Reproducibility Project: Replication Report

Replication of Experiment 2 by Read & Scholten (2012, *Journal of Experimental Psychology: Language, Memory and Cognition*)

Ayse Zeynep Enkavi
zenkavi@stanford.edu

Introduction

The aim of this study was to investigate systematic biases in intertemporal decisions when they are framed as sequences instead of unitary outcomes using the tradeoff model as proposed by Read & Scholten (2010). The tradeoff model proposes that decisions between intertemporal choices with unitary outcomes are made by comparing the difference between valued outcomes with the difference between weighted delays. When the difference in weighted delays outweighs the difference between valued outcomes one chooses the smaller sooner amount and vice versa. A tradeoff parameter scales the difference between the two attributes and the functions to weigh time and amounts are concave to account for diminishing sensitivity. This is different than discounting models prevalent in the literature, where the value of the delayed outcome is weighted by the delay. The main idea of the original study is that the tradeoff model accounts for the anomalies in behavior better than the any of the existing models.

Intertemporal choices, as studied in the lab, can be framed as two-outcome sequences without normatively changing their value, in the sense that the addition of common consequences should cancel out in the comparison according to traditional discounting models. More formally, this is called the independence axiom and is a fundamental assumption of most, if not all, expected utility theories. In the specific experiment I am interested this addition of common consequences is done in three ways:

1. By explicitly stating the zero-outcome at each time point. For example a choice between $10 today versus $30 in two months becomes one between $10 today and $0 in 2 months versus $0 today and $30 in two months. The interest rates with this manipulation stay the same and this is manipulated between subjects. Preference for larger later rewards (LL) is expected to be higher with this manipulation (hidden-zero effect).

2. By adding the same amount to the earlier time point for both options (A_0). For example the choice between $10 today versus $30 in two months becomes one between $510 today versus $500 today and $30 in two months. The interest rates with this manipulation stay the same again and this is manipulated within subjects using $500 and $1000. Preference for LL is expected to decrease with this manipulation (front-end amount effect) when the LL is small compared to A_0 but might reverse and increase as LL gets larger.

3. By adding the same amount to the earlier time point for the sooner reward and and the later time point for the later reward ($A_{0:2}$). For example the choice between $10 today versus $30 in two months becomes one between $510 today versus $530 in two months. The interest rates with this manipulation decrease and this is manipulated within subjects using $500 and $1000. Preference for LL is expected to decrease with this manipulation (relative magnitude effect).

The extended tradeoff models suggests that intertemporal decisions with non-unitary outcomes (sequences with two outcomes) are simplified by treating the sequences as unitary outcome decisions with the total absolute value of the sequence to be received at an average delay depending on the magnitude of the outcomes to be received at different time points. As the example in the paper clarifies: A sequence of receiving $500 today and $30 in 2 months is treated as receiving $530 in 2 weeks, where the sooner delay weighs more compared to the later one because the amount to be received at that time point is larger. Additionally the model accounts for preference for spreading (σ), an anomaly discovered in earlier studies using sequences: accordingly deviations from a uniform distribution ($d(x_1, x_2)$) in the rewards of a sequence are undesirable and detract from the total value of the sequence. Overall then the trade-off rule suggests that choices with non-unitary outcomes are made by comparing the difference between valued absolute totals of sequences ($v(x_1 + x_2)$), that decrease in value the more they deviate from a uniform distribution, with the difference between weighted adjusted delays ($\omega(\widehat{t_L}) - \omega(\widehat{t_s})$). This has been described mathematically as below in the original paper, where the details of the amount valuation and delay adjustment functions can be found:

$$v(x_{L1} + x_{L2}) + v(x_{S1} + x_{S2}) - \sigma(d(x_{L1}, x_{L2})) - (d(x_{S1}, x_{S2})) = \kappa(\omega(\widehat{t_L}) - \omega(\widehat{t_s}))$$

There are four target findings of the experiment:

1. Relative magnitude effect: decrease of probability of LL as $A_{0:2}$ increases due to diminishing marginal utility in the valuation of absolute amounts;

2. Hidden zero effect: increase of probability of LL as zero outcomes are stated explicitly because the adjusted delay (and the deviance from uniformity in certain cases) of the smaller sooner sequence outweighs the deviance from uniformity of the larger later sequence;

3. Front-end amount effect: decrease in probability of LL as A_0 increases when diminishing sensitivity and preference for spreading outweigh the delay adjustment of the smaller sooner sequence in the explicit zero condition and when diminishing sensitivity outweighs the preference for spreading and delay adjustment in the explicit zero condition.

4. Reversal of front end amount effect: increase in probability of LL as $A_{0:2}$ increases in addition to A_0 because the increase in A_2 offsets the advantage of diminishing sensitivity for the smaller sooner sequence.

These findings violate the independence axiom of conventional discounting models but are accounted for by the extended tradeoff model.

Methods

Power Analysis

Post-hoc power analyses completed by G*Power software for the reported effect sizes and degrees of freedom for the repeated measures anova as well as simulations indicated that the study was overpowered with even 277 subjects. We therefore propose collecting 300 subjects.

Planned Sample

Planned sample size is 300 and study will be ended once this sample is collected. MTurk workers from the U.S. only with approval ratings above 97% will be recruited. Demographics data will be collected and compared to the original sample.

Materials

Subjects respond to 9 questions consisting of intertemporal decisions such as receiving $10 today versus receiving $30 in two months in one group or receiving $1,010 today and $0 in 2 months versus $500 today and $30 in months when A_0.= 500, $A_{0:2}$ = 500 and the condition is explicit-zero in another group. Participants are randomly assigned to the implicit-zero condition or the explicit-zero condition. The items are depicted in the table from the original paper below. The order of the stimuli in Table 1 was randomized across participants. Following these questions participants complete a demographics questionnaire asking their age, gender, ethnicity, their education level (in second sample), whether they smoke or use recreational drugs and a question on their decision strategy asking whether the amounts or the delays were more important in their decision process.

Table 1: Adding a Common Amount ($A_{0:2}$) to the Immediate Outcome of Smaller But Sooner (SS) and the Delayed Outcome of Larger but Longer (LL) and Adding a Common Amount (A_0) to the Immediate Outcomes of Both Options in the Explicit-Zero Condition. Note: Implicit-zero condition is obtained by suppressing zero outcomes. Delays are in months. (Reproduced from Read & Scholten, 2012)

	$A_{0:2}$		
	$0	$500	$1000
A_0	**A**	**D**	**G**
$0	($10, 0; $0, 2) ($0,0; $30,2)	($510, 0; $0, 2) ($0,0; $530,2)	($1010, 0; $0, 2) ($0,0; $1030,2)
	B	**E**	**H**
$500	($510, 0; $0, 2) ($500,0; $30,2)	($1010, 0; $0, 2) ($500,0; $530,2)	($1510, 0; $0, 2) ($500,0; $1030,2)

	C	F	I
$1000	($1010, 0; $0, 2) ($1000,0; $30,2)	($1510, 0; $0, 2) ($1000,0; $530,2)	($2010, 0; $0, 2) ($1000,0; $1030,2)

Procedure

The original study was completed by: “A total of 277 Portuguese residents (42% male, average age 30 years, 65% having at least completed college or university, and 74% being employed or a student)”. The current replication will be completed by a total of 300 Amazon Mechanical Turk workers participating by completing an online questionnaire. We chose not to recruit a larger sample because the reported effect sizes were large and both post-hoc power analyses and simulation indicated that we should be able to detect our effects of interest with this sample size. Demographics data will be compared to the original sample.

Analysis Plan

We will conduct “a 3 (common amount added to the immediate outcome of SS and the delayed outcome of LL, denoted as $A_{0:2}$) X 3 (common amount added to the immediate outcomes of both options, denoted as A_0) X 2 (implicit- or explicit-zero outcomes) mixed analysis of variance.” This will be the main analysis to be completed as it is the main behavioral analysis for this experiment. The expected interaction between $A_{0:2}$ and A_0 will be followed up with one way ANOVAs on data split by the $A_{0:2}$ amount.

Following up on this we will also fit the tradeoff model described above (and in the equation 3 of the original paper) to the aggregate data of all subjects (i.e. 18 data points for each test item) to obtain predicted proportions of the larger later reward replicating figure 3 in the original paper.

Differences from Original Study

This sample will consist of U.S. citizens and residents instead of Portuguese residents. We are not certain whether these questions were embedded in a larger study in the original paper but our replication will be stand alone followed by demographics questions. It is unclear whether any subjects were excluded. We are also unclear on exact compensation amount but our replication paid subject $0.15. Notably, however, none of these are expected to change any of the experimental results.

(Post Data Collection) Methods Addendum

Actual Sample

First replication sample consisted of 298 subjects after excluding two subjects for incomplete data due to technical errors. Second replication sample consisted of 292 subjects after excluding eight subjects for incomplete data due to technical errors. Detailed demographics can be found in Table 2.

Differences from pre-data collection methods plan

Forgot to collect education data in demographics questionnaire of first sample. Collected another sample with education data. Details on demographics for this additional sample can also be found in Table 2.

Results

Table 2: Summary of demographics and mixed anova results of interest as reported in the original paper and reproduced in the first column and estimated in two independent replication samples.

	Scholten & Read (2012)	**Batch 1**	**Batch 2**
Sample	N= 277, 42% male, age = 30, education = 65% college grad	N = 298, 63% male, age = 32, education = NA	N = 292, 65% male, age = 32, education = 55% college grad
Relative magnitude effect	$F(2, 550) = 125.46$, $p < 0.005$ $\eta_p^2 = 0.31$	$F(2, 592) = 203.397$, $p < 0.001$ $\eta_p^2 = 0.40$	$F(2, 580) = 151.851$, $p < 0.001$ $\eta_p^2 = 0.34$
Hidden zero effect	$F(1, 275) = 41.07$, $p < 0.005$ $\eta_p^2 = 0.13$	$F(1, 296) = 1.899$, $p = 0.169$ $\eta_p^2 = 0.006$	$F(1, 290) = 6.015$, $p = 0.015$ $\eta_p^2 = 0.02$
A_0 interaction with $A_{0:2}$	$F(4, 1100) = 48.76$, $p < 0.005$ $\eta_p^2 = 0.15$	$F(4, 1184) = 27.12$, $p<0.001$ $\eta_p^2 = 0.08$	$F(4, 1160) = 33.81$, $p < 0.001$ $\eta_p^2 = 0.10$
Front end amount effect	$F(2, 550) = 17.79$, $p < 0.005$ $\eta_p^2 = 0.06$	$F(2, 592) = 14.198$, $p < 0.001$ $\eta_p^2 = 0.05$	$F(2, 580) = 19.452$, $p < 0.001$ $\eta_p^2 = 0.06$
Reversal of front end amount effect	$F(2, 550) = 61.12$, $p < 0.005$ $\eta_p^2 = 0.18$	$F(2, 592) = 32.739$, $p < 0.001$ $\eta_p^2 = 0.10$	$F(2, 580) = 38.84$, $p < 0.001$ $\eta_p^2 = 0.12$

Data preparation

Data were collected on lab servers in individual log files for each subject. Log files consisted of lines for subjects MTurk ID's, date and time they began the HIT, test data (coding the amounts, delays, locations of all options, condition, choice and reaction times) and demographics information. Script to import, parse and merge all subjects' data into a single dataframe with one trial in each row can be found at https://github.com/zenkavi/PSYCH254/blob/master/rsrep-experiment/DataImport_rsrep.Rmd.

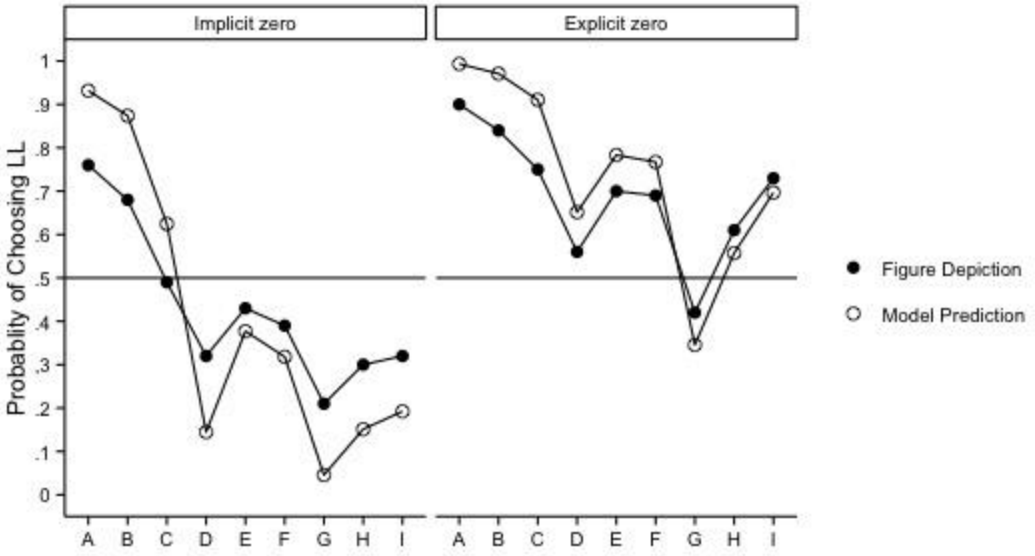
Confirmatory analysis

The results of the 3 X 3 X 2 mixed analysis of variance and the follow-up one-way ANOVAs disentangling the interaction can be found in Table 2. We replicated three of the four expected effects with comparable effect sizes. Namely, participants were less patient as the relative difference between the smaller sooner reward and larger later reward increased (relative magnitude effect as $A_{0:2}$ increases $F(2, 592) = 203.397$, $p < 0.001$, $\eta_p^2 = 0.40$); less patient when the smaller sooner reward of both sequences increased (front-end amount effect $F(2, 592) = 14.198$, $p < 0.001$, $\eta_p^2 = 0.05$) but more patient when the larger later reward was increased concurrently ($F(2, 592) = 32.739$, $p < 0.001$, $\eta_p^2 = 0.10$) .

Confusingly we failed to replicate one of the better-documented contextual effects: the hidden-zero effect, i.e. increase in patience in the explicit zero condition. In the first sample we failed to find a main effect of the between subjects condition on the proportion of patient choices ($F(1, 296) = 1.899$, $p = 0.169$, $\eta_p^2 = 0.006$). We followed up on this in multiple ways as detailed in “Exploratory Analyses” below.

The second part of the analyses consisted of fitting the tradeoff model to the aggregate data of all subjects. Aside from replicating their results the motivation for this came from the ease to interpret the parameters of the model as they were “psychologically-informed”. Yet, this was not a very straightforward procedure. The first problem we ran into was in estimating the predicted proportions reported in the original paper figure 3 with the parameter estimations reported in table 3. When we tried using their parameter estimates to estimate the proportion of patient choices for each item we got a similar pattern of responses but the percentages varied as depicted in Figure 1. In fact the RMSD comparing the figure depictions to the model predictions with the reported parameter estimates was 0.123. We were unable, however, to determine what adjustment might have lead to this slight difference in predictions. For the remainder of the modeling efforts we used the model that yielded the discrepant results keeping in mind that it might not be identical to what the authors have used.

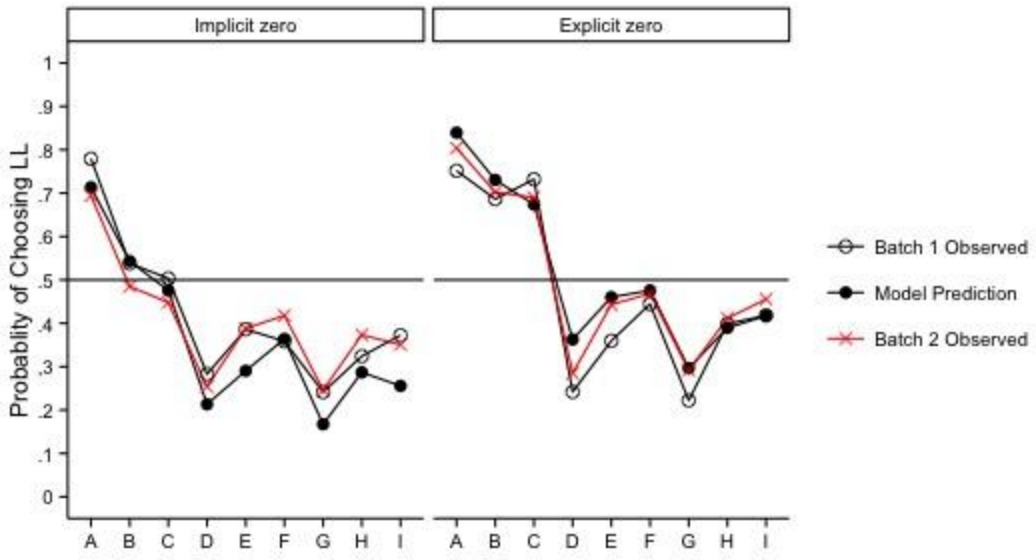
Figure 1: Comparison of percentage of patient responses as predicted by the parameter estimates reported in the original paper with the values reported in figure 3 of the original paper. The general pattern is similar but the percentages are not indicating an unreported adjustment to the implementation of the model.



Secondly we used their parameter estimates (reproduced in Table 3 first column) to see how well it predicted our data. The RMSD of this attempt was 0.237 for the first batch and 0.230 for the second batch. Finally, we estimated the parameters for our data using the RMSD for the combination of parameters as the performance measure. Our best fit is reproduced in Table 3 with an RMSD 0.068. The error parameter was allowed to vary between 1 and 7,

diminishing outcome sensitivity varied between 0 and 0.008, preference for spreading for large differences between 0 and 0.03, preference for spreading small differences between 0 and 1, tradeoff parameter 0 and 10, diminishing delay sensitivity was fixed at 10^{-11} to represent convergence to 0 as the delays in the experimental setup were held constant and the weighted delay adjustment varied between 0 and 1. Over 25000 combinations were tried before settling on this combination. The second replication batch served as a test set for the model and the RMSD was 0.054 using our best fit parameters.

Figure 2: Observed and predicted probability of choosing LL for each test item of the first replication sample. Model predictions are based on our best fit (i.e. not using their parameter estimates) to the first batch of data. RMSD was 0.068. Batch 2 data was used as a test set. The parameters from batch 1 gave an RMSD of 0.054 for batch 2.



There are a few things to note when comparing our parameter estimates to the ones reported in the original paper. First we should keep in mind that our fitting procedure might have encountered a local minima and is less than optimal. Secondly there most likely is an adjustment to the model that the authors have used that is unbeknownst to us. Third, our estimation of the error parameter is much larger than their estimate indicating larger noise in estimation. Large discrepancies were also found with the other parameters as well indicating a lack of stability of the model. Our estimate for diminishing sensitivity to accumulated outcomes was almost three times as large as their estimate. This would imply that our samples were less sensitive to changes in accumulated outcomes. Furthermore our estimate of preference for spreading of constituent outcomes for large differences converged to zero which simplified the model for certain items and implied that preference for spreading only mattered for small differences. In other words when the deviations from the uniform distribution was very different for the two sequences other factors determined the decision. An example of this is item E: when choosing between $1,010 now or $500 now and $530 later while the SS deviates $505 from a

uniform distribution LL deviates only $15. Yet subjects preferred the SS sequence more often in all conditions. When the difference between the deviations were small, however, our estimates indicated that this factor mattered much more than the original paper. Despite all these differences there were also convergences in the values of certain parameters. Both the tradeoff parameter and the delay averaging parameters were relatively similar to the values reported in the original paper. These similarities suggest that the model might be capturing indeed a common tendency in how people aggregate multiple delays and how important one attribute is compared to the other at least in the limited presentation structure of such experimental stimuli.

Table 3: Adjusted tradeoff model estimates. First column reproduces estimates reported in original paper. Second column is our estimates on batch 1 data.

	Scholten & Read (2012)	**Batch 1**
ε - Error	2.3954	6.0000
γ **- Diminishing sensitivity to accumulated outcomes**	0.0026	0.0073
σ_U **- Preference for spreading of constituent outcomes (large difference)**	0.0023	0.0000
σ_u **- Preference for spreading of constituent outcomes (small difference)**	0.0138	0.1090
α **- Tradeoff between time and outcome advantages**	6.3775	6.0000
τ **- Diminishing sensitivity to adjusted delays**	0.0000	0.0000
***q* - Departure from a weighted averaging of delays**	0.2995	0.3000
R^2 - Goodness of fit	*0.91*	*0.98**
R^2_{Adj} - Adjusted goodness of fit	*0.88*	*0.98**
RMSD - Badness of fit	*0.06*	*0.068*

**R^2 estimates are based on regressions of the observed probabilities on the predicted probabilities without an intercept. Therefore the adjusted R^2 statistic is unlikely to be correcting for the correct number of parameters. We used this method because restricting the slope to be 1 yielded negative results and this was how we were able to obtain similar values for their observed and predicted values reported in the paper so it was our best guess as to how they might have been calculated. Parameter selection was based not on these but on RMSD which was independent of any linear relationship constraint.*

Exploratory analyses

The lack of explicit zero effect in the first batch in the mixed ANOVA was first followed up with a one-way ANOVA. This yielded a significant explicit zero effect ($F(1,2680) = 7.36$, $p = 0.007$, $\eta_p^2 = 0.003$). This analysis ignores, however, the repeated measures nature of the data and inflates the degrees of freedom. It was only used to prompt thinking about the potential interactions of the explicit zero effect with the within subjects manipulations (additions to the front-end A_0 and diagonal additions $A_{0:2}$). In fact, these interactions were part of the initial mixed ANOVA (parts of which are reported in Table 1) and both were significant ($F(2, 592) = 4.639$, $p = 0.01$, $\eta_p^2 = 0.015$ for the interaction with the diagonal addition and $F(2, 592) = 8.97$, $p < 0.001$, $\eta_p^2 = 0.029$ for the interaction with the front end amount). These results implied a dependence of the explicit zero effect on the within subjects manipulations but did not clarify the exact relationship. To get a better understanding of this we looked at each item more carefully.

We found that the explicit zero effect depended both on the A_0 and $A_{0:2}$ in differing ways. As Table 4 demonstrates the hidden zero effect size in batch 1 increases as the front end amount A_0 increases and decreases as the diagonal addition $A_{0:2}$ increases.

Notably, the explicit zero effect appears to consistently fail in batch 1 when the absolute difference between the smaller sooner amount and the larger later amount is constant (i.e. $30 when A_0 = $0) but it does not fail in the same way depending on the diagonal addition $A_{0:2}$. Specifically, while subjects were mostly willing to wait for item A in both the implicit (77.93%) and explicit (75.16%) condition, the opposite was true for items D and G in both the implicit (28.28% and 24.14% respectively) and explicit (24.18% and 22.22% respectively) conditions. In other words the explicit zero effect is outweighed by the effect of the absolute magnitude difference (people don't care how you frame it when the difference is only $30).

As for item E, where the explicit zero effect also failed in batch 1, people are overwhelmingly impatient (61.38% in the implicit condition and 64.05% in the explicit condition), which means that subjects prefer receiving $1,010 today to receiving $500 today and $530 in 2 months. Our data indicates a lack of a preference for spreading (for large differences) though this would not explain the lack of the explicit zero effect for the item.

Table 4: Size of explicit zero effect in (proportion of LL in explicit condition - proportion of LL in implicit condition). Batch 1 is in first row of each cell and batch 2 is in second row. Positive numbers indicate increasing patience in explicit zero condition.

	$A_{0:2}$		
	$0	$500	$1000
A_0	**A**	**D**	**G**
$0	-2.77% 10.98%	-4.09% 3.11%	-1.92% 4.49%
	B	**E**	**H**
$500	14.83% 21.75%	-2.67% 5.50%	7.46% 3.83%

	C	F	I
$1000	22.86% 24.21%	8.58% 5.04%	4.59% 10.49%

Another reason why we have not gotten a main effect of the hidden zero effect in the first replication attempt might be due to gender differences. There is over a 20% difference between the percentage of males in the original sample versus our first sample. We checked for whether this might be the case in two ways. First we randomly sampled males from our sample to have an overall sample with 42% males and ran the mixed ANOVA on this sample. The hidden-zero effect was again not significant ($F(1, 188) = 0.717$, $p = 0.398$, $\eta_p^2 = 0.004$). Alternatively we could also use our whole sample and include gender as another between subjects factor. The gender hidden-zero effect interaction was not significant ($F(1, 294) = 0.157$, $p = 0.692$, $\eta_p^2 = 0.001$). Importantly, in both of these methods the other effects of interest remain significant. Based on these analyses gender differences are not likely to be the reason why we did not get the hidden zero effect.

Perhaps most importantly we tried replicating the results in yet another independent sample (batch 2). In this sample we found a significant hidden-zero effect as reported in Table 2. But the effect size was not nearly as large as the original study and we also found a significant interaction of the effect with the diagonal addition $A_{0:2}$ ($F(2, 580) = 9.092$, $p < 0.001$, $\eta_p^2 = 0.02$). As seen in Table 4, though there seems to be a positive hidden-zero effect for each item the size of the effect is largest when $A_{0:2} = \$0$ but decreases as $A_{0:2}$ grows.

Discussion

Summary of Replication Attempt

In this replication we were interested in seeing whether adding common consequences to intertemporal decisions that should be irrelevant changed preference. Specifically we wanted to see whether and how framing the same decision as a unitary outcome or as a sequence by explicitly stating the zero amounts received at each time point, the addition of the same amount to both outcomes and the addition of the same amount to the sooner time point of each option changed the proportion of patient choices. We expected patient choices to increase when unitary outcomes were framed as sequences (hidden-zero effect), to decrease with the addition of the same amount to both outcomes (relative magnitude effect) and to decrease with the addition of the same outcome to the sooner time point only when the later outcome was small (front-end amount effect) and to reverse as the later reward grew (reversal of the front-end amount effect). We replicated three out of four expected contextual effects in the first attempt and all of them in the second attempt.

Though we ruled out some reasons (e.g. gender differences in samples) we were unable to determine why we did not get the hidden-zero effect in the first sample. More importantly, however, we found an interesting interaction of the hidden-zero effect with the addition of a common amount to both outcomes that replicated in both samples. Specifically the hidden-zero

effect was stronger for smaller amounts but decreased as the amounts increased.

Finally when we fit the extended tradeoff model to our data we found that it was most likely overfit to their data which was demonstrated both by the RMSD in our data using their estimates, the lack of stability of most of the parameters and by the larger error parameter in our estimations. Still we were able to estimate parameters that had a relatively low RMSD in the training sample and also had a low RMSD in the test sample suggesting that the model might indeed be tapping into parts of the decision process for these stimuli. The stability of some of the parameters that are conceptually particularly important for this model, as they distinguish it from most discounting models in the expected utility framework, added further support to this as well.

Commentary

Though it was worrying not to get the hidden-zero effect in the first sample finding the interaction of the hidden zero effect with the magnitude of the amounts was interesting and implies a more complicated story than a simple heuristic where the relative magnitude effect outweighs the hidden-zero effect.

The stability of q and $\boldsymbol{\alpha}$ hint at a general preference/strategy in adjusting delays depending on outcomes when there are multiple outcomes and how much one values one attribute over the other. The stability of these two parameters is of particular interest because these are the most novel contributions of the tradeoff model in thinking about intertemporal decisions compared to most other temporal discounting models where the values of delayed outcomes are valued “by” the delay and not “against” them.

In trying to fit the model to our data we seem to have gotten better fits for the explicit zero condition than implicit zero condition. This might be indicative of the model being too sensitive to the current design and we hope to examine this further my testing the model in another experiment for my first year project.

Overall we showed how intertemporal decisions were susceptible to contextual effects once again. The tradeoff model appears to be a good start for psychologically informed parameters but its implementation on aggregate level data is relatively crude and it is too sensitive to the experimental design