (lab)
Beshears et al. (2020)	institutional features	endowment allocation to commitment account	Higher early-withdrawal penalties attract more commitment account deposits.	1045									Rand American Life Panel
Blaufus and Milde (2021)	**behavioural biases (framing)**	**savings rate**	**Matching contributions attract higher savings than deferred or immediate taxation regimes.**	**306**	**X**				**X**	**X**	**X**		**students (lab)**
Bohr et al. (2019)	institutional features	optimal consumption	Mandatory (vs. voluntary) savings improves total lifetime consumption.	45	X				X	X	X	X	students (lab)
Brown et al. (2009)	learning	deviation from optimal consumption	Subjects save too little at first, but learn to save optimally over repeated life-cycles.	72	X		X			X		X	students (lab)
Brown et al. (2008)	biases	choice of life annuity	Individuals prefer an annuity over alternative products when presented in a consumption frame; non-annuitized products are preferred when presented in an investment frame.	1342									Internet survey (participants age > 50)
Carbone (2005)	heterogeneity	consumption	There is only a minor link between the strategies employed by the subjects and their demographic characteristics.	495	X		X		X				CentER family expenditure panel
Carbone (2006)	behavioural biases	consumption	Discounting model gives the best explanation, but subjects are myopic.	594	X		X		X			X	CentER panel and students
Carbone and Duffy (2014)	learning	consumption	Provision of social information on past average levels of consumption results in a greater deviation of consumption from optimal paths.	60	X				X			X	students (lab)
Carbone and Hey (2004)	decision problem feature	deviation from optimal consumption	Over-sensitivity of consumption to income changes due to unemployment.	96	X				X			X	students (lab)

(*continued on next page*)

Table 1 (*continued*)

Study	Focus	Main Dependant Variable(s)	Main Finding(s)	N	MP	LR	InU	ME	IoS	ELB	SWRP	IU	Sample
Carbone and Infante (2014)	decision problem feature	consumption-to-wealth	Ambiguity (vs. risk) triggers savings	30	X		X		X			X	students (lab)
Duffy and Li (2019)	institutional features	optimal consumption	100% pension replacement rate yields the highest experimental payoff.	119	X				X		X	X	students (lab)
Fatas et al. (2007)	**institutional features**	**choice of retirement period**	**Subjects retire later with lump sum payoffs instead of annuities or combination thereof.**	**82**		**X**							**students (lab)**
Feltovich and Ejebu (2014)	learning	optimal saving	Inter-personal comparisons (by assigning subjects to groups and displaying rankings based partly on consumption) increases under-saving and leads to lower money earnings.	170	X		X				X	X	students (lab)
Gechert and Siebert (2020)	behavioural biases	savings	Participants on average form and maintain a stock of wealth although not optimal.	180	X		X						students (lab)
Hey and Dardanoni (1988)	decision problem feature	consumption	Actual behaviour differs significantly from optimal behaviour; the comparative static implications of actual behaviour appear to be optimal	128	X	X			X			X	students (lab)
Hurwitz et al. (2020)	institutional features	division of savings between annuity and lump sum	Providing a mandatory minimum annuity rule creates an anchoring effect that reduces annuitization.	277				X					students (lab)
Koehler et al. (2015)	**decision problem feature**	**accumulated savings at start of retirement**	**Most subjects save enough, and longer retirement attracts higher savings.**	**149**	**X**			**X**		**X**	**X**		**M-Turk**
Levy and Tasoff (2020)	biases	overconsumption	Observed behaviour consistent with behaviour predicted by exponential growth bias.	399	X				X			X	students (lab)
Meissner (2016)	**decision problem feature**	**deviation from consumption smoothing**	**Consumption smoothing is worse when subjects need to borrow from the future than save from the present.**	**76**	**X**		**X**			**X**		**X**	**students (lab)**
Meissner and Rostam-Afschar (2017)	learning	consumption / save (borrow)	Some subjects learn to comply with Ricardian Equivalence.	176	X		X			X			students (lab)

Notes: N is the number of observations. *MP* is whether there are Multiple Period decisions per Life. *LR* is Longevity Risk. *InU* is Income Uncertainty. *ME* is Mandatory Expenses. *IoS* is Interest on Savings. *ELB* is Enforced Lifetime Budget. *SWRP* is Separate Work and Retirement Phase. *IU* is Induced Utility. *Sample* includes the population where the sample is drawn and the platform (lab or online). The rows in bold font are the five studies that are included in this replication study.

found that people overreact to the risk of unemployment. In a study, in which they varied the length of the retirement phase, Koehler et al. (2015) found that most participants responded sensibly by saving more of their current income when faced with a long compared to a short retirement phase. Meissner (2016) studied optimal consumption on an increasing and decreasing income path and found that when people are required to borrow to smooth consumption (i.e., when their income path is increasing), deviations from optimal behaviour are more likely. Anderhub et al. (2000) relaxed the assumption in most experiments that the survival probabilities are constant and found that the average subject reacts in a qualitatively correct way to “good” and “bad” news concerning survival risk. While most studies have considered decisions under income distribution risk, Carbone and Infante (2014) studied decision-making under risk and ambiguity and found that behaviour under ambiguity is characterized by a significant pattern of under-consumption compared to behaviour under risk. In terms of the quality of the general decision behaviour of the subjects, Hey and Dardanoni (1988) found that the subjects respond optimally to changes in discount factors and the return on savings.

The retirement decision problem has features that can also be determined by the institutional environment. Bohr et al. (2019) studied the introduction of automatic savings schemes and found that individuals save less with such schemes, but the reduction is only partial in that the total lifetime consumption measures are higher. Duffy and Li (2019) considered different pension replacement rates and found that subjects achieve the highest experimental payoff when offered a constant life-cycle endowment profile (100% pension replacement rate). Hurwitz et al. (2020) investigated the benefits of implementing a minimum annuity rule and found that this does not guarantee an increase in the demand for annuities, and may even reduce it. Beshears et al. (2020) evaluated the benefits of introducing higher withdrawal penalties in retirement savings schemes and found that higher early withdrawal penalties attract more commitment account deposits. Fatas et al. (2007) examined whether the pension benefits scheme (lump sum payments or annuities) affects retirement decisions in the face of longevity risk and found that concentrating payments (shifting from annuity to lump sum) can motivate subjects to postpone retirement.

However, the complexity of this decision problem also raises the question of whether people learn to deal with the problem from experience or from the choices of others. Brown et al. (2009) found that subjects save too little at first, but learn to save close to optimal amounts after three or four rounds (of one simulated life-cycle each). Meissner and Rostam-Afschar (2017) found that people learn to operate under a Ricardian tax scheme (a tax cut in early periods of the experiment, followed by a tax increase of the same magnitude in later periods), but the aggregate effect of taxation on consumption persists even after eight rounds. Because the subjects in the field made decisions for only one life, important insights can arise from social learning. Carbone and Duffy (2014) found that the provision of social information on past average levels of consumption results in a greater deviation of consumption from optimal paths. Similarly, Feltovich and Ejebu (2014) allowed for interpersonal comparison and found that providing this information leads to worse outcomes in the form of more under-saving and lower money earnings. In contrast, Ballinger et al. (2003) analysed learning effects using an intergenerational structure and found that subsequent generations perform significantly better in terms of savings than previous generations.

Few studies have analysed the effect of specific behavioural biases on retirement financial decisions. Levy and Tasoff (2020) found that the subjects’ decision behaviour is affected by the exponential growth bias. Agnew et al. (2008) found that an excessive extrapolation of the past performance of the financial market influences the demand for annuities. Blaufus and Milde (2021) found that different frames of tax-related pension incentives can influence retirement savings, while Brown et al. (2008) also found that the use of different frames can affect the demand for annuities. Several experiments have reported evidence that subjects behave myopically (Carbone and Hey, 2004; Ballinger et al., 2003; Carbone, 2005; 2006) and have dynamically inconsistent preferences (Brown et al., 2009). In terms of general decision-making behaviour, Carbone (2005) found that subjects apply common rules of thumb to solve the optimization problem. Subjects also exhibit preferences for building wealth, even if it is not optimal to do so (Gechert and Siebert, 2020).

Finally, some studies have aimed to explain the heterogeneity in behaviour based on dynamic decision-making tasks. Ballinger et al. (2011) found that cognitive abilities (but not personality measures) are good predictors of heterogeneity in saving behaviour observed as a result of using shorter than optimal planning horizons. Carbone (2005) concluded that demographic characteristics have minor effects on the planning horizon of the subjects and on the strategies applied to solve the optimization problem. Carbone (2006) found that hyperbolic discounting affects the behaviour of students more strongly than that of the general population, which cannot be explained solely by age differences, as younger people are generally considered to be more hyperbolic discounters.

2.2. Design features of the experiments

Most experimental studies on retirement decision-making require sequential decisions over several periods of simulated life (a round). The number of periods can be either fixed or determined by some random process. There is an implicit longevity risk when the number of periods is not fixed, which brings interesting complications into the optimization problem facing the subjects (Agnew et al., 2015; Anderhub et al., 2000; Fatas et al., 2007; Hey and Dardanoni, 1988).

Another source of uncertainty in the optimization problem that can be introduced is stochastic income. This type of uncertainty can be used in different ways. It can be linked to the probability of becoming unemployed or later re-employed (Carbone and Hey, 2004). It can also be represented by a simple i.i.d. process (Ballinger et al., 2003) or by a fluctuating stream of either high or low income (Feltovich and Ejebu, 2014; Carbone, 2005; Carbone and Infante, 2014; Meissner and Rostam-Afschar, 2017). Alternatively, it can be implemented by adding or subtracting a constant error term from an otherwise linear income process (Meissner, 2016). Introducing an uncertain income as an experimental feature is certainly realistic. However, when analysing deviations from optimal consumption paths, it can be difficult to distinguish between deviations caused by a misperception of probabilities and deviations caused by the general cognitive difficulty of finding the optimal solution. For this reason, some studies have used deterministic income paths (e.g., Duffy and Li, 2019).

In some experiments, subjects are required to cover some mandatory expenses during the simulated life-cycle in order to incentivize savings (Hurwitz et al., 2020; Koehler et al., 2015; Agnew et al., 2015). This feature can also determine their survival in experiments.

In approximately half of the studies reviewed, savings were incentivized through an interest-bearing savings account. While offering interest increases the attractiveness of saving versus immediate consumption, this can increase the computational burden to participants and lead to suboptimal decisions.

Some studies have introduced a retirement phase as part of the inter-temporal optimization problem (Blaufus and Milde, 2021; Bohr et al., 2019; Duffy and Li, 2019; Feltovich and Ejebu, 2014; Koehler et al., 2015). In the retirement phase, there is no uncertainty about exogenous income, which is set to zero, meaning that subjects will only be able to consume and/or pay expenses in the retirement phase from their savings that they accumulate during the working phase. The solution to inter-temporal optimization problems with and without such a retirement phase may differ depending on whether subjects misinterpret the probabilities concerned, for instance by overreacting to events occurring with certainty (periods with zero income) as compared to events occurring with very high/low probability (periods with unemployment or income shock risk).

Only a few studies have enforced a lifetime budget, whereby any wealth left at the last period is automatically spent (Blaufus and Milde, 2021; Bohr et al., 2019; Brown et al., 2009; Koehler et al., 2015; Meissner, 2016; Meissner and Rostam-Afschar, 2017). This feature simplifies the analysis of experimental decisions and facilitates calibration of several theoretical models underpinning the experimental designs, but it may potentially obfuscate instances of suboptimal behaviour or misunderstanding of the experimental tasks.

Finally, to motivate subjects to optimize their consumption paths, most studies have linked subjects' consumption choices to their payoffs. Some studies have specified the link by using a particular (induced) utility function. When there is no interest earned on savings, and payoffs are based on lifetime outcomes, inducing a utility function is essential. Otherwise, subjects might just assign most of their lifetime consumption to one or some of the periods, then consume little (or save just enough for expenses, if applicable), as many possible combinations of period consumption would yield the same lifetime outcome. Experiments without an induced utility can motivate consumption smoothing by linking compensation to choices in one random period. This latter task design is much simpler for subjects to understand, although it carries the small drawback of allowing risk-seeking subjects to gamble by concentrating most consumption in just one period in the hope that this period is eventually selected for payoff.

3. Replications of adapted experimental designs

Taking into account the existing body of previous experimental studies on retirement decision-making (see Table 1), we selected four experiments that spanned a heterogeneous set of research topics and experimental design features. In terms of research topics, we selected two studies investigating the impact of different decision characteristics, such as ambiguous survival probabilities (Anderhub et al., 2000), and different income paths (Meissner, 2016) on the consumption behaviour over time. Dealing with these characteristics of the decision problem requires a certain level of cognitive abilities and experiences, such as experience with debt management. The variability with respect to these characteristics is usually low in traditional student samples. The third study evaluated the relevance of institutional features related to the design of retirement benefits on the decision when to retire in the presence of survival risk (Fatas et al., 2007). The decision problem requires dealing with probabilities of survival, which could be a cognitively demanding task, with important implications for policymakers designing the form of retirement benefits. The fourth study addressed the relevance of behavioural effects, and specifically framing effects, on the decision of how much to save for retirement (Blaufus and Milde, 2021). Depending on the task, older people might not be subject to framing effects as observed by Pu et al. (2017).

The selected experiments also differ with respect to the task features summarized in Table 1. We consider diversity in the task features as a selection criterion because these features might cause different levels of inefficiencies in decisions between the original samples and our replications. These features also correspond to the many flavours of life-cycle models (for an overview, see Browning and Crossley, 2001), and could not be plausibly investigated in a single experimental study that simultaneously considers all these decision features using a single parsimonious model.

Finally, our selection of the studies was motivated by the technical feasibility (or lack thereof) of certain experimental designs features using online unassisted samples. Under this consideration, some experimental designs, such as the design used by Brown et al. (2009), could not be deployed.

Subjects for all replications were recruited from the Germany recruitment pool of the market research company Bilendi. Since this pool of subjects is not very well known among experimental researchers, we also replicated the study of Koehler et al. (2015), which uses a simplified retirement decision-making task with Amazon Mechanical Turk workers to evaluate how income availability over the life-cycle affects consumption behaviour. The main goal of this replication is to see whether our online pool of subjects from the general population can manage such experimental tasks, and whether they respond to financial incentives, which we introduced in addition and which were part of the other replications.

Replications and, in some cases, additional analyses of individual experiments were pre-registered on AsPredicted.[1] Each of these studies addresses a different research question; hence, we do not propose any joint analysis of individual replication results with respect to their original hypotheses. While discussing some replications, in light of the results we found, we offer some additional non-preregistered analyses that are clearly noted as such.

In the replications, we focused on one or two main effects of each study. We intended to replicate the studies using subjects from the general population, who would perform the tasks online without any assistance from experimenters at hand. For this purpose, we modified the original experimental designs and adjusted their tasks as needed. We drew the subjects from the same large pool, and used the same deployment method, quality control mechanism, and common design and interface features in all replications to avoid differences in the results between the studies arising from such differences in the implementation.

In addition to replicating the main effects of the original studies, in the Appendix we present (non-preregistered) analyses of the main effects broken down by subsamples based on sociodemographic characteristics of the subjects (age, gender, income, education, and financial training). The main effects are not always statistically significant in all subsamples, and the significance of the main effects across the subsamples differs between the studies. However, across all studies, the main effects held true in the three subsamples: the subjects who are older than 50 years, those who do not have higher education, and those who had not participated in any financial training. The subsamples with these characteristics clearly do not overlap with the student subsamples used in the original studies.

In the following subsections, we first discuss the approach and procedures we used to modify and adjust the experimental designs and their tasks, and the general engagement and performance metrics of subject participation. We then discuss the specific replication results for each study. For parsimony, we will skip most or all of the discussions of the models and hypotheses used and developed by the authors of the original studies and refer inter-

[1] See Pre-registration (1), (2), (3), (4) and (5).

ested readers to the respective original published research papers instead.

3.1. Redesign and adaptation of experimental tasks

The original experimental sessions included extensive subject education and training. In addition, some experiments had a very complex set of instructions, including direct mathematical formulae presented to subjects to explain the induced utility and complex payoff mechanisms. These features of the original studies would make any attempt to closely replicate all the original experiments unfeasible. To address this challenge, while aiming to preserve the main mechanisms we wanted to replicate, we modified and redesigned the experimental tasks to varying degrees.

In three experiments, we reduced the number of rounds and/or periods per round, preserving the structure of lifetime budget constraints and the relative scale of income paths, expenses, and other environmental variables where applicable. There is a long-standing concern in the literature about the elicitation of decision-making sets for subjects that need to engage in dynamic programming and the minimum necessary number of periods over which optimization is to be done. However, we believe that a partial reduction in the length of each round, or the number of rounds, is not as much of an issue in our replications as it would have been in experiments that rely on stochastic environmental variables that persist over many periods (such as in the first task of Brown et al., 2009).[2]

Three experiments originally used numerous sequential computer screens for feedback on results, reassurance of procedures, and indirect attention checks. Compounded over dozens of periods and several rounds per subject, this approach greatly lengthens the total session time. In our replications, we streamlined the interface so that the information and decision screens and action buttons for each round (i.e., one experimental life) could fit on one screen.

We used dynamic tables, one per round, that were progressively filled with each period's decision and populated from the beginning with information on constant or predetermined environmental variables (such as a predetermined income path).[3] Where not obvious, we implemented hovering text balloons that quickly expanded the concept of variables at the top of the dynamic tables.

For input on consumption and savings decisions in all relevant experiments, we used sliders (automatically adjusted to the boundaries of budget constraints, if any) instead of text fields. Changing the decision slider(s) would also reveal the simple accounting mechanics on savings and cash balances, where relevant, and give feedback on expected payoffs in future periods (as in Blaufus and Milde, 2021). Together with the one-dynamic-screen-per-round approach, this greatly reduced the need to navigate through different screens, substantially reducing the time required to complete the otherwise repetitive multi-period decision tasks.

Other experimental design features that substantially contribute to the session's completion time in the original studies are instructions and training on the task. Although at the beginning of the session we showed the instructions and asked subjects to read them, we let the subjects know that the instructions would always be available during the main task. This was implemented using clickable tabs at the bottom of the dynamic screens. Each tab had a small, self-contained piece of information that addressed only one aspect of the experimental task. To further improve the accessibility of instructions, we replaced explicit complex mathematical formulae (such as the induced utility in Meissner, 2016) with graphs that showed, more intuitively, the relevant functional relationship between variables.

The session flow in all replications is illustrated in Fig. 1. Once the subjects completed reading the instructions, they started a trial round.[4] This allowed them to learn by doing the main experimental task, with ready access to the instructions in tabs at the bottom of the screen.[5] The subjects then answered a quiz with four or five questions on the basic mechanics or features of the task before moving on to the rounds of the main task. Random elements of the payoff determination, such as the selection of one period of one round for compensation, were only revealed at the very end of the session. After the main task, subjects were asked basic demographic information (age, gender, education, income range, and financial training/experience). We elicited their risk preference with an assignment task (of their main task earnings) from Gneezy and Potters (1997), and elicited their time preferences (patience) as their willingness to delay their variable payoff by 1, 2 or 3 months for 5% monthly interest.[6] The final payment was determined by the earnings with the main experimental task, the outcome of the risk-taking task and the choice of the time preference task. Subjects were only informed at the end of the experiment about their final payoff and its components.

The experiments were deployed in German, which was the default interface language. Less than 2% of the subjects decided to use English, which was offered as an alternative language. The experiments were programmed in oTree (Chen et al., 2016). Power analyses were computed with GPower (Erdfelder et al., 2009).

3.2. Subject engagement, quality control, and decision efficiency

The experimental sessions were conducted in individual batches for each experiment between September 2021 and March 2022. A total of 6,213 subjects clicked on e-mail invitations sent[7] from the market research panel.[8]

We implemented strict quality control on responses. Subjects were dropped if they skipped too fast through the instruction screens at the beginning of the sessions (thresholds of 10 to 60 seconds). During the quiz, the subjects were dropped if they answered more than two wrong questions on a first attempt or gave any wrong answer in a second attempt.[9] They were also automatically removed from the experiment if they did not finish the session more than 60 minutes after the quiz had been completed.[10]

Panel A of Table 2 details the attrition at each step for all the replications. The completion rate ranges from 21.5% to 50.6% of in-

[2] The reduction in the number of rounds would have affected the analyses of within-subject learning across rounds. We did not study learning across rounds, except in the pre-registered additional analysis in Fatas et al. (2007).

[3] All screenshots for all treatments of the replications are available in the Online Repository.

[4] In experiments adopting a within-subject treatment, the trial round was always identical to the treatment the subjects would undergo in the first live round.

[5] The trial round was not relevant for the payoff.

[6] All payments were credited to the subject accounts directly by the market research company, upon receipt of a master payment file from us. Since subjects in their pool often participate in a few surveys or activities per month and are used to being paid regularly, it is unlikely that the options for delayed payment would have been avoided due to concerns about administrative and time costs to recover delayed payments.

[7] Any subject that gave consent and started the trial of one replication experiment was automatically excluded from participating in any other.

[8] The invitation emails are brief, informing subjects mostly of the expected length of the task and expected payoff.

[9] A second quiz attempt highlighted the questions they got wrong and displayed a reminder with the relevant snippet from the instructions that had the relevant information needed to correct the wrong answer(s). We shuffled the order of the options of the quiz questions in the second attempt.

[10] Very few subjects appear to have been removed from the experiment for taking too long while continuously engaged in the tasks. In all cases, this removal procedure ensured that subjects who abandoned their screens and browser tabs would not be able to resume the experiment many hours or days later.

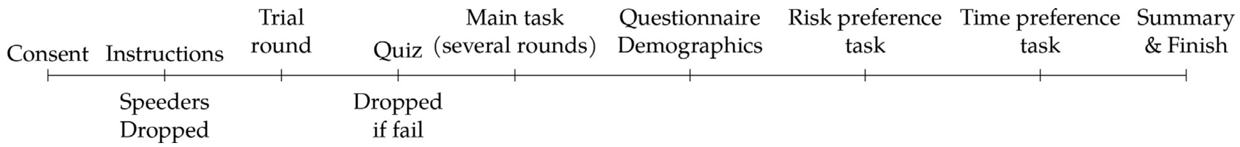

Fig. 1. Sequence of the steps for each replication.

Table 2
Overview of Attrition, Payoff and Completion Time.

Panel A: Subject Attrition

	Anderhub *et al.* (2000)		Fatas *et al.* (2007)		Koehler *et al.* (2015)		Meissner (2016)		Blaufus & Milde (2021)	
	Obs	%	Obs	%	Obs	%	Obs	%	Obs	%
No consent	89	9.3	80	7.6	64	5.5	77	6.0	130	7.5
Drop out at instructions	104	10.9	92	8.8	160	13.9	189	14.6	290	16.7
Drop out at trial round	85	8.9	45	4.3	101	8.8	87	6.7	149	8.6
Drop out / failed quiz	193	20.2	113	10.8	138	12.0	109	8.4	90	5.2
Drop out during tasks	146	15.3	187	17.9	347	30.1	554	42.8	556	32.0
Finished	339	35.5	530	50.6	344	29.8	278	21.5	522	30.1
Total	956	100.0	1047	100.0	1154	100.0	1294	100.0	1737	100.0

Panel B: Payoff (Euro) and Completion Time

	Anderhub *et al.* (2000)		Fatas *et al.* (2007)		Koehler *et al.* (2015)		Meissner (2016)		Blaufus & Milde (2021)	
	Payoff	Total time	Payoff	Total time	Payoff	Total time	Payoff	Total time	Payoff	Total time
Min	0.00	5.45	0.00	3.28	0.00	8.35	0.00	8.07	0.00	12.37
50^{th}-percentile	3.12	14.55	3.43	9.75	1.85	26.91	20.14	22.73	8.38	25.41
95^{th}-percentile	11.42	47.02	16.36	25.53	9.98	62.98	46.44	54.70	25.01	65.62
Max	27.20	14224.83	57.10	7483.52	24.97	2728.22	101.16	4668.42	74.89	2273.38
(Obs. > 65 min)		(12)		(5)		(14)		(8)		(26)

Notes: Panel A shows subject participation according to their furthest stage reached per experiment. *Dropped out at instructions* include the subjects who were rejected for having gone through instruction screens too fast (10s to 60s threshold depending on experiment). *Dropped out at quiz* include the subjects who were rejected for failing to answer a quiz with five or six multiple-choice questions about the experimental instructions, after the trial round. The summary of attrition includes all the subjects who clicked the invitation link and landed on the first web-page of the experiment. In Panel B, *Payoff,* in Euro, is the sum of variable incentive payoff for the main experiment and the payoffs of the risk-taking and patience tasks, and it does not include the non-variable fee of € 4.76 for completing the experiment. *Total time* is the total time (in minutes) that the subjects spent to finish the experiment. The large number of total time in the row *Max* comes from the subjects who finished the experiment but did not click Finish in the end. The last row summarizes the number of observations where the total time is longer than 65 minutes. The summary of payoff and completion time includes only the subjects who completed the experiment.

vitation clicks,[11] but the completion rate is not significantly different between the treatments within each replication. We tested the equality of completion rate between treatments for each replication with a proportion test (ANOVA analysis) if the experiment had 2 (3) treatment groups. The p-value is 0.26 for the replication of Anderhub et al. (2000), 0.07 for Fatas et al. (2007), 0.73 for Koehler et al. (2015), 0.77 for Meissner (2016), and 0.81 for Blaufus and Milde (2021). The p-values remain the same when running logistic regressions and testing if the treatment indicators are equal to zero.

The payment to a subject includes a payment for finishing the study and an incentive payoff based on the outcome of the replication.[12] Panel B of Table 2 summarizes the incentive component of the payoff of the subject and the completion time for the experiments. All replications could produce zero incentive payoff for the subjects, and the largest incentive payoff was € 101.16. The panel also summarizes the completion time of the subjects who answered all questions.[13]

The randomization of subjects to treatment cells in all experiments seems satisfactory with respect to the demographics of the subjects, as seen in Table 3. For most characteristics and treatments of each experiment, there are no significant differences within each experiment at the 5% level, except a few instances. The means of *education* are different ($p = 0.03$) between the treatments of the replication of Koehler et al. (2015). ANOVA tests show that the means of the variable *patience* are different ($p = 0.02$) in the replication treatments of Fatas et al. (2007).

It should be noted that variables *risk-taking* and *patience* were generated after the main tasks, so the subjects' expectations about their earnings from the main task could affect their decisions on the risk-taking task and the time preference task.[14]

Finally, we evaluated the effects of individual subject characteristics on their economic efficiency of decisions across the experiments, with results shown in Table 4.[15] Across four experiments,[16] *female* subjects made less efficient decisions than males, and such gender effect is only statistically significant in the replication of Koehler et al. (2015). In three experiments, higher *risk-taking* subjects performed significantly worse in most studies except the replication of Koehler et al. (2015).[17] The subjects who have participated in *financial training* performed better than those

[11] Data collection for the reproduction of Blaufus and Milde (2021) was affected by a database load surge that slowed down the interface for some hours of the second day of data collection, which motivated some subjects to abandon the task.

[12] In addition to a variable incentive payoff, subjects who finished the experiment earned € 4.76 for participating in the study.

[13] A few subjects who answered all the questions but forgot to click 'Finish' skew the maximum completion time shown in the table.

[14] Even though the uncertainty would only be resolved at the end of the experiment, subjects who performed poorly in the main task on all rounds could consider their low expected payoff when deciding on the risk-taking task.

[15] This analysis was not preregistered.

[16] Fatas et al. (2007) does not have a within-subject dynamic endogenous (to the main task) benchmark for decision efficiency, given its task design.

[17] We cannot exclude an instance of gambling, as the risk preference elicitation follows the main task: subjects who know to have performed badly in the main tasks might well decide to take more risk in the following risk-taking task to recover perceived "losses" in the main task.

Table 3
Subject Characteristics and Treatment Assignments.

	Anderhub *et al.* (2000)		Fatas *et al.* (2007)			Koehler *et al.* (2015)		Meissner (2016)		Blaufus & Milde (2021)		
	Product	Summation	Annuity	Combined	Lump	Long first	Short first	Borrow first	Save first	Deferred	Immediate	Matching
Age	44.57	48.53	48.37	49.59	48.98	41.92	41.05	43.64	42.56	47.60	48.98	48.88
	(1.39)	(1.61)	(1.09)	(1.19)	(1.06)	(1.32)	(1.20)	(1.44)	(1.60)	(1.20)	(1.11)	(1.12)
Observations	176	163	177	170	183	166	178	147	131	162	178	182
Test statistic	-1.88 ($p = 0.06$)		0.30 ($p = 0.74$)			0.48 ($p = 0.63$)		0.50 ($p = 0.62$)		0.44 ($p = 0.65$)		
Female	0.42	0.52	0.42	0.45	0.51	0.56	0.49	0.56	0.49	0.41	0.40	0.38
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Observations	176	163	177	170	182	165	178	144	130	162	178	182
Test statistic	-1.75 ($p = 0.08$)		1.28 ($p = 0.28$)			1.28 ($p = 0.20$)		1.68 ($p = 0.09$)		0.11 ($p = 0.89$)		
Education	2.34	2.45	2.83	2.67	2.64	2.25	2.49	2.25	2.49	2.36	2.37	2.49
	(0.07)	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.08)	(0.07)	(0.08)	(0.07)	(0.07)	(0.07)
Observations	169	160	176	170	181	164	171	144	130	157	177	179
Test statistic	-1.12 ($p = 0.26$)		1.92 ($p = 0.15$)			**-2.25 ($p = 0.03$)**		0.66 ($p = 0.51$)		1.21 ($p = 0.30$)		
Financial training	0.24	0.26	0.23	0.29	0.28	0.20	0.25	0.20	0.25	0.19	0.21	0.24
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Observations	171	159	177	168	181	163	176	143	129	157	175	179
Test statistic	-0.38 ($p = 0.70$)		0.75 ($p = 0.47$)			-1.18 ($p = 0.24$)		-1.33 ($p = 0.18$)		0.61 ($p = 0.55$)		
Income level	6.54	6.64	7.40	7.26	7.45	6.72	6.94	6.72	6.94	7.28	7.40	7.37
	(0.24)	(0.25)	(0.21)	(0.24)	(0.22)	(0.24)	(0.24)	(0.24)	(0.24)	(0.23)	(0.22)	(0.21)
Observations	166	152	172	159	168	156	165	133	119	148	171	174
Test statistic	-0.28 ($p = 0.78$)		0.19 ($p = 0.82$)			-0.66 ($p = 0.51$)		-1.51 ($p = 0.13$)		0.07 ($p = 0.93$)		
Risk-taking	35.52	31.94	31.37	31.98	29.65	24.60	27.42	24.60	27.42	27.66	26.17	26.55
	(1.77)	(1.51)	(1.67)	(1.50)	(1.59)	(1.44)	(1.48)	(1.44)	(1.48)	(1.46)	(1.37)	(1.47)
Observations	176	163	177	170	183	166	178	147	131	162	178	182
Test statistic	0.52 ($p = 0.13$)		0.58 ($p = 0.56$)			-1.36 ($p = 0.17$)		-0.29($p = 0.78$)		0.28 ($p = 0.76$)		
Patience	2.64	2.51	2.69	2.75	2.40	2.66	2.85	2.66	2.85	2.70	2.80	2.73
	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)	(0.10)
Observations	176	163	177	170	183	166	178	147	131	162	178	182
Test statistic	0.88 ($p = 0.38$)		**3.80 ($p = 0.02$)**			-1.37 ($p = 0.17$)		1.28 ($p = 0.20$)		0.31 ($p = 0.73$)		

Notes: Standard errors are in parentheses. *Test statistic* is from the test that checks the equality of means between the treatments. For the replications with 2 treatment groups, *Test statistic* is t-statistic of t-test if the variable is non-binary and z-statistic of proportion test if the variable is binary. For the replications with 3 treatment groups, *Test statistic* is F-statistic of ANOVA analysis. The p-value is in the parenthesis. *Age* is in years old. *Female* is an indicator for female. *Education* equals 1 if the subjects have no qualification, 2 if vocational education, 3 if Bachelor degree, 4 if Master degree and 5 if Doctoral degree. *Financial training* is a dummy indicating that subjects state that they had participated in courses on financial decision-making. *Income level* equals 1 if the monthly household disposable income is below € 400, 2 if the income is between € 400 and € 800, and the value increases with the interval of € 400 to 11 that indicates the income is more than € 4,000. *Risk-taking* is the decision in the risk taking task at the end of the survey where the subjects chose how many percentage points (0-100) of their earnings they would like to put into a lotto. *Patience* is the decision at the end of the survey where the subjects decided how much they were willing to delay the payment to earn interest and equal to 1, 2, 3 and 4 for the choice of no delay, 1 month, 2 months and 3 months, respectively. The observations who chose to give no answer to the questions of gender, education, income or financial training are not included. The cells in bold and blue font are with $p < 0.05$.

who have not, and the effect of financial training is statistically significant only in the replication of Koehler et al. (2015). These differences do not seem to arise from different effect levels between the treatment assignments, as the specifications incorporate round × treatment fixed-effects for the relevant replicated experiments.

3.3. Replication results of Anderhub, Güth, Müller and Strobel (2000)

The experiment analysed how ambiguity (and its resolution) in the probability of survival affects consumption decisions over time. In the main task, subjects started a round – comprising four to six periods – facing three possible chances of being terminated in one period (1/6, 1/3, or 1/2).[18] In the first and second periods, subjects did not face termination while one of the probabilities was removed, reducing ambiguity until only one of the probabilities was left. Then, in the third, fourth, and fifth periods, subjects faced the termination probability that remained. Subjects made consumption decisions out of an initial endowment until they were terminated. Upon termination, any unspent amount from the initial endowment was lost. The round ended automatically after six periods if subjects were not terminated earlier. A subject went through six rounds, and each round had a different sequence of the removal of the three termination probabilities.

To see how the behaviour changes with different risk structures, the treatment conditions implement two different forms of the induced lifetime utility based on period consumption c for subject i at periods t from the first until termination period T. In the *Summation* condition, the payoff is given by the sum of the square root of period consumption $\left(U_i = \sum_{t=1}^{T} \sqrt{c_{i,t}}\right)$. In the *Product* condition, the payoff is given by the product of period consumption $\left(U_i = \prod_{t=1}^{T} c_{i,t}\right)$. The smoothing incentives are larger in the *Product* condition, since the expected payoff in that condition is substantially reduced if subjects spend all their endowment before termination (as one of the periods would have zero consumption and, thus, the lifetime utility for that round would be zero).

Analysis of the behaviour with respect to a risk-neutral optimal benchmark suggests two distinct behavioural patterns. First, subjects need to dynamically adjust their spending based on the sequential resolution of the ambiguity of the termination probabilities. When a larger or smaller termination probability is removed, the expected remaining length of the round of the subject increases or decreases, respectively; this should lead subjects to increase or decrease spending in the following period accordingly.

[18] In our study, the different levels of termination risk were implemented using different distributions of colours in card decks (the original study used numerical ranges of a dice).

Table 4
Effects of Individual Characteristics on the Efficiency of Decisions.

	Anderhub *et al.* (2000)		Koehler *et al.* (2015)		Meissner (2016)		Blaufus & Milde (2021)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Age	-0.000	-0.000	-0.006***	-0.006***	-0.000	-0.000	-0.300*	-0.291*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.123)	(0.121)
Female	0.091	0.092	-0.052*	-0.052*	-0.036	-0.024	1.572	1.510
	(0.049)	(0.049)	(0.023)	(0.023)	(0.019)	(0.019)	(4.037)	(4.019)
Education	0.041	0.042	0.033**	0.032**	0.009	0.010	-4.323	-3.863
	(0.029)	(0.029)	(0.012)	(0.012)	(0.010)	(0.010)	(2.211)	(2.184)
Financial training	-0.072	-0.072	0.058*	0.058*	0.031	0.018	-5.410	-4.945
	(0.056)	(0.056)	(0.028)	(0.028)	(0.024)	(0.023)	(4.710)	(4.698)
Income	-0.016	-0.016	-0.003	-0.003	-0.006*	-0.007*	-0.011	-0.028
	(0.008)	(0.008)	(0.004)	(0.004)	(0.003)	(0.003)	(0.696)	(0.684)
Risk-taking	0.003**	0.003*	-0.000	-0.000	-0.002***	-0.002***	0.416***	0.417***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.102)	(0.100)
Patience	-0.032	-0.033	0.040***	0.040***	-0.011	-0.011	1.583	1.556
	(0.019)	(0.019)	(0.008)	(0.008)	(0.008)	(0.007)	(1.394)	(1.392)
Constant	1.181***	1.175***					66.596***	54.145***
	(0.111)	(0.116)					(9.768)	(10.397)
Round/treatment FE	No	Yes	No	Yes	No	Yes	No	Yes
Observations	309	309	1252	1252	968	968	954	954

Notes: The dependent variables are the measurements of the efficiency of decisions: for Anderhub et al. (2000), it is the mean of deviations of the observed decisions from the optimal decisions; for Koehler et al. (2015), the dependent variable is the dummy indicating that there is no unspent money in the last period and no bankruptcy happened; for Meissner (2016), it is the dummy indicating that there is no overspending; for Blaufus and Milde (2021), it is the mean of absolute deviations from the optimal saving of the periods in a round. The optimal saving is the saving that maximizes the expected payoff. Given that the periods have an equal chance to determine the final payoff, the optimal saving is same for each period and it is 74 points for treatment Immediate and Matching and 124 points for treatment Deferred. The results of the first and last columns are OLS estimations, and the results of the second and third columns are marginal effects of logistic regressions. *Age* is in years old. *Female* is an indicator for female. *Education* equals 1 if the subjects have no qualification, 2 if vocational education, 3 if Bachelor degree, 4 if Master degree and 5 if Doctoral degree. *Financial training* is a dummy indicating that subjects state that they had participated in courses on financial decision-making. *Income level* equals 1 if the monthly household disposable income is below € 400, 2 if the income is between € 400 and € 800, and the value increases with the interval of € 400 to 11 that indicates the income is more than € 4,000. *Risk-taking* is the decision in the risk taking task at the end of the survey where the subjects chose how many percentage points (0-100) of their earnings they would like to put into a lotto. *Patience* is the decision at the end of the survey where the subjects decided how much they were willing to delay the payment to earn interest and equal to 1, 2, 3 and 4 for the choice of no delay, 1 month, 2 months and 3 months, respectively. The round/treatment control covariates include the round number and the treatment dummies. The observations who chose to give no answer to the questions of gender, education, income or financial training are not included. The number of observations equals the number of subjects for the first study and the number of the decisions made by all subjects in all the rounds for the last three studies. Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Second, the consumption should monotonically decrease from the third period on.

Subjects participated in six rounds (twelve in the original study), comprising all permutations (twice each in the original study) of the sequence of resolution of the uncertainty of termination risk (the order in which the card decks – with different termination probabilities – are removed). We used the same initial endowment for both treatments (11.92 ECU) as in the original study, and adjusted the exchange rates of lifetime utility-induced points into currency to account for the different conditional expectation of payoffs given the induced utility function of both treatments.

In our replication, 339 subjects completed the experiment: 176 in the treatment *Product* and 163 in *Summation* (in the original study, 50 subjects participated in each treatment). In the *Product* treatment, there were 1056 cases (176 subjects × 6 rounds) and the average reward was € 3.33; in the *Summation* treatment, there were 978 cases (163 subjects × 6 rounds) and the average reward was € 4.16.[19]

Following the original study, we show the univariate statistics for the average consumption decision per period according to the ambiguity resolution path in Figs. 2 and 3. The red outline of the nodes indicates that on average, the subjects of that treatment, period, and uncertainty resolution path spent above the optimal consumption levels for that node; the blue indicates spending below these optimal levels. The average efficiency is defined as U/U^*, where U is the average payoff in all six rounds, and U^* is the expected optimal payoff.

In our sample, the efficiency rate in the condition *Summation* is higher than in the condition *Product*, as in the original study, but the average efficiency of consumption decisions is smaller in both treatments as compared to the original study. Additionally, we found that there is substantially less differential adjustment to the resolution of uncertainty of termination probabilities than in the original study. For example, in the second period (X2), for the *Summation* condition the average spending ranges from 2.47 to 2.54 points, while in the original study the averages ranged between 2.56 and 3.23. We also found that, contrary to the original study, when ambiguity of termination risk is eliminated, the fraction of endowment consumed did not vary substantially according to the optimal levels under either treatment condition, as seen in Table 5. The fraction of consumption under the *Summation* condition ranged only from 0.69 to 0.72 (compared to optimal levels of 0.70 to 0.89 and observed levels in the original study of 0.70 to 0.83). Overall, these observations suggest that the subjects in our sample reacted much less to changes in their termination risk than subjects in the original study, of which the main result was that “subjects’ reactions to information about termination probabilities are qualitatively correct.”

To evaluate whether subjects respond qualitatively correctly to the resolution of uncertainty, we first checked the reactions to the removal of the first termination probability (card deck). The

[19] The average reward is among all the cases in a treatment (all subjects in all the rounds). For a subject, one out of the six rounds is randomly chosen to determine the incentive payoff and the average incentive payoff among the subjects is € 3.48 for the *Product* treatment and € 4.74 for the *Summation* treatment.

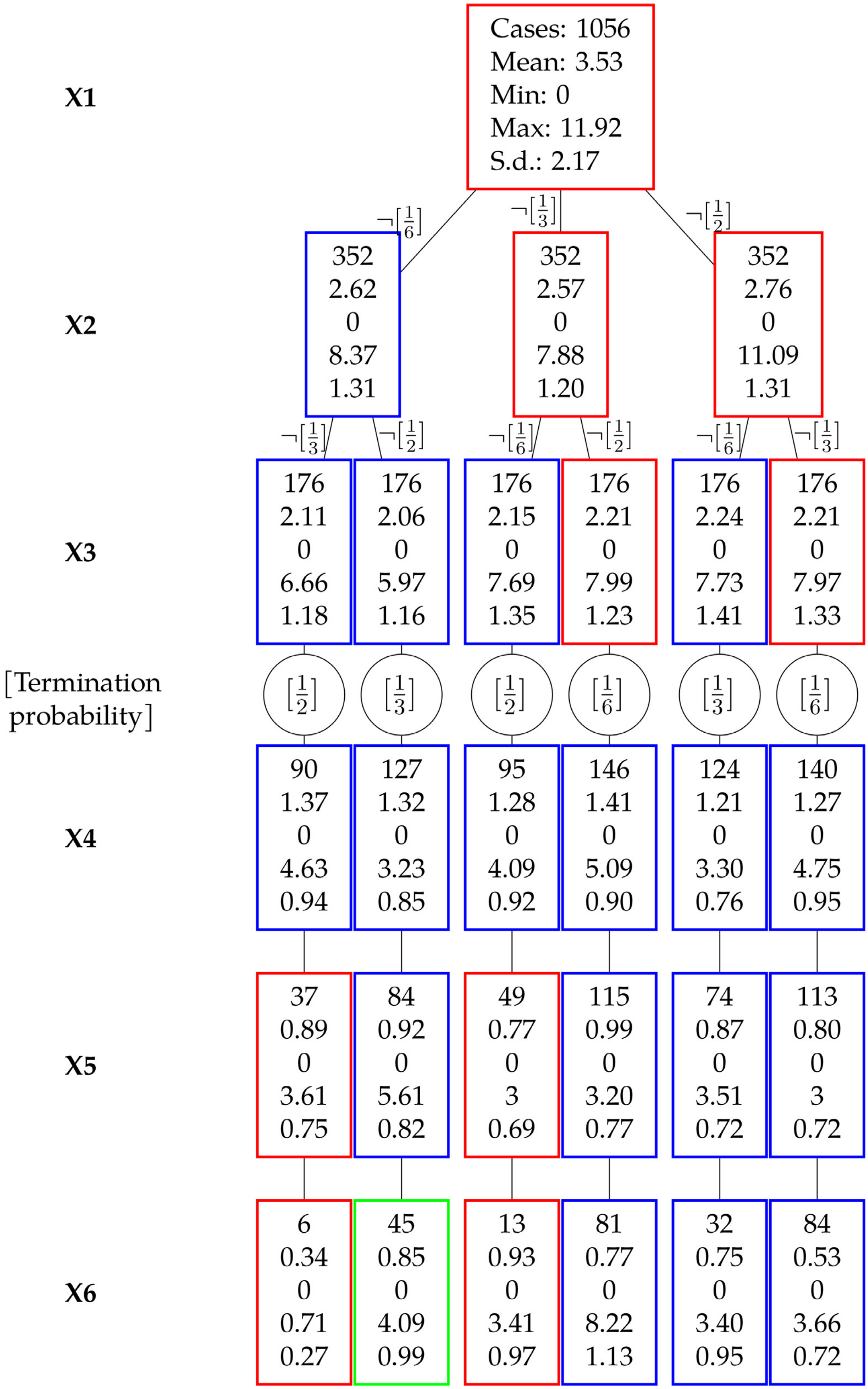

Fig. 2. Anderhub et al. (2000) Replication: Average Behaviour on Product Treatment. *Notes:* The labels on the left X1 to X6 indicate the period of the spending decision for *Product* treatment. In each node box, the five values (from the top to the bottom) indicate the number of cases, mean, maximal, minimal, and standard deviation. $\neg[\frac{1}{2}]$ ($\neg[\frac{1}{3}]$, $\neg[\frac{1}{6}]$) indicates the card deck indicating termination probability $\frac{1}{2}$ ($\frac{1}{3}$, $\frac{1}{6}$) is removed. After Period 3 (X3), the finally stayed card deck determines the termination probability in the next periods and it can be $\frac{1}{2}$, $\frac{1}{3}$, or $\frac{1}{6}$. The color of the node box border indicates how the observed mean of spending decision is compared to the optimal spending paths: red means over-spending, blue means under-spending, and green means same as the optimal spending.

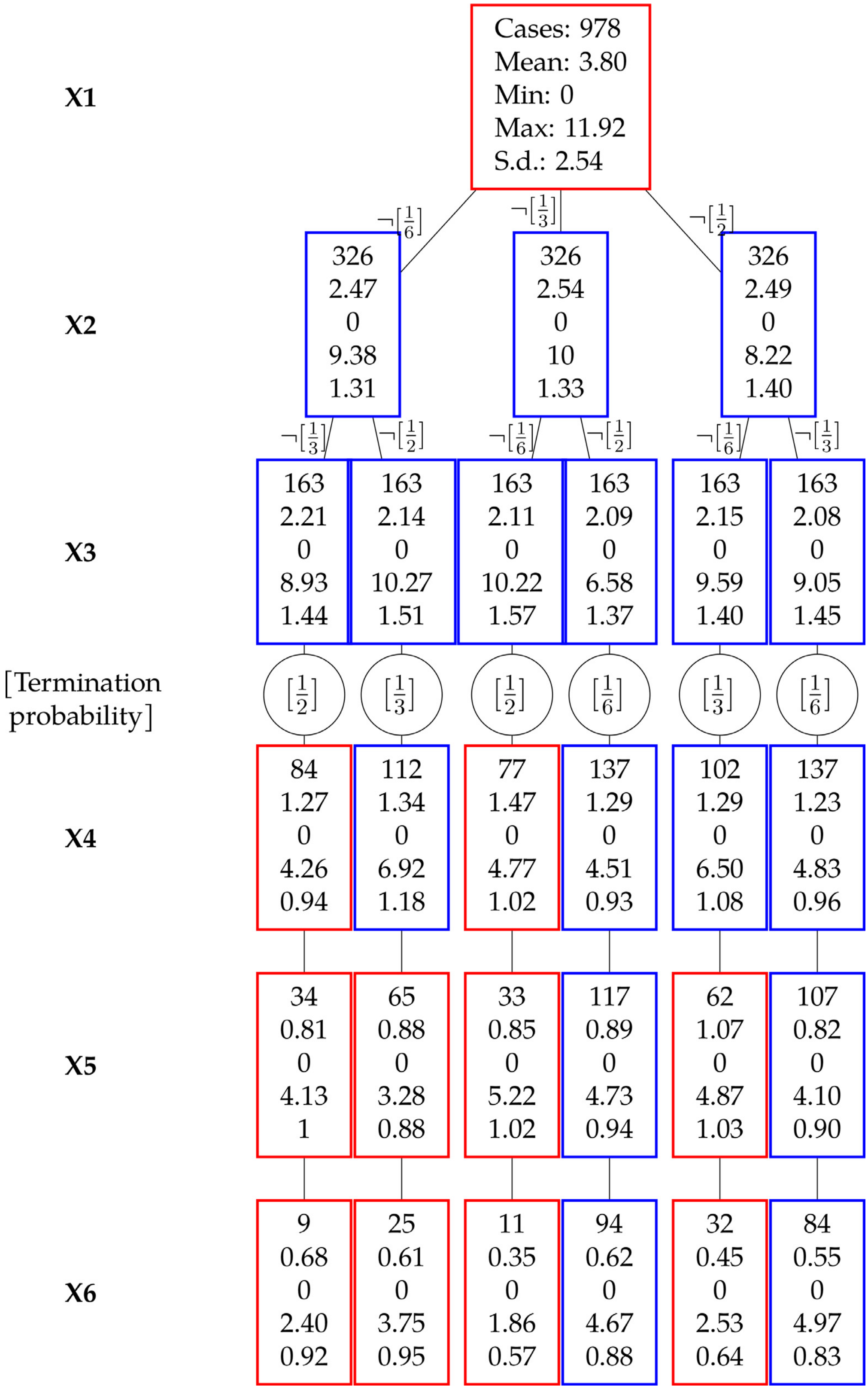

Fig. 3. Anderhub et al. (2000) Replication: Average Behaviour in Summation Treatment. *Notes:* The labels on the left X1 to X6 indicate the period of the spending decision for *Summation* treatment. In each node box, the five values (from the top to the bottom) indicate the number of cases, mean, maximal, minimal, and standard deviation. $\neg[\frac{1}{2}]$ ($\neg[\frac{1}{3}]$, $\neg[\frac{1}{6}]$) indicates the card deck indicating termination probability $\frac{1}{2}$ ($\frac{1}{3}$, $\frac{1}{6}$) is removed. After Period 3 (X3), the finally stayed card deck determines the termination probability in the next periods and it can be $\frac{1}{2}$, $\frac{1}{3}$, or $\frac{1}{6}$. The color of the node box border indicates how the observed mean of spending decision is compared to the optimal spending paths: red means over-spending and blue means under-spending.

Table 5
Anderhub et al. (2000) Replication: Consumption and Resolution of the Survival Ambiguity.

Sequence Rank	1st period removal	2nd period removal	Product	Summation
1	¬[1/6]	¬[1/3]	0.67 (0.80)	0.72 (0.89)
2	¬[1/3]	¬[1/6]	0.69 (0.76)	0.69 (0.88)
3	¬[1/6]	¬[1/2]	0.70 (0.66)	0.71 (0.81)
4	¬[1/2]	¬[1/6]	0.72 (0.59)	0.69 (0.79)
5	¬[1/3]	¬[1/2]	0.70 (0.58)	0.71 (0.71)
6	¬[1/2]	¬[1/3]	0.72 (0.56)	0.72 (0.70)

Notes: Average fraction of initial wealth consumed in the first three periods, according to path of resolution of ambiguity on longevity risk. *1st* and *2nd period removal* are the card decks removed in the first and second period, which eventually eliminates ambiguity of the actual survival probabilities subjects will face (the remaining card deck being then used to determine survival after periods 3, 4 and 5). Sequences are ranked in descending order of optimal consumption fraction (in parentheses). *Product* and *Summation* are the treatments. ¬ means the removal of a card deck and the following fraction indicates the termination probability of the removed card deck. E.g., ¬[1/2] means the card deck with termination probability 1/2 is removed.

Table 6
Anderhub et al. (2000) Replication: Reactions to the first removed card deck and finally stayed card deck.

Panel A: **Mean consumption share in period 2 and 3**

Treatment	Mean consumption share in period 2 (x_2 / S_2)			Mean consumption share in period 3 (x_3 / S_3)			Obs.	Obs. fulfilling Condition 1	Obs. fulfilling Condition 2
	First period removal			Finally stayed deck					
	¬[1/6]	¬[1/3]	¬[1/2]	[1/6]	[1/3]	[1/2]			
Product	0.33	0.34	0.35	0.44	0.42	0.40	176	19	29
Summation	0.35	0.35	0.34	0.43	0.41	0.42	163	39	24

Panel B: ***P*-value of *t*-test of the inequality between the consumption shares**

Treatment	*P*-value of *t*-test					
	Null hypothesis for Condition 1			Null hypothesis for Condition 2		
	$\left(\frac{x_2}{S_2}\middle|\neg[1/6]\right) \le \left(\frac{x_2}{S_2}\middle|\neg[1/2]\right)$	$\left(\frac{x_2}{S_2}\middle|\neg[1/3]\right) \le \left(\frac{x_2}{S_2}\middle|\neg[1/2]\right)$	$\left(\frac{x_2}{S_2}\middle|\neg[1/6]\right) \le \left(\frac{x_2}{S_2}\middle|\neg[1/3]\right)$	$\left(\frac{x_3}{S_3}\middle|[1/6]\right) \ge \left(\frac{x_3}{S_3}\middle|[1/2]\right)$	$\left(\frac{x_3}{S_3}\middle|[1/3]\right) \ge \left(\frac{x_3}{S_3}\middle|[1/2]\right)$	$\left(\frac{x_3}{S_3}\middle|[1/6]\right) \ge \left(\frac{x_3}{S_3}\middle|[1/3]\right)$
Product	0.886	0.863	0.545	0.978	0.807	0.874
Summation	0.387	0.336	0.554	0.664	0.260	0.857

Notes: The mean consumption share is computed from all the subjects in period 2 ($\frac{x_2}{S_2}$) and 3 ($\frac{x_2}{S_2}$) correspondingly. For each subject, there are two rounds out of the six rounds where the first removed card deck (finally stayed card deck) is of a same termination probability. The consumption share in period 2 (period 3) for each subject is the mean of the shares of the two rounds with the same termination probability of first removed card deck (finally stayed card deck). Condition 1 refers to $(\frac{x_2}{S_2}|\neg[1/6]) > (\frac{x_2}{S_2}|\neg[1/3]) > (\frac{x_2}{S_2}|\neg[1/2])$, implying that the spending in period 2 is larger when a card deck with a low termination probability is removed than a card deck with a high termination probability is removed. Condition 2 refers to $(\frac{x_3}{S_3}|[1/2]) > (\frac{x_3}{S_3}|[1/3]) > (\frac{x_3}{S_3}|[1/6])$, implying that the spending in period 3 is larger when a card deck with a high termination probability finally stays than a deck with a low termination probability stays. The *p*-value is from *t*-tests of the null hypotheses.

removal of a card deck for a low termination probability after the first period decreases the expected length of the round for the subject, and thus he/she should consume more in the second period; conversely, removal of a card deck with a high termination probability increases the expected length of the round, and incentivizes a reduction in consumption. This implies the condition $(\frac{x_2}{S_2}|\neg[1/6]) > (\frac{x_2}{S_2}|\neg[1/3]) > (\frac{x_2}{S_2}|\neg[1/2])$, where ¬ is the removal of a card deck (set for one termination probability), x_2 is the spending decided in the second period, and S_2 is the disposable amount in that period. E.g., $(\frac{x_2}{S_2}|\neg[1/2])$ means the proportion of the decided spending to the disposable endowment in Period 2 when the card deck with termination probability 1/2 is removed. Likewise, in the third period, when all ambiguity has been resolved and one probability remains, subjects should consume more when the final termination probability is high and less when that probability is low. This implies the condition $(\frac{x_3}{S_3}|[1/2]) > (\frac{x_3}{S_3}|[1/3]) > (\frac{x_3}{S_3}|[1/6])$, where $(\frac{x_3}{S_3}|[1/2])$, for example, means the proportion of the decided spending to the disposable endowment in Period 3 when the card deck with termination probability 1/2 remains. The mean consumption shares in Panel A of Table 6 exhibit no obvious difference when different card decks are removed. The statistical tests in Panel B of Table 6 also show that, on average, the subjects in our sample do not fulfil either of the two conditions. For each condition, the original study rejects the hypothesis that the subjects do not fulfil the condition with a binomial test for both treatments. A summary of these analyses for socio-demographic subsamples is reported in Tables A2 and A3 in the Appendix.

Another check of the quality of decisions relies on the observation that consumption should always be larger in the earlier periods when the period of termination (the length of the round) is still undetermined. In Table 7, we tabulate the percentage of cases where this condition is met, according to the termination periods of each subject in each round. Our results confirm the observation in the original study that a large fraction of cases does not adhere to relatively relaxed conditions. For example, in the right column of panel A for the *Product* condition, only 35.6% of the cases who reached period 6 had monotonically decreasing consumption between periods 3 and 6 (35.5% in the original study). In Panel B for *Summation*, 35.3% of the cases met the same conditions (48.7% in the original study).[20]

[20] In this analysis, most of the discrepancies between our results and the original study are due to our relatively smaller differences, in each termination period and treatment, between the fraction of subjects who satisfy the condition strictly (as before) or weakly ($x_3 \geqslant x_4 \geqslant x_5 \geqslant x_6$). By contrast, in the original study many subjects violated the strict condition, but kept consumption numerically constant between two rounds. This specific difference between violations of strictly and weak conditions is, arguably, due in part to our use of a slider precise to increments of 0.01, rather than requiring a numerical input.

Table 7
Anderhub et al. (2000) Replication: Facing an uncertain future.

Panel A: **Treatment Product**

	Cases	%		Cases	%		Cases	%
$T \geq 4$	729	100.0	$T \geq 5$	477	100.0	$T = 6$	264	100.0
$x_3 > x_4$	525	72.0	$x_3 > x_4 > x_5$	241	50.5	$x_3 > x_4 > x_5 > x_6$	94	35.6
$x_3 \geq x_4$	569	78.1	$x_3 \geq x_4 \geq x_5$	281	58.9	$x_3 \geq x_4 \geq x_5 \geq x_6$	126	47.7
$T \geq 5$	477	100.0	$T = 6$	264	100.0			
$x_4 > x_5$	333	69.8	$x_4 > x_5 > x_6$	125	47.3			
$x_4 \geq x_5$	368	77.1	$x_4 \geq x_5 \geq x_6$	157	59.5			
$T = 6$	264	100.0						
$x_5 > x_6$	180	68.2						
$x_5 \geq x_6$	209	79.2						

Panel B: **Treatment Summation**

	Cases	%		Cases	%		Cases	%
$T \geq 4$	655	100.0	$T \geq 5$	422	100.0	$T = 6$	258	100.0
$x_3 > x_4$	482	73.6	$x_3 > x_4 > x_5$	213	50.5	$x_3 > x_4 > x_5 > x_6$	91	35.3
$x_3 \geq x_4$	514	78.5	$x_3 \geq x_4 \geq x_5$	256	60.7	$x_3 \geq x_4 \geq x_5 \geq x_6$	132	51.2
$T \geq 5$	422	100.0	$T = 6$	258	100.0			
$x_4 > x_5$	289	68.5	$x_4 > x_5 > x_6$	112	43.4			
$x_4 \geq x_5$	328	77.7	$x_4 \geq x_5 \geq x_6$	155	60.1			
$T = 6$	258	100.0						
$x_5 > x_6$	161	62.4						
$x_5 \geq x_6$	197	76.4						

Notes: Cases is the number of decisions, all of the decisions of all the subjects. $T \geq k$ (k=4, 5, 6) means that the subject reaches at least period k. x_k (k=4, 5, 6) is the consumption decision in period k.

3.4. Replication results of Fatas, Lacomba and Lagos (2007)

This study investigated the impact of the structure of retirement payouts on the choice of when to retire when the subjects face longevity risks. The three considered structures are *Annuity, Lump sum*, and *Combined*. At the start of a round, subjects chose the period in which they wanted to start collecting retirement benefits. In every period, there was risk of being terminated, and subjects only earned payoffs in a round while they are still active.

In the *Annuity* treatment, subjects received a fixed payout per period, starting at their chosen retirement period. In the *Lump sum* treatment, subjects earned a single payout at their chosen retirement period and nothing in any other active period. In the *Combined* treatment, they earned both a lump sum and an annuity, as in the previous treatments. In all the treatments, the payout was higher if the subjects chose a later period to retire (i.e., start collecting the payout). However, the subjects received payouts only if they were active when the retirement period arrived. The expected value of the payoff per round was equal for all the treatments and chosen periods of retirement. In our study, subjects underwent three rounds (the original study comprised a single round) to allow evaluating learning effects, of which one was randomly chosen for compensation based on the total payoff accrued in the chosen round.

Termination in a round was determined by a random draw of cards without replacement at each period – starting with 14 green cards and one red card (in the original study, coloured balls were used instead). Each period, a card was selected from the stack, and the subject was terminated in a round when a red card is drawn. This procedure generates a survival function with interesting properties: known maximum length of experimental life (as a red card will be eventually drawn with certainty), decreasing one-period survival probabilities, and increasing rate-of-change of survival probabilities. These are properties also found on stylized human survival curves for individuals approaching the typical age of retirement.

The three treatments have equal expected lifetime payout for any period of retirement chosen by subjects (when adjusted for the implicit survival probabilities) such that in theory there should be no difference between the treatments in the choice of the timing of retirement if subjects were neutral to the structure of the payoffs.

In our replication, 530 subjects completed the experiment (177 in the *Annuity*, 170 in the *Combined*, and 183 in the *Lump sum* condition – in the original study, these numbers were 28, 26, and 22, respectively). Similarly to the results of the original study, we found that subjects earning *Lump sum* payments chose to retire later than those earning *Annuity* or *Combined* (with all payoffs actuarially equivalent), as shown in Fig. 4. On average, subjects in the *Annuity* condition chose to retire after 5.49 periods, those in the *Lump sum* condition retired latest (after 6.32 periods), and those in *Combined* after 6.13 periods (in the original experiment, they chose to retire after 5.0, 9.0, and 7.0 periods, respectively).

As in the original study, we found significant treatment effects between the treatments of *Lump sum* and *Annuity*. Following the original study, using *Lump sum* as a baseline, we regressed the chosen retirement period on the treatment indicator variables while using our own measures of *risk-taking* and *patience* as controls. The results are shown in Table 8. In the full specification (4), subjects in the *Annuity* condition chose to retire 0.916 periods earlier than those in *Lump sum*. The difference was smaller (0.863 periods) but still significant before controlling for patience in (5). The difference in the estimated coefficients of *Annuity* and *Combined*, shown in the bottom panel (0.635 and 0.645 periods in (4) and (5)), is significant, as it was in the original study. However, contrary to the original study, the difference in the coefficients between *Lump sum* and *Combined* was not significant in either specification. Table A4, in the Appendix, shows results for the same analysis repeated for socio-demographic subsamples.

Similarly to the original study, we found that higher risk taking is significantly associated with a later choice of retirement timing: each additional percentage point allocated to a risky asset in a Gneezy and Potters (1997) task was associated with a delayed retirement timing of 0.026 to 0.028 periods (the original study used a different risk-preference elicitation method). Patience was also positively associated with a delay in retirement. Each month that subjects chose to wait for their payoff in exchange for 5% interest (per month) is associated with a delay in the choice of retirement period between 0.184 and 0.255 periods.

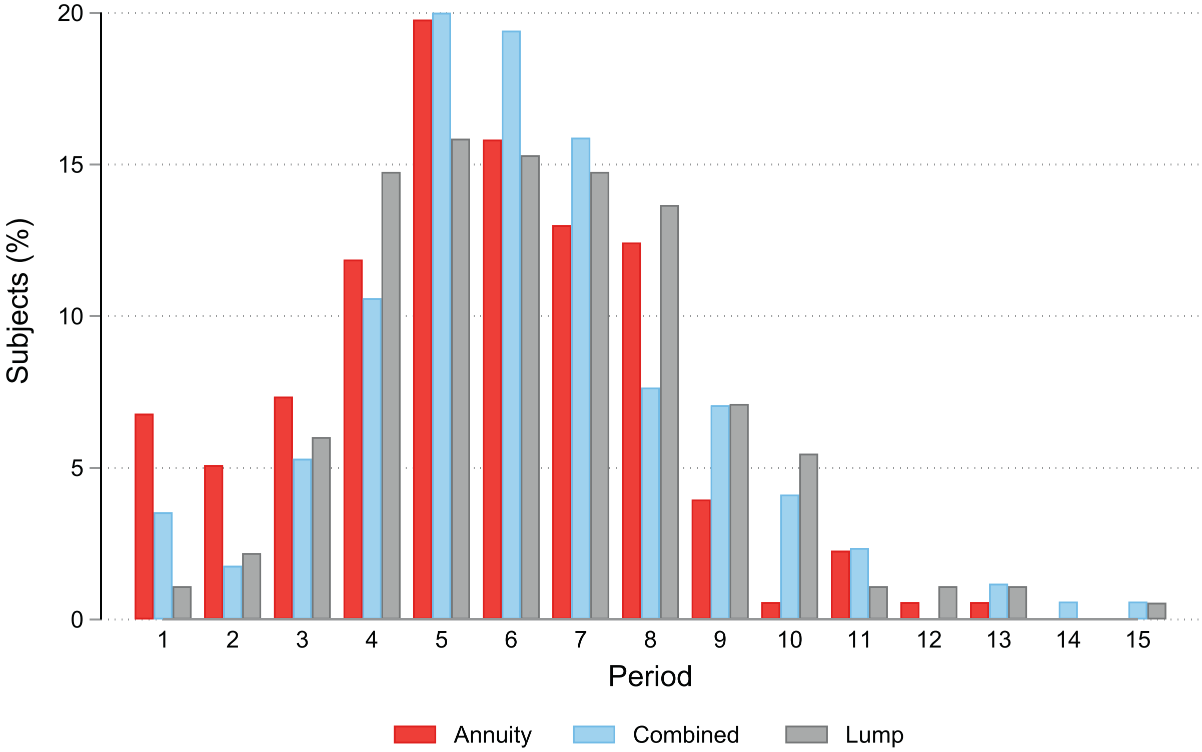

Fig. 4. Fatas et al. (2007) Replication: Timing of Retirement. *Notes:* Period chosen by subjects to (start) collecting payoffs, conditional on not having been terminated.

Table 8
Fatas et al. (2007) Replication: Timing of Retirement Treatment Effects.

	(1)	(2)	(3)	(4)	(5)
Risk-taking	0.028***		0.026***	0.026***	0.028***
	(0.005)		(0.005)	(0.005)	(0.005)
Patience		0.255**	0.184*	0.193*	
		(0.080)	(0.079)	(0.079)	
Annuity		-0.779***	-0.916***	-0.863***	
			(0.215)	(0.247)	(0.247)
Combined				-0.281	-0.218
				(0.250)	(0.250)
Constant	5.136***	5.336***	4.980***	5.087***	5.481***
	(0.182)	(0.234)	(0.259)	(0.276)	(0.224)
(Annuity–Combined)				-0.635*	-0.645*
				(0.250)	(0.251)
R2	0.059	0.019	0.090	0.092	0.082
Prob. >F	0.000	0.002	0.000	0.000	0.000
Observations	530	530	530	530	530

Notes: The results are from OLS estimations. The dependent variable is the mean retirement period chosen in the three rounds. *Annuity* and *Combined* are dummies for subjects assigned to such treatment conditions; *Lump sum* is the baseline. *Risk-taking* is the decision in the risk taking task at the end of the survey where the subjects chose how many percentage points (0-100) of their earnings they would like to put into a lotto. *Patience* is the decision at the end of the survey where the subjects decided how much they were willing to delay the payment to earn interest and equal to 1, 2, 3 and 4 for the choice of no delay, 1 month, 2 months and 3 months, respectively. Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In an additional preregistered analysis that was not part of the original study, we analysed how termination at round 1 and/or 2 affected the choice of timing of retirement in later rounds. Termination is the most salient event in a round, and being terminated before one's chosen retirement period means that no payoff is accrued in that round. The experience of termination in earlier rounds might influence the subsequent decisions of subjects in later rounds, as they learn throughout the rounds. The results of this additional analysis are presented in Table 9.

We found that generally, a later termination in earlier round(s) was associated with a significantly delayed choice of retirement in subsequent round(s). In specification (3), controlling for the treatment, a first round that lasted one period longer delayed the retirement timing chosen in the second round by 0.06 periods. A much more salient event is that the subjects survived at least until the period they had chosen to earn (or start earning) their payoffs. In specification (4), we regressed the choice of timing of retirement in round 2 on whether the subject survived until their chosen timing of retirement during round 1. In round 1, surviving at least until the chosen period delayed the subsequent choice of timing of retirement chosen in round 2 by 2.782 periods. The direct effect of one later period for termination was then a further delay of 0.277

Table 9
Fatas et al. (2007) Further Analysis: Effects of Experienced Termination Period on Later Decisions.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
End period round 1	0.060*	0.066*	0.060*	0.277***	0.077**	0.083**	0.078**	0.254***
	(0.029)	(0.029)	(0.028)	(0.037)	(0.030)	(0.029)	(0.029)	(0.037)
End period round 2					0.051	0.050	0.062*	0.191***
					(0.030)	(0.029)	(0.029)	(0.037)
Annuity		-0.733*	-0.832**	-0.788**		-1.033***	-1.139***	-0.966***
		(0.304)	(0.298)	(0.280)		(0.309)	(0.300)	(0.280)
Combined		0.177	0.047	-0.054		-0.362	-0.497	-0.545
		(0.307)	(0.301)	(0.283)		(0.312)	(0.304)	(0.282)
Non-zero pay round 1				2.782***				2.290***
				(0.332)				(0.335)
Non-zero pay round 2								1.763***
								(0.340)
Risk-taking			0.027***	0.022***			0.030***	0.023***
			(0.006)	(0.006)			(0.006)	(0.006)
Patience			0.188	0.152			0.201*	0.118
			(0.096)	(0.090)			(0.097)	(0.090)
Constant	5.888***	6.029***	4.811***	2.329***	5.193***	5.614***	4.181***	0.666
	(0.264)	(0.310)	(0.390)	(0.471)	(0.362)	(0.401)	(0.471)	(0.578)
Observations	530	530	530	530	530	530	530	530

Notes: The results are from OLS estimations. The dependent variable is the decision of retirement starting period in round 2 in columns (1-4) and the decision of retirement starting period in round 3 in columns (5-8). *End period round 1* (*2*) is the termination period in round 1 (2). *Annuity* and *Combined* are dummies for subjects assigned to such treatment conditions; *Lump sum* is the baseline. *Non-zero pay round 1* (*2*) is a dummy indicating the payoff in round 1 (2) is non-zero (one of the three rounds is randomly chosen at the end of the study to determine the final payoff). *Risk-taking* is the decision in the risk taking task at the end of the survey where the subjects chose how many percentage points (0-100) of their earning they would like put into a lotto. *Patience* is the decision at the end of the survey where the subjects decided how much they were willing to delay the payment to earn interest and equal to 1, 2, 3 and 4 for the choice of no delay, 1 month, 2 months and 3 months, respectively. Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

periods for round 2. In specifications (5-8), we tested retirement timing in round 3 given outcomes of the first two rounds: the effect of termination period and survival until the chosen period remained significant in all the specifications.

3.5. Replication results of Meissner (2016)

This study evaluated the consumption smoothing behaviour when debt is treated differently than savings. To study this question, the study allowed interest-free borrowing. Over a set number of periods in a life (round), subjects decided on savings and consumption while facing different broad income paths, increasing or decreasing throughout a round, with local stochastic perturbations (such that the income paths were not strictly monotonically increasing or decreasing). The study tested the hypothesis that with an induced CARA utility reward function of the consumption in a period, subjects should smooth their consumption throughout all periods in a round. On a downward income path, smoothing lifetime consumption requires *saving* from earlier periods for later consumption (Saving condition). On an upward income path, lifetime consumption smoothing requires borrowing from later periods when income will be higher (*Borrowing* condition).

Treatment groups differed in the sequence of conditions that the subjects faced. In the treatment *Savings First*, subjects played two rounds in the *Saving* condition, then switched to *Borrowing* for another two rounds, while in the *Borrowing First* treatment this order was reversed. In the last period of either condition, no decision was made, and its consumption (spending) was set such that lifetime consumption would equal lifetime income. Our replication focused on the treatment effect of symmetric financial decisions (saving or borrowing) on lifetime consumption smoothing.[21]

The null hypothesis of the study is that regardless of the income path, the consumption paths should be equally smooth if debt is not evaluated differently than saving. The behaviour under the first of the conditions to which subjects were randomly assigned should not differ from the behaviour under the second condition in the last half of the session. The expected payoff is maximal when subjects smooth their consumption regardless of the income path.

To simplify the task and make it viable to implement with our sample, we first reduced the length of the experimental life (from 20 to 16 periods) and the repetitions (from three to two rounds per condition). We also modified the variables in the experimental environment of the original study. In our replication, subjects earned income in points and variable incentives, per period, in the form of induced CARA utility over their consumption, which was then converted into euros (‘Eurocent Rewards’). In the original experiment, subjects earned “Talers” instead, which they converted into utility-induced “points,” summed across each round and then converted into monetary units. We bypassed this intermediate utility computational variable and presented the CARA-induced utility conversion as both a static graph and as dynamic text information per period, as subjects manipulated a slider prior to confirming their decisions. We also simplified the variable incentive to be the lifetime sum of ‘Eurocent Rewards’ in one randomly chosen round (the original study used the average of total payoffs of one round of the first treatment and one round of the second treatment, per subject). To reduce the task complexity, we also did not allow for negative spending (to be distinguished from negative savings, i.e., borrowing) in periods other than the last.[22]

In total, 278 subjects completed the experiment, of whom 147 in the *Borrowing First* treatment and 131 in *Savings First* (the original experiment recruited 38 subjects for each treatment).[23]

[21] The original study further investigated the roles of myopia and learning on these consumption decisions.

[22] Voluntary negative spending, as allowed in the original study, is a very hard feature to conceptualize for subjects, and it would have required a significant expansion of the instruction set. In the original study, which allowed negative spending as an induced CARA utility function that could be defined in the negative domain, only 24 of 9120 (subjects × period × round) spending decisions were negative.

[23] In the original preregistered plan, we had proposed excluding subjects who, in a first attempt, got more than one mistake in the instruction quiz. This resulted in an unexpectedly high rejection rate that was not acceptable for our market research

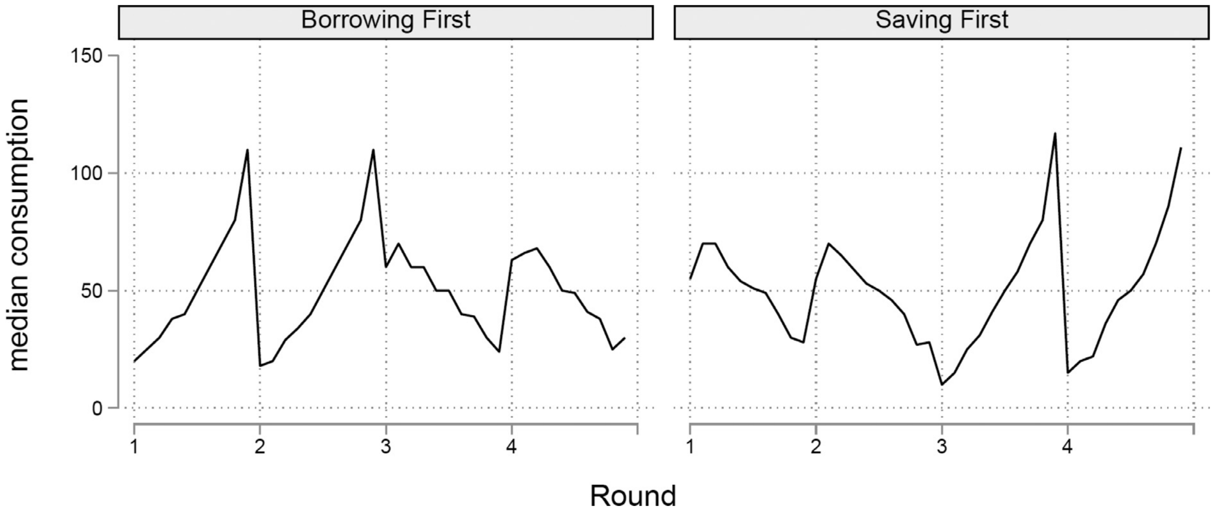

Fig. 5. Meissner (2016) Replication: Median Consumption per Period over Sequential Rounds by Treatments. *Notes:* Borrowing (saving) first subjects play rounds 1 and 2 in the borrow (saving) condition; and rounds 3 and 4 in the saving (borrow) condition.

In Fig. 5, we see – as in the original study – that subjects in the *Borrowing* condition have a greater variance in their consumption path than in *Saving* and do not borrow from future income to smooth consumption in earlier periods. Lifetime consumption is smoother under the *Saving* condition, as its subjects have to save a part of the income they have already earned at present instead of borrowing from future expected higher income. However, compared to the original study, the median consumption among subjects in the *Saving* condition is not as smooth in our study. Order effects of the income paths from treatments *Borrowing First* or *Savings First* did not appear to significantly affect the results in each treatment of our experiment, as in the original study.

Following the original study, we use three measures to evaluate deviations from optimal consumption. *Measure 1* is the lifetime sum (within a round) of the period deviations between observed consumption and the optimal consumption at each period, conditioned on the wealth (unspent endowment) of the subject at the start of each period. *Measure 2* is the lifetime sum of the absolute value of those same period deviations.[24] As in the original study, subjects deviated more from conditionally optimal consumption paths in the *Borrowing* condition than in the *Saving* condition (i.e., rounds 1-2 for *Borrowing First* and 3-4 for *Savings First* as seen in Fig. 6). In turn, *Measure 3* is the lifetime sum of the period utility losses between observed consumption and optimal consumption at ex-ante (start of a round) optimal wealth levels.

In an additional preregistered analysis, we controlled for the impact of *risk-taking* and *patience*, and found that the results remained qualitatively unchanged: treatment has significant effects on *Measure 1* and *Measure 2*, but not on *Measure 3*.[25]

Table 10
Meissner (2016) Replication: Sub-Optimal Consumption Paths.

		round 1	round 2	round 3	round 4
median (m1)	BF	303.35	311.87	-107.55	-93.18
	SF	-120.10	-99.11	342.19	314.18
mean (m1)	BF	152.53	190.96	-95.24	-63.49
	SF	-135.78	-104.80	307.66	316.12
p-value		<0.001	<0.001	<0.001	<0.001
median (m2)	BF	387.54	389.32	235.53	240.37
	SF	192.50	200.64	372.84	334.04
mean (m2)	BF	514.69	520.90	362.06	348.55
	SF	323.59	327.94	444.65	405.65
p-value		<0.001	<0.001	<0.001	<0.001
median (m3)	BF	252.73	265.50	195.66	179.39
	SF	118.27	151.92	238.23	203.78
mean (m3)	BF	>100,000	>100,000	>100,000	>100,000
	SF	>100,000	>100,000	>100,000	>100,000
p-value		<0.001	<0.001	0.1164	0.9488

Notes: Deviations and absolute deviations from conditional optimal consumption, following the original study's *m1* and *m2*, respectively; and utility losses from deviations from unconditional optimal consumption *(m3)* at the subject X round level. *BF* and *SF* are Borrowing First and Saving First treatment conditions. *P-values* are calculated for Mann-Whitney-U tests of difference of means between both treatments. $N = 278$.

In bivariate analyses with Mann-Whitney U tests, reported in Table 10, we found that *Measure 1* and *Measure 2* differ statistically significantly between treatments in all rounds (effect size – in the first round – 0.470 and 0.412; statistical power (5% level) 0.973 and 0.916 for *Measure 1* and *2* respectively). In the original study, *Measure 1* was statistically significant in all rounds, and *Measure 2* was significant in three of the six rounds (5% level). Thus, deviations from conditionally optimal consumption paths are higher for the *Borrowing* condition than for the *Savings* condition, regardless of the within-subject order of both conditions. This lends supports to the debt-aversion hypothesis, as subjects are less willing to borrow from the future to consume now than to save from the present to consume in the future in order to smooth consumption. The utility loss from the deviation from the unconditionally optimal consumption path (*Measure 3*) is significant only for the first two rounds before the switch of the conditions, making it resemble the results from the original study, in which it was significant only for the first three rounds before the switch. In the Appendix, Tables A5 and A6 show a similar analysis over different socio-demographic subsamples of subjects for *Measure 1* and *2*, respectively.

panel vendor. After the experimental data collection had been live for less than one day, and only ten subjects had completed the experiment, we suspended data collection, discarded these observations altogether, and restarted data collection the following day with a relaxed restriction to allow two initial mistakes in a first attempt at the quiz while maintaining the requirement of no mistakes in a second attempt; see Subsection 3.1.

[24] This implies that Measure 1 and Measure 2 recalculate the optimal consumption path for the remaining periods of each round, for each subject, considering both the past income path and the previous decisions the subject already made in previous periods of that round.

[25] ANOVA analysis was used in the additional analysis. The independent variables in ANOVA include treatment dummy (if *Borrowing First*), the condition (if *Borrowing*), the risk-taking choice, the delay choice (patience), and the interaction between the treatment and risk-taking choice, treatment and delay choice, condition and risk-taking choice, and condition and delay choice.

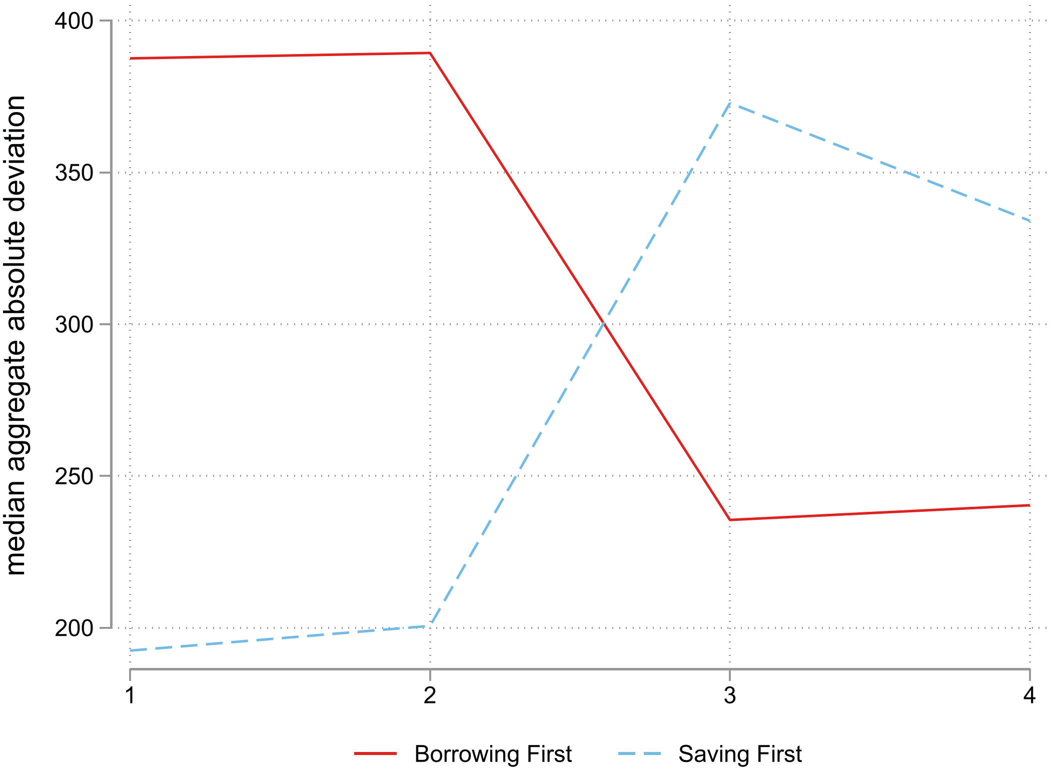

Fig. 6. Meissner (2016) Replication: Sub-Optimal Consumption. *Notes:* Medians of *Measure 2* (mean absolute deviation of consumption from optimal path at each round, per subject × round) by treatment condition. Borrowing (saving) first subjects play rounds 1 and 2 in the borrow (saving) condition; and rounds 3 and 4 in the saving (borrow) condition.

3.6. Replication results of Blaufus and Milde (2021)

For this replication, we were interested in the main treatment effects of different but economically equivalent taxation regimes on retirement savings decisions. The experiment consisted of a "working" phase and a "rest" phase. During the working phase, subjects decided between saving and spending. Each round had ten working periods (with fixed wages) and five resting periods. Each subject completed two rounds in a treatment condition that did not change for these first two rounds. The treatment conditions varied the taxation regime for savings. In *Immediate* taxation, subjects paid income taxes immediately, but their savings were tax-free upon withdrawal during retirement. In *Deferred* taxation, subjects did not pay income taxes on their savings (they got a tax rebate from income taxes) but were taxed later when they withdrew savings during retirement. Finally, in the *Matching* condition, subjects received matching contributions to their savings and paid taxes later, upon withdrawal, during retirement. The balance in all savings accounts earned an interest of 5% per period, with interest earned being taxable or tax-exempt according to the tax rule applied to the principal amount of savings. Withdrawals after retirement were automatically calculated and made equal for all periods of the rest phase.[26]

Subjects completed two rounds, and were compensated based on their consumption decision in one randomly chosen period of one round, to incentivize them to smooth their consumption. As the three treatment conditions yield economically-equivalent returns on savings, they should command equal after-tax effective savings rates.

To simplify the experimental design, we removed an attention check and reassurance screen of tax return filings and integrated the projections of retirement income directly into the main interface screen. Further, we replaced the real effort task generating income in the working phase (a time-consuming transcribing task requiring printed handouts) with a simplified version of the Gill and Prowse (2012) sliders task. In terms of control variables, we retained age and gender, but used our own risk-taking measure for identification of *High risk-taking* subjects taking the 75th percentile cut-off here from the original study. Furthermore, we used our measure of *financial training*as a replacement for the original study measure of financial knowledge. Due to session time constraints, we did not collect information on tax aversion or procrastination.

As in the original study, our main dependent variables were *savings rate* (naïve rates compared to wages) and *effective savings rate* (which accounts for the different taxation regimes on withdrawal). With the tax rate ζ, the (naïve) savings rate for all treatment conditions is defined as $\left(\frac{savings}{wage(1-\zeta)}\right)$. The effective savings rate that makes the (after-tax) withdrawals economically equivalent to those in the *Immediate* condition is defined as $\left(\frac{savings}{wage(1-\zeta)}\right) \times (1-\zeta)$ for the *Deferred* condition. With the matching contribution rate ϕ, for the *Matching* condition, the effective savings rate is defined as $\left[\frac{savings(1+\phi)}{wage(1-\zeta)}\right] \times (1-\zeta)$.

For our replication, we collected 522 valid responses (306 in the original study), of which 182 in the *Immediate* treatment condition, 162 in *Deferred*, and 178 in *Matching* (in the original study, 104, 105 and 97, respectively).

We first calculated the unconditional means of the compatible savings rates across treatments, with 95% confidence intervals (see Fig. 7). As in the original study, we observed that the savings rates

[26] Interest was still paid on the savings balance during retirement, and accrued interest was considered when calculating the fixed withdrawal amount for all rest periods.

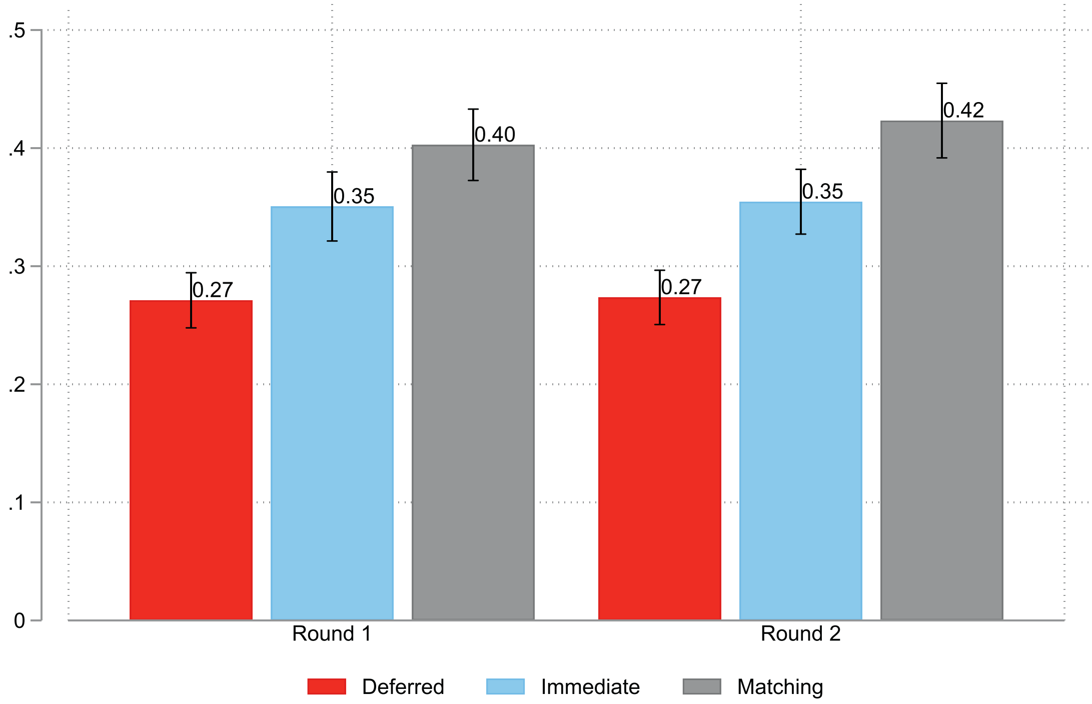

Fig. 7. Blaufus and Milde (2021) Replication: Average Savings Rates (95% confidence interval). *Notes:* Direct (total) saving rates used for *Immediate* condition, and effective savings rates for *Deferred* and *Matching*, per round.

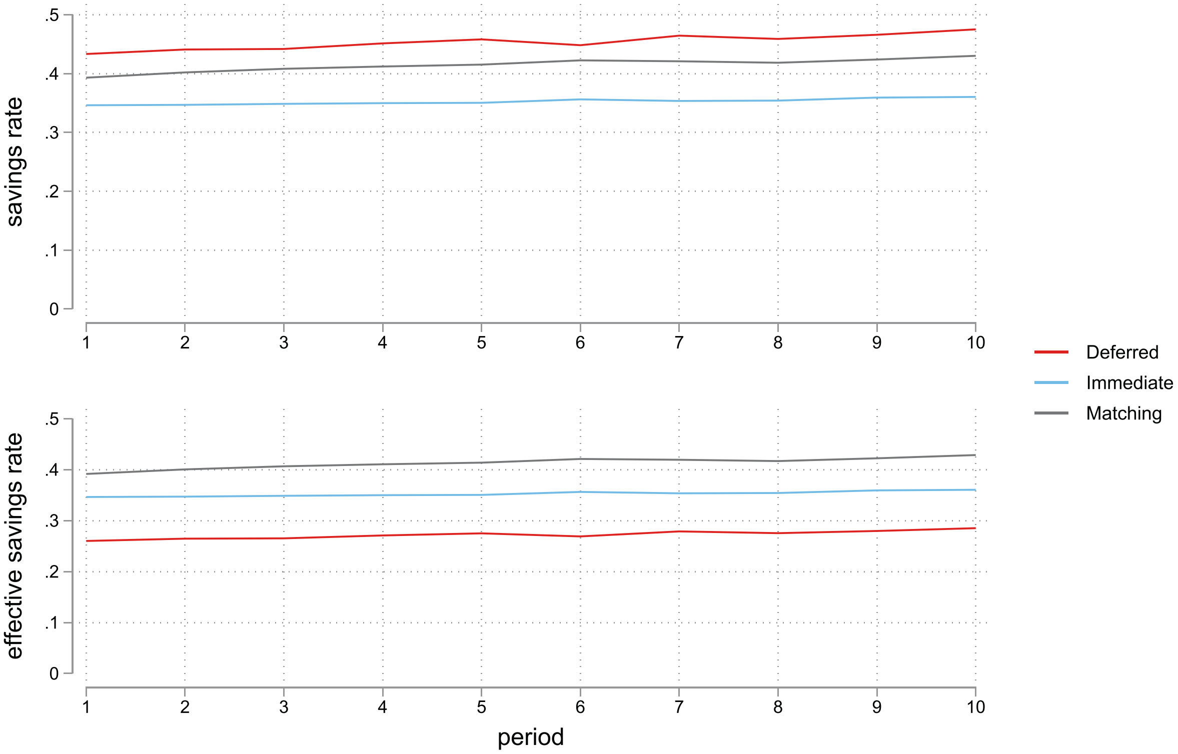

Fig. 8. Blaufus and Milde (2021) Replication: Savings Persistence. *Notes:* Average (effective) saving rates per period across rounds.

did not change significantly between the first and second round, and *Immediate* savings rates were higher than *Deferred* effective savings rate.

Both savings measures were reasonably stable over periods, as their aggregate levels per period and round show in Fig. 8.

Following the analysis of the original study, we regressed savings rates and effective savings rates, observed at the subject × period × round level,[27] on the binary indicators of treatment and the aforementioned covariates. The results of the estimation are presented in Table 11. All models include subjects of the *Immediate*

[27] Therefore, we have 10 observations per subject per round, covering its working periods.

Table 11
Blaufus and Milde (2021) Replication: Drivers of Saving Behaviour.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	SR	SR	SR	ESR	ESR	SR	SR	SR	ESR	ESR
sequence	1	1	2	1	2	1	1	2	1	2
Deferred	0.101***	0.092***	0.096***	-0.086***	-0.085***					
	(0.025)	(0.024)	(0.024)	(0.018)	(0.018)					
Matching						0.054*	0.052*	0.069**	0.050*	0.067**
						(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
Period	0.003**	0.003**	0.003**	0.002**	0.002*	0.003***	0.003***	0.002*	0.003***	0.002*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
High age		0.103***	0.034	0.078***	0.030		0.043	0.022	0.043	0.022
		(0.027)	(0.027)	(0.021)	(0.021)		(0.023)	(0.024)	(0.023)	(0.023)
Male		-0.050*	-0.048	-0.040*	-0.036		-0.019	-0.019	-0.019	-0.019
		(0.025)	(0.025)	(0.020)	(0.020)		(0.023)	(0.023)	(0.023)	(0.023)
Financial training		-0.008	0.019	0.002	0.020		0.005	0.019	0.005	0.019
		(0.027)	(0.028)	(0.022)	(0.022)		(0.024)	(0.026)	(0.024)	(0.026)
High risk-taking		0.116***	0.085**	0.089***	0.066**		0.088***	0.085***	0.088***	0.085***
		(0.028)	(0.028)	(0.022)	(0.021)		(0.026)	(0.025)	(0.026)	(0.025)
Constant	0.335***	0.316***	0.337***	0.323***	0.340***	0.334***	0.314***	0.326***	0.314***	0.326***
	(0.016)	(0.024)	(0.024)	(0.021)	(0.021)	(0.016)	(0.022)	(0.022)	(0.022)	(0.022)
Observations	3,440	3,440	3,440	3,440	3,440	3,600	3,600	3,600	3,600	3,600
Subjects	344	344	344	344	344	360	360	360	360	360
R2	0.0429	0.1298	0.0813	0.1240	0.0835	0.0164	0.0560	0.0578	0.0553	0.0568

Notes: The table presents regression results of random-effects models explaining subject's (effective) savings rates. The savings rate (SR) is defined as the saving amount in a given period divided by the income in this period. The effective savings rate (ESR) is the savings rate multiplied by $(1 - tax\ rate)$. *Deferred* is a dummy variable equal to one if the observation belongs to the deferred-tax treatment. *Matching* is a dummy variable equal to one if the observation belongs to the matching treatment. *Male* is a dummy variable equal to one if the subject is male. *High age* and *High risk-taking* is a dummy variable taking the value of one if the subject's answer to the underlying question is above the 75th percentile of all the observations. *Period* is a time variable equal to decision period in each sequence (from 1 to 10). *Financial training* is a dummy variables taking the value of one if subjects state that they had participated in courses on financial decision-making. Standard errors clustered at subject level are reported in parentheses. *R2* is the R-squared for overall model. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

treatment. For treatment contrasts, models (1-5) include *Deferred* subjects only, while models (6-10) add *Matching* subjects only.

Both treatment coefficients are statistically significant in all estimation models, and the magnitudes of our estimated coefficients are similar to those of the original study. In our replication, both the *Deferred* and the *Matching* savings schemes increased the base savings rate from the *Immediate* condition (models 1-3 and 6-8 in Table 11). In model (2), the base savings rate of the *Deferred* subjects was on average 9.2 percentage points higher than that of the *Immediate* subjects in the first round. In model (8), the base savings rate of the *Matching* subjects was on average 6.9 percentage points higher than that of the *Immediate* subjects. Tax rebates and matching contributions appeared to attract savings in nominal terms, as in the original study.

However, this comparison of base savings rates ignores the fact that, in both *Deferred* and *Matching* conditions, withdrawals will be taxed, whereas *Immediate* withdrawals are tax-exempt. Like the original study, our analysis of effective savings rates shows that the economically equivalent savings rate of the *Deferred* subjects is on average 8.6 percentage points lower than that of the *Immediate* subjects (see model (4) in Table 11). However, the effective savings rate of the *Matching* subjects is on average 5.0 percentage points higher than that of the *Immediate* subjects (see model (9) in Table 11). In other words, the *Matching* contribution tax regime generates higher average post-tax net pension savings than the baseline *Immediate* taxation scheme. We repeated these analyses for socio-demographic subsamples of subjects and report the results in the Appendix, in Tables A7 and A8 for savings rate and effective savings rate, respectively.

In contrast to the original study, we found that *male* was a significant negative predictor of savings rates in the *Immediate* and *Deferred* treatment group. Furthermore, in our replication, *High Risk Taking*'s coefficient was significant and positive in all specifications, while in the original study, this variable was not statistically significant. Furthermore, we found that *Period* has a positive and significant coefficient in our sample, while in the original study it had a significant negative coefficient. However, the effect magnitude of *Period* is small. In period 10, subjects in our sample would save 2% to 3% more from their income than in period 1. In the original study, savings and effective savings rates decreased over periods.

3.7. Replication results of Koehler, Langstaff and Liu (2015)

This study analyses whether subjects adjust their consumption behaviour to changes in the availability of income over a life-cycle. To evaluate this behaviour, subjects were asked to make decisions over several rounds of multiple periods. Each round had a working phase and a retirement phase. During the working phase, subjects earned a predetermined income, which increased over periods. They decided how much to spend and how much to save in a simple, interest-free cash account. During the retirement phase, income was zero. We focused on the main treatment effect of the relative length of the retirement phase (*Short* or *Long* retirement) to the total life length (in periods). In our replication, out of 16 periods per round, subjects were 'retired' for four periods in the *Short* retirement condition and for eight periods in the *Long* condition. In the original study, each round lasted 24 periods, with *Short* retirement consisting of 6 periods and *Long* retirement of 12 periods. In our study, subjects played two rounds under one condition, then changed to the other for another two rounds, with a random assignment of the starting condition (in the original study, subjects played four rounds, switched conditions and then played another four rounds). The compensation in our replication depended on the spending in one randomly selected period. The original study did not use variable incentives.

In every period, subjects had to pay expenses, which were automatically deducted from their income. At the start of a round, a card deck with the value of all possible expenses for every period was shown. Then, at each period one card was randomly chosen and removed (without replacement), determining the actual expenses of that period. During the working phase, income was always larger than mandatory expenses, such that even subjects who always consumed all their income in all periods would still be able to meet their expenses. In the retirement phase, subjects who did not save enough in the working phase would be unable to meet expenses, or go 'bankrupt.'[28] In the original, non-incentivized study, bankruptcy did not have any further repercussions for the subject; in our replication, however, a bankrupt subject would earn zero variable payoff if a round in which he/she was bankrupt was selected for compensation.[29] Compared to the original design, this bankruptcy penalty strengthens the incentive for subjects to smooth consumption and, at the very least, save enough during the working periods to meet the mandatory expenses known to await them during retirement.

We collected valid responses from 344 subjects (149 in the original study), of whom 166 started the session under the retirement condition *Long* and 178 started under the condition *Short* before switching.

Following the original study, we analysed (1) whether the subjects saved enough for retirement (i.e., whether they made sufficient adjustments in saving in response to the manipulation of retirement length), and (2) whether the subjects smoothed their consumption over periods. With respect to the first question, we found that participants saved more when faced with a *Long* retirement period than when faced with a *Short* retirement period, as in the original study. In ANOVA analyses, the retirement length treatment has a significant effect on retirement savings, with $F(1375) = 1495,\ adjusted\ R^2 = 0.752,\ p < 0.001$ (effect size 0.52, statistical power > 0.999), whereas the original study found $F(1147) = 379,\ adj.\ R^2 = 0.72,\ p < 0.001$. With respect to the second question, we found that consumption smoothing as measured by the variability of spending did not differ between conditions, with $F(1375) = 0.52,\ adj.R^2 = 0.648,\ p = 0.471$ (effect size 0.01, statistical power 0.071). This observation is in contrast to the original study, which found a significantly greater mean spending variation (lower consumption smoothing, on average) in the *Long* condition than in the *Short* condition. These results of our replication do not change qualitatively after controlling for *risk-taking* and *patience*.[30] The observation that consumption smoothing activities do not differ between treatments could be related to the incentives for consumption smoothing that we introduced.

In the Appendix, Table A9, in a non-preregistered analysis, we repeated these analyses for subsamples split according to socio-demographic characteristic of the subjects. We also used this replication to evaluate whether having a sample drawn from the general population, that is on average older and has a lower level of education than the sample of the original study,[31] matters for the significance of the treatment effects. The results suggest that neither age nor education level affects the significance of the main treatment effects.

As in the original study, we did not force subjects to automatically spend all points they had in the last period and allowed them to end a round with points remaining in the savings account.[32] As part of our additional analysis, we investigated the implication of this feature on the subjects' decision behaviour.

Table 12 presents the means of several experimental environmental and decision variables. Lifetime income is fixed at 1620 points and lifetime expenses at 720, leaving a budget of 900 points for lifetime consumption. However, the average observed lifetime spending ranges between 591 and 693 points only. This means that on average, subjects left substantial amounts of savings unspent at the end of their experimental life. We therefore classify subjects into three types according to their lifetime savings and spending pattern: 'bankrupt,' 'endlife non-spenders,' and 'effective planners.' Bankrupt subjects did not save enough to cover the remaining mandatory expenses during retirement. 'Endlife non-spenders' did not spend all their points in the last period of a round, wasting them. All the other subjects are effective planners.

The subjects saved, on average, 52.7% of their income and 72% of their available budget in the first round in the *Long* retirement condition and 36.9% and 57.1%, respectively, in the *Short* condition. In these same first rounds, 13.9% of the subjects went bankrupt, and of those who did, their average deficit was 96 points in condition *Long*. Likewise, 9% of the subjects in the first round *Short* did the same, for an average deficit of 54 points. Furthermore, 68.7% of the subjects in condition *Long* ended the first round with an unspent savings balance (average savings lost of 383 points among those who did), as did 70.2% of the subjects in the condition *Short*. The fraction of subjects who did lose savings by not spending them appears high, but also did not change noticeably between rounds. We do not have original study results to compare the prevalence of this outcome for each type of subject there.

The average savings and consumption paths for each type and for the entire sample are shown in Fig. 9.

Since income increases along the periods during the working phase, while expenses do not, savings and spending are naturally less constrained over time. In both treatments, 'bankrupt' subjects increase spending at a faster rate and save much less than other subjects. They also take too long, on average, to reduce consumption after retirement given their low savings. Subjects who leave unspent savings seem to spend too little (and save too much) throughout the periods, without other obvious decision patterns that might explain why they leave so much unspent savings behind.

4. Discussion and implications for future research

In this section, combining insights from our replications and the current state of various strands of experimental research, we discuss possible implications for future experimental design for studies on individual retirement decision-making. We also highlight the limitations of our replication study and offer a modest suggestion for an agenda for future experimental research on retirement decision-making. Finally, we briefly discuss some policy implications of our findings.

4.1. Replication of modified tasks, task design features, and implementation challenges

We replicated most of the main effects of the five studies we reviewed. We compressed or reduced the scope of the original studies to fit a short session time limit, and we used simplified instructions for online general population samples. These modifications allowed the use of a heterogeneous unassisted online sam-

[28] Mechanically, this was represented by negative involuntary savings forced upon subjects when their savings balance was smaller than the current period's mandatory expenses.

[29] This is implemented to prevent strategic but unwanted behaviour on consumption decisions. For instance, consider a subject who, as periods advance, sees that the random realization of expenses will backload the high expense periods during the retirement phase. This subject could decide to spend more during the lower expenses period, while his budget slack to spend is higher, even while knowing that he would eventually go bankrupt, in order to maximize lifetime spending before bankruptcy.

[30] The results regarding the retirement savings and the spending variability remain the same when the observations with savings left unspent at the end of a round are excluded from the ANOVA analyses.

[31] The average age in our sample is 42 years old, the average age in the original study is 29 years old. The average qualification in our sample is 3.4 (3 is vocational qualification, 4 is Bachelor level), the average qualification in the original study is at a Bachelor level.

[32] However, we informed the subjects about this feature in the instructions. In addition, in the quiz that subjects had to pass before the main task, we tested whether they understood that the payoff would be determined by a randomly chosen period.

Table 12
Koehler et al. (2015) Replication: Decision Constraints and Outcomes.

Condition Round	long 1	long 2	long 3	long 4	short 1	short 2	short 3	short 4
Treatment sequencing	long first	long first	short first	short first	short first	short first	long first	long first
Lifetime income	1620	1620	1620	1620	1620	1620	1620	1620
Lifetime expenses	720	720	720	720	720	720	720	720
Lifetime spending	650	639	688	693	676	678	591	616
Retirement expenses	360	362	367	364	183	180	183	181
Retirement savings	912	969	952	949	609	608	742	707
Saving rate (from income)	0.527	0.561	0.548	0.546	0.369	0.368	0.444	0.424
Saving rate (from budget)	0.720	0.760	0.738	0.738	0.571	0.569	0.684	0.653
Bankruptcy prevalence	0.139	0.072	0.096	0.039	0.09	0.067	0.024	0.018
Undersaving / deficit	-96	-102	-43	-106	-54	-39	-99	-77
Lost savings prevalence	0.687	0.675	0.646	0.663	0.702	0.657	0.717	0.717
Lost savings	383	398	335	319	326	342	434	398
Spending variability	32.5	29.83	30.10	28.10	30.93	29.58	32.43	29.76
Difference savings to previous round		58	344	-3		0	-227	-34
N	166	166	178	178	178	178	166	166

Notes: Treatment sequencing is the subjects' condition in first two rounds. *Lifetime income*, *lifetime expenses* and *retirement expenses* are environmental variables. *Lifetime spending* is the sum of all decisions in all periods. *Retirement savings* is the savings balance after the last period of the working phase. *Saving rates* are the fraction saved from income of discretionary budget at each period. *Bankruptcy rate* is fraction of subjects who did not save enough to cover mandatory expenses in retirement, and *Undersaving/deficit* is the sum of expenses that exceeds retirement savings in all retirement periods for this group of subjects. *Lost saving prevalence* is fraction of subjects who had unspent savings at the end, for whom *Lost savings* is savings left after last period. *Spending variability* is the standard deviation of spending. *Difference savings to previous period* is average change in retirement savings from the previous round.

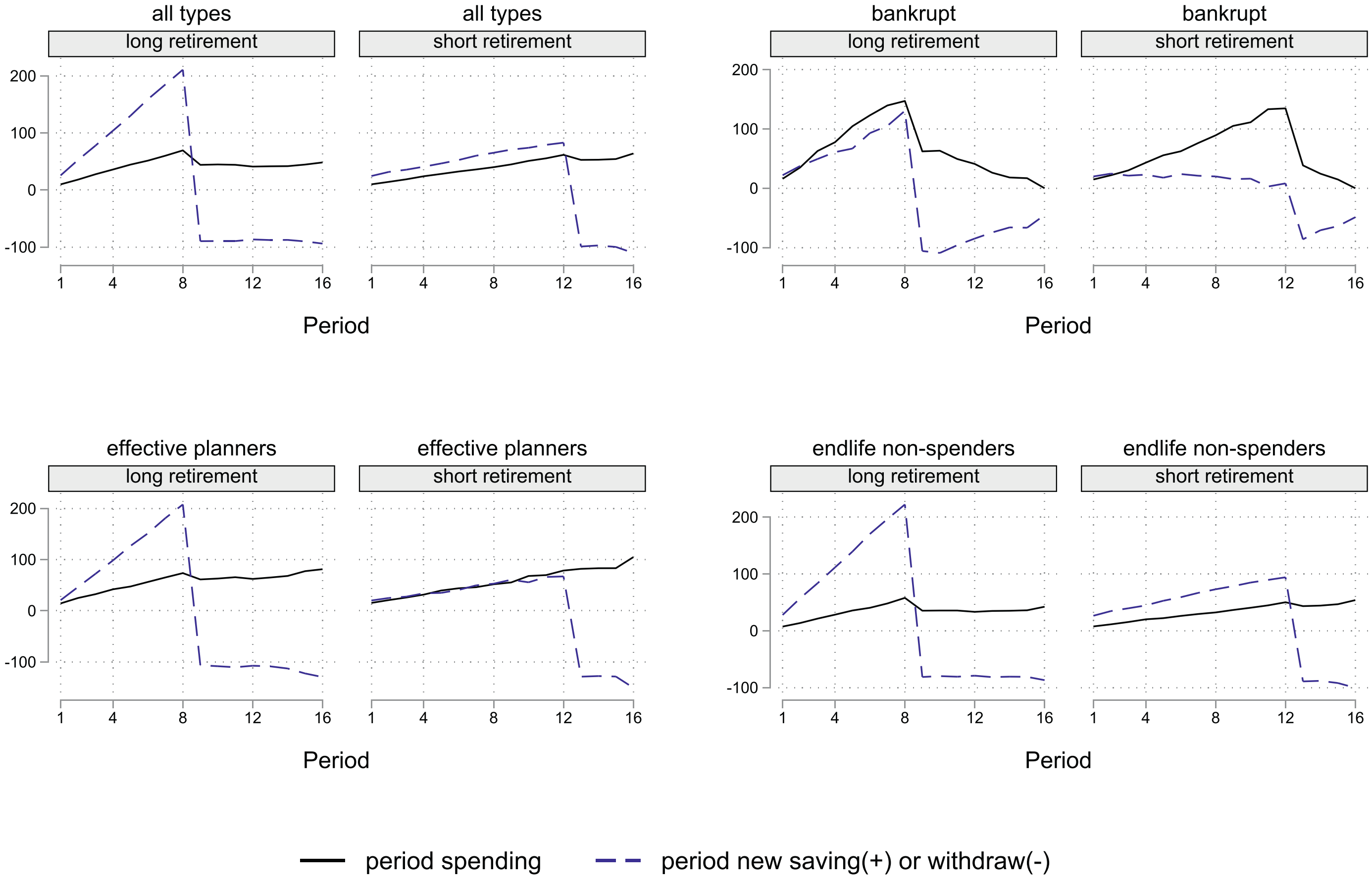

Fig. 9. Koehler et al. (2015) Replication: Spending and Saving per Period. *Notes:* Negative savings are withdraws in the rest phase. *Bankrupt* subject types saved below mandatory expenses, *endlife non-spenders* left unspent savings at last period, and *effective planners* did neither.

ple without yielding excessive noise in the observed results. This highlights the potential for adopting general features of simplified life-cycle experimental tasks, like those we used, in future experimental work, echoing the proposition of Koehler et al. (2015). However, some important considerations and precautions, which we discuss in the following, may be the concern of future experiments.

We observed that in general, subjects' consumption smoothing still is fairly suboptimal, regardless of whether incentives for smoothing are presented in the form of lifetime induced utility or selection of one period per round. With respect to the latter, we did not observe consistent high-stakes gambling behaviour, i.e., subjects did not concentrate consumption or spending in just one period, creating a low chance of a high-value payoff.

One task design feature of concern is to impose a lifetime budget constraint, such that lifetime income matches lifetime consumption (with interest if applicable). In experiments that do not impose the constraint, subjects might spend too little throughout the periods, and leave unspent experimental currency units that are of no value after the end of a round. In particular, underconsumption (or oversaving) in later life periods has been identified in other studies using intertemporal allocation tasks, outside the context of retirement-like decision-making (e.g., Yamamori et al., 2018). Future experiments that impose lifetime budget constraints and then study lifetime outcomes (such as induced utility from spending or consumption in all periods) should look at the impacts of such constraints that self-resolve in the last period. Simultaneous aggregation of lifetime utility from subjects who on the one hand, in violation of a lifetime budget constraint, leave money unspent at the end of a round, and on the other who consume everything before the last period(s) does not allow distinguishing between these different decision-making phenomena. If both groups of subjects are present in a sample, while some concave utility is induced, and the task imposes automatic decisions in the last period to meet a lifetime budget constraint, then aggregated results might not identify such inefficient decisions. Additionally, the estimates of the treatment effects could be downward-biased.

Our strict subject retention criteria eliminated more than half of all subjects initially recruited through our market panel vendor (see Subsection 3.2). Departing from the usual practices, we allowed subjects to proceed immediately from instructions to a practice round and a quiz afterward. We did not pay any compensation (not even a show-up fee) for subjects who did not pass the post-trial quiz. With such procedures, we imposed a minimum engagement that resembles the requirement in an in-person lab session of answering all questions of a quiz correctly before being allowed to proceed. At the same time, we allowed the subjects to revisit the instructions throughout the quiz and all subsequent tasks.

We reduced the number of discrete periods and/or rounds. Such changes did not materially affect the panel structure of the data collected on relevant points. More severe reductions to fewer periods should be implemented with caution to avoid degenerating the natural computational and sequencing complexity present in life cycle optimization decisions (through dynamic programming) in the field or in the laboratory.[33]

Further experiments might help learn the particular impacts of other features on life cycle experimental tasks. These often sidestep any implementation of time-discounting factors across periods, other than interest on savings. Relatively complex utility forms can be imposed through incentive-reward functions. However, we still have limited knowledge of how subjects would react if decisions were measured non-parametrically (as in Abdellaoui et al., 2010), when, for example, longevity uncertainty and changes in the institutional environment are simultaneously introduced into the same task.

4.2. Limitations

First, we replicate studies on different topics related to the retirement decision-making problems, but we cannot jointly evaluate the success of our replications. This is the case because there are not enough studies on any single topic in the experimental retirement decision-making literature. This field of experimental research is still relatively young and encompasses several topics covered by only few studies each, and even then the outcome measures are not clearly defined as to allow for such joint analyses through the possible replication of several studies on the same topic and research question.

Second, our replication study covers only a subset of the topics addressed in experimental studies on retirement decision-making as summarized in Table 1. There were practical and operational restrictions imposed by the use of an online sample from a research panel of the general population, with limited attention, no possibility for interaction between subjects, and for real-time experimenter assistance. We thus were not able to cover other relevant topics, such as social learning and social interactions, that would require experimental tasks unfeasible for deployment in our sample. These restrictions also limited the scope of topics from which we could select studies to replicate, as certain topics had no feasible experiments for replication with our general population online sample.

Third, the characteristics of our sample required adaptation of certain features of the experimental designs of the original studies. Although these adaptations did not prevent the evaluation of the main treatment effects, we can only speculate about the reasons for which we were not able to replicate some of the original results. It is not particularly reasonable, although possible, that certain simplifications of the original experimental designs led to the non-replications that we observed.

Finally, in an effort to use the limited time and attention of our respondents efficiently, we assessed only a set of individual characteristics that is common in all original studies and that we additionally consider as important for the underlying decision problems. It is possible that some other personal characteristics – unrelated to those that we considered in our replications – also have an impact on the main outcome measures.

4.3. Future Research

Our general results suggest that most main treatment effects on individual decisions in the life-cycle can be studied with much simpler task designs, apt for deployment in online samples from the general population. This should open up opportunities for future experimental research that broadens our understanding of possible heterogeneous treatment effects in a more systematic framework, once the simpler designs reduce the hurdles for recruitment of broader and more heterogeneous samples.

Apart from questions related to the experimental findings, future research should consider more systematic studies on specific topics. The overall complexity of life-cycle optimization and the cognitive demands it places on the average person who actually makes retirement decisions should attract more systematic studies on the specific heuristics and rules of thumb adopted by subjects with respect to the different features of those decisions. The use of heuristics in individual decisions and the possible biases embedded in these decisions could extend beyond the issue of whether voluntary retirement savings levels adhere to some normative model of optimal behaviour (as in Benartzi and Thaler, 2007). Winter et al. (2012) showed that utility losses relative to the combined adoption of simple heuristics do not accrue substantially in relation to optimal solutions from a normative perspective of standard intertemporal preferences. There is also some survey evidence (Binswanger and Carman, 2012) implying that engagement with retirement financial preparation through rules of thumb can substitute for strategic planning, producing better outcomes in retirement savings wealth than those who do not adopt any structured approach. The potential of stylized simple rules to improve retirement planning in interaction with different characteristics of retirement decisions should be investigated in more depth.

Furthermore, experimental work should contribute to assessing how individuals break down the complex inputs of decisions (such

[33] Discrete life length of less than 15 periods is uncommon both in the experimental and numerical optimization literature on optimization over the life cycle.

as the annuitization choice) and the interaction between the inputs and other factors that determine decision behaviour in controlled settings. This is necessary since the theoretical or simulation-based literature does not sufficiently agree on what the necessary assumptions are for the unsettled and unsolved annuity puzzle. With simulations, Peijnenburg et al. (2016) questioned some previous assumptions about the attractiveness (or lack thereof) of pension wealth annuitization for many subjects, which implies that normative prescriptions for rational annuitization decisions are less likely to break down than in the earlier work of Davidoff et al. (2005) or Brown et al. (2008). More experiments are needed to simultaneously implement several key features of the annuitization decision. This could allow descriptive models to emerge and better explain whether, why, and to what extent subjects should (or should not) annuitize their pension wealth.

4.4. Policy implications

Our results, taken together, suggest that individuals have limitations in their capacity to solve dynamic programming problems even in stylized and simplified form as in our replications. In the field, these decision problems are much more complex and, for the most part, do not allow subjects to learn from their own mistakes.

In particular, pension reforms over the last two decades have often focused on increasing individual control over certain financial choices in retirement, relaxing compulsory elements, creating opt-outs, and introducing flexible financial arrangements. A large empirical literature evaluates their impact (see Gough and Niza, 2011, for an overview). Our results show the evidence that the systematic patterns in retirement decisions lead to suboptimal outcomes in the behaviour of the general population. These patterns and the observation that the participants of the general population are not sensitive enough to changes of the decision environment should be considered when designing pension reforms.

5. Conclusion

Individual retirement financial decisions are complex, which makes them prone to magnification of biases and cognitive mistakes with adverse effects on the decision outcomes. The suboptimal outcomes are likely persistent, since retirement saving decisions also offer limited learning opportunities due to long lags between the moment of a decision and its outcome. Experimental research on retirement decisions and on how heterogeneous individuals engage in these decisions is therefore acutely needed to advance our understanding of many empirical field outcomes that cannot be easily reconciled with theoretical normative models addressing these decisions.

To that end, we redesigned four experimental studies, each addressing different topics and incorporating different features of the retirement decision problem, and attempted to replicate their main findings. We used reduced-scope tasks and/or a simplified decision environment to make the tasks suitable for implementation with online samples of a general adult population in incentivized settings. We replicated most of the main effects of the original studies we selected for this exercise, which might raise the external validity of the findings.

Finally, we note that limitations remain in the extant simulation-based and field empirical literature on several topics concerning retirement decision-making. These present opportunities for a promising future agenda for experimental research.

Declaration of Competing Interest

None for all the authors.

CRediT authorship contribution statement

Kremena Bachmann: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Andre Lot:** Methodology, Investigation, Writing – original draft, Writing – review & editing. **Xiaogeng Xu:** Software, Formal analysis, Investigation, Data curation, Writing – review & editing. **Thorsten Hens:** Supervision, Funding acquisition.

Data availability

Data, analysis code, screenshot of experiments' interfaces and other materials are available in the Online OSF repository at https://osf.io/jbkwz/.

Appendix A. Main Effects and Socio-Demographic Characteristics

Table A1
Summary of main effects in socio-demographic subsamples.

		Anderhub *et al.* (2000)				Fatas *et al.* (2007)	
		Fulfil Condition 1		Fulfil Condition 2		Choice of retirement timing	
		Product	Summation	Product	Summation	Lump>Annuity	Combined>Annuity
Full sample		✗	✗	✗	✗	√	√
Age	≤=35	✗	✗	✗	✗	✗	✗
	>35 & ≤=50	✗	✗	✗	✗	√	✗
	>50	✗	✗	✗	✗	√	√
Gender	Non-female	✗	✗	✗	✗	√	√
	Female	✗	✗	✗	✗	√	✗
High education	No	✗	✗	✗	✗	√	√
	Yes	✗	✗	✗	✗	✗	✗
Income (Euro)	<2000	✗	✗	✗	✗	√	√
	2000 to 3200	✗	✗	✗	✗	✗	✗
	>3200	✗	✗	✗	✗	√	√
Financial training	No	✗	✗	✗	✗	√	√
	Yes	✗	✗	✗	✗	✗	✗

		Koehler *et al.* (2015)		Meissner (2016)		Blaufus & Milde (2021)	
		Retirement savings	Spending variablility	*m1*	*m2*	Savings rate (Effective savings rate)	
		Long>Short	Long>Short	BF>SF	BF>SF	Def.> (<)Imme.	Mat.> (>)Imme.
Full sample		√	✗	√	√	√(√)	√(√)
Age	<=35	√	✗	√	√	√(√)	✗(✗)
	>35 & <=50	√	✗	√	√	✗(√)	✗(✗)
	>50	√	✗	√	√	√(√)	√(√)
Gender	Non-female	√	✗	√	√	√(√)	√(√)
	Female	√	✗	√	√	√(√)	✗(✗)
High education	No	√	✗	√	√	√(√)	√(√)
	Yes	√	✗	√	√	√(√)	✗(✗)
Income (Euro)	<2000	√	✗	√	√	√(✗)	✗(✗)
	2000 to 3200	√	✗	√	√	√(√)	✗(✗)
	>3200	√	✗	√	√	✗(√)	√(√)
Financial training	No	√	✗	√	√	√(√)	√(√)
	Yes	√	√	√	√	✗(√)	✗(✗)

Notes: For the replication of Anderhub *et al.* (2000), Condition 1 $((\frac{x_2}{S_2}|\neg[1/6]) > (\frac{x_2}{S_2}|\neg[1/3]) > (\frac{x_2}{S_2}|\neg[1/2]))$ indicates that the spending is larger when a card deck with a low termination probability is removed than a card deck with a high termination probability is removed. Condition 2 $((\frac{x_3}{S_3}|[1/2]) > (\frac{x_3}{S_3}|[1/3]) > (\frac{x_3}{S_3}|[1/6]))$ indicates that the spending is larger when a card deck with a high termination probability finally stays than a deck with a low termination probability stays. For the replication of Fatas *et al.* (2007), the larger the choice of retirement timing is, the later a subject chooses to retire. For the replication of Koehler *et al.* (2015), the retirement savings is the savings balance after the last period of the working phase. The spending variability is the standard deviation of spending. For the replication of Meissner (2016), *m1* and *m2* are the deviations and absolute deviations from conditional optimal consumption, respectively, and the results are from Round 1. For the replication of Blaufus & Milde (2021), *Def.* is treatment Deferred, *Imme.* is treatment Immediate, and *Mat.* is treatment Matching. The savings rate is defined as the saving amount divided by the income in a period and the effective savings rate is defined as the savings rate multiplied by $(1 - tax\ rate)$. √ indicates that an effect is confirmed in the full sample or a subsample, and ✗ indicates that an effect is not found.

Table A2
Anderhub et al. (2000) Replication: Reactions to the first removed card deck and finally stayed card deck in socio-demographic subsamples.

		Treatment	Mean consumption share in period 2 ($\frac{x_2}{S_2}$)			Mean consumption share in period 3 ($\frac{x_3}{S_3}$)			Obs.	Obs. fulfilling Condition 1	Obs. fulfilling Condition 2
			$(\frac{x_2}{S_2}\|\neg[1/6])$	$(\frac{x_2}{S_2}\|\neg[1/3])$	$(\frac{x_2}{S_2}\|\neg[1/2])$	$(\frac{x_3}{S_3}\|[1/6])$	$(\frac{x_3}{S_3}\|[1/3])$	$(\frac{x_3}{S_3}\|[1/2])$			
Age (years old)	≤35	Product	0.34	0.31	0.32	0.43	0.45	0.41	67	13	12
		Summation	0.32	0.32	0.32	0.47	0.42	0.44	60	13	9
	>35 & ≤50	Product	0.34	0.35	0.39	0.49	0.41	0.43	46	4	7
		Summation	0.35	0.36	0.34	0.46	0.44	0.45	17	4	4
	>50	Product	0.32	0.35	0.36	0.42	0.38	0.37	63	2	10
		Summation	0.37	0.37	0.36	0.40	0.39	0.41	86	22	11
Gender	Non-female	Product	0.32	0.31	0.32	0.40	0.40	0.38	102	14	18
		Summation	0.36	0.35	0.34	0.44	0.42	0.46	79	20	16
	Female	Product	0.35	0.37	0.40	0.49	0.44	0.43	74	5	11
		Summation	0.34	0.35	0.34	0.43	0.40	0.39	84	19	8
High education	No	Product	0.33	0.33	0.35	0.44	0.42	0.40	115	14	17
		Summation	0.34	0.36	0.34	0.43	0.41	0.43	101	21	14
	Yes	Product	0.34	0.36	0.36	0.43	0.43	0.41	54	4	11
		Summation	0.36	0.35	0.35	0.44	0.41	0.42	59	16	10
Income (Euro)	<2000	Product	0.36	0.34	0.33	0.42	0.42	0.40	63	9	10
		Summation	0.35	0.36	0.33	0.44	0.40	0.41	57	11	7
	2000 to 3200	Product	0.33	0.33	0.38	0.45	0.44	0.42	50	5	11
		Summation	0.34	0.37	0.34	0.42	0.38	0.42	44	11	7
	>3200	Product	0.33	0.34	0.36	0.46	0.41	0.39	53	5	5
		Summation	0.37	0.35	0.37	0.46	0.45	0.47	51	14	9
Financial training	No	Product	0.34	0.34	0.36	0.45	0.43	0.40	130	15	22
		Summation	0.35	0.35	0.33	0.44	0.43	0.45	118	30	19
	Yes	Product	0.32	0.32	0.33	0.40	0.40	0.39	41	2	7
		Summation	0.36	0.36	0.38	0.43	0.37	0.37	41	9	4

Notes: The mean consumption share is computed from all the subjects in period 2 ($\frac{x_2}{S_2}$) and 3 ($\frac{x_2}{S_2}$) correspondingly. For each subject, there are two out of the six rounds where the first removed card decks (the finally stayed card deck) have the same termination probability. The consumption share in period 2 (period 3) for each subject is the mean of the shares of the two rounds with the same termination probability of first removed card deck (finally stayed card deck). Condition 1 refers to $(\frac{x_2}{S_2}|\neg[1/6]) > (\frac{x_2}{S_2}|\neg[1/3]) > (\frac{x_2}{S_2}|\neg[1/2])$, implying that the spending is larger when a card deck with a low termination probability is removed than a card deck with a high termination probability is removed. Condition 2 refers to $(\frac{x_3}{S_3}|[1/2]) > (\frac{x_3}{S_3}|[1/3]) > (\frac{x_3}{S_3}|[1/6])$, implying that the spending is larger when a card deck with a high termination probability finally stays than a deck with a low termination probability stays.

Table A3
Anderhub et al. (2000) Replication: Statistical tests on the reactions to the first removed card deck and finally stayed card deck in socio-demographic subsamples.

		Treatment	*P*-value of *t*-test						Obs.
			Null hypothesis for Condition 1			Null hypothesis for Condition 2			
			$\left(\frac{x_2}{S_2}\|\neg[1/6]\right) \le \left(\frac{x_2}{S_2}\|\neg[1/2]\right)$	$\left(\frac{x_2}{S_2}\|\neg[1/3]\right) \le \left(\frac{x_2}{S_2}\|\neg[1/2]\right)$	$\left(\frac{x_2}{S_2}\|\neg[1/6]\right) \le \left(\frac{x_2}{S_2}\|\neg[1/3]\right)$	$\left(\frac{x_3}{S_3}\|[1/6]\right) \ge \left(\frac{x_3}{S_3}\|[1/2]\right)$	$\left(\frac{x_3}{S_3}\|[1/3]\right) \ge \left(\frac{x_3}{S_3}\|[1/2]\right)$	$\left(\frac{x_3}{S_3}\|[1/6]\right) \ge \left(\frac{x_3}{S_3}\|[1/3]\right)$	
Age (years old)	≤35	Product	0.239	0.610	0.162	0.288	0.109	0.749	67
		Summation	0.385	0.424	0.460	0.232	0.695	0.108	60
	>35 & ≤50	Product	0.956	0.913	0.639	0.064	0.612	**0.036**	46
		Summation	0.466	0.331	0.637	0.457	0.558	0.400	17
	>50	Product	0.900	0.654	0.813	0.072	0.318	0.161	63
		Summation	0.448	0.394	0.554	0.526	0.655	0.369	86
Gender	Non-female	Product	0.453	0.688	0.271	0.152	0.173	0.465	102
		Summation	0.329	0.383	0.442	0.719	0.874	0.285	79
	Female	Product	0.965	0.856	0.773	**0.030**	0.398	0.051	74
		Summation	0.498	0.381	0.617	0.131	0.424	0.176	84
High education	No	Product	0.856	0.912	0.385	**0.036**	0.233	0.141	115
		Summation	0.480	0.289	0.694	0.455	0.749	0.217	101
	Yes	Product	0.730	0.517	0.715	0.241	0.287	0.443	54
		Summation	0.357	0.504	0.353	0.351	0.631	0.237	59
Income (Euro)	<2000	Product	0.218	0.354	0.343	0.260	0.302	0.451	63
		Summation	0.300	0.177	0.657	0.181	0.522	0.167	57
	2000 to 3200	Product	0.950	0.936	0.551	0.175	0.328	0.311	50
		Summation	0.461	0.211	0.760	0.541	0.787	0.244	44
	>3200	Product	0.831	0.752	0.609	**0.024**	0.257	0.094	53
		Summation	0.440	0.716	0.235	0.557	0.626	0.430	51
Financial training	No	Product	0.892	0.843	0.593	**0.030**	0.175	0.173	130
		Summation	0.216	0.154	0.593	0.672	0.813	0.329	118
	Yes	Product	0.656	0.685	0.468	0.382	0.405	0.476	41
		Summation	0.654	0.641	0.514	0.087	0.495	0.089	41

Notes: The mean consumption share is computed from all the subjects in period 2 ($\frac{x_2}{S_2}$) and 3 ($\frac{x_2}{S_2}$) correspondingly. For each subject, there are two out of the six rounds where the first removed card decks (the finally stayed card deck) have the same termination probability. The consumption share in period 2 (period 3) for each subject is the mean of the shares of the two rounds with the same color of first removed card deck (finally stayed card deck). Condition 1 refers to $(\frac{x_2}{S_2}|\neg[1/6]) > (\frac{x_2}{S_2}|\neg[1/3]) > (\frac{x_2}{S_2}|\neg[1/2])$, implying that the spending is larger when a card deck with a low termination probability is removed than a card deck with a high termination probability is removed. Condition 2 refers to $(\frac{x_3}{S_3}|[1/2]) > (\frac{x_3}{S_3}|[1/3]) > (\frac{x_3}{S_3}|[1/6])$, implying that the spending is larger when a card deck with a high termination probability finally stays than a deck with a low termination probability stays. The *p*-values in bold and blue font are those that are smaller than 0.05.

Table A4
Fatas et al. (2007) Replication: Timing of Retirement Treatment Effects in Socio-demographic Subsamples.

	Age (years old)			Gender		High education		Income (Euro)			Financial training	
	≤35	>35 & ≤50	>50	Non-female	Female	No	Yes	<2000	2000 to 3200	>3200	No	Yes
Risk-taking	0.023**	0.021*	0.029***	0.027***	0.024***	0.024***	0.028***	0.009	0.010	0.036***	0.024***	0.034***
	(0.009)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.011)	(0.011)	(0.007)	(0.006)	(0.009)
Patience	0.235	0.436**	0.058	0.171	0.236*	0.186	0.194	0.155	0.352	0.131	0.242*	0.076
	(0.140)	(0.160)	(0.117)	(0.115)	(0.110)	(0.110)	(0.115)	(0.158)	(0.182)	(0.117)	(0.096)	(0.145)
Annuity	-0.682	-1.044*	-0.923*	-0.853*	-0.998**	-1.330***	-0.562	-1.499**	-0.845	-0.805*	-0.976**	-0.753
	(0.436)	(0.463)	(0.378)	(0.363)	(0.337)	(0.349)	(0.350)	(0.510)	(0.520)	(0.376)	(0.297)	(0.456)
Combined	-0.083	-0.660	-0.118	-0.004	-0.602	-0.224	-0.357	-0.483	-1.506**	0.030	-0.325	-0.196
	(0.433)	(0.495)	(0.373)	(0.370)	(0.331)	(0.334)	(0.376)	(0.493)	(0.566)	(0.368)	(0.306)	(0.437)
Constant	4.567***	4.753***	5.426***	5.095***	5.034***	5.132***	5.073***	5.645***	5.203***	5.039***	5.080***	4.952***
	(0.495)	(0.557)	(0.400)	(0.437)	(0.344)	(0.380)	(0.403)	(0.529)	(0.565)	(0.422)	(0.323)	(0.553)
(Annuity–Combined)	-0.598	-0.384	-0.805*	-0.849*	-0.396	-1.106**	-0.205	-1.016*	0.661	-0.835*	-0.651*	-0.557
	(0.443)	(0.494)	(0.372)	(0.359)	(0.345)	(0.350)	(0.362)	(0.498)	(0.560)	(0.385)	(0.300)	(0.461)
R2	0.133	0.130	0.076	0.086	0.111	0.104	0.093	0.076	0.111	0.147	0.088	0.128
Prob. >F	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.020	0.000	0.000	0.001
Observations	114	154	262	286	243	284	246	131	103	234	387	139

Notes: The results are from OLS estimations. The dependent variable is the mean retirement period chosen in the three rounds. *High education* means Bachelor, Master or Doctoral degree. *Income* means the monthly household disposable income. *Annuity* and *Combined* are dummies for subjects assigned to such treatment conditions; *Lump sum* is the baseline. *Risk-taking* is the decision in the risk taking task at the end of the survey where the subjects chose how many percentage points (0-100) of their earnings they would like to put into a lotto. *Patience* is the decision at the end of the survey where the subjects decided how much they were willing to delay the payment to earn interest and equal to 1, 2, 3 and 4 for the choice of no delay, 1 month, 2 months and 3 months, respectively. Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A5
Meissner (2016) Replication: Sub-Optimal Consumption Paths (Measure 1) in Socio-demographic subsamples.

				round 1	round 2	round 3	round 4	Obs.
Age (years old)	≤35	mean (m1)	BF	201.65	188.98	-61.75	13.38	57
			SF	-147.82	-95.37	320.52	303.33	60
		p-value		<0.001	<0.001	<0.001	<0.001	
	>35 & ≤50	mean (m1)	BF	61.98	71.42	-344.87	-220.76	40
			SF	-7.73	97.46	404.02	378.25	23
		p-value		0.003	0.013	<0.001	<0.001	
	>50	mean (m1)	BF	168.98	288.83	66.28	-25.31	50
			SF	-182.09	-213.50	245.41	302.34	48
		p-value		0.003	0.013	<0.001	<0.001	
Gender	Non-female	mean (m1)	BF	132.30	171.13	-83.13	-81.56	73
			SF	-127.78	-98.66	312.44	300.34	79
		p-value		<0.001	<0.001	<0.001	<0.001	
	Female	mean (m1)	BF	170.52	240.25	-114.40	-57.66	71
			SF	-147.29	-116.68	301.54	342.06	51
		p-value		<0.001	<0.001	<0.001	<0.001	
High education	No	mean (m1)	BF	181.63	229.63	-112.34	-61.36	75
			SF	-128.41	-116.47	305.74	328.43	73
		p-value		<0.001	<0.001	<0.001	<0.001	
	Yes	mean (m1)	BF	112.50	143.92	-76.28	-66.71	69
			SF	-142.27	-87.98	310.04	300.40	57
		p-value		<0.001	<0.001	<0.001	<0.001	
Income (Euro)	<2000	mean (m1)	BF	247.40	210.75	-39.78	-115.63	46
			SF	-173.80	-36.02	407.10	398.98	36
		p-value		<0.001	<0.001	<0.001	<0.001	
	2000 to 3200	mean (m1)	BF	79.49	272.58	-88.32	-11.40	41
			SF	-84.55	-89.88	350.92	327.57	31
		p-value		<0.001	<0.001	<0.001	<0.001	
	>3200	mean (m1)	BF	94.38	89.70	-169.12	-69.72	46
			SF	-149.48	-174.37	242.94	268.68	50
		p-value		<0.001	<0.001	<0.001	<0.001	
Financial training	No	mean (m1)	BF	125.46	190.89	-118.58	-72.56	115
			SF	-123.68	-102.43	308.58	338.58	95
		p-value		<0.001	<0.001	<0.001	<0.001	
	Yes	mean (m1)	BF	240.80	263.57	36.99	-7.83	28
			SF	-177.07	-115.21	297.82	251.92	34
		p-value		<0.001	<0.001	<0.001	<0.001	

Notes: Deviations from conditional optimal consumption, following the original study's *m1. High education* means Bachelor, Master or Doctoral degree. *Income* means the monthly household disposable income. *BF* and *SF* are Borrowing First and Saving First treatment conditions. *P-values* are calculated for Mann-Whitney-U tests of difference of means between both treatments.

Table A6
Meissner (2016) Replication: Sub-Optimal Consumption Paths (Measure 2) in Socio-demographic subsamples.

				round 1	round 2	round 3	round 4	Obs.
Age (years old)	≤35	mean (m2)	BF	510.52	531.84	326.63	364.56	57
			SF	314.38	256.40	387.06	370.01	60
		p-value		<0.001	<0.001	<0.001	0.271	
	>35 & ≤50	mean (m2)	BF	518.07	571.53	481.64	385.62	40
			SF	189.36	287.66	409.68	380.70	23
		p-value		<0.001	<0.001	0.003	0.037	
	>50	mean (m2)	BF	516.74	467.93	306.78	300.63	50
			SF	399.43	436.67	533.38	462.17	48
		p-value		<0.001	<0.001	<0.001	<0.001	
Gender	Non-female	mean (m2)	BF	549.51	526.71	367.45	376.13	73
			SF	301.69	289.76	397.29	387.16	79
		p-value		<0.001	<0.001	<0.001	<0.001	
	Female	mean (m2)	BF	473.24	482.51	334.88	305.95	71
			SF	360.06	390.15	521.98	437.55	51
		p-value		<0.001	<0.001	<0.001	<0.001	
High education	No	mean (m2)	BF	512.03	496.54	407.47	357.75	75
			SF	354.20	388.88	485.55	440.98	73
		p-value		<0.001	<0.001	<0.001	<0.001	
	Yes	mean (m2)	BF	524.92	556.51	320.51	339.47	69
			SF	284.72	251.94	394.58	362.01	57
		p-value		<0.001	<0.001	<0.001	0.028	
Income (Euro)	<2000	mean (m2)	BF	449.93	500.55	288.54	332.57	46
			SF	426.27	368.14	489.96	482.55	36
		p-value		<0.001	<0.001	<0.001	<0.001	
	2000 to 3200	mean (m2)	BF	576.14	526.44	470.50	359.82	41
			SF	223.21	237.29	356.32	336.12	31
		p-value		<0.001	<0.001	0.019	0.049	
	>3200	mean (m2)	BF	505.56	517.98	350.38	363.76	46
			SF	313.77	364.47	446.72	396.15	50
		p-value		<0.001	<0.001	<0.001	0.148	
Financial training	No	mean (m2)	BF	541.19	512.36	371.00	351.32	115
			SF	332.22	329.23	452.80	400.17	95
		p-value		<0.001	<0.001	<0.001	<0.001	
	Yes	mean (m2)	BF	429.68	485.62	303.25	318.44	28
			SF	310.37	336.83	422.65	424.82	34
		p-value		<0.001	<0.001	0.002	0.012	

Notes: Absolute deviations from conditional optimal consumption, following the original study's *m2. High education* means Bachelor, Master or Doctoral degree. *Income* means the monthly household disposable income. *BF* and *SF* are Borrowing First and Saving First treatment conditions. *P-values* are calculated for Mann-Whitney-U tests of difference of means between both treatments.

Table A7
Blaufus and Milde (2021) Replication: Drivers of Saving Behaviour (Savings Rate) in Socio-demographic Subsamples.

Panel A: **Treatment Deferred vs. Immediate**

	Age (years old)			Gender		High education		Income (Euro)			Financial training	
	≤35	>35 & ≤50	>50	Non-female	Female	No	Yes	<2000	2000 to 3200	>3200	No	Yes
Deferred	0.163**	0.055	0.090**	0.092**	0.112**	0.115***	0.082*	0.111*	0.129**	0.068	0.119***	0.034
	(0.051)	(0.047)	(0.033)	(0.031)	(0.040)	(0.032)	(0.038)	(0.049)	(0.043)	(0.041)	(0.029)	(0.052)
Observations	670	950	1,820	2,080	1,360	2,280	1,080	940	1,040	1,240	2,630	730
Subjects	67	95	182	208	136	228	108	94	104	124	263	73
R2	0.1031	0.0177	0.0376	0.0376	0.0500	0.0510	0.0363	0.0471	0.0720	0.0216	0.0568	0.0103

Panel B: **Treatment Matching vs. Immediate**

	Age (years old)			Gender		High education		Income (Euro)			Financial training	
	≤35	>35 & ≤50	>50	Non-female	Female	No	Yes	<2000	2000 to 3200	>3200	No	Yes
Matching	-0.014	0.071	0.072*	0.070*	0.027	0.066*	0.038	0.041	0.034	0.094**	0.072**	0.011
	(0.042)	(0.047)	(0.029)	(0.028)	(0.034)	(0.028)	(0.034)	(0.043)	(0.041)	(0.033)	(0.025)	(0.039)
Observations	720	810	2,070	2,180	1,420	2,390	1,170	980	1,100	1,370	2,740	800
Subjects	72	81	207	218	142	239	117	98	110	137	274	80
R2	0.0033	0.0304	0.0300	0.0273	0.0049	0.0226	0.0090	0.0092	0.0062	0.0511	0.0260	0.0057

Notes: The table presents regression results of random-effects models explaining subject's saving rates. The savings rate (SR) is defined as the saving amount in a given period divided by the income in this period. *Deferred* is a dummy variable equal to one if the observation belongs to the deferred-tax treatment. The other covariates include *Period* and the constant term. *Matching* is a dummy variable equal to one if the observation belongs to the matching treatment. *High education* means Bachelor, Master or Doctoral degree. *Income* means the monthly household disposable income. Standard errors clustered at subject level are reported in parentheses. *R2* is the R-squared for overall model. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A8
Blaufus and Milde (2021) Replication: Drivers of Saving Behaviour (Effective Savings Rate) in Socio-demographic Subsamples.

Panel A: **Treatment Deferred vs. Immediate**

	Age (years old)			Gender		High education		Income (Euro)			Financial training	
	≤35	>35 & ≤50	>50	Non-female	Female	No	Yes	<2000	2000 to 3200	>3200	No	Yes
Deferred	-0.075*	-0.117**	-0.071**	-0.078**	-0.085**	-0.067**	-0.097***	-0.071	-0.067*	-0.093**	-0.065**	-0.129***
	(0.038)	(0.037)	(0.025)	(0.024)	(0.031)	(0.025)	(0.029)	(0.040)	(0.033)	(0.031)	(0.022)	(0.037)
Observations	670	950	1,820	2,080	1,360	2,280	1,080	940	1,040	1,240	2,630	730
Subjects	67	95	182	208	136	228	108	94	104	124	263	73
R2	0.0405	0.0952	0.0354	0.0423	0.0460	0.0291	0.0748	0.0298	0.0320	0.0612	0.0284	0.1183

Panel B: **Treatment Matching vs. Immediate**

	Age (years old)			Gender		High education		Income (Euro)			Financial training	
	≤35	>35 & ≤50	>50	Non-female	Female	No	Yes	<2000	2000 to 3200	>3200	No	Yes
Matching	-0.016	0.069	0.070*	0.069*	0.026	0.064*	0.036	0.040	0.033	0.092**	0.070**	0.010
	(0.041)	(0.047)	(0.029)	(0.028)	(0.034)	(0.028)	(0.034)	(0.043)	(0.041)	(0.033)	(0.025)	(0.039)
Observations	720	810	2,070	2,180	1,420	2,390	1,170	980	1,100	1,370	2,740	800
Subjects	72	81	207	218	142	239	117	98	110	137	274	80
R2	0.0035	0.0293	0.0290	0.0263	0.0044	0.0217	0.0083	0.0087	0.0057	0.0497	0.0250	0.0055

Notes: The table presents regression results of random-effects models explaining subject's effective saving rates. The effective savings rate (ESR) is defined as the saving amount in a given period divided by the income in this period and multiplied by $(1 - tax\ rate)$. *Deferred* is a dummy variable equal to one if the observation belongs to the deferred-tax treatment. The other covariates include *Period* and the constant term. *Matching* is a dummy variable equal to one if the observation belongs to the matching treatment. *High education* means Bachelor, Master or Doctoral degree. *Income* means the monthly household disposable income. Standard errors clustered at subject level are reported in parentheses. *R2* is the R-squared for overall model. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A9
Koehler et al. (2015) Replication: Effects of Retirement Length in Socio-demographic Subsamples.

Panel A: **Retirement Savings**

	Age (years old)			Gender		High education		Income (Euro)			Financial training	
	≤35	>35 & ≤50	>50	Non-female	Female	No	Yes	<2000	2000 to 3200	>3200	No	Yes
F-statistic	*F*(559)= 788	*F*(335)= 366	*F*(479)= 392	*F*(647)= 934	*F*(723)= 615	*F*(859)= 793	*F*(479)= 692	*F*(447)= 390	*F*(339)= 475	*F*(435)= 648	*F*(1051)=1110	*F*(303)= 380
adjusted R^2	0.772	0.760	0.736	0.781	0.729	0.719	0.799	0.675	0.780	0.806	0.749	0.760
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Observations	560	336	480	648	724	860	480	448	400	436	1052	304
Subjects	140	84	120	162	181	215	120	112	100	218	263	76

Panel B: **Variability of Spending**

	Age (years old)			Gender		High education		Income (Euro)			Financial training	
	≤35	>35 & ≤50	>50	Non-female	Female	No	Yes	<2000	2000 to 3200	>3200	No	Yes
F-statistic	*F*(559)=2.96	*F*(335)=0.49	*F*(479)=0.09	*F*(647)=3.19	*F*(723)=0.35	*F*(859)=0.42	*F*(479)=0.28	*F*(447)=0.17	*F*(399)=0.00	*F*(435)=0.45	*F*(1051)=0.01	*F*(303)=4.82
adjusted R^2	0.646	0.649	0.641	0.653	0.645	0.661	0.626	0.572	0.541	0.763	0.658	0.677
p-value	0.086	0.486	0.769	0.075	0.552	0.518	0.599	0.678	1.000	0.504	0.904	0.029
Observations	560	336	480	648	724	860	480	448	400	436	1052	304
Subjects	140	84	120	162	181	215	120	112	100	109	263	76

Notes: The table shows the effects of retirement length treatment on retirement savings and spending variability from ANOVA analyses. *High education* means Bachelor, Master or Doctoral degree. *Income* means the monthly household disposable income.

References

Abdellaoui, M., Attema, A.E., Bleichrodt, H., 2010. Intertemporal Tradeoffs for Gains and Losses: An Experimental Measurement of Discounted Utility. Economic Journal 120 (545), 845–866. doi:10.1111/j.1468-0297.2009.02308.x.

Agnew, J.R., Anderson, L.R., Gerlach, J.R., Szykman, L.R., 2008. Who Chooses Annuities? An Experimental Investigation of the Role of Gender, Framing, and Defaults. American Economic Review 98 (2), 418–422. doi:10.1257/aer.98.2.418.

Agnew, J.R., Anderson, L.R., Szykman, L.R., 2015. An Experimental Study of the Effect of Market Performance on Annuitization and Equity Allocations. Journal of Behavioral Finance 16 (2), 120–129. doi:10.1080/15427560.2015.1034857.

Anderhub, V., Güth, W., Müller, W., Strobel, M., 2000. An experimental analysis of intertemporal allocation behavior. Experimental Economics 3 (2), 137–152. doi:10.1007/BF01669305.

Ballinger, T.P., Hudson, E., Karkoviata, L., Wilcox, N.T., 2011. Saving behavior and cognitive abilities. Experimental Economics 14 (3), 349–374. doi:10.1007/s10683-010-9271-3.

Ballinger, T.P., Palumbo, M.G., Wilcox, N.T., 2003. Precautionary Saving and Social Learning Across Generations: an Experiment. The Economic Journal 113 (490), 920–947. doi:10.1111/1468-0297.t01-1-00158.

Benartzi, S., Thaler, R.H., 2007. Heuristics and Biases in Retirement Savings Behavior. Journal of Economic Perspectives 21 (3), 81–104. doi:10.1257/jep.21.3.81.

Beshears, J., Choi, J.J., Harris, C., Laibson, D., Madrian, B.C., Sakong, J., 2020. Which early withdrawal penalty attracts the most deposits to a commitment savings account? Journal of Public Economics 183, 104144. doi:10.1016/j.jpubeco.2020.104144.

Binswanger, J., Carman, K.G., 2012. How real people make long-term decisions: The case of retirement preparation. Journal of Economic Behavior and Organization 81 (1), 39–60. doi:10.1016/j.jebo.2011.08.010.

Blaufus, K., Milde, M., 2021. Tax Misperceptions and the Effect of Informational Tax Nudges on Retirement Savings. Management Science 67 (8), 5011–5031. doi:10.1287/mnsc.2020.3761.

Bohr, C.E., Holt, C.A., Schubert, A.V., 2019. Assisted savings for retirement: An experimental analysis. European Economic Review 119, 42–54. doi:10.1016/j.euroecorev.2019.05.020.

Brown, A.L., Chua, Z.E., Camerer, C.F., Brown, A.L., 2009. Learning and Visceral Temptation in Dynamic Saving Experiments. Quarterly Journal of Economics 124 (1), 197–231. doi:10.1162/qjec.2009.124.1.197.

Brown, J.R., Kling, J.R., Mullainathan, S., Wrobel, M.V., 2008. Why Don't People Insure Late-Life Consumption? A Framing Explanation of the Under-Annuitization Puzzle. American Economic Review 98 (2), 304–309. doi:10.1257/aer.98.2.304.

Browning, M., Crossley, T.F., 2001. The Life-Cycle Model of Consumption and Saving. Journal of Economic Perspectives 15 (3), 3–22. doi:10.1257/jep.15.3.3.

Carbone, E., 2005. Demographics and Behaviour. Experimental Economics 8 (3), 217–232. doi:10.1007/s10683-005-1464-9.

Carbone, E., 2006. Understanding intertemporal choices. Applied Economics 38 (8), 889–898. doi:10.1080/00036840500399313.

Carbone, E., Duffy, J., 2014. Lifecycle consumption plans, social learning and external habits: Experimental evidence. Journal of Economic Behavior and Organization 106, 413–427. doi:10.1016/j.jebo.2014.07.010.

Carbone, E., Hey, J.D., 2004. The Effect of Unemployment on Consumption: An Experimental Analysis. Economic Journal 114 (497), 660–683. doi:10.1111/j.1468-0297.2004.00236.x.

Carbone, E., Infante, G., 2014. Comparing behavior under risk and under ambiguity in a lifecycle experiment. Theory and Decision 77 (3), 313–322. doi:10.1007/s11238-014-9443-2.

Chen, D.L., Schonger, M., Wickens, C., 2016. oTree-An open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance 9, 88–97. doi:10.1016/J.JBEF.2015.12.001.

Davidoff, T., Brown, J.R., Diamond, P.A., 2005. Annuities and Individual Welfare. American Economic Review 95 (5), 1573–1590. doi:10.1257/000282805775014281.

Druckman, J.N., Kam, C.D., 2011. Students as Experimental Participants: A Defense of the "Narrow Data Base". Cambridge University Press, pp. 41–57.

Duffy, J., Li, Y., 2019. Lifecycle consumption under different income profiles: Evidence and theory. Journal of Economic Dynamics and Control 104, 74–94. doi:10.1016/j.jedc.2019.05.006.

Erdfelder, E., Faul, F., Buchner, A., Lang, A.G., 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods 41 (4), 1149–1160. doi:10.3758/BRM.41.4.1149.

Fatas, E., Lacomba, J.A., Lagos, F., 2007. An experimental test on retirement decisions. Economic Inquiry 45 (3), 602–614. doi:10.1111/j.1465-7295.2007.00027.x.

Feigenbaum, J., Gahramanov, E., Tang, X., 2013. Is it really good to annuitize? Journal of Economic Behavior and Organization 93, 116–140. doi:10.1016/j.jebo.2013.07.005.

Feltovich, N., Ejebu, O.-Z.Z., 2014. Do positional goods inhibit saving? Evidence from a life-cycle experiment. Journal of Economic Behavior & Organization 107 (PB), 440–454. doi:10.1016/j.jebo.2014.01.015.

Gechert, S., Siebert, J., 2020. Preferences over wealth: Experimental evidence. Journal of Economic Behavior & Organization 200, 1297–1317. doi:10.1016/j.jebo.2020.06.006.

Gill, D., Prowse, V., 2012. A Structural Analysis of Disappointment Aversion in a Real Effort Competition. American Economic Review 102 (1), 469–503. doi:10.1257/aer.102.1.469.

Gneezy, U., Potters, J., 1997. An experiment on risk taking and evaluation periods. Quarterly Journal of Economics 112 (2), 631–645. doi:10.1162/003355397555217.

Gough, O., Niza, C., 2011. Retirement Saving Choices: Review of the Literature and Policy Implications. Journal of Population Ageing 4 (1-2), 97–117. doi:10.1007/s12062-011-9046-4.

Hanel, P.H., Vione, K.C., 2016. Do student samples provide an accurate estimate of the general public? PLoS ONE 11 (12), e0168354. doi:10.1371/journal.pone.0168354.

Heimer, R.Z., Myrseth, K.O.R., Schoenle, R.S., 2019. YOLO: Mortality Beliefs and Household Finance Puzzles. Journal of Finance 74 (6), 2957–2996. doi:10.1111/jofi.12828.

Hey, J.D., Dardanoni, V., 1988. Optimal Consumption Under Uncertainty: An Experimental Investigation. The Economic Journal 98 (390), 105. doi:10.2307/2233308.

Horton, J.J., Rand, D.G., Zeckhauser, R.J., 2011. The online laboratory: conducting experiments in a real labor market. Experimental Economics 14 (3), 399–425. doi:10.1007/s10683-011-9273-9.

Hurwitz, A., Sade, O., Winter, E., 2020. Unintended consequences of minimum annuity laws: An experimental study. Journal of Economic Behavior and Organization 169, 208–222. doi:10.1016/j.jebo.2019.11.008.

Koehler, D.J., Langstaff, J., Liu, W.Q., 2015. A simulated financial savings task for studying consumption and retirement decision making. Journal of Economic Psychology 46, 89–97. doi:10.1016/j.joep.2014.12.004.

Krupnikov, Y., Levine, A.S., 2014. Cross-Sample Comparisons and External Validity. Journal of Experimental Political Science 1 (1), 59–80. doi:10.1017/xps.2014.7.

Levy, M.R., Tasoff, J., 2020. Exponential-growth Bias in Experimental Consumption Decisions. Economica 87 (345), 52–80. doi:10.1111/ecca.12306.

Lugilde, A., Bande, R., Riveiro, D., 2019. Precautionary Saving: a Review of the Empirical Literature. Journal of Economic Surveys 33 (2), 481–515. doi:10.1111/joes.12284.

Meissner, T., 2016. Intertemporal consumption and debt aversion: an experimental study. Experimental Economics 19 (2), 281–298. doi:10.1007/s10683-015-9437-0.

Meissner, T., Rostam-Afschar, D., 2017. Learning Ricardian Equivalence. Journal of Economic Dynamics and Control 82, 273–288. doi:10.1016/j.jedc.2017.07.004.

Peijnenburg, K., Nijman, T., Werker, B.J., 2016. The annuity puzzle remains a puzzle. Journal of Economic Dynamics and Control 70, 18–35. doi:10.1016/j.jedc.2016.05.023.

Peterson, R.A., 2001. On the use of college students in social science research: Insights from a second-order meta-analysis. Journal of Consumer Research 28 (3), 450–461. doi:10.1086/323732.

Pu, B., Peng, H., Xia, S., 2017. Role of Emotion and Cognition on Age Differences in the Framing Effect. The International Journal of Aging and Human Development 85 (3), 305–325. doi:10.1177/0091415017691284.

Winter, J.K., Schlafmann, K., Rodepeter, R., 2012. Rules of Thumb in Life-Cycle Saving Decisions. Economic Journal 122 (560), 479–501. doi:10.1111/j.1468-0297.2012.02502.x.

Yamamori, T., Iwata, K., Ogawa, A., 2018. Does money illusion matter in intertemporal decision making? Journal of Economic Behavior and Organization 145, 465–473. doi:10.1016/j.jebo.2017.11.019.