Interpreting Time Horizon Effects in Inter-Temporal Choice

Thomas Dohmen[1]

1 Introduction

Understanding inter-temporal decision making is crucial for economics, and choice experiments have been an important tool in this undertaking. Early experiments in psychology had a particularly important impact on economics, showing that varying the time horizon affects choices in a way that suggests declining rather than constant discounting (for a survey see Frederick et al., 2002). This evidence provided part of the motivation for new economic models incorporating hyperbolic, or quasi-hyperbolic, discount functions (e.g., Loewenstein and Prelec, 1992; Laibson, 1997; O'Donoghue and Rabin, 1999), with important implications in terms of the possibility of dynamically inconsistent preference and self-control problems. Similar choice experiments have also been used by economists as a way to potentially measure the extent of dynamic-inconsistency among different groups of people, or predict economic outcomes thought to be related to self-control (e.g., Sutter et al., 2010, and many others).

Taking into account more recent evidence, however, it is less clear that time horizon effects in experiments are in fact capturing the shape of an underlying discount function. While initial studies did find declining discounting, this was predominantly (though not exclusively) using one type of design.[1] Other studies using different but theoretically equivalent designs have found constant discounting, increasing discounting, or some other type of choice pattern, on average (e.g., Anderhub et al., 2001; Read, 2001; Read and Roelofsma, 2003; Rubinstein, 2003; Read et al., 2005; Harrison et al., 2005; Zauberman et al., 2009; Ebert and Prelec, 2007; Benhabib et al., 2010; Andrioni and Sprenger, 2010; Andersen et al., 2010; Sutter et al., 2010; Halevy, 2011). Also, studies relating experimental measures of dynamic inconsistency to outcomes have found mixed results, and other mechanisms besides self-control problems could possibly explain the relationships that are observed.[2] Thus, it is not clear what discounting assumption is a good description of

[1] Frederick et al.(2002) survey ten studies that compare discounting over overlapping time horizons of different lengths (what we call an "overlapping design" below), and find behavior consistent with declining discounting. They also find the same pattern comparing *across* roughly thirty studies that use either short or long time horizons. By contrast, four studies use a different design, involving comparing non-overlapping time horizons of equal lengths (what we call a "shifted design" below). These find evidence that can be interpreted as declining discounting, but various design features of these latter studies might offer alternative interpretations, and make comparison to most subsequent experiments conducted by economists more difficult.

[2] In some cases studies have found no difference in outcomes based on usual measures of dynamic inconsistency (Sutter et al., 2010; Harrison et al., 2010), or mixed evidence depending on the outcome

the evidence as a whole, or if any discounting model explains the data. Resolving the controversy has been difficult, however, partly due to design and subject pool differences across studies.

This paper provides new evidence on how time horizon affects measured impatience, comparing different types of measures within the same framework. We adhere to the workhorse, "price-list" framework that is the industry standard for economists, but our approach is more comprehensive than most previous studies, in that we consider multiple time horizon lengths, starting dates, orders, and stake sizes, with nine treatments in all. To address concerns about subject pool and generalizability, we use two relatively large, representative samples of the adult population (N=500 and N=1,500), and investigate choice patterns for the overall population as well as robustness across various sub-populations of particular interest. Our experiments involve real money, and credibility of payments is very strong, for example because participants are in a long-term relationship with the surveying company and also the individual surveyor. We can check whether qualitative results are replicated across the two data sets despite different subjects, parameters, and timing of data collection. Rich information on life outcomes also makes it possible to test the predictive power of the experimental measures.

We designed the experiment to explicitly compare three different approaches to testing discounting assumptions. The most commonly used approach, which we denote the "overlapping design" (OD), involves measuring discounting over two or more overlapping time horizons for the same individual, where all time horizons start on the same date. For example, one of our OD measures involves comparing discounting between the immediate future, denoted 0, and 6 months, to discounting between 0 and 12 months.[3] Studies using this design often find a pattern consistent with declining (hyperbolic) discounting (Frederick et al., 2002). Another approach, which we denote the "shifted design" (SD), involves two non-overlapping time horizons of the same length, where the start date of one time horizon is shifted forward in time. For example, one of our SD measures compares

being studied (Meier and Sprenger, 2008; Meier and Sprenger, 2010; Burks et al., 2011) or gender of the subject (Ashraf et al., 2006). Chabris et al. (2008) find a relationship between discounting measures and outcomes, but using the level of the discount rate rather than measures of sensitivity to time horizon.

[3] Use of 0 to denote the immediate payments is a slight abuse of notation, because payments were never truly immediate; they arrived in the immediate future, typically two days from the date of the experiment. We discuss the reasons for this (commonly used) front end delay feature of the design in more detail below.

discounting between 0 and 6 months to discounting between 6 months and 12 months. While the shifted design corresponds to the commonly cited "apple" thought experiment by Thaler (1981), it has actually been less commonly used than the overlapping design.[4] Studies that implement this approach sometimes find that behavior is consistent with declining discounting (e.g., Kirby and Herrnstein, 1995, and others), but in other cases find constant or even increasing discounting (e.g, Anderhub et al., 2001; Sutter et al., 2010). We also designed the experiment to allow what we call the "overlapping, shifted design" (OSD). This involves comparing discounting over a long time horizon to discounting over a shorter time horizon, which overlaps but is shifted forward so as to start later in time. For example, we compare discounting between 0 and 12 months to discounting between 6 months and 12 months. This latter approach was initially tried by Baron (2000) and Read (2001), and yielded little support for declining discounting.

It is important to note these approaches to testing discounting assumptions typically involve a set of maintained assumptions, for example the assumption that people treat monetary payments as consumption. Relaxing these assumptions can affect the predictions of different models in important ways. We assess the ability of different models to explain the data, both with and without such assumptions.

The main stylized fact from our analysis is that time horizon matters for inter-temporal choice, in a way that is hard to explain with standard discounting models: People are more impatient for short time horizons than longer time horizons, inconsistent with constant discounting, but are relatively insensitive to when a given time horizon starts, contrary to non-constant discounting.[5] For example, people are more impatient from 0 to 6 months than 0 to 12 months, but similarly impatient for 6 to 12 months compared to 0 to 6 months. This pattern is also inconsistent with a broader class of models, in that it implies a violation of transitivity.[6] The same qualitative pattern is observed in

[4] The thought experiment involves a choice between 1 apple today or 2 apples tomorrow, and between 1 apple in a year or 2 apples in a year and a day. The time horizon is a single day in both choices, but the start date is shifted into the future for the second choice. Declining discounting could explain a "preference reversal", such that the person prefers to have 1 apple in the choice involving today, but 2 apples in the choice about the more distant future.

[5] We discuss how, under some alternative assumptions, non-constant discounting models tend to make the same predictions as constant discounting. In this case, the models have trouble explaining why time horizon length matters. The same argument applies to quasi-hyperbolic models, if the "present-bias" falls in the brief time window before early payments arrive in our experiments.

[6] Transitivity is violated maintaining canonical assumptions of stationarity and separability. See Roelofsma and Read (2000) for some of the first evidence that inter-temporal choices can violate transitivity.

our second data set, despite using different parameters, time horizon lengths, and starting dates. The pattern is also robust in that it holds for different sub-populations that we consider (by gender, age, education, cognitive ability, and financial situation). Conducting the analysis at the individual-level also shows that few people are consistently in line with any discounting model, while the majority exhibit the pattern observed at the aggregate level, which violates transitivity. The results thus pose an important puzzle for modeling and understanding inter-temporal choice.

The findings also provide a potentially useful lens through which to view previous evidence in the literature. They show that it can be misleading to focus on only one type of measure, and they raise questions about interpreting the results of studies that have this feature. They also show that "mixed evidence" can be generated by varying a particular aspect of the design, holding subject pool and other design characteristics constant. Indeed, the particular method variance we observe maps, broadly speaking, onto the patterns of inconsistent results observed across previous studies: we reliably find declining discounting with OD measures, similar to most previous studies using this approach, and less evidence for declining discounting using other types of measures, similar to many (but not all) previous studies using SD or OSD approaches.

One alternative explanation for time horizon effects is that these reflect biased perceptions, rather than preferences (this type of argument has been proposed in various forms by Read, 2001, Rubinstein, 2003, Ebert and Pelec, 2007, and Zauberman et al., 2009). For example, comparing 0 to 6 months to 0 to 12 months, the objective time duration doubles, but if the subjective duration less than doubles, this makes it more attractive to wait for the longer time horizon all else equal. A similar argument could lead to greater impatience for 6 to 12 compared to 0 to 12. One testable prediction of this explanation is that salient reference points might influence inter-temporal choice, given that subjective perceptions are well-known to be sensitive to such framing effects (e.g., Tversky and Kahneman, 1981).[7] Using one of our data sets, where treatment order was randomized, we find that the effect of time horizon length is smaller and not statistically significant, when comparing first treatments across subjects, while it is more pronounced and significant in

[7] In lab experiments with hypothetical payments, Zauberman et al. (2009) measure subjective perceptions of time duration directly, and find compression of perceived time duration. Ebert and Prelec (2007) show that framing effects alter perceptions of time duration in inter-temporal choice experiments (in the lab with hypothetical rewards).

later treatments. Thus, relative comparisons to a salient reference point, in the form of a previous treatment, do affect choice, and furthermore, time horizon effects are largely due to such relative comparisons. Almost all previous studies use within-subject comparisons, and thus our results indicate that the effect of time horizon may be more about relative comparisons than has previously been realized.[8]

Our results are complementary to a strand of the literature in psychology, and also some recent work by economists, that questions whether time horizon effects reflect discounting. Previous studies (e.g., Barron, 2001; Read, 2001; Read and Roelofsma, 2003; Read et al., 2005) found, as in our study, "subadditivity" of discounting on average, such that discounting is more extreme when measured using sub-intervals.[9] Our study is complementary in that we use representative samples combined with incentivized experiments, assess the robustness of the results for different sub-populations, compare across a wide variety of different parameter values and time horizons lengths, show how relative comparison and anchors might be important for driving time horizon effects, and replicate results across data sets. Other studies have found evidence pointing to a role for subjective perceptions in explaining time horizon effects (Ebert and Prelec, 2007; Zauberman et al., 2009), but differ from ours in that they use hypothetical rewards and student subjects. Harrison et al. (2002) and Andersen et al. (2010) also study time horizon effects in representative samples, but focusing on OD comparisons.[10] Some recent studies have also deviated from the typical approach of eliciting all choices, about the present and about the future, at one point in time; returning to subjects multiple times, they test whether the pattern of declining discounting implied by choices made in the initial interview actually translate into the predicted direction of choice reversal when decisions are offered again in the future (Harrison et al., 2005; Airoldi et al., 2011; Halevy, 2011), and so far conclude

[8] Two studies that we are aware of allow comparing within-subject and between-subject designs, either implicitly or explicitly, and find results that are in line with ours, although they use hypothetical incentives. Ebert and Prelec (2007) find increased time sensitivity in a within-subjects design, and the data reported in Zauberman et al. (2009) indicate that the within-subject decline in discounting is about twice as large as the between-subject decline in discounting, although the authors do not comment on the difference.

[9] The tendency for people to be more impatient for both 0 to 6 months, and 6 to 12 months, than 0 to 12 months, implies subadditivity: The total discounting from compounding the discount factors for the two sub-intervals is less than the discount factor obtained from the longer time horizon measure. Additivity is a prediction of all discounting models, regardless of the shape of the discount function, maintaining the canonical assumptions of time stationarity and separability of the period utility function.

[10] Assuming a specific functional form for utility, they use experimental measures to calibrate concavity of utility, and conclude that the resulting implied discount function is constant.

in the negative. Our approach is different in that it shows that even the initial choices are not consistent with discounting predictions, and thus helps explain why returning to subjects in the future does not yield behavior consistent with the predictions of discounting models.

In summary, the findings have three main implications. (1) The results pose an important puzzle for understanding inter-temporal choice, as the usual candidate models of time preference, and a broader class of models that requires transitivity, do not explain the modal choice pattern; existing models will of course continue to be useful in many settings, but given that the experiments involve real stakes and relatively simple trade-offs, the inability to explain behavior is a notable limitation and calls for further research. (2) Designs that focus on only one type of measure might be misleading, by seeming to support one discounting assumption or another, and relatedly, variation in time horizon effects across populations, or correlations of such measures with economic outcomes, may not in fact reflect dynamic inconsistency or self-control problems, but rather something else. (3) The fact that the menu of savings options offered to an individual may influence inter-temporal choice has potentially important policy implications, even suggesting possible interventions designed to encourage saving; this reference-dependence may also help shed light on why firms present information about financial assets, or payment plans, to customers in the ways that they do.

The rest of the paper is organized as follows. Section 2 describes the data sets, treatments, and behavioral predictions. Section 3 presents the main results, and discusses ability of different models to explain the stylized facts. Section 4 offers evidence on alternative explanations. Section 5 concludes.

2 Design of the Experiment

2.1 Data collection and experimental procedures

Our analysis uses two data sets. One data set, denoted the SOEP data, involves a subsample of participants in the German Socio-Economic Panel (SOEP), a large panel data set for Germany (for a detailed description of the SOEP see Haisken-DeNew and Frick, 2003; Schupp and Wagner, 2002; Wagner et al., 2007). The second data set, denoted the Pretest data, involves a separate sample of individuals, which was collected by the SOEP administration as part of the annual process of “pretesting” questions for potential use in the SOEP survey. The Pretest data, and the SOEP data, were collected during Spring of 2005, and Spring and Summer of 2006, respectively. Collection in each case was done by the same professional surveying company that administers the SOEP every year. Sampling for each data set was done according to the same procedure used to generate the SOEP sample, and individuals were visited by interviewers in their own homes.[11] Both the Pretest and SOEP samples were constructed so as to be representative of the adult population, age 17 and older, living in Germany.[12] In total the Pretest data include 532 subjects, and the SOEP data include 1,503 subjects.

Participants in our studies went through a computer assisted personal interview (CAPI) conducted with a laptop. The interview consisted of two parts. First, subjects answered a detailed questionnaire. The items in the questionnaire were presented in the standard format used by the SOEP. Topics included demographic characteristics, financial situation, health, and attitudes. The questionnaire also included two brief tests of cognitive ability. The full questionnaire, in German and translated into English, is available upon request. At the end of the questionnaire, subjects were invited to participate in the second part of the interview, which consisted of a paid experiment.

The first step in the experimental procedure involved the experimenter presenting subjects with some example choices. The experimenter explained the types of choices that the subject would make, and how payment would work. In particular, subjects were informed that the experiment would involve multiple choices, that one choice situation would be randomly selected after all choices had been made, and that the choice they made in this situation would potentially be relevant for their payoff. Subjects knew that at the end of the experiment a random device would determine whether they were actually paid, with the probability of being paid equal to 1/7 in the Pretest, and 1/9 in the SOEP data. This procedure gives subjects an incentive to choose according to their true preferences in each choice situation, and thus is incentive compatible. After explaining the nature of the

[11] For each of 179 randomly chosen primary sampling units (voting districts), an interviewer was given a randomly chosen starting address. Starting at that specific local address, the interviewer contacted every third household and had to motivate one adult person aged 17 or older to participate. For a detailed discussion of the random walk method of sampling see Thompson (2006).

[12] Respondents had to turn 18 during the year of the interview to be eligible.

experiment and the rules for payment, the experimenter asked subjects whether they were willing to participate. Subjects who agreed to participate were given further instructions, and then allowed to ask questions. Once there were no more questions, the experiment began, and subjects were asked to make their actual choices. An example of the script and instructions used in the experiments is presented in Appendix A below, translated from German into English.

Our experiments were designed to give a measure of the annual internal rate of return (IRR) needed to induce an individual to wait, for a given time horizon. A time horizon Tts is defined by starting date t, and ending date s. For a given horizon an individual chose between an early payment, X_t, available at the start of the horizon, or a larger, later payment, Z_s, available at the end of the horizon. In all choices for a given horizon, the amount of X_t was held constant, but the later payment, Z_s, was larger in each subsequent choice. For most time horizons, the value of Z_s in the first choice was calibrated to be consistent with an annual IRR of 2.5 percent, assuming semi-annual compounding, and each subsequent value of Z_s implied an additional 2.5 percentage point increase in the annual rate of return, up to a maximum of 50 percent.[13] Some time horizons had a coarser measurement of the IRR, in steps of 5 percentage points, but allowed measuring annual IRRs as high as 100 percent. We obtain an incentive compatible measure of impatience for a given time horizon by observing the value of Z_s, or equivalently the annual IRR, needed to induce the individual to wait. Across treatments, we varied t and s, and the amounts X_t and Z_s.

During the experiment subjects were presented with the different choices one at a time on the computer screen. The first time that a subject switched from the early to the delayed payment, the subject was asked whether he or she also preferred to wait for any larger payment, and all subjects agreed.[14] Subjects knew that one row would be randomly

[13] We chose semi-annual compounding of the annual interest rate because this is a natural compromise between the two types of compounding German subjects are most familiar with: quarterly compounding on typical bank accounts, and annual reports on the rate of return from savings accounts, pension funds, or stock holdings. Using semi-annual compounding also helps avoid prominent round numbers in the choices, which could potentially influence switching choices.

[14] As a consequence, there are no non-monotonic choices in the data. This procedure helped to mitigate potential confusion among subjects about the structure of the choice problem, by re-iterating the tendency for the late payment to only increase across further choices. One concern could be that this procedure tends to "lock-in" mistakes, if subjects make a mistake and then do not want to tell the experimenter that they have changed their mind. However, even if this were the case, to the extent that it affects all measures it should not matter for our conclusions, because they are based on differences across measures. Furthermore, studies that use alternative approaches and end up with non-monotonic

selected at the end of the experiment, and that their decision in that row could be relevant for their payoff. Subjects also knew that all payments would be sent by mail following the interview, and thus would arrive within at most two days due to the well-known two-day guarantee for delivery by the German postal service.[15] Certificates for immediate payments could be cashed immediately, once they arrived, while certificates for payments in the future would be cashable only at the specified time.

In our design, there are several factors that helped to ensure equal credibility of early and late payments. This is desirable in order to prevent, e.g., that early payments enjoy relatively higher credibility, and thus choices are distorted in the direction of overstating impatience, or even overstating the degree of declining discounting.[16] First, the design involved a "front-end delay" (see Coller and Williams,1999) in the sense that all payments arrived by mail between 1 and 2 days after the experiment, regardless of when the payment would be cashable. This is a standard procedure in discounting experiments to help make early and late payments equally credible. Second, experiments for both data sets were conducted by the professional agency used by the SOEP, which is highly credible and well-known because of its role in conducting election polls for German public television. Interviewers also left their contact details at the end of the experiment, making it easy for subjects to contact the institute. There were no reports, from any of the interviewers, about subjects expressing concerns regarding credibility of payments. Third, all participants in the SOEP data were members of the SOEP panel itself. Thus, these individuals were in a long-term relationship with the surveying agency, and typically even with the same individual surveyor, so that there should have been no doubts about credibility. Despite the fact that our setting created very strong credibility on the side of the agency

choices are put in the position of needing to drop (non-randomly) a part of the data.

[15] The Deutsche Post is highly successful at achieving an explicit goal to delivery mail within two days, for packages with origin and destination within Germany.

[16] As pointed out by Andrioni and Sprenger (2010) and others, unequal credibility of payments could generate the appearance of declining discounting in SD comparisons, because the earlier time horizon includes an immediate, especially credible payment while the later horizon does not. For a series of reasons, discussed below, credibility is especially good in our design, and thus this argument would be unlikely to apply to our data. Furthermore, it would not explain patterns that resemble declining discounting in OD comparisons, constant discounting in SD comparisons, or increasing discounting in OSD comparisons, all of which we find. Thus, while unequal credibility of payments could in some settings generate "spurious" time horizon effects, this does not seem to explain the time horizon effects we observe. While Andrioni and Sprenger (2009) do not replicate declining discounting in an OD comparison, and note that this might reflect strong credibility of payments, an alternative explanation could be their novel (and innovative) elicitation method using convex time budgets.

conducting the experiment, a remaining potential concern might be that subjects worried that they themselves would misplace their payment before it became cashable, say, one year in the future. Empirically, however, we find no evidence that individuals had such concerns.[17]

A final note on the design concerns the length of the front-end-delay. Many previous studies that test discounting assumptions have had front-end-delays ranging from one day to as long as one month (e.g., Meier and Sprenger, 2010; Harrison et al., 2002). While useful for equalizing credibility, a potential drawback of such a delay is reduced "immediacy" of early payments. Indeed, some non-constant discounting models, which assume a discrete drop in discount rates between the present and future, are based on the intuition that present-bias reflects psychological processes that respond to immediate rewards. To the extent that the front-end-delay eliminates immediacy of early payments, such a design feature may "miss" present-bias. Unfortunately, the verdict is still out on how immediate payments would need to be, in order to trigger the aforementioned psychological processes. It is quite intuitive that the present is today, and the future is tomorrow. On the other hand, one could think that immediacy is actually about having a consumption opportunity literally in one's hand, and that even a delay of ten minutes might already be too long for rewards to feel immediate.[18] Making a trade-off between credibility and immediacy, we chose the shortest possible front-end delay compatible with avoiding a same-day credibility problem.[19] Importantly, models where present bias occurs within the window of the front-end-delay still make clear testable predictions in the experiment, namely constant IRRs across time horizons (we discuss this in detail below).

[17] For example, the preference for early versus later payments is the same, regardless of whether the early payment is cashable immediately, or only in one year.

[18] Notably, almost without exception studies have had a delay of at least some hours before early rewards are usable, so rewards are not truly immediate (e.g., Burks et al., 2009). One exception is McClure et al. (2007), where rewards are squirts of juice over the course of minutes.

[19] Mainly for the purposes of concreteness, so as to avoid having to say 1 or 2 days all the time, but also to heighten immediacy (without affecting credibility), the experimental instructions told subjects that immediate rewards would be referred to as being received "Today"[quotes included]. At the same time, the instructions were very clear that all rewards would arrive after the experiment by post, and that: "Today means you can cash the check you receive by post immediately". Ultimately, we see little evidence that this wording lead to differential behavior in early versus later time horizons.

2.2 Treatments

Table 1 summarizes the various treatments. As shown in the table, the Pretest data involved 500 subjects participating in the experiments, with three different measures of annual IRR for each subject. The measures come from three different time horizons, 0 to 6 months (T06), 0 to 12 months (T012), and 6 months to 12 months (T612).[20] The order of the treatments in the Pretest data was randomized across individuals. The early payment was always 100 Euros, and the largest delayed payment always implied an IRR of 50 percent for waiting the specified length of time. If individuals never chose the later payment, their IRR was right-censored, and coded as having a (lower-bound) value of 52.5 percent.

The SOEP data differ in that there are only two treatments for each individual (see Table 1), and treatments also varied across sub-samples. For a first sub-sample of 490 individuals, impatience was measured for 0 to 6 months (T06) and 0 to 12 months (T012), giving two measures of annual IRR for each subject. The second sub-sample of 487 were asked about 0 to 1 month (T01) and 0 to 12 months (T012). For the third sub-sample of 526 we measured discounting for 0 to 1 month (T01b) and 12 to 13 months (T1213). The measures for the third sub-sample were different, because IRRs were measured in steps of 5 percent rather than 2.5 percent, and the upper-bound IRR in each horizon was 105 percent rather than 52.5percent. We denote the one-month measure in this sub-sample T01b, to distinguish it from T01 in the second sub-sample. Order was predetermined in the SOEP data: for the first two sub-samples, the T012 measure was always elicited first; for the third sub-sample, T01b was elicited first. A random device on the computer selected whether an individual was assigned to the first, second, or third sub-sample experiments. In the SOEP data the early payment was always 200 Euros, and thus stakes were higher than in the Pretest, even accounting for the lower payment probability of 1/9 rather than 1/7.

[20] As discussed above, 0 is a slight abuse of notation, since immediate payments arrived up to two days in the future.

2.3 Behavioral Predictions

The type of experimental design we implemented has been used extensively to test discounting models, based on a certain set of maintained assumptions. We first derive predictions under the usual assumptions, and then discuss how qualitative predictions change if the assumptions are relaxed.

We illustrate the predictions by considering, without loss of generality, an example with three time horizons, T06, T012, and T612, corresponding to the structure of the Pretest data. Assume compounding occurs once every 6 months, and for simplicity that a period is 6-months long. Assume for now that early payments at time 0 are literally available on the day of the experiment, when preferences are measured. Also, adopt the usual maintained assumptions that utility is locally linear, that people treat monetary payments as consumption, and that utility is additively separable and time stationary.

When making decisions in T06, T012, and T612, subjects decide for the early or late payment depending on whether or not the offered *annual* rate of return r in a given choice is sufficiently attractive to induce waiting. Thus, decisions involve the following comparisons:

$$(1+\frac{r}{2})X_0 \lesseqgtr Z_6 \qquad (1+\frac{r}{2})^2 X_0 \lesseqgtr Z_{12} \qquad (1+\frac{r}{2})X_6 \lesseqgtr Z_{612}$$

The lowest delayed payment such that the individual prefers to wait establishes the points of indifference for each horizon, and defines the internal rates of return:[21]

$$(1+\frac{IRR_{T06}}{2})X_0 = Z_6 \qquad (1+\frac{IRR_{T012}}{2})^2 X_0 = Z_{12} \qquad (1+\frac{IRR_{T612}}{2})X_6 = Z_{612} \quad (1)$$

Solving for IRRs yields:

$$IRR_{T06} = 2(\frac{Z_6}{X_0}-1) \qquad IRR_{T012} = 2((\frac{Z_{12}}{X_0})^{\frac{1}{2}}-1) \qquad IRR_{T612} = 2(\frac{Z_{12}}{X_6}-1) \quad (2)$$

We now consider how different assumptions about time preference affect predictions for IRRs.

[21] Because the delayed payment is a discrete variable in the experiment, the lowest delayed payment that is preferred actually establishes an upper bound for the IRR, while the largest delayed payment that is not preferred establishes the lower bound. Without any consequences for the qualitative results, we neglect this issue and use the upper bound when deriving predictions.

2.3.1 Constant discounting

In the case of constant discounting, an individual is indifferent between the early and delayed payments in T06, T012, and T612 when

$$(1+\frac{\gamma}{2})X_0 = Z_6 \qquad (1+\frac{\gamma}{2})^2 X_0 = Z_{12} \qquad (1+\frac{\gamma}{2})X_6 = Z_{12} \tag{3}$$

Where γ is the constant rate of time preference. As this is the same as condition (1), if IRRs are the same in all three horizons, it follows directly that a constant discounter chooses Z_s such that the measured IRR is invariant with respect to time horizon: $IRR_{T06} = IRR_{T012} = IRR_{T612}$.

2.3.2 Declining and increasing discounting

In the case of declining discounting, there are different discount rates, γ_1 and γ_2, for periods 1 and 2, such that $\gamma_1 > \gamma_2$. In this case the points of indifference in T06, T012, and T612 are given by

$$(1+\gamma_1)X_0 = Z_6 \qquad (1+\gamma_1)(1+\gamma_2)X_0 = Z_{12} \qquad (1+\gamma_2)X_6 = Z_{12}. \tag{4}$$

Substituting into (2) shows that, If choices are generated by this model, measured IRRs will have the form

$$IRR_{T06} = 2((1+\gamma_1)-1) \quad IRR_{T012} = 2(((1+\gamma_1)(1+r_2))^{\frac{1}{2}}-1) \quad IRR_{T612} = 2((1+\gamma_2)-1) \tag{5}$$

Given $\gamma_1 > \gamma_2$ this implies $IRR_{T06} > IRR_{T012} > IRR_{T612}$. Intuitively, impatience should be greatest in T06 with declining discounting, because it is the most focused on periods close to the present. Behavior in T612 should be the most patient, because it only concerns future payments. Behavior in T012 should be in-between, as it includes the present but also extends substantially into the future. An analogous argument establishes that with increasing discounting, the model predicts the opposite ranking of IRRs by time horizon, $IRR_{T06} < IRR_{T012} < IRR_{T612}$.

2.3.3 Predictions with different/weaker assumptions

Relaxing the assumption that early payments arrive immediately may matter for the predictions of some types of non-constant discounting models. Specifically, the fact that early payments actually arrive up to two days in the future matters for models where there are discrete changes in the discount rate during those two days, followed by constant discounting thereafter. A particularly important example is the quasi-hyperbolic model (e.g., Phelps and Pollak, (1968); Laibson, 1997; O'Donoghue and Rabin, 1999), in the case that the period is assumed to be less than two days (the model is silent about the correct period assumption). Such a model could imply a high discount rate between today and tomorrow, and then a constant discount rate between any two future periods.[22] Given the experimental design, this type of model thus predicts $IRR_{T06} = IRR_{T012} = IRR_{T612}$, the same as with constant discounting, i.e., invariance of IRRs with respect to time horizon is needed for the model to explain the data.[23]

Studies that use monetary payments to test discounting assumptions assume (usually implicitly) that people treat money rewards as consumption, but if this assumption is not valid, it changes the qualitative predictions for all non-constant discounting models. If payments are viewed as fully fungible, as predicted by standard theory, then subjects should use the experiment as an opportunity for arbitrage, and in all time horizons choose to wait when the annual rate of return exceeds the going market interest rate. Outside of the experiment the subjects can then use the credit market as appropriate, for example, borrowing against future experimental earnings using the lower market interest rate in order to finance their desire for present-biased consumption. All discounting models in this case predict $IRR_{T06} = IRR_{T012} = IRR_{T612}$, at the relevant interest rate faced by each individual. Thus, the ability of discounting models to explain the data assuming full arbitrage depends on whether IRRs are invariant to time horizon.

If the assumption of linear utility is not accurate, and the unobserved utility function is instead concave, this also affects predictions, in a way that depends on the particular time horizons being compared. Given that later payments are always larger, diminishing marginal utility of money would be another reason, besides time preference, why later payments are discounted relative to early payments. Unobserved concavity therefore leads to overestimating the degree of impatience, but importantly, the bias should be larger

[22] The model is quasi-hyperbolic in the sense that it provides a step-function approximation to a hyperbolic discount function (see, e.g., Laibson, 1997).

[23] If the present bias is assumed to extend more than 2 days into the future, but less than 6 months, the quasi-hyperbolic model makes the same predictions as models with hyperbolic, or other forms, of continuously declining discounting.

for long time horizons, because in that case payments are larger in absolute magnitude. Thus, for OD comparisons (where we find the strongest declining discounting), it works against finding declining discounting.[24] Analogously, for OSD comparisons (where we find the strongest increasing discounting) it tends to understate increasing discounting. SD comparisons should be unaffected, since the decline in marginal utility affects both measures in exactly the same way and is differenced out. Thus, while we do not identify concavity of the utility function in our design, the direction of the bias is clear, and it turns out that correcting for concavity would only strengthen our conclusions about aggregate patterns.[25]

If the assumptions about time separability and time stationarity of the utility function are not valid, then measured IRRs may vary across time horizons for reasons unrelated to non-constant discounting and dynamic inconsistency. For example, people may anticipate that the marginal utility of consumption in the future depends not just on distance from the present, due to discounting, but also on state-contingent preferences, for example an upcoming vacation. In this case, one could observe many different patterns of IRRs across time horizons, for any given discount function. Our approach is to assess whether, even maintaining these standard assumptions, the data may violate standard discounting models, or for that matter any model that requires satisfying basic axioms such as transitivity.

3 Results on IRR and time horizon

Figure 1 presents the cumulative distribution functions of the annual IRR for each of the different time horizons. The top panel shows results from the Pretest data, and the bottom panel shows results from the SOEP data. For the SOEP data, we pool the T012

[24] If the estimates of impatience are affected by concavity of the unobserved utility function, we have $I\tilde{R}R_{T06} = 2(\frac{u(Z_6)}{u(X_0)} - 1)$, $I\tilde{R}R_{T012} = 2((\frac{u(Z_{12})}{u(X_0)})^{\frac{1}{2}} - 1)$, and $I\tilde{R}R_{T06} = 2(\frac{u(Z_{12})}{u(X_6)} - 1)$. By concavity, it is clear that $I\tilde{R}R_{T06} < IRR_{T06}$, $I\tilde{R}R_{T012} < IRR_{T012}$, and $I\tilde{R}R_{T612} < IRR_{T612}$, so that we overestimate the level of impatience in each case. It is also clear, however, that overestimation is more severe for T012 than the shorter horizons. Thus, we have $I\tilde{R}R_{T06} - I\tilde{R}R_{T012} < IRR_{T06} - IRR_{T012}$, and $I\tilde{R}R_{T612} - I\tilde{R}R_{T012} < IRR_{T612} - IRR_{T012}$, and we understate the difference in IRRs between short and long time horizons if utility is concave.

[25] When the goal is to estimate the precise level of the discount rate for a given horizon, rather than just an upper bound, one approach is to assume a specific utility function, and calibrate the curvature using risk aversion experiments (see Andersen et al., 2008; Andersen et al., 2010). For other approaches see, e.g., Attema et al. (2010), and Andrioni and Sprenger (2009).

measures across the two sub-samples that have this treatment, as the order and stake sizes are identical.[26]

In the top panel of Figure 1, we see that people in the Pretest data are substantially more patient for T012 than for T06, or T612. People are quite similar in impatience, however, for the T06 and T612 measures.[27] Thus, people are more impatient for shorter time horizons than longer time horizons, but similarly impatient over the short horizons regardless of starting date. The bottom panel shows the same qualitative patterns in the SOEP data. Observed IRRs increase monotonically as time horizon length decreases, with greater impatience for T06 than T012, and even greater impatience for T01 than T06. At the same time, IRRs are relatively insensitive to the starting date of the time horizon, in that people are similarly impatient for T01b and T1213.[28] The distribution for T01 is also similar to T01b and T1213, except for a deviation in the direction of greater patience starting around the middle of the range for T01, potentially due to either an order effect, or an effect of the different upper bound for the IRR, or both.[29] Despite this framing difference between the one-month measures, they are all clearly more similar to each other, than to measures with longer horizons.

Table 2 reports descriptive statistics regarding IRRs for each time horizon measure. We focus our discussion on median IRRs, as the right-censoring of the data makes interpretation of means more difficult, and likely understates the true impact of time horizon length on mean IRRs.[30] As shown in the table, the medians are significantly different at

[26] The cumulatives for the two measures considered separately are very similar, and are not significantly different. Means (20.48 and 20.29) and medians (17.50 and 17.50) are also very similar.

[27] Distribution tests confirm that the 6-month measures were significantly different from the 12-month measure ($p < 0.001$;$p < 0.001$; Kolmogorov-Smirnov), but that the 6-month measures were not significantly different from one another ($p < 0.90$; Kolmogorov-Smirnov).

[28] The distributions for T06 and T012 are significantly different, as are distributions for T01 and T06 ($p < 0.001$;$p < 0.001$; Kolmogorov-Smirnov). The distributions for T01b and T1213 are not significantly different ($p < 0.98$; Kolmogorov-Smirnov).

[29] The distributions for T01b and T1213 are each significantly different from the distribution for T01 ($p < 0.001$;$p < 0.001$; Kolmogorov-Smirnov).

[30] Means are calculated coding right-censored IRRs with a value of 52.25, or 105, depending on the upper bound for the time horizon measure. This top-coding likely understates the level of the true mean IRR for a given horizon. Importantly, however, as seen in the table, the fraction censored decreases monotonically with greater time horizon length, which suggests that the mean IRR is more strongly biased downwards for short time horizons. This implies that differences in mean IRRs tend to understate the true differences between mean IRRs for short and long time horizons. Indeed, the differences in mean IRRs across short versus time horizons are almost always smaller than differences in median IRRs, and this is especially true when comparing the longest to the shortest horizons, exactly as one would expect with greater downward bias in mean IRRs for shorter horizons.

the one-percent level, or not significantly different, in line with the contrasts suggested by the cumulative distribution functions.[31] A single exception is that the median for T612 is significantly higher than for T06, implying increasing discounting in median terms rather than constant discounting. The distribution as a whole, however, is not significantly different for these measures, and the difference in medians is small in relative terms.[32] Thus, discounting is still relatively similar across time horizons for this comparison. Also as expected, given the differing upper bounds, medians for T01 in the second sub-sample of the SOEP data are substantially smaller than those for the T01b and T1213 measures in the third sub-sample.[33]

Tables 3 and Table 4 provide another way to look at the results on IRR and time horizon, using interval regressions that correct for right- and left-censoring of the dependent variable. The dependent variable is measured IRR, and independent variables are dummy variables for time horizon length. Standard errors are robust and adjusted to allow for potential correlation of the error term across observations from the same individual.

In Column (1) of Table 3 we see that IRRs in the Pretest data are significantly lower for the T012 measure compared to T06, while there is not a significant difference between T612 and T06. Table 4 presents similar regression analysis based on the SOEP data. Results are reported separately for the three sub-samples. Looking at Column (1) we again see a pattern of lower IRRs for longer horizons, but similar IRRs across horizons of the same length regardless of starting date: T012 is significantly lower than T06 and T01, but T1213 is not significantly different from T01b.

The regression analysis also allows checking robustness with respect to order effects, in the Pretest data where order was randomized. We add dummy variables for the different possible treatment orders, and interactions of T012 and T612 with all of the different orders, to the specification used in Column (1) of Table 3, and again find a significant

[31] Significance levels are based on paired t-tests, or (non-parametric) signed-rank tests, for equality of means and medians, respectively.

[32] We tested for equality of the distributions using the two-sample Kolmogorov-Smirnov test.

[33] We may also understate the difference in median IRRs when comparing T01 to T012 in the SOEP data, because the median IRR in T01 is right-censored. We also looked at the impact of time horizons on IRR by estimating a Cox proportional hazard model for deciding to wait in a given treatment. This explicitly takes into account the right-censoring of the data which otherwise tends to understate mean differences across horizons. We found that lengths of the time horizon had highly significant impacts on the hazard of deciding to wait, while starting date of time horizons did not. Results are available upon request.

difference between T012 and T06, but not T06 and T612. Furthermore, all interaction terms are not statistically significant, individually or jointly.[34] Thus, the qualitative results are systematic across all different possible treatment orders, and are not driven by any particular treatment ordering.

In subsequent columns, Table 3 and Table 4 explore the robustness of the qualitative results observed for the whole sample, across different sub-populations. We consider sub-populations defined by gender, age, cognitive ability, education level, income level, and credit constraints. Such variables might matter for how individuals respond to different measures of IRR, for a variety of reasons. For example, cognitive ability or education might be related to the level of financial sophistication of an individual, and familiarity with market interest rates, and potentially affect sensitivity to varying time horizon.[35] Income or credit constraints could potentially matter, if people cannot borrow and have expenses that will arrive within the scope of the time horizons considered in the experiment.[36] Looking at the regressions estimates, it is apparent that the qualitative results on time horizon effects are quite robust, and do not appear to be driven by degree of financial sophistication or financial situation: The difference between long and short time horizons is observed for every sub-sample, in both data sets, regardless of the measures used, while horizons of similar length, but different starting dates, are not significantly different.[37] Thus, our data indicate that the patterns we find are not isolated to specific populations, but rather are a quite robust and general tendency.

Similar conclusions arise from an individual-level analysis. To identify individual "types", we focus on the Pretest data, because we have seen that it is important to consider more than just one type of measure, and in that case there are multiple measures (OD, SD, and OSD) for each person. Table 5 reports fractions of individuals exhibiting

[34] Results available upon request.

[35] Cognitive ability is measured using two different tests, which are then standardized and average to provide a measure of overall cognitive ability. For a description of the tests see Dohmen et al., (2010).

[36] To create the dummy variable indicating being credit constrained, we use a question that asks: "If you suddenly encountered an unforeseen situation, and had to pay an expense of 1,000 Euros within the next two weeks, would it be possible for you to make that payment?"

[37] In previous work (Dohmen et al., 2010) we found a significant correlation between cognitive ability and impatience measured over an annual time horizon (T012). We replicate this finding in the Pretest data, even when averaging across all three time horizons, and also in the SOEP for each sub-sample, averaging across time horizons. Thus, there is accumulating evidence that the *level* of IRR for an individual is related to cognitive ability. By contrast, the overall pattern of *differences* in IRR across time horizons that we observe are equally pronounced for low and high ability individuals.

different possible qualitative choice patterns across the different measures. Adhering to the usual maintained assumptions for testing discounting assumptions, we classify individuals who consistently have the same IRR across all time horizons as constant discounters (this fraction would include quasi-hyperbolic, or subjects engaging in arbitrage, under weaker assumptions). People with $IRR_{T06} > IRR_{T012} > IRR_{T612}$ are classified as declining discounters, and people with the opposite pattern are classified as increasing.

One feature of the individual-level data that is immediately apparent is that fractions of individuals who can be classified as being fully consistent with the predictions of constant, declining, or increasing discounters are relatively small. The first row of Table 5 reports these fractions, which are 13.37, 8.80, and 7.04 percent, respectively. The modal pattern, as reflected in the aggregate results, is for individuals to be more impatient in both short time horizons than the long time horizon; almost 50 percent of individual exhibit this behavior. Among these, roughly half have equal IRRs for the two short time horizons, exhibiting zero sensitivity to starting date, while the others have different IRRs for the two short horizons. The remaining fraction of the sample, in the "Other" category, is also inconsistent with discounting assumptions. In includes individuals who have lower IRRs in both short horizons than the long horizon, and those having at least one short horizon that is different from the long horizon, but the same IRR for the long horizon and the other short horizon. Overall, about 70 percent of individuals are inconsistent with all discounting types. The second row of the table shows how the fractions change allowing for "random errors": Each observed IRR is converted to an interval 5 percentage points wide, and IRRs are classified as being different only if their respective intervals do not overlap. Not surprisingly, the fraction of constant discounters increases substantially, to 36.92 percent, using this conservative approach that favors finding constant discounting. The fractions of declining and increasing discounters are only 5.02 and 4.30 percent, respectively. The fraction of individuals who are inconsistent with standard discounting models remains substantial, at 53.77 percent of the sample.

3.1 Discussion: Models and the stylized facts

To summarize, the empirical analysis yields two main stylized facts: (1) People are more impatient for short than long time horizons; (2) people are relatively insensitive to when a given time horizon starts. In the Pretest data this is shown by the pattern $IRR_{T06} =$

$IRR_{T612} > IRR_{T012}$, and in the SOEP data by $IRR_{T01b} = IRR_{T1213} \approx IRR_{T01} > IRR_{T06} > IRR_{T012}$.

In light of the predictions derived in Section 2.3, it is clear that this pattern is not well explained by either constant, declining, or increasing discounting, maintaining the usual identifying assumptions. The sensitivity of IRRs to time horizon length is inconsistent with constant discounting, while the insensitivity of IRRs to starting date of a given horizon is contrary to the key prediction of declining or increasing discounting.

The pattern is also inconsistent with non-constant discounting models where the discount rate changes discretely in the first one or two days (if the change occurs farther in the future, qualitative predictions are the same as declining discounting models discussed above). Such models would predict constant IRRs across time horizons, because even the earliest payments are available starting two days in the future. This class of models includes the quasi-hyperbolic model, two-system models that make similar predictions (Thaler and Shefrin, 1981; Fudenberg and Levine, 2006), and the "fixed cost" version of non-constant discounting discussed by Benhabib et al. (2010), which combines constant discounting with a fixed cost of receiving payments in the future.[38] The finding that time horizon does strongly influence IRRs is thus inconsistent with this class of non-constant discounting models.

Relaxing the assumption that people treat money as consumption, or allowing for concavity of utility, the usual candidate models are still not consistent with the data. If people treat money as fully fungible, and therefore use the experiment as an opportunity for arbitrage, IRRs should be invariant to time horizon, contrary to the findings. Accounting for concavity of the unobserved utility function, the pattern we observe would become even more pronounced, leaving conclusions unchanged.

We summarize the ability of different models to explain individual time horizon comparisons, as well as the data as a whole, in Table 6. The most important message of the table is that none of the models can accommodate the whole set of facts. The table is also potentially useful in that it suggests which types of designs may tend to find

[38] This fixed cost discount function makes qualitatively similar predictions to the quasi-hyperbolic model, except that it predicts greater patience for larger stakes. The recent two-system model of Fundenberg and Levine (2010) makes similar predictions to the hyperbolic model, allowing for a continuous decline in discount rates moving into the present. Thus, the model is not fully consistent with the data, for same reasons discussed in the context of declining discounting models.

which types of patterns. This shows how focusing on a particular type of time horizon comparison may be misleading, by seeming to support one class of models or another.

Maintaining only the usual workhorse assumptions of separability and time stationarity of the period utility function, the data are inconsistent with a broader class of models because they imply a violation of transitivity. For example, T06, T012, and T612 can all be represented as choices between triples, where elements denote utilities at time 0, 6, and 12.

$$\begin{array}{llllll} & T06 & & T012 & & T612 \\ A & (\Delta_0 u(X), 0, 0) & A & (\Delta_0 u(X), 0, 0) & D & (0, \Delta_6 u(X), 0) \\ B & (0, \Delta_6 u(Y), 0) & C & (0, 0, \Delta_{12} u(Z)) & E & (0, 0, \Delta_{12} u(Y)) \end{array} \tag{6}$$

where Δ_t is a potentially time-varying discount factor, and $u()$ is the (stationary) instantaneous utility function. In the data, for $X < Y < Z < 2Y$ (more precisely, for X=100, Y=116, Z=128) we observe $A \succ B$, $C \succ A$, and $D \succ E$. Because we observe $D \succ E$, we know that $\Delta_6 u(X) > \Delta_{12} u(Z)$. We now consider what this implies for hypothetical choice T612', corresponding to

$$\begin{array}{ll} & T612' \\ B & (0, \Delta_6 u(Y), 0) \\ C & (0, 0, \Delta_{12} u(Z)) \end{array} \tag{7}$$

Assuming that $u()$ is weakly concave, $\Delta_6 u(X) > \Delta_{12} u(Y)$ implies $\Delta_6 u(Y) > \Delta_{12} u(Z)$. This follows directly from (potentially weak) concavity and the fact that $X < Y < Z$, and $Y - X > Z - Y$. In other words, in T612' we will have $B \succ C$, which implies $B \succ C \succ A \succ B$ and a violation of transitivity. This violation is present for *any* arbitrary discounting function.[39]

In summary, the data pose a puzzle for modeling inter-temporal choice. The choice pattern is not well explained by the usual candidate models of inter-temporal choice, or by a broader class of models that may have any form of discount function, and any weakly concave utility function, but require transitivity.[40] Furthermore, the choice pattern is very

[39] The violation of transitivity stems from the choice pattern labeled "subadditivity" by Read (2001), although in that study the connection between additivity and transitivity is not made explicit.

[40] Without the usual restriction of time stationarity, the period utility function can vary systematically with distance from the present, and the data need not imply a violation of transitivity. To see this, note that $\Delta_6 u_6(X) > \Delta_{12} u_{12}(Y)$ need not imply $\Delta_6 u_6(Y) > \Delta_{12} u_{12}(Z)$, for example if $u_6(Y) - u_6(X) < u_{12}(Z) - u_{12}(Y)$. Relaxing the usual assumption of a separable discount function would also allow

robust: We see the same patterns in the population as a whole, as well as within different sub-populations, we replicate the results across different data sets, and we find the results despite using real incentives.

4 Subjective perceptions and reference comparisons in intertemporal choice

One potential explanation for time horizon effects, which does not have to do with time preference, is a mis-match between objective magnitudes and subjective perceptions. For example, if doubling objective time duration leads to less than double the subjectively perceived duration, this creates a tendency, all else equal, to be more patient for long time horizons that short horizons, independent of starting date (this type of explanation has been proposed in various forms by, e.g., Read et al., 2001; Rubenstein, 2003; Ebert and Prelec, 2007; Zauberman et al., 2009). We do not measure subjective perceptions directly, but this explanation generates testable predictions, in the sense that salient reference points are well known to influence subjective magnitude judgments (Tversky and Kahneman, 1981)). We exploit the randomization of treatment order in the Pretest data to investigate whether the absence or presence of a salient reference point, in the form of an earlier treatment, influences subsequent choice behavior. Especially if sensitivity to time horizon length is mainly observed for later treatments, rather than the first, this suggests that subjective perceptions of *relative* magnitudes are an important underlying mechanism (explanations based on the shape of an underlying discount function would not predict any effect of reference points).

Figure 2 reproduces the cumulative distribution functions for the Pretest data, but using only first treatments in the left-hand panel, and second and third treatments in the right-hand panel. We see that, in fact, the stylized facts discussed in relation to Figure 1 are present mainly in later treatments. Looking at first treatments, across subjects, there is less evidence of an impact of time horizon on IRRs, and differences are not statistically significant ($p < 0.52$), while there is is a larger and statistically significant time horizon effect for both second and third treatments ($p < 0.01$; $p < 0.03$).[41] Notably, we found

rationalizing the data without a violation of transitivity.

[41] Results are from regressions of IRR on dummies for T012 and T612, conditioning on first, second, and

previously that the effect of time horizon length is still observed when controlling for different possible orders. Thus, while the tendency to be more patient for long horizons appears to be mainly a phenomenon of relative comparisons, the pattern is systematic and emerges for all treatment orders. Thus, the results indicate that, in our setting, people are more patient for long time horizons if they have already seen the "terms" offered for waiting over a shorter horizon, and conversely that people become more impatient for short horizons if they have previously faced choices over a long time horizon. These findings are consistent with time horizon effects reflecting subjective perceptions, which are influenced by relative comparisons.

5 Conclusion

The results of this paper pose an important puzzle for understanding the nature of inter-temporal choice. People are observed to be more impatient for short than long time horizons, contrary to constant discounting, but insensitive to starting date, in contrast to non-constant discounting models where the starting date is crucial. This pattern of inter-temporal choice is also inconsistent with a broader class of models, in that it implies a violation of transitivity. The phenomenon is robust to real incentives, it is pervasive in the population as a whole, as well as different sub-groups, and it replicates across different data sets and different types of measures. Additional evidence points to reference-dependent subjective perceptions as a potential explanation for time horizon effects.

For theory, the results are a call for further research on the nature of inter-temporal choice. Standard models of time preference, both with constant and non-constant discounting, will remain parsimonious, intuitive, and highly useful tools for many applications. The results from the experiments do pose an important challenge, however, as they involve incentivized choices, and relatively simple inter-temporal trade-offs, and yet these models fail to capture important aspects of behavior. Our findings support one particular avenue for future research, which is to incorporate subjective, and reference-dependent, perceptions of time durations and money magnitude into models of inter-temporal choice.

The influence of salient reference points on inter-temporal choice also has potentially important policy implications, and may also provide insights into firm behaviors. Policy

third treatments.

makers may need to take into account how menus of savings options may influence savings behavior, and may even use reference-dependence as a type of policy intervention. The influence of reference comparisons might also help shed light on why firms choose certain ways of presenting information to consumers, about different possible financial assets.

Another implication of our findings is that experimental measures capturing time horizon effects need to be interpreted cautiously. Focusing on only one type of measure can be misleading, and a different interpretation may need to be given to methodologies that use time horizon effects to assess the degree of, e.g., present-bias and self-control problems, in a population of subjects, or to predict life outcomes thought to partly reflect lack of self-control. Indeed, the mixed evidence on relating time horizon effect measures to outcomes (discussed in the introduction) is exactly as one would expect in light of our results. To the extent that time horizon sensitivity mainly reflects reference dependence, one would tend to find a strong relationship between experimental measures and life outcomes only to the extent that the set of reference comparisons available at the time of choice had a close correspondence between the experiment and the field, something that has not been an explicit design goal of most experiments.

Using our own data sets, we have also related experimental measures commonly interpreted as capturing dynamic inconsistency to outcomes such as smoking, poor nutrition, body mass index, low wealth, overdrawn checking account, etc.. (results are available upon request). We occasionally find significant correlations with outcomes, but sometimes signs that are significant and in the "wrong" direction. With corrections for multiple hypothesis testing, almost no results remain significant. Again, this null result is as expected, given our findings that the measures are probably not capturing self-control problems. We did find, however, that the average IRR for an individual across time horizons (or other measures that capture the level of discounting rather than cross-horizon differences) was a reasonably successful predictor of outcomes in both data sets, after corrections for multiple hypotheses. For example, poor nutrition, smoking, having an overdrawn checking account, and low wealth, are all more likely for someone with a higher IRR.[42] These findings are quite consistent with those of, e.g., Chabris et al. (2008), who also use the level of discounting (instantaneous discount rate at time zero) and find similar relationships to a

[42] In Vischer et al. (2011) we reach similar conclusions about the predictive power of the level of impatience, based on relating the T012 measure in the SOEP to various economic outcomes.

similar set of outcomes, or to Harrison et al. (2002). Thus, the results on outcomes are encouraging in that experimental measures do appear to provide a behaviorally meaningful index of the *level* of impatience.

Developing alternative ways to measure self-control problems, besides the traditional paradigm of monetary choice experiments, is clearly an important direction for future research. One possibility is to use actual consumption opportunities, rather than monetary payments, in case this provides a more direct assessment of self-control problems. Some studies have found patterns consistent with declining discounting in an SD design using food rewards over the course of minutes (McClure et al., 2007). It has not yet been shown, however, whether a food-based experiment with a complete design in terms of multiple time horizon comparisons, yields the same puzzling patterns of intransitivity. Other approaches might involve consumptions goods, and varying degree of salience, in terms of, e.g., being able to see the good, or even smell or have a taste (see, e.g., Bushong et al., 2010). Sensitivity to such cues of immediate availability might prove to be a good predictor of self-control problems.[43]

A final point is that the results of this paper do not, by any means, show that self-control problems are not important or prevalent in economic decisions. That people struggle for self-control is intuitive, and in our opinion, self-evident. Furthermore, there are other types of evidence, from the lab and the field, consistent with people struggling with temptation (Read and van Leeuwen, 1998; Read et al., 1999;), or adopting strategies to prevent these tendencies (e.g., Della Vigna and Malmendier, 2006; Ariely and Wertenbroch, 2002; Ashraf et al., 2006; Milkman et al., 2009). Our paper does show, however, that the types of time horizon effects in monetary choice experiments that have often been cited as evidence for self-control problems, may not be capturing self-control problems after all. Again, this calls for further research on developing simple and accurate ways to measure the extent of self-control problems.

[43] See also an interesting, alternative paradigm for monetary choice experiments, holding monetary magnitudes constant and varying delay length, proposed by Noor (2011).

Figures

Figure 1: Cumulative Distributions of IRR

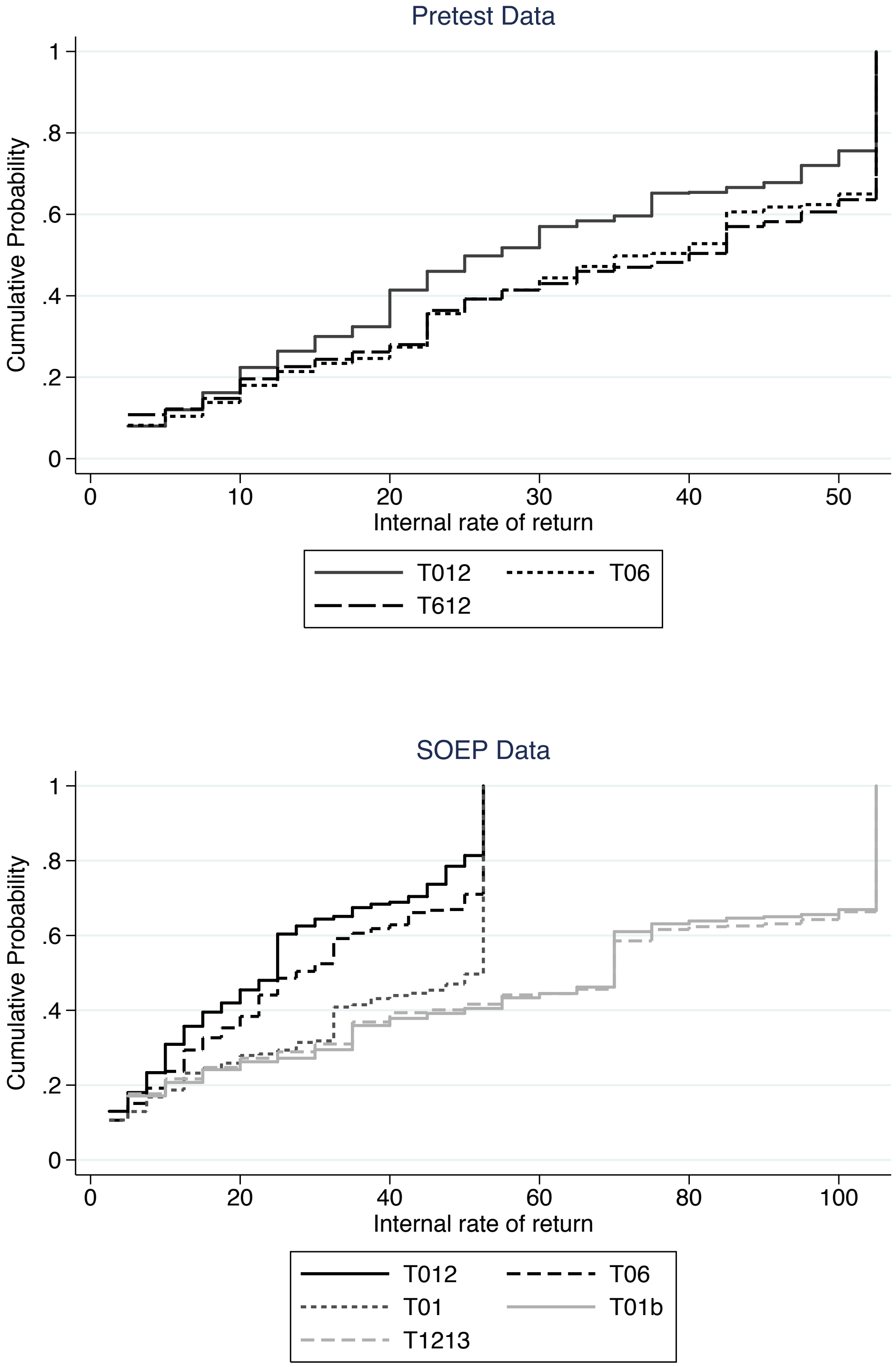
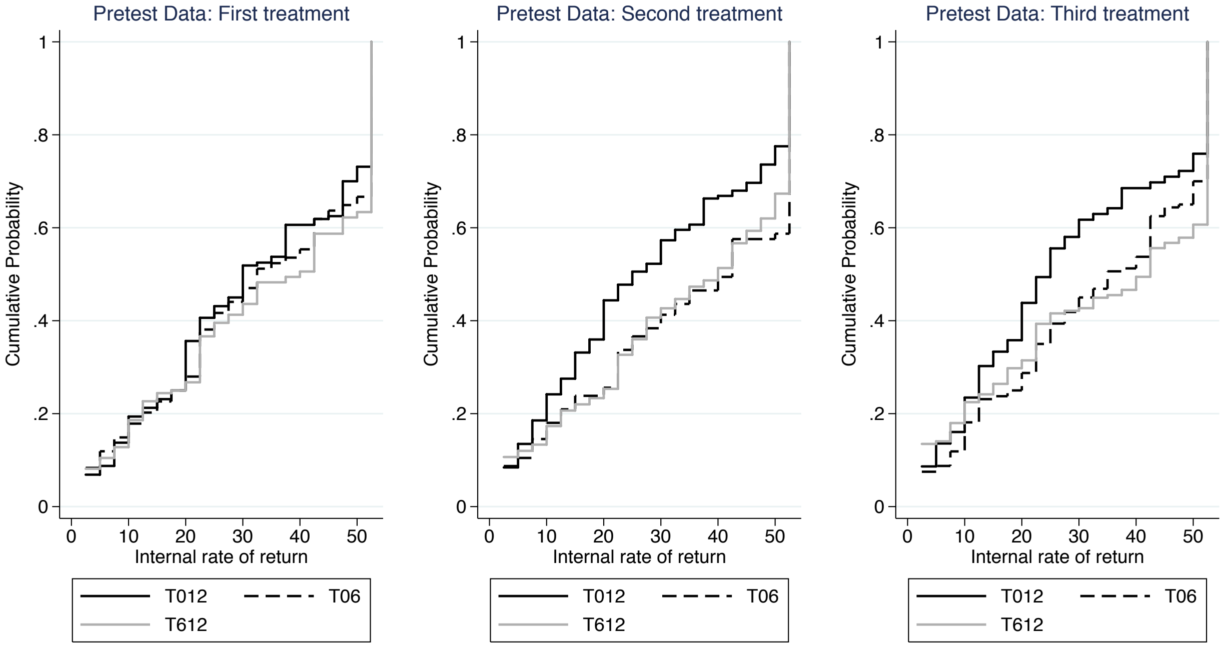


Figure 2: Treatment Order and IRR: Pretest Data



Tables

Table 1: Summary of Treatments

Measure	Data set	Sub-	Early payment	Upper-bound	Obs.
		sample	(in Euro)	IRR	
T012	Pretest	n.a.	100	52.5%	500
T06	Pretest	n.a.	100	52.5%	500
T612	Pretest	n.a.	100	52.5%	500
T012	SOEP	1 & 2	200	52.5%	977
T06	SOEP	1	200	52.5%	490
T01	SOEP	2	200	52.5%	487
T01b	SOEP	3	200	105%	526
T1213	SOEP	3	200	105%	526

Table 2: Descriptive Statistics for IRR by Time Horizon

Measure	Data set	Mean	Median	S.D. of	Fraction		Mean test	Med. test
		IRR	IRR	IRR	right-cens.			
T012	Pretest	29.40	27.50	18.19	0.24	T012 vs. T06	$p < 0.01$	$p < 0.01$
T06	Pretest	33.56	37.50	18.10	0.35	T06 vs. T612	$p < 0.78$	$p < 0.05$
T612	Pretest	33.76	40.00	18.78	0.36	T612 vs. T012	$p < 0.01$	$p < 0.01$
T012	SOEP	26.07	25.00	18.41	0.19	T012 vs. T06	$p < 0.01$	$p < 0.01$
T06	SOEP	29.62	27.50	18.89	0.29	T06 vs. T01	$p < 0.01$	$p < 0.01$
T01	SOEP	36.56	52.50	19.41	0.53	T01 vs. T012	$p < 0.01$	$p < 0.01$
T01b	SOEP	60.87	70.00	39.08	0.33	T01 vs.T01b	$p < 0.01$	$p < 0.01$
T1213	SOEP	60.89	70.00	39.83	0.34	T01b vs. T1213	$p < 0.99$	$p < 0.88$

Table 3: IRR as a Function of Time Horizon, by Demographic Groups: Pretest Data

	All	Males	Females	Age	Age	IQ	IQ	Less	More	Income	Income	Not credit	Credit
				$\leq$med.	>med.	$\leq$med.	>med.	educated	educated	$\leq$med.	>med.	Constrained	Constrained
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
T12	-6.42***	-8.20***	-4.82***	-6.56***	-6.34***	-6.09***	-6.66***	-6.63***	-5.83***	-5.75***	-6.85***	-6.64***	-5.03**
	(0.85)	(1.20)	(1.19)	(1.05)	(1.39)	(1.53)	(0.96)	(1.04)	(1.42)	(1.32)	(1.10)	(0.92)	(2.17)
T612	0.14	0.25	0.07	-1.26	1.75	2.22	-1.38	0.71	-1.30	2.05	-1.11	-0.77	3.80
	(1.11)	(1.50)	(1.63)	(1.47)	(1.70)	(1.68)	(1.48)	(1.32)	(2.08)	(1.72)	(1.45)	(1.21)	(2.91)
Constant	3.32***	3.28***	3.36***	3.22***	3.44***	3.37***	3.28***	3.35***	3.23***	3.25***	3.36***	3.29***	3.36***
	(90.12)	(61.59)	(65.06)	(66.83)	(60.65)	(58.95)	(68.20)	(74.40)	(48.49)	(57.18)	(68.92)	(78.41)	(40.60)
Observations	1500	693	807	768	732	657	843	1113	387	591	909	1131	330

Notes: Interval regression estimates. Dependent variable is the IRR for a given time horizon, with three time horizons (observations) per individual. Age groups are defined as less than or equal to the median age, and greater than median age, respectively. Low and high IQ indicate below and above median cognitive ability respectively. Less and more educated indicate whether an individual did not, or did, complete the Abitur, a college entrance exam in Germany. Low and high income indicate below and above median household income, respectively. Credit constraints are measured by a question asking about ability to borrow money in the event of an unexpected expense. In parentheses, robust s.e., adjusted for clustering on individual. *, ** indicates significance at 10 and 5 percent level.

Table 4: IRR as a Function of Time Horizon, by Demographic Groups: SOEP Data

	All	Males	Females	Age	Age	IQ	IQ	Less	More	Income	Income	Not credit	Credit
				≤med.	>med.	≤med.	>med.	educated	educated	≤med.	>med.	Constrained	Constrained
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Sample 1: T012 vs. T06													
T012	-5.26***	-4.86***	-5.66***	-5.90***	-4.58***	-3.08***	-6.55***	-5.33***	-4.97***	-3.20**	-5.77***	-5.33***	-5.07***
	(0.68)	(0.89)	(1.03)	(0.94)	(0.98)	(0.93)	(0.93)	(0.77)	(1.44)	(1.49)	(0.76)	(0.72)	(1.87)
Constant	31.17***	28.31***	34.15***	30.48***	31.85***	31.87***	30.74***	34.11***	21.46***	29.91***	31.49***	27.63***	43.85***
	(1.34)	(1.86)	(1.91)	(1.87)	(1.91)	(2.13)	(1.72)	(1.62)	(2.17)	(2.98)	(1.49)	(1.41)	(3.22)
Observations	980	504	476	486	494	358	622	730	218	198	782	744	230
Sample 2: T012 vs. T01													
T012	-19.67***	-21.05***	-18.46***	-19.10***	-20.28***	-18.63***	-20.26***	-20.15***	-16.64***	-19.90***	-19.60***	-17.87***	-27.41***
	(1.35)	(2.04)	(1.80)	(1.97)	(1.86)	(2.12)	(1.75)	(1.62)	(2.39)	(2.64)	(1.57)	(1.44)	(3.70)
Constant	44.96***	46.61***	43.51***	41.08***	48.84***	46.24***	44.17***	47.50***	32.35***	42.77***	45.80***	41.10***	60.49***
	(1.80)	(2.64)	(2.47)	(2.42)	(2.70)	(2.83)	(2.34)	(2.11)	(3.51)	(3.18)	(2.18)	(1.94)	(4.49)
Observations	974	444	530	468	506	352	622	766	180	238	736	762	212
Sample 3: T1213 vs. T01													
T1213	-1.45	-0.22	-2.42	-0.10	-3.14	2.32	-3.15	1.78	-7.48	-6.56	0.23	-0.93	-3.60
	(3.36)	(5.03)	(4.52)	(4.33)	(5.27)	(6.28)	(3.95)	(4.15)	(6.03)	(6.94)	(3.84)	(3.70)	(7.12)
Constant	97.25***	96.53***	97.83***	91.92***	103.49***	118.21***	85.41***	102.48***	82.34***	96.81***	97.37***	89.92***	139.69***
	(5.16)	(7.72)	(6.94)	(6.79)	(7.89)	(9.91)	(5.86)	(6.36)	(8.85)	(10.48)	(5.93)	(5.25)	(18.21)
Observations	1052	464	588	552	500	424	628	784	240	254	798	860	188

Notes: Interval regression estimates, separately by sub-sample of the SOEP data. Dependent variable is the IRR for a given time horizon, with two time horizons (observations) per individual. Low and high IQ indicate below and above median cognitive ability respectively. Less and more educated indicate whether an individual did not, or did, complete the Abitur, a college entrance exam in Germany. Low and high income indicate below and above median net personal income, respectively. Credit constraints are measured by a question asking about ability to borrow money in the event of an unexpected expense. In parentheses, robust s.e., adjusted for clustering on individual. *, ** indicates significance at 10 and 5 percent level.

Table 5: Individual types, Pretest data

Types	Constant	Declining	Increasing	$IRR_{T06} = IRR_{T612}$ & $IRR_{T612} > IRR_{T012}$	$IRR_{T06} \neq IRR_{T612}$ & $IRR_{T06} > IRR_{T012}$ & $IRR_{T612} > IRR_{T012}$	Other
Percent	11.34	11.63	10.76	20.06	25.58	20.64
Types allowing for "error"	Constant	Declining	Increasing	$IRR_{T06} = IRR_{T612}$ & $IRR_{T612} > IRR_{T012}$	$IRR_{T06} \neq IRR_{T612}$ & $IRR_{T06} > IRR_{T012}$ & $IRR_{T612} > IRR_{T012}$	Other
Percent	31.21	8.48	7.88	18.48	9.70	24.24

Notes: Other includes individuals who violate discounting predictions in the various other ways: IRRs for both short intervals are less than for the long interval; IRR for one short horizon is greater (less) than IRR for the longer horizon, while IRR for the other short horizon is equal to IRR for the long horizon. In the first row, the sample (N=344) excludes individuals for whom right-censoring prevents unambiguous classification: Individuals with censoring in two or more time horizons. In the second row, the same restriction applies, but 14 additional individuals are excluded (N=330). This occurs for two reasons: 4 are excluded because of "cycling" or intransitivity in the relations between the three horizon measures . E.g., interval for T06 overlaps with interval for T012 but is strictly above interval for T612, while interval for T012 overlaps with interval for T612, implying intransitivity because $IRR_{06} = IRR_{012}$, $IRR_{06} > IRR_{612}$, but $IRR_{012} = IRR_{612}$. The remaining 10 that are excluded involve the IRR for one time horizon being right censored, while the IRRs for both of the others are lower. Once we allow for errors, one or more of the lower horizons overlaps with the censored observation, and thus are no longer unambiguously classifiable.

Table 6: Stylized Facts on IRR and Time Horizon, and Different Models

	Constant discounting	Declining discounting	Increasing discounting	$\beta - \delta$, two-system, or “fixed cost” discounting, with present$\leq$ 2 days	Any discounting with arbitrage
Pretest data					
Finding 1					
T06>T012		Yes			
Finding 2					
T06=T612	Yes			Yes	Yes
Finding 3					
T612>T012			Yes		
SOEP data					
Finding 4					
T06>T012		Yes			
Finding 5					
T01>T012		Yes			
Finding 6					
T01>T06		Yes			
Finding 7					
T01b=T1213	Yes			Yes	Yes

Notes: Findings compare mean (median) IRRs for different time horizons. Table entries of “Yes” indicate stylized facts that a given model can explain. Declining discounting includes, but is not limited to, hyperbolic discounting, quasi-hyperbolic discounting with the present > 2 days, and fixed cost discounting assuming present > 2 days. Discounting with arbitrage refers to predictions when individuals do not treat monetary payments as consumption.

A Experiment Instructions

In the following we present a translation of the German instructions. Instructions were presented to the interviewer on the screen of the laptop computer, and were read aloud to the subjects by the interviewer.

Screen 1

Now that the interview is over we invite you to participate in a behavioral experiment, which is important for economic science. The experiment involves financial decisions, which you can make in any way you want to. The questions are similar to those asked in the questionnaire with the exception that THIS TIME YOU CAN EARN REAL MONEY!

I will first explain the decision problem to you. Then you will make your decisions. A chance move will then determine whether you actually earn money.

Every 7th participant wins!

HOW MUCH MONEY YOU WILL EARN AND AT WHICH POINT IN TIME WILL DEPEND ON YOUR DECISIONS IN THE EXPERIMENT.

If you are among the winners, your amount will be paid by check. In this case the check will be sent to you by post.

Screen 2

Participants were then shown a choice table for the respective experiment as an example. The table was printed on a green piece of paper and was handed to participants for them to study.

The experimenter continued explaining how the experiment would work.

The interviewer gave the following explanation:
In each row you see two alternatives. You can choose between

- A fixed amount of 100 Euro (column A "today")
- and a somewhat higher amount, which will be paid to you only "in 12 months" (column B).

Payment "today" means that the check you get by post can be cashed immediately.
Payment "in 12 months" means that the check you get can be cashed only in 12 months.

You start with row 1 and then you go down from row to row. In each row you decide between 100 Euro today (column A) and a higher amount (column B); please always keep the timing of the payments in mind. The amount on the left side always remains the same, only the amount on the right side increases from row to row.

Which row on one of the tables will be relevant for your earnings will be determined by a random device later.

Screen 3

As you can see, you can earn a considerable amount of money. Therefore, please carefully consider your decisions.

Can we start now?

If the participant agreed, the experiment started. If not, the experimenter said the following:

The experiment is the part of the interview where you can earn money! Are you sure that you DO NOT WANT TO PARTICIPATE?

If the participant still did not want to participate, the experiment was not conducted and the participant answered a few final questions. In case the subject wanted to participate the experiment began.

Participants studied their table. The experimenter asked for the subject's decision in each row, whether they preferred the option in Column A or B, starting with the first row. In case a participant preferred the higher, delayed amount the experimenter asked:

You have decided in favor of the higher amount of X in X months. We assume that this implies that for all higher amounts you also prefer the later payment, meaning that for all remaining rows all higher amounts will be selected (i.e., Column B).
If the participant did not agree, he kept on deciding between columns A and B.

Once the first table was completed, the second table was presented to the participant. The experimenter then said:

Now there is a second table. Please look at the table. You will do the same as before but please note that the dates of payment and also the payments on the right side of the table have changed.
For the second and third tables, the same procedure as with the first table was followed.

When the tables were completed, participants were asked whether they had thought about interest rates during the experiment and if so, which interest rate they had in mind and whether they had compared this interest rate with those implied in the decision tables. They were then asked what they would do with the 100 Euro from the experiment within the next weeks. Alternative answers were, "spend everything", "spend most of the money and save something", "save most of it and spend something" and "save everything", or "no reply".

Then it was determined whether the participant was among those who would be paid. Participants could choose their "lucky number" between 1 and 7. They could then press on one out of seven fields on the computer, which represented numbers from 1 to 7. If they hit "their" number they won, otherwise they did not win. In case they won, it was *determined which of the tables was selected and which row of the respective table. This was done again by pressing on fields presented to participants on the computer screen. In the end subjects who had won were informed that they would be sent the check by mail.*