Title
Essays in Behavioral Economics

Author
Henderson, Austin

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays in Behavioral Economics

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Management

by

Austin McMaster Henderson

Committee in charge:

Professor Uri Gneezy, Chair
Professor Sanjiv Erat
Professor Karthik Muralidharan
Professor Paul Niehaus
Professor Sally Sadoff

2020

The Dissertation of Austin McMaster Henderson is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2020

EPIGRAPH

It's hard to imagine a surer sign that one is dealing with an irrational economic system than the fact that the prospect of eliminating drudgery is considered to be a problem.

David Graeber

I am trying to make the deadly serious point that, as of today, an economic utopia is not wishful thinking but, in some substantial degree, the necessary alternative to self-destruction. The moral challenge and the grim problem we face is that the life of affluence and pleasure requires exact discipline and high imagination.

Alan Watts

In my darker moments I reflect on the fact that caring about politics is a mental illness, kind of— what it is is essentially a mad and desperate search for a sense of control. It's to feel like you have some sort of say over what's going to happen, even though deep down in your darkest heart you know it's not true. You know that you are totally at the whims of history and fate, the machinations of man and beast, and that you have very little say about where you end up. Politics is a way to feel like, "maybe I can influence it." For the most part that's what it is, a big displacement of one's sense of helplessness. And that goes beyond caring about politics, it goes to voting, to being involved and everything, because it really is an illusion—politics really is beyond us for the most part. Huge forces like capitalism shape things beyond a scale that we could ever even comprehend because we're so small. But—and I've been trying to fight against this for a year now because I don't want to lose my way, I don't want to get stars in my eyes—but I cannot get over this feeling, even when I'm feeling at my darkest and most pessimistic, that this is a moment that is the beginning of a chance to actually get closer to putting your hand somewhere close to the levers of destiny. Not with this election, not with this primary, not even with Bernie getting in—that's just the opening of the door. You have to step through and you have to walk. But I really feel like that if we have an ability as a species to come together in a common recognition of humanity, that this is maybe our last chance to do it. But we might actually be able to.

Matt Christman

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank all the members of my committee for their advice and support, without which this would not be possible.

I would also like to thank my peers, in particular Alex Kellogg, Frances Lu, Wanqun Zhao, Allie Lieberman, Ariel Fridman, Kristen Duke, Pol Campos Mercadé, Florian Schneider, Konstantin Lucks, Duygu Ozdemir, and Binnur Balkan for their friendship and endless help.

Lastly, I would like to thank my co-authors Uri Gneezy, Gary Charness, Jorge Barraza, Gideon Nave, Amos Nadler, Corry McDonald, Patrick Barlow, and Jerrod Keith.

Chapter 3 is joint work with Corry McDonald, Patrick Barlow, and Jerrod Keith. Written permission has been granted by the co-authors for the use of this chapter.

Chapter 4, in full, is a reprint of material as it appears in Scientific Reports 2018. Written permission has been granted by the co-authors for the use of this chapter.

Chapter 5, in full, is a reprint of material as it appears in the Journal of Economic Behavior and Organization 2018. Written permission has been granted by the co-authors for the use of this chapter.

VITA

2010–2014 Bachelor of Arts, Pomona College

2015-2020 Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Austin Henderson, Garrett Thoelen, Amos Nadler, Jorge Barraza, and Gideon Nave. Testing the influence of testosterone administration on men's honesty in a large laboratory experiment. Scientific Reports 8, 11556 (2018).
Gary Charness, Uri Gneezy, and Austin Henderson. Experimental methods: measuring effort in economics experiments. Journal of Economic Behavior and Organization 149, 74-87 (2018).

FIELDS OF STUDY

Major Field: Behavioral Economics

Studies in Development Economics

ABSTRACT OF THE DISSERTATION

Essays in Behavioral Economics

by

Austin McMaster Henderson

Doctor of Philosophy in Management

University of California San Diego, 2020

Professor Uri Gneezy, Chair

In this dissertation, I present several economic research papers. In my first paper, I test novel strategies for using incentives to reduce misbehavior, as well as test the role of social norms in determining the efficacy of audits. In my second paper, I examine preferences for and responses to redistributionary taxation in a real-effort laboratory experiment. In my third paper, I examine the characteristics and experiences of medical students which predict whether they enter primary care specialties. In my fourth paper I test the influence of exogenous testosterone on the tendency to cheat in a laboratory experiment. In my fifth paper I review and contrast various methodologies for measuring effort in economics experiments.

Chapter 1

Auditing with Rewards for Honesty

1.1 Abstract

Strategies to deter misbehavior often use auditing and punishments, but punishments may be undesirable due to practical or ethical concerns. I test the effect of rewards for honest behavior on cheating. Participants in an experiment were given an incentive to cheat, some probability of audit, and a reward for honest behavior, a punishment for cheating, or both. I find that small probabilities of large rewards are effective when combined with punishments, but small rewards are ineffective or backfire. In a second experiment, I elicit beliefs about the social norms of behavior in the first experiment through a coordination game. I find that auditing tends to reduce the social norm against cheating when compared to control, but that large rewards combined with punishments significantly increase the perceived appropriateness of behaving honestly. These findings provide guidance for implementing rewards, and contribute to knowledge of how auditing affects behavior.

1.2 Introduction

Cheating is an important behavior in situations wherein a party can exploit an information asymmetry to their advantage, such as in reporting income to a tax authority or hours worked to an off-site manager. Strategies to deter cheating have frequently involved the use of audits, which with some probability detect cheating and punish the cheater. Economics research has

typically examined the impact of auditing on cheating through the lens of the economics-of-crime approach, in which the gains from cheating are weighed against the probability and cost of detection (Becker, 1968). This logic suggests, for example, that if the probability of detection is low and costly to increase, then the size of the punishment should be increased, an insight which has been influential in shaping real policing approaches.

In this paper, I experimentally test two questions that emerge from the incentives perspective of auditing. The first and central question: how do rewards for good behavior affect cheating? The second: what effect do negligible changes in expected value from auditing schemes, as produced from a low detection probability and small incentive, have on cheating?

The first motivation for studying these questions is straightforward; cheating can be costly, and identifying incentive schemes that can reduce cheating, particularly cost-effective ones, has clear policy value. Importantly, an optimal incentive structure is not necessarily one that maximally deters cheating, but one that balances the level of deterrence against its cost. Consider the example of fare evasion: if more is spent on enforcing fare payment than the enforcement induces riders to pay, there is over-enforcement. There is a lack of clear causal evidence on the cost-efficiency of auditing systems with negligible changes in expected value, as is the case when fare evaders are very unlikely to be caught and fines are small. This cost-efficiency reasoning is especially relevant however as a potential drawback of rewards, as they can be treated as additional costs of enforcement, as opposed to punishments which generate revenue.

The second motivation is that in some circumstances punishments are not attractive for moral reasons, and so even if rewards are less cost-effective than punishments they could still be an attractive tool. Continuing the fare evasion example, those who have the strongest incentive to free-ride are in poverty and could be greatly harmed by a fine. Rewarding good behavior attenuates this dilemma. While this could also mean rewards going to those not in poverty, or people who would've paid the fare regardless, these considerations should be weighed against the ethical imperative to minimize suffering. Individuals may also prefer to face incentive structures framed as rewards rather than as punishments, even if the expected payoffs of actions are the

same (e.g. Brink and Rankin, 2013). This may potentially improve the subjective welfare of those who may be audited, or even allow for auditing systems to be implemented that would otherwise engender push-back.

In a first experiment, I examine cheating behavior, with known audit probabilities and rewards and/or punishments. The study was designed such that the expected value of cheating or honesty was equal across various combinations of incentives (i.e. small punishments, small rewards, and/or large rewards) and audit probabilities (i.e. a low 1% or a high 10%). The equivalency of expected values provides clear comparisons of efficacy and is also theoretically important, as the workhorse Becker (1968) model predicts that the effect of rewarding honesty will be equivalent to punishing cheating so long as the expected utilities of each action are identical. This model also provides the baseline hypothesis that small rewards and punishments will have a negligible effect in comparison to no auditing at all. While this relatively straightforward expected utility model is useful to generate baseline predictions and to link this work to other auditing research, these predictions are not shared by other behavioral models.

Loss aversion (Kahneman and Tversky, 1979), which describes the tendency of individuals to prefer avoiding losses to earning equivalent gains, has been found in a wide variety of economically important contexts, including financial decision-making (Benartzi and Thaler, 1995), seller behavior in the housing market (Genesove and Mayer, 2001), and in abstract laboratory settings (Gneezy and Potters, 1997). In this context, loss aversion predicts that equivalently sized punishments will have a larger behavioral effect than rewards. Relatedly, risk-as-feelings (Loewenstein et al., 2001) predicts that individuals may process risks fearfully, thus making them more effective than rewards, a hypothesis supported by a real-world tax auditing study (Bergolo et al., 2017).

Several models also predict that small incentives may have an outsized impact. The inverse s-shaped probability weighting function of Kahneman and Tversky (1979) captures that individuals tend to treat low probability outcomes as if they were more likely, thus potentially increasing the efficacy of unlikely incentives, both rewards and punishments. Previous studies

have also found that specific features of a reward structure are more heavily weighted, for instance, people are more concerned with the size of a top prize rather than the expected value of a lottery ticket (Garrett and Sobel, 1999; Forest, Simmons, and Chesters, 2000). An emphasis on the potential upside of a reward, rather than just an evaluation of expected value, might make a small probability of a large reward more effective than a larger probability of a small reward.

The results from this experiment show that rewards can be an effective aspect of incentive systems, but that they may also be ineffective or even backfire. Furthermore, the efficacy of rewards is contingent upon the particular manner in which the incentive system is structured, and not simply upon changes in expected value of actions. Treatments in which participants faced a small reward actually had a higher rate of cheating than control or their mirrored punishment treatments, and the rate of cheating when participants faced a small reward actually increased as the auditing rate increased. Conversely, treatments in which participants faced a large reward led to a non-significantly lower rate of cheating than control, and combining a large reward with a small punishment led to the lowest rate of cheating of any treatment tested.

Regarding negligible incentive changes, I find that a negligible change in expected value from punishment led to a marginally significantly lower rate of cheating compared to control, with an economically meaningful drop of 40% in the rate of cheating. While a higher chance of audit and punishment produced an even lower rate of cheating, the difference between this and a negligible punishment was not significant.

To further investigate the mechanism behind these results, I examine whether introducing auditing may change the perceived social norms of cheating or behaving honestly, following recent economics research which has emphasized the role of social-image concerns in the decision to cheat (Abeler, Nosenzo, and Raymond, 2019; Gneezy, Kajackaite, and Sobel, 2018; Khalmetski and Sliwka, 2019). A potential explanation for small rewards backfiring is crowding out–adding an undersized monetary incentive can sometimes be counterproductive, as it crowds out the more powerful intrinsic motivations (e.g. Gneezy & Rustichini, 2000). If individuals value adherence to perceived social norms, then strengthening or weakening those norms could

change behaviors. In the vein of Krupka and Weber (2011), in a second experiment I use a coordination game to elicit beliefs of the social appropriateness of cheating or honest actions.

The results of this second experiment showed that most incentives, except for a negligible change from rewards, reduced the perceived social norm against cheating when compared with control. This is consistent with a larger chance of a small reward backfiring, but not consistent with treatments that had punishments being effective. Large rewards combined with punishments were also associated with an increase in the perceived appropriateness of behaving honestly, significant when compared against 4 of the other 6 treatments.

The primary contribution of this paper is in demonstrating the behavioral changes from rewards and from negligible auditing incentives. Small rewards were either ineffective or actually backfired, which has important implications for the implementation of such incentives in real-world settings, as well as for theories of the effects of auditing on behavior. Large but improbable rewards, however, were more effective in reducing cheating, particularly when combined with small punishments, and might be a useful tool for policymakers. My results also support implementing punishments, even if they are small in magnitude and unlikely. An implication of this finding is that it may be cost-effective to lower the auditing rate, as much of the deterrence may be achieved with a low rate.

Secondly, this paper contributes to knowledge about the mechanisms through which audits are effective. My findings contributes to research which shows that the effect of audits is complex and sensitive to the particulars of the incentive structure, rather than directly driven by changes in expected value (e.g. Laske, Saccardo, and Gneezy, 2018). In addition to changing the expected value of actions, my second experiment provides evidence audits change the perceived social norms of behavior, and can either reinforce or detract from the desired behavioral outcome. It also shows though that behavior may not move in the same direction as perceived social norms, leaving the importance of social norms in this context an open question.

The paper proceeds as follows. In section 2 I present Experiment 1, on cheating behavior. In section 3 I present Experiment 2, on social norms. I discuss and conclude in section 4.

1.3 Experiment 1: Cheating Behavior

1.3.1 Model, Experimental Design, and Hypotheses

Model

In this study, I use a model based on Becker (1968) in order to generate baseline expectations of behavior using a tractable framework. I expand the Becker model by including a fixed intrinsic cost of cheating. This inclusion is based on a large body of literature in behavioral economics which finds that most individuals are to some degree averse to lying or cheating (e.g. Gneezy, 2005; Battigalli, Charness, and Dufwenberg, 2013; Fischbacher and Föllmi-Heusi, 2013; Cappelen, Sørensen, and Tungodden, 2013), and recent research which suggests that this lying aversion is insensitive to the stakes at hand (Kajackaite and Gneezy, 2017). Evidence also suggests that there is a distribution of types of cheaters, with some who will never cheat, some who will cheat given sufficiently large incentives to do so, and some who will always cheat given any positive incentive (Gneezy et al., 2013), which I model here as a distribution of fixed cheating costs.

I use the assumption of a distribution of fixed intrinsic cheating costs to estimate broad power calculations for Experiment 1 (see Appendix A.1), which I then use to help produce and support my hypotheses.

Utility functions in this model have standard properties, and outcomes are dollar denominated, with a strictly positive utility value of each dollar. Risk aversion is assumed to not play a significant role in such small stakes (Rabin, 2000), a common assumption in related studies (e.g. Laske, Saccardo, and Gneezy 2018). When the expected utility of being honest, EU_h, is larger than the expected utility of cheating, EU_c, the individual is honest, and vice versa. The monetary components of this decision are the baseline (i.e. not related to audits) monetary outcomes if they are honest, X, and if dishonest, Y. By construction in this study, $Y > X$. The probability that their action is audited and whether they were truthful detected is p, and the consequence for detected behavior is M. The intrinsic cost of cheating is L.

First, consider the model with participants facing a chance of rewards (r): the expected utility of being honest (1) and cheating (2) are:

$$EU_{h,r} = (1-p) * U(X) + p * U(X+M) \tag{1.1}$$

$$EU_{c,r} = U(Y-L) \tag{1.2}$$

Similarly, here is the model with a chance of punishment (k), with the expected utility of being honest in (3) and cheating in (4)

$$EU_{h,k} = U(X) \tag{1.3}$$

$$EU_{c,k} = (1-p) * U(Y) + p * U(Y-M-L) \tag{1.4}$$

These equations hold that if the EU of honesty and cheating are the same across two incentive structures, then behavior should be the same.

I test the following three hypotheses, which are produced by the model and power calculations:

H1: The rate of cheating will be significantly lower when the gap in expected value between cheating and honesty is substantially reduced.

H2: Very small changes in the expected value of cheating will not lead to a significant difference in the observed rate of cheating.

H3: Holding expected values constant, rewards for honest behavior will yield the same proportion of subjects who cheat as punishments for cheating.

Experimental Design

Participants in the experiment play a one shot cheating game. In the game, the decision maker observes the outcome of what is essentially a coin flip, with outcome [H,T] equally likely with $p = 0.5$. The decision maker then faces a choice: whether to report H, which has a low payoff, and T, which has a high payoff. In either case, they face some probability p, say $p = 0.1$, of being audited, in which the true flip is revealed to the auditor and they are either rewarded or punished according to the treatment.

This experiment was designed to facilitate comparisons are between punishments and rewards. Accordingly, the amounts participants stood to earn from honesty or cheating were calibrated across reward and punishment treatments such that the expected values of actions were either identical or as similar as possible.

In a practical example from the experiment, in the 10% chance of reward treatment honestly reporting an H earns \$0.50 plus the chance of reward, and reporting a T earns \$1.00 guaranteed. The expected value of being honest with a \$1.00 reward is \$0.60, so the difference in expected value is still \$0.40 in favor of misreporting. In the equivalent 10% chance of punishment treatment, honestly reporting an H has a guaranteed value of \$0.60, and dishonesty reporting a T had a value of \$1.10 minus the chance of being punished by one dollar. Thus in both treatments the expected value of honesty is \$0.60 and dishonesty \$1.00, with a gap of \$0.40. A comparison of expected values of actions is presented in Table 1.

In the control treatment, participants were explicitly told they would not face any audit[1].

Experimental Procedure

Participants were recruited via Amazon mTurk to participate in a 2-minute task, with a posted pay rate of \$0.10 for completion. Amazon MTurk is one of the largest online platforms

[1] A treatment in which participants were not informed at all about the possibility of audit was initially run to be used as a control, and 14 of 80 (17.5%) of participants cheated. The treatment in which participants were explicitly told they would not be audited was run in order to reduce the possibility that participants might assume there was a surprise audit that they were not informed of, which would add a layer of ambiguity to interpreting results.

Table 1.1. Expected values of honesty and cheating in each treatment

Treatment	EV of Honesty	EV of Cheating
Control	\$0.50	\$1.00
1% Punishment	\$0.50	\$0.99
1% Reward	\$0.51	\$1.00
10% Punishment	\$0.60	\$1.00
10% Reward	\$0.60	\$1.00
1% Reward \$10	\$0.60	\$1.00
1% Reward \$10 & P	\$0.60	\$0.99

for task-based work, in which individuals choose to accept posted tasks for some promised payment upon satisfactory completion (as determined by the task lister). MTurk has recently become widely used in the social sciences, and is used to study a variety of phenomena, such as preferences for redistribution (Almås, Cappelen, and Tungodden, 2016) and the responsiveness of deception to detection probabilities and the size of the fine (Laske, Saccardo, and Gneezy, 2018). Upon accepting the task on MTurk, participants were directed to the online survey platform Qualtrics where they participated in the experiment. After completing the task, participants were given a matching code to enter into Qualtrics and thus receive payment according to their actions and whether or not they had been audited.

Participants were first informed that they would see on the next screen with 0.5 probability an image of a large letter H or a T, and that they would be asked to report which letter they saw. For reporting an H, they would receive a low payment (either \$0.50 or \$0.60), and for reporting a T they would receive a high payment (either \$1.00 or \$1.10). Amounts varied for a H or T in order to make the expected gain of cheating equivalent between treatments. There were seven treatments, and participants were randomly assigned to one just one, in that they never saw other conditions. This determined the instructions which followed, which are summarized in Table 2 along with the number of participants in each treatment.

When ready, they clicked to proceed to the next screen, which displayed either an H or a T, and a multiple choice button asking which letter was displayed. After responding, participants

Table 1.2. Summary of treatments

Treatment	Description	N
Control	Participants are told that their answers would not be checked for veracity.	122
1% Chance of Punishment	Participants are told that there is a 1% chance that their report will be audited, and if they were found to be dishonest they would be penalized by $1.00 (thus reducing their earnings to $0.10 for participating).	123
10% Chance of Punishment	Participants are told that there is a 10% chance that their report will be audited, and if they were found to be dishonest they would be penalized by $1.00 (thus reducing their earnings to $0.10 for participating).	123
1% Chance of Small Reward	Participants are told that there is a 1% chance that their report will be audited, and if they were found to be honest they would receive a bonus of $1.00.	118
10% Chance of Small Reward	Participants are told that there is a 10% chance that their report will be audited, and if they were found to be honest they would receive a bonus of $1.00.	126
1% Chance of Large Reward	Participants are told that there is a 1% chance that their report will be audited, and if found to be honest they would receive a bonus of $10.00.	102
1% Chance Large Reward or Punishment	Participants are told that there is a 1% chance that their report will be audited, and if found to be honest they would receive a bonus of $10.00, or if dishonest, a penalty of $1.00.	103

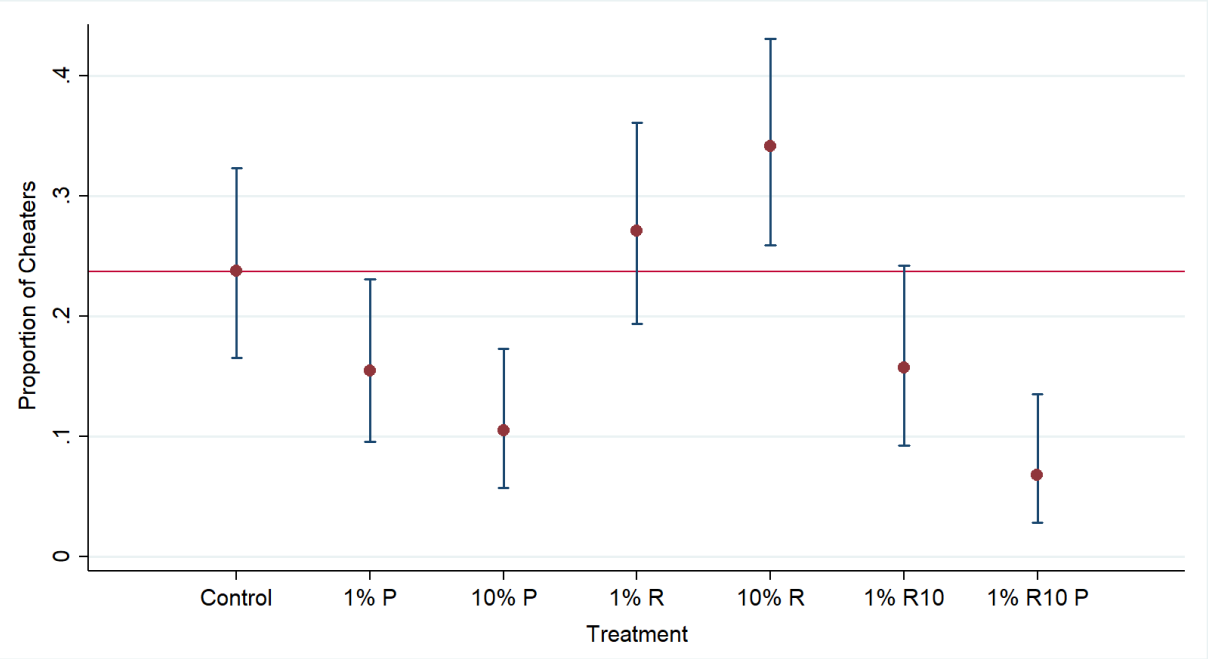

Figure 1.1. Error bars represent 95% binomial exact confidence interval. % is audit probability, P indicates punishment for cheating, R indicates reward for honesty, R10 indicates $10 reward. Horizontal line is rate of cheating in control (no audit).

proceeded to the next screen, which gave them a number to enter on MTurk to coordinate payment, and the experiment was concluded.

1.3.2 Results

A total of N = 815 participants faced an incentive to cheat in this experiment. Of those, 118 faced a 1% chance of reward and 32 (27.12%) cheated, 126 faced a 10% chance of reward and 43 (34.13%) cheated, 122 faced no chance of audit or consequences (control) and 29 (23.77%) cheated, 123 faced a 1% chance of punishment and 19 (15.45%) cheated, and 124 faced a 10% chance of punishment of whom 13 (10.48%) cheated. Facing a 1% chance of a $10 reward, 16 of 102 cheated (15.69%), and facing a 1% chance of $10 reward as well as a $1 punishment 7 of 103 cheated (6.80%). Figure 1 displays the proportion of individuals who cheated, separated by treatment, with the control treatment setting a baseline. Table 3 presents all pairwise comparisons of proportion of participants who cheated in each treatment.

Table 1.3. Displayed is the p-value from a χ^2-test of equal proportions. Number in parentheses next to each treatment is the proportion of participants who cheated. % is audit probability, P indicates punishment for cheating, R indicates reward for honesty, R10 indicates $10 reward. * indicates p-values less than or equal to 0.1, ** indicates p-values less than or equal to 0.05, *** indicates p-values less than or equal to 0.01.

Treatment (Observed Proportion)								
Treatment		10% P (0.10)	1% P (0.15)	0 (0.24)	1% R (0.27)	10% R (0.34)	1% R10 (0.16)	1% R10P (0.07)
	10% P	X	0.25	0.01***	<0.01***	<0.01***	0.25	0.33
	1% P	X	X	0.10*	0.03**	<0.01***	0.96	0.04**
	0	X	X	X	0.55	0.07*	0.13	<0.01***
	1% R	X	X	X	X	0.24	0.04**	<0.01***
	10% R	X	X	X	X	X	<0.01***	<0.01***
	1% R10	X	X	X	X	X	X	0.04**
	1% R10P	X	X	X	X	X	X	X

Minimum detectable effect

Based on the observed proportion of cheaters in the control treatment (23.77%) I calculate the MDE, which is the smallest possible change in the proportion of cheaters detectable at a given power at the 0.05 significance level using a Pearson's χ^2-test, the statistical test used throughout the results to compare the rate of cheating between treatment groups. I assume N = 125 per group for this section, with power calculations based on true N per group below. At power equal to 0.8, an increase in cheating of 16.45pp is the MDE, and at power equal to 0.5 an increase in cheating of 11.30pp is the MDE. At power equal to 0.8, a decrease in cheating of 13.29pp is the MDE, and at power equal to 0.5 a decrease in cheating of 9.71pp is the MDE. A discussion of the power of this study and rates of cheating relative to similar experiments, as well as power calculations for the subsequently presented results that are not included in the main text, are presented in Appendix A.1.

Hypothesis 1

H1: The rate of cheating will be significantly lower when the gap in expected value between cheating and honesty is substantially reduced.

Result: mixed. The rate of cheating was lower with a 10% chance of audit and punishment, but higher with a 10% chance of audit and reward. The rate of cheating was also lower with a 1% chance of a $10 reward, and lowest when a $10 reward was combined with a punishment.

When the chance of punishment was 10%, the rate of cheating was 10.48%, with the difference in rates of cheating between this and control (26.04%) significantly different at the 0.05 level ($\chi^2(1) = 7.67$, $p < 0.01$). The achieved power for a difference between a 10% chance of punishment and control significant at the 0.05 level is 0.84, indicating that this finding was sufficiently powered. This finding is potentially economically significant as well, as the reduction to roughly 40% of the control rate of cheating could be meaningful in many contexts

However, with a 10% chance of reward, 34.13% of participants cheated, which was significantly different at the 0.10 level from control ($\chi^2(1) = 3.23$, $p = 0.07$) but not a 1% chance of reward ($\chi^2(1) = 1.41$, $p = 0.24$). This was not only not a reduction of cheating, as predicted, but actually an increase, and similarly a potentially economically meaningful one.

This failure of did not carry over to alternative reward structures. With a 1% chance of audit and a $10 reward, 15.69% of participants cheated, which although was not significantly below the rate of cheating in control ($\chi^2(1) = 2.26$, $p = 0.13$), was in the predicted direction. This result was not highly powered, however, and it may be that with a larger sample size it could be a significant effect. This finding warrants further research.

Lastly, a 1% chance of audit and a $10 reward combined with a punishment led to 6.80% of participants cheating, which was significantly below control ($\chi^2(1) = 11.97$, $p < 0.01$). This was the lowest observed rate of cheating, which should engender future research on mixed incentive structures.

Overall, 2 of the 4 treatments in which the expected value of cheating was reduced by $0.10 or $0.11 produced rates of cheating significantly lower than control, and a 10% chance of punishment and 1% chance of large reward and punishment both did so with achieved power greater than 0.8. The outlier, a 10% chance of audit and small reward, went significantly in the opposite direction of the incentive and of the prediction.

Hypothesis 2

H2: Very small changes in the expected value of cheating, as from a 1% chance of audit, will not lead to a significant difference in the observed rate of cheating.

Result: mixed. A 1% chance of audit and reward led to an insignificantly higher rate of cheating than control, but a 1% chance of audit and punishment led to a marginally significantly lower rate of cheating.

With a 1% chance of detection and punishment, the rate of cheating was 15.45%, which is lower than the rate of cheating in control, 26.04%. This decrease was statistically significant at the 0.10 level ($\chi^2(1) = 2.7$, $p = 0.10$). This result is notable because it suggests that a small audit probability and punishment led to a 60% rate of cheating compared to control, but larger studies will be necessary to confirm this result.

With a 1% chance of reward, the rate of cheating (27.12%) was higher than in control, but this was not statistically significant ($\chi^2(1) = 0.65$, $p = 0.42$).

Hypothesis 3

H3: Equal differences in expected values between honesty and cheating will lead to equal rates of cheating between treatment groups who face punishment and reward incentives.

Result: negative. There was no equivalence in the rates of cheating between punishment and reward treatments with equal expected values of actions, or even between reward treatments with different incentive structures.

The rate of cheating in the 1% reward treatment was significantly higher than in the 1% punishment treatment at the 0.05 level ($\chi^2(1) = 4.92$, $p = 0.03$). This finding was reasonably powered, with an achieved power for a difference significant at the 0.05 level of 0.60.

More distinctly, the rate of cheating in the 10% reward treatment was higher than the 10% punishment treatment at the 0.01 level ($\chi^2(1) = 20.10$, $p < 0.01$). This result achieved power for a difference significant at the 0.05 level of 0.99.

The rates of cheating between the 1% chance of a $10 reward and a 10% chance of punishment was not significant ($\chi^2(1) < 0.01$, $p = 0.96$). However, the rate of cheating with a 1% chance of audit and a $10 reward and punishment and a 10% chance of punishment was significant ($\chi^2(1) = 4.12$, $p = 0.04$), but achieved a low power equal to 0.16.

Taken together, these results suggest that gaps in expected value were a poor predictor of the effect of auditing systems on cheating behavior. Mirrored treatments (1% and 10% chance of audit and small rewards or punishments) led to particularly different behavior. Implications of this finding are considered further in the discussion.

1.4 Experiment 2: Social Norms

Experiment 1 demonstrated a number of findings not well explained by my model, or by alternative models of decision making under risk. One potential explanation is that behavior was influenced by a desire to adhere to social norms, which themselves were affected by the presence of audits in idiosyncratic ways. The consideration of social norms is supported by recent research, which has emphasized the role of non-material factors including self-image, morality, and social-status (e.g. Dufwenberg, 2016; Shalvi, Eldar, and Bereby-Meyer, 2012; Gneezy, Kajackaite and Sobel, 2018; Khalmetski and Sliwka, 2019) in the decision to cheat. If participants gain utility from adherence to social norms, then changes in perceived social norms, or the weight an individual places on adhering to them, could change behavior. For instance, if introducing rewards either makes honest behavior less socially desirable or cheating less socially undesirable, then more individuals will cheat.

Following Krupka and Weber (2013), who use a coordination game to elicit social norms on behavior in the dictator game and its variants, Experiment 2 elicits social norms of the behavioral options in Experiment 1.

1.4.1 Model, Experimental Design, and Procedure

Model

Krupka and Weber employ the following simple utility function describing social norm compliance, in which the utility of an action a given a set of possible actions k is determined by the value V of the monetary payoff π of that action and the value γ of adhering to the social norm N of action.

$$U(a_k) = V(\pi(a_k)) + \gamma N(a_k) \tag{1.5}$$

I assume that individuals all value social norm adherence, and thus $\gamma > 0$. I also assume that participants have a stable γ such that they are consistent in their valuation of adhering to norms, and that it is perceived norms which may change. Norms N are [-1,1] and describe the extent to which it is perceived that individuals should or should not take an action, with 1 indicating maximal social desirability, 0 indicating neutrality, and -1 indicating maximal social undesirability. Norms are defined by the pair of the action under consideration, honesty h or cheating c, and the incentive structure, reward r, punishment p, or control (no audit) z. Expanding Equation 5 into the framework of an auditing structure with rewards yields the following.

$$EU_{h,r} = (1-p) * (U(X) + \gamma N(h|r)) + p(U(X + M + \gamma N(h|r)) \tag{1.6}$$

$$EU_{c,r} = U(Y - L + \gamma N(c|r)) \tag{1.7}$$

In this context, the expected utility of honesty is increasing in N (i.e. as the strength of the norm increases), and conversely, the hit to the expected utility of cheating is decreasing in N. For example, in the case $N(h,r) < N(h,z)$, then an individual would actually perceive the norm of behaving honestly when facing a reward as less than that when facing no audit, and thus the expected utility of behaving honestly would be lowered in this dimension.

Note that the perceived social norms of behaving honestly and cheating are evaluated separately, rather than as one construct. For example, an individual may appraise that an auditing structure changes the perceived appropriateness of behaving honestly without changing the perceived badness of behaving badly.

Experimental Design and Procedure

As in Experiment 1, participants were recruited through Amazon MTurk and directed to a survey on Qualtrics. Participants were paid $0.25 for participating, and then once on Qualtrics told that they could earn more depending on their decisions. On Qualtrics, they were given instructions which illustrated the task: they would be presented with a hypothetical scenario in which a decision maker A could choose to either lie or tell the truth about the outcome of a coin flip.

In this between-subjects design, participants each viewed 1 hypothetical scenario, each corresponding to one of the audit regimes in Experiment 1. In all cases, participants were given the situation in which the hypothetical decision-maker observed the low-outcome coin flip, and thus had a material incentive to cheat. Participants were told they should mark for each action available to decision maker A whether the action was very socially desirable, socially desirable, neither socially desirable or undesirable, socially undesirable, or very socially undesirable. Lastly, they were told that they would be given an additional bonus of $1 if the rating they gave to an action was the most frequent response among other participants in the study.

1.4.2 Results

A total of N = 836 participants participated in this study, with roughly 120 participants per treatment. Compared to control, most treatments led to a decrease in the rated inappropriateness of cheating, with the differences generally becoming larger as incentives grew either more probable or larger (with the exception of a large reward and punishment). With honesty, only offering a large reward and punishment was marginally more appropriate than control. Full

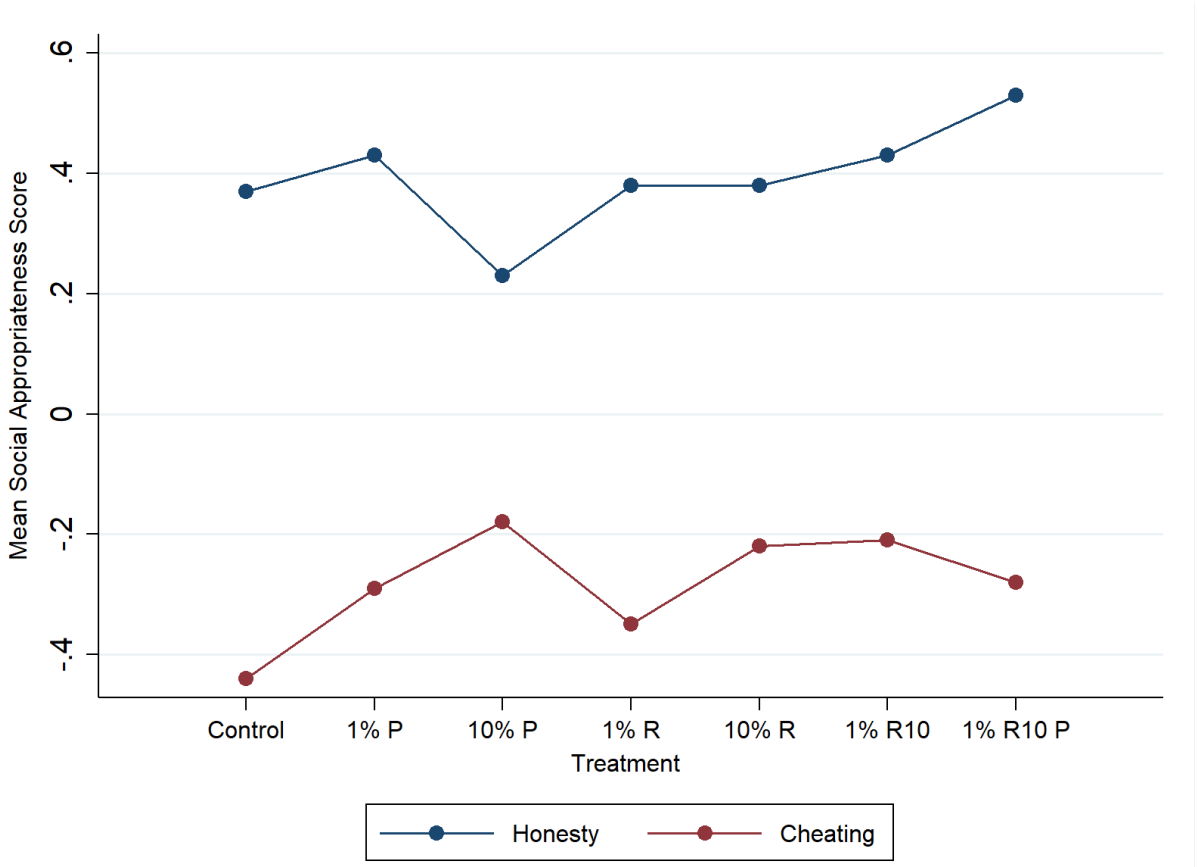

Figure 1.2. Mean social appropriateness score of cheating and honesty in each treatment
% is audit probability, P indicates punishment for cheating, R indicates reward for honesty, R10 indicates $10 reward. Mean scores calculated by assigning a 1 point for very socially appropriate, 1/3 point for socially appropriate, 0 for neutral, -1/3 for socially inappropriate, -1 for very socially inappropriate.

Table 1.4. Percieved social appropriateness of cheating and honesty Number and proportion of respondents who rated each action at a given level of social appropriateness, split by audit probability and punishment or reward. Reporting H corresponds to reporting honestly, reporting T corresponds to reporting dishonestly. % is audit probability, P indicates punishment for cheating, R indicates reward for honesty, R10 indicates $10 reward. Mean scores calculated by assigning a 1 point for very socially appropriate (++), 1/3 point for socially appropriate (+), 0 for neutral, -1/3 for socially inappropriate (-), -1 for very socially inappropriate (–).

Treatment	Outcome Reported	Mean	++ (&)	+ (%)	0 (%)	- (%)	–(%)
Control (N = 119)	H	0.37	59 (49.58)	12 (10.08)	8 (6.72)	31 (26.05)	9 (7.56)
	T	-0.44	10 (8.40)	7 (5.88)	16 (13.45)	32 (26.89)	54 (45.38)
1% P (N = 119)	H	0.43	62 (52.10)	15 (12.61)	13 (10.92)	19 (15.97)	10 (8.40)
	T	-0.29	15 (12.61)	10 (8.40)	13 (10.92)	42 (35.29)	39 (32.77)
10% P (N = 120)	H	0.23	50 (41.67)	13 (10.83)	17 (14.17)	20 (16.67)	20 (16.67)
	T	-0.18	25 (20.83)	11 (9.17)	19 (15.83)	22 (18.33)	43 (35.83)
1% R (N = 120)	H	0.38	57 (47.50)	14 (11.67)	20 (16.67)	19 (15.83)	10 (8.33)
	T	-0.35	7 (5.83)	18 (15.00)	19 (15.83)	32 (26.67)	44 (36.67)
10% R (N = 118)	H	0.38	59 (50.00)	9 (7.63)	16 (13.56)	25 (21.19)	9 (7.63)
	T	-0.22	16 (13.56)	11 (9.32)	19 (16.10)	40 (33.90)	32 (27.12)
1% R10 (N = 121)	H	0.43	62 (51.24)	23 (19.01)	8 (6.61)	15 (12.40)	13 (10.74)
	T	-0.21	18 (14.88)	17 (14.05)	13 (10.74)	36 (29.75)	37 (30.58)
1% R10P (N = 119)	H	0.53	69 (57.98)	20 (16.81)	10 (8.40)	11 (9.24)	9 (7.56)
	T	-0.28	18 (15.13)	16 (13.45)	16 (13.45)	19 (15.97)	50 (42.02)

tables presenting rank-sum tests of equal proportions between all treatments are presented in Appendix A.3.

Participants evaluating the appropriateness of cheating in the cases of a 10% chance of audit and punishment or reward both rated cheating as less inappropriate than control (punishment, rank-sum test, Z = 2.66, $p < 0.01$; reward, rank-sum test, Z = 2.79, $p = 0.01$). Offering just a \$10 reward also led to a decrease in the inappropriateness of cheating compared to control (Z = 2.69, $p = 0.01$), and a 1% chance of a \$10 reward combined with a punishment led to a marginally lower rating of the inappropriateness of cheating (Z = 1.67, $p = 0.09$). These shifts in norms are consistent with the higher observed rate of cheating with a 10% chance of audit and reward, but counter to the observed behavior with auditing and punishments and/or a large reward. The social appropriateness of cheating was marginally higher with a 1% chance of audit and punishment (Z = 1.78, $p = 0.08$). This would be consistent with a higher rate of cheating, and is thus also contrary to the observed cheating behavior.

Conversely, the social appropriateness of reporting honestly was mostly unaffected by small incentives alone, with the only significant difference that reporting honestly was seen as less socially desirable in the case of a 10% chance of audit and punishment compared to a 1% chance of audit and punishment (Z = -2.00, $p = 0.05$). A large reward combined with a punishment, however, lead to a higher level of appropriateness of behaving honestly when compared to almost all other treatments, significant when compared to a 10% chance of small punishment (Z = 3.19, $p < 0.01$, control (Z = 1.91, $p = 0.06$, 1% chance of small reward (Z = 1.92, $p = 0.06$, and 10% chance of small reward (Z = 1.71, $p = 0.07$.

1.5 Discussion

This primary goals of this paper are to examine the effects of rewards and small punishments on cheating, and better understand how auditing affects behavior.

The answer with regard to rewards is complex. Small rewards were either ineffective

or actually backfired. However, large rewards showed more promise, as there was suggestive evidence a large reward by itself could reduce cheating, and a large reward combined with a punishment drastically reduced cheating. Punishments, on the other hand, were generally effective, even when they had a negligible effect on the expected values of choices. These findings are potentially useful for designing practical incentive structures for reducing cheating, or other misbehaviors.

Experiment 2, which tested the influence of auditing incentives on the social norms of cheating behavior, examined a mechanism to potentially explain why some incentive structures work and some do not. I found that perceived social norms did change in the presence of audits, and that norms in favor of honesty and against cheating did not move in tandem. The relationship between social norms and auditing schemes was also idiosyncratic, and the EU model with social norms did not explain the observed cheating behavior.

The most effective treatment however, a large reward combined with a punishment, did actually led to a stronger perceived norm in favor of honesty (but not cheating), the only norm change relative to control of any treatment that could be expected to reduce cheating. A further question is why this happened. One possibility is that the large rewards and punishments were merely a particularly salient feature which facilitated participants correctly coordinating on the same norm (e.g. Mehta, Starmer, and Sugden, 1994), but did not actually change the genuine norm. A second is that this treatment may have genuinely changed how participants conceived of the norm of being honest, and that this contributed to it's efficacy. Future research might better leverage social norms as a component of auditing schemes.

These data do not lend themselves to a neat theoretical explanation. Neither my EU model, Prospect Theory, salience, or social norms with crowding out are consistent with all the results. A more complex explanation that is consistent with the data (but not itself conclusively demonstrated) is that auditing schemes provoke different emotional responses depending upon the presence of punishments, which may provoke fear (Loewenstein et al., 2001). Individuals in a fearful state may process risky decisions qualitatively differently than those who are not,

allowing for different models to describe behavior according to emotional state. I consider this dual-process explanation, as well as the inconsistencies of other explanations more in depth in Appendix A.3.

There is a need for more laboratory experiments to better understand the dynamics of rewards, which can both aid in theory development as well as provide more pragmatic guidance for field implementation. The straightforward systematic laboratory exploration of combinations of rewards and punishments might provide more evidence useful to developing theory, as well as cost effective practical guidance for field implementation. Examining rewards in a repeated games context may be particularly useful in determining the long-term efficacy of such incentives before field implementation.

Testing rewards in field settings is ideal for evaluating a key motivation of this line of research: identifying cost-effective auditing schemes. Promising environments may be fare evasion on public transport, by rewarding those who can prove having purchased a ticked, or in rewarding safe driving behavior. Several studies have found rewards can be effective when combined with continuous monitoring equipment in reducing bad driving (Mazureck and van Hattem, 2006; Lahrmann et al., 2012). An extension of this would be to incentivize good driving when continuous monitoring is not feasible, for instance police who observe good driving could send rewards to drivers. It is in these field contexts possible to estimate the value of reduced misbehavior (e.g. foregone car accidents) and compare it to the expenditure on rewards.

1.6 References

1. Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. Econometrica, 87(4), 1115-1153.

2. Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. Journal of public economics, 1(3-4), 323-338.

3. Alm, J., Bloomquist, K. M., & McKee, M. (2015). On the external validity of laboratory tax compliance experiments. Economic Inquiry, 53(2), 1170-1186.

4. Alm, J., Deskins, J., & McKee, M. (2009). Do individuals comply on income not reported by their employer?. Public Finance Review, 37(2), 120-141.

5. Alm, J., Cherry, T., Jones, M., & McKee, M. (2010). Taxpayer information assistance services and tax compliance behavior. Journal of Economic Psychology, 31(4), 577-586.

6. Alm, J. (2012). Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies. International tax and public finance, 19(1), 54-77.

7. Alm, J., Jackson, B. R., & McKee, M. (2009). Getting the word out: Enforcement information dissemination and compliance behavior. Journal of Public Economics, 93(3-4), 392-402.

8. Alm, J., & McKee, M. (2004). Tax compliance as a coordination game. Journal of Economic Behavior & Organization, 54(3), 297-312.

9. Almås, I., Cappelen, A. W., & Tungodden, B. (2020). Cutthroat capitalism versus cuddly socialism: Are Americans more meritocratic and efficiency-seeking than Scandinavians?. Journal of Political Economy, 128(5), 1753-1788.

10. Battigalli, P., Charness, G., & Dufwenberg, M. (2013). Deception: The role of guilt. Journal of Economic Behavior & Organization, 93, 227-232.

11. Becker, G. S. (1968). Crime and punishment: An economic approach. In The economic dimensions of crime (pp. 13-68). Palgrave Macmillan, London.

12. Benartzi, S., & Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. The quarterly journal of Economics, 110(1), 73-92.

13. Bérgolo, M. L., Ceni, R., Cruces, G., Giaccobasso, M., & Perez-Truglia, R. (2017). Tax audits as scarecrows: Evidence from a large-scale field experiment (No. w23631). National Bureau of Economic Research.

14. Brink, A. G., & Rankin, F. W. (2013). The effects of risk preference and loss aversion on individual behavior under bonus, penalty, and combined contract frames. Behavioral Research in Accounting, 25(2), 145-170.

15. Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2013). When do we lie?. Journal of Economic Behavior & Organization, 93, 258-265.

16. Coricelli, G., Joffily, M., Montmarquette, C., & Villeval, M. C. (2010). Cheating, emotions, and rationality: an experiment on tax evasion. Experimental Economics, 13(2), 226-247.

17. Cornwell, C., & Trumbull, W. N. (1994). Estimating the economic model of crime with panel data. The Review of economics and Statistics, 360-366.

18. Dishman, R. K., Nakamura, Y., Garcia, M. E., Thompson, R. W., Dunn, A. L., & Blair, S. N. (2000). Heart rate variability, trait anxiety, and perceived stress among physically fit men and women. International Journal of Psychophysiology, 37(2), 121-133.

19. Dufwenberg, M., & Dufwenberg, M. A. (2018). Lies in disguise–A theoretical analysis of cheating. Journal of Economic Theory, 175, 248-264.

20. Dulleck, U., Fooken, J., Newton, C., Ristl, A., Schaffner, M., & Torgler, B. (2016). Tax compliance and psychic costs: behavioral experimental evidence using a physiological marker. Journal of Public Economics, 134, 9-18.

21. Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. Journal of the European Economic Association, 11(3), 525-547.

22. Forrest, D., David Gulley, O., & Simmons, R. (2004). Substitution between games in the UK national lottery. Applied Economics, 36(7), 645-651.

23. Garrett, T. A., & Sobel, R. S. (1999). Gamblers favor skewness, not risk: Further evidence from United States' lottery games. Economics Letters, 63(1), 85-90.

24. Genesove, D., & Mayer, C. (2001). Loss aversion and seller behavior: Evidence from the housing market. The quarterly journal of economics, 116(4), 1233-1260.

25. Gneezy, U. (2003). The W effect of incentives. University of Chicago Graduate School of Business.

26. Gneezy, U. (2005). Deception: The role of consequences. American Economic Review, 95(1), 384-394.

27. Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying Aversion and the Size of the Lie. American Economic Review, 108(2), 419-53.

28. Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. The quarterly journal of economics, 112(2), 631-645.

29. Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013). Measuring lying aversion. Journal of Economic Behavior & Organization, 93, 293-300.

30. Gneezy, U., & Rustichini, A. (2000). A fine is a price. The Journal of Legal Studies, 29(1), 1-17.

31. Gneezy, U., & Rustichini, A. (2004). Incentives, punishment and behavior. Advances in behavioral economics, 572-89.

32. Internal Revenue Service (2019). Data Book, 2018. Publication 55B.

33. Kajackaite, A., & Gneezy, U. (2017). Incentives and cheating. Games and Economic Behavior, 102, 433-444.

34. Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In Handbook of the fundamentals of financial decision making: Part I (pp. 99-127).

35. Khalmetski, K., & Sliwka, D. (2019). Disguising lies—Image concerns and partial lying in cheating games. American Economic Journal: Microeconomics, 11(4), 79-110.

36. Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. Econometrica, 79(3), 651-692.

37. Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary?. Journal of the European Economic Association, 11(3), 495-524.

38. Lahrmann, H., Agerholm, N., Tradisauskas, N., Berthelsen, K. K., & Harms, L. (2012). Pay as You Speed, ISA with incentives for not speeding: Results and interpretation of speed data. Accident Analysis & Prevention, 48, 17-28.

39. Laske, K., Saccardo, S., & Gneezy, U. (2018). Do fines deter unethical behavior? The effect of systematically varying the size and probability of punishment. The Effect of Systematically Varying the Size and Probability of Punishment (April 5, 2018).

40. Lochner, L., & Moretti, E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. American economic review, 94(1), 155-189.

41. Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. Psychological bulletin, 127(2), 267.

42. Machin, S., Marie, O., & Vujić, S. (2011). The crime reducing effect of education. The Economic Journal, 121(552), 463-484.

43. Mazureck, U., & van Hattem, J. (2006). Rewards for safe driving behavior: Influence on following distance and speed. Transportation research record, 1980(1), 31-38.

44. McKee, M., Alm, J., Cherry, T., & Jones, M. (2008). Final Report for TIRNO–07–P–00683 on Behavioral Tax Research. Washington, DC.

45. Mehta, J., Starmer, C., & Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination games. The American Economic Review, 84(3), 658-673.

46. Prendergast, D. & Donohue, P. (2011). ”Subway turnstile jumpers saving loads of money, even when they are caught, study shows.” New York Daily News.

47. Slemrod, J., Blumenthal, M., & Christian, C. (2001). Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota. Journal of public economics, 79(3), 455-483.

48. Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). Psychological science, 23(10), 1264-1270.

1.7 Appendix A.1 Power and minimum detectable effects

Detecting treatment effects in this experiment is dependent upon the fraction of participants who are responsive to changes in their incentives to cheat. Moreover, generating hypotheses of the detectability of treatment effects require making reasonable assumptions about the distribution of lying costs. To illustrate, as the expected gains from cheating decrease, the model predicts that a smaller fraction of participants will choose to cheat. To justify hypotheses based on a prediction of whether the change in the fraction of participants who choose to cheat will be detectable given a change in the incentives, I refer to previous experimental findings and argue for changes of plausible magnitudes, as well as posit a uniform distribution of lying costs within the incentive space (i.e. the outcome space between reporting heads and tails). These experiments were not designed to make explicit, generalized claims about the true rate of individuals who will cheat at a given incentive, however, they do present a range of observed cheating rates at given incentives, and thus provide a reasonable context to Experiment 1.

Three reports find that in laboratory experiments 33-53% of participants will not lie at the maximum experimental incentive, and 3-8% of participants will lie given any positive incentive (Gneezy et al. 2013; Kajackaite and Gneezy, 2017; Laske, Saccardo, and Gneezy, 2018). The participants of interest in this study are those who will respond to the level of incentives, and based on this prior experimental data, may make up between 39% and 64% of the total. Taking a middle ground, let us assume 50% of the participants in the study are sensitive to the probability and consequences of detection.

In this model, changes in the propensity to cheat in participants facing a 1% chance of audit would be from those participants with a positive, but very small cost of lying, essentially those participants for whom the original expected gain from cheating of \$0.50 placed them exactly on the margin. The observed proportion of cheaters in the control treatment was 23.77%, with a sample size of N = 125. At power equal to 0.8, a decrease in cheating of 13.29pp is the Minimum Detectable Effect (MDE), and at power equal to 0.5 a decrease in cheating of

9.71pp is the MDE. Thus if participants are purely making their decisions based on expected value with standard risk preferences, it must be that at minimum 9.71pp of participants were on the margin in order to detect the effect. Observing such an effect would be inconsistent with prior research, which has found that larger proportional and absolute changes (e.g. the incentive to cheat increases by 10pp) rarely produce changes of such magnitude. Assuming a uniform distribution and that 50% of participants are sensitive to the level of incentives, the number of participants who would change from cheating to honesty is equal at each $0.01 interval between $0.50 and $0.60. That is, for each decrease in the expected gains from cheating of $0.01, 1pp fewer participants will choose to cheat. This suggests that a 1% chance of audit there should not be a detectable change according to the model.

Conversely, a 20% drop in the expected gains from cheating (i.e. the expected gains from cheating relative to honesty is $0.50 in control, and $0.40 when participants face a 10% chance of audit and punishment) is intuitively more probably associated with producing at least 10pp fewer cheaters. Assuming again a uniform distribution with 50% of participants sensitive to the level of incentives, 10pp fewer participants will cheat, which should be (and was) detectable. This was roughly in line with the actual observed reduction of 13.29%.

Hypothesis 1: Power

H1: The rate of cheating will be significantly lower when the gap in expected value between cheating and honesty is substantially reduced.

Result: mixed. The rate of cheating was lower with a 10% chance of audit and punishment, but higher with a 10% chance of audit and reward. The rate of cheating was also lower with a 1% chance of a $10 reward, and lowest when a $10 reward was combined with a punishment.

When the chance of punishment was 10%, the rate of cheating was 10.48%, with the difference in rates of cheating between this and control (26.04%) significantly different at the 0.05 level ($\chi^2(1) = 7.67$, $p < 0.01$). The achieved power for a difference between a 10% chance

of punishment and control significant at the 0.05 level is 0.84, indicating that this finding was sufficiently powered.

With a 10% chance of reward, 34.13% of participants cheated, which was significantly different at the 0.10 level from control ($\chi^2(1) = 3.23$, $p = 0.07$) but not a 1% chance of reward ($\chi^2(1) = 1.41$, $p = 0.24$). The achieved power for detecting a difference between a 10% chance of reward and control is 0.36, and the necessary sample size for detecting this difference at a 0.05 significance level with power equal to 0.8 is 372.

With a 1% chance of audit and a $10 reward, 15.69% of participants cheated, whichc was not significantly below the rate of cheating in control ($\chi^2(1) = 2.26$, $p = 0.13$). The achieved power at the 0.05 significance level is 0.38.

Lastly, a 1% chance of audit and a $10 reward combined with a punishment led to 6.80% of participants cheating, which was significantly below control ($\chi^2(1) = 11.97$, $p < 0.01$). This finding achieved power equal to 0.96 at the 0.05 significance level.

Hypothesis 2: Power

H2: Very small changes in the expected value of cheating, as from a 1% chance of audit, will not lead to a significant difference in the observed rate of cheating.

Result: mixed. A 1% chance of audit and reward led to an insignificantly higher rate of cheating than control, but a 1% chance of audit and punishment led to a marginally significantly lower rate of cheating.

With a 1% chance of detection and punishment, the rate of cheating was 15.45%, which is lower than the rate of cheating in control, 26.04%. This decrease was statistically significant at the 0.10 level ($\chi^2(1) = 2.7$, $p = 0.10$). The achieved power for detecting a significant difference at the 0.05 level at these proportions is 0.44, with a required sample size to achieve power = 0.8 of N = 290 per group. This result is notable because the roughly 40% reduction in the rate of cheating is likely to be economically meaningful in many circumstances, and thus establishing whether it is a generally true finding or anomalous is valuable.

With a 1% chance of reward, the rate of cheating (27.12%) was higher than in control, but this was not statistically significant ($\chi^2(1) = 0.65$, $p = 0.42$). The achieved power for detecting this difference at the 0.05 level was 0.07.

Hypothesis 3: Power

H3: Equal differences in expected values between honesty and cheating will lead to equal rates of cheating between treatment groups who face punishment and reward incentives.

Result: negative. There was no equivalence in the rates of cheating between punishment and reward treatments with equal expected values of actions, or even between reward treatments with different incentive structures.

The rate of cheating in the 1% reward treatment was significantly higher than in the 1% punishment treatment at the 0.05 level ($\chi^2(1) = 4.92$, $p = 0.03$). The achieved power for a difference significant at the 0.05 level is 0.60, with a required N = 192 per group to detect a significant difference at the 0.05 level with power equal to 0.8.

More distinctly, the rate of cheating in the 10% reward treatment was higher than the 10% punishment treatment at the 0.01 level ($\chi^2(1) = 20.10$, $p < 0.01$). The achieved power for a difference significant at the 0.05 level is 0.99.

The rates of cheating between the 1% chance of a \$10 reward and a 10% chance of punishment was not significant ($\chi^2(1) < 0.01$, $p = 0.96$). However, the rate of cheating with a 1% chance of audit and a \$10 reward and punishment and a 10% chance of punishment was significant ($\chi^2(1) = 4.12$, $p = 0.04$), but achieved a low power equal to 0.16.

1.8 Appendix A.2 Pairwise comparisons of social norms

Table 1.5. Perceived social norms of reporting T (cheating). Rank-sum test of equal proportion of participants who cheated, split by audit probability and punishment (P) or reward (R). Z-score is reported, number in parentheses is p-value. * indicates p-values less than or equal to 0.1, ** indicates p-values less than or equal to 0.05, *** indicates p-values less than or equal to 0.01.

		Treatment						
		0.1P	0.01P	0	0.01R	0.1 R	R10	R10P
Treatment	0.1P	X	1.08 (0.28)	2.66 (<0.01***)	-1.44 (0.15)	-0.22 (0.83)	-0.18 (0.86)	-0.93 (0.35)
	0.01P	X	X	1.78 (0.08*)	-0.27 (0.79)	1.06 (0.29)	0.98 (0.33)	0.05 (0.96)
	0	X	X	X	1.15 (0.13)	2.79 (0.01***)	2.69 (0.01***)	1.67 (0.09*)
	0.01R	X	X	X	X	1.28 (0.20)	1.32 (0.19)	0.45 (0.66)
	0.1R	X	X	X	X	X	0.01 (0.99)	-0.79 (0.43)
	R10	X	X	X	X	X	X	-0.83 (0.41)
	R10P	X	X	X	X	X	X	X

Table 1.6. Perceived social norms of reporting H (honest report). Rank-sum test of equal proportion of participants who cheated, split by audit probability and punishment (P) or reward (R). Z-score is reported, number in parentheses is p-value. * indicates p-values less than or equal to 0.1, ** indicates p-values less than or equal to 0.05, *** indicates p-values less than or equal to 0.01.

		Treatment						
		0.1P	0.01P	0	0.01R	0.1 R	R10	R10P
Treatment	0.1P	X	-2.00 (0.05**)	-1.26 (0.21)	1.46 (0.14)	1.42 (0.16)	0.07 (0.94)	3.19(<0.01***)
	0.01P	X	X	0.73 (0.46)	-0.60 (0.55)	-0.58 (0.56)	0.07 (0.94)	1.23 (0.22)
	0	X	X	X	0.22 (0.83)	0.21 (0.83)	0.74 (0.45)	1.91 (0.06*)
	0.01R	X	X	X	X	-0.04 (0.97)	0.73 (0.46)	1.92 (0.06*)
	0.1R	X	X	X	X	X	0.64 (0.52)	1.81 (0.07*)
	R10	X	X	X	X	X	X	1.15 (0.25)
	R10P	X	X	X	X	X	X	X

1.9 Appendix A.3 Dual-process explanation

An important question remains from the results of these two experiments: what can account for the entire observed pattern of cheating behavior?

Two characteristics of prospect theory are consistent with some of the results here: both loss aversion (although notably strong loss aversion) and overweighting of small probabilities are consistent with the largely lower rate of cheating with a 1% chance of audit and punishment. Overweighting of small probabilities though would predict similarly large effects from a 1% chance of audit and reward, which was not observed. Moreover, the higher rate of cheating with small rewards is not predicted by prospect theory.

Another potential explanation is that aspects of the different incentive structures differed in how salient they were to the cheating decision. Punishments may be a particularly salient feature of decisions with an ethical choice, compared to the less-common feature of a reward for good behavior. Participants may, due to their past experience of being punished for unethical actions, place a higher weight on the negative feelings associated with a punishment than the unfamiliar positive feelings of a reward, and in anticipating these feelings, be more motivated by punishments. This would be consistent with punishments generally being more effective than similarly structured rewards, and punishments combining with large rewards to be most effective. Similarly, the size of the reward may be a more salient feature of the decision making process than the probability of audit (see Laske, Saccardo, and Gneezy 2018 for an example of the size of a punishment being more salient than the probability). Salience alone however does not seem to conveniently explain why crowding out was observed with small rewards but not large rewards.

An explanation which builds on potential differences in how incentive frames are processed emotionally is that audits partially influence behavior according to whether the audits induce fear. When auditing introduces the potential for punishment, it could induce fear or anxiety in some participants, who then behaved to minimize that fear, rather than maximize expected utility, à la risk-as feelings (Loewenstein et al., 2001). When there is no auditing or

auditing and rewards, individuals are not induced into fear or anxiety, and their decision-making process can be captured by an expected utility model. Thus, the asymmetry in the efficacy of rewards and punishments can be explained by an asymmetric emotional response which leads to alternative decision making processes driving behavior. This interpretation, along with the empirical results of these experiments, leads to two lines of discussion: the first, some potential limits of standard economic reasoning in analyzing the impact of audits, and the second, policy implications and avenues for future research.

Risk-as-feelings (Loewenstein et al., 2001) has been raised as a potentially important explanation for behavior in tax-compliance behavior, including in a real-world tax auditing study (Bergolo et al., 2017). That fear or anxiety is induced is supported by empirical evidence: in related experiments, Coricelli et al. (2010) find that emotional arousal is related to punishments and predictive of tax evasion in the lab, and Dulleck et al. (2016) find correlation between stress markers and compliance during a laboratory tax study. The model holds, among other contentions, that in risky situations which provoke a fear response, behavior will differ from the predictions of expected utility models such as that of Becker (1968) that are predicated upon cognitive evaluation of potential consequences. More precisely, fear is not merely anticipated and brought into the decision making function (e.g. an individual predicts regret after being punished, which could be modeled as a cost), but changes the evaluative apparatus entirely. This model predicts that individuals in a fearful or anxious state will act to avoid the fear- or anxiety-inducing outcome, while being relatively insensitive to the true probability. Probability neglect, or the insensitivity of individuals to the true probabilities of fearful outcomes is consistent with data from Experiment 1, in the lack of a significant difference between a 1% and 10% chance of audit and punishment, as well as data from the Bergolo et al. (2017) study. This dual-process explanation can accommodate the general efficacy of punishments, as well as social norms being important in the efficacy of rewards. It does have the drawback however of being a complex explanation that post-hoc fits the observations, and needs more evidence.

Determining which (if any, or some combination) of these explanations can best explain

the impact of auditing incentives is an important avenue for future research. Research similar to Coricelli et al. (2010) or Dulleck et al. (2016) could, through using physiological tools, better establish the proposed dual-process mechanism for cheating behavior in this study. Particularly, functional magnetic resonance imaging could particularly identify differences in brain region activation, to move beyond more general measures of emotional arousal and stress. It is also worth exploring whether emotional or dual-process models of decision making outperform expected utility models in predicting behavior in circumstances where emotional responses may be even stronger, such as when punishment types differ (e.g. prison time vs. fines), or when characteristics of the environment with an emotional potential are more or less salient.

Chapter 2

Preferences for and Responses to Redistribution

2.1 Abstract

Preferences for and responses to redistributionary taxation are central topics within political economics and public finance. An important question is how inequality changes redistributionary preferences, with canonical models predicting greater pre-tax income inequality leading to greater demand for redistribution, and low demand for redistribution in middle-income voters. In a laboratory experiment, participants in pairs voted on a redistributionary tax before engaging in a piece-rate real-effort task, a methodology new to this experimental literature. The amount earned per completed task was also varied across participants. When participants were of equal productivity, there was strong demand for redistributive taxation, in contradiction to models which predict demand for redistribution to be low when expected pre-tax earnings will be close. Effort levels were not affected by tax or wage rates, even when participants faced full redistribution that was equivalent to a 50% marginal tax rate.

2.2 Introduction

Redistributionary taxation is a central and defining feature of the political landscape of many modern democracies (McCarty, Poole, and Rosenthal, 2016). The income tax in the

United States, for example, accounted for over 47% of total tax revenues in 2016 (Office of Management and Budget, 2017), is progressive, redistributive, and the subject of substantial political debate. Much research has gone into exploring the individual characteristics associated with demand for redistributionary taxation (e.g. Alesina & Giuliano, 2011) and modeling demand for redistribution.

The widely influential Meltzer and Richard (MR) model, built upon work by Romer (1975), explains demand for redistribution as the result of self-interested preferences (Meltzer & Richard, 1981). In the MR model, an individual's demand for a linear redistributive tax is inversely based on their position in the income distribution, tempered by the disincentive effect of taxation. One of the appealing features of the MR model is that the distribution of demand for redistribution along an expected income ranking, in combination with the median voter theorem (Downs, 1957), makes predictions about the political implementation of redistributionary policies. This feature also draws attention to the importance of voters around the median or mean income in playing a decisive role in politically important aggregations of demand for redistribution.

A primary prediction of the MR model is that greater inequality will lead to greater redistribution due to the gap between mean and median incomes, but there is scant macro field evidence that this is true (Alesina & Giuliano, 2009). Some of the empirical failures of the MR model may be due to weaknesses in the median voter theorem (Milanovic, 2000) or the general difficulties of detecting preferences through the noise of complicated political processes.

Beyond this, another issue is that some demand for redistribution is not just self-interested, with a large body of economics research supporting that social preferences, the source of initial inequality (such as luck or effort), efficiency concerns, and more are important considerations (e.g. Fehr & Schmidt, 1999; Fong, 2001; Almås, Cappelen, Sørensen, & Tungodden, 2010; Erkal, Gangadharan, & Nikiforakis 2011; Starmans, Sheskin, and Bloom, 2016).

A third issue, central to this study, is that even individuals with purely self-interested preferences for redistribution face difficult choices when it is not clear whether redistribution will lead to a net increase or decrease in income. Highlighted by the MR model's emphasis

on middle-income voters is that individuals around the middle of an income distribution are those most likely to be uncertain over their future position in the income distribution relative to the mean, and subsequently, face a challenge in deciding which level of redistribution is actually in their self-interest[1]. The disincentive effect of taxation is also particularly important for middle income voters uncertain of the effects of taxation, as they will either not benefit from redistribution, see their small benefits from redistribution diminish with a shrinking tax base, or even end up worse off because of the combination of a shrunken tax base and their own lowered productivity.

Laboratory experiments offer, if not a holistic lens through which to view the entire political economy, an attractive avenue to study the specific circumstances in which a model of redistributionary preferences makes valid predictions (Tausch, Potters, and Riedl, 2013). Specifically, two recent experiments, Agranov and Palfrey (2015) and Durante et al. (2014), directly test the MR model, along with other models of social preferences. These two studies indeed find support for the contention that greater pre-tax income inequality leads to higher demand for redistribution.

I follow this literature, and examine preferences for and responses to redistributionary taxation in a laboratory experiment. In the experiment, participants in pairs were told they would work on a task and given either an equal or unequal wage rate for each task completed. They then voted for the linear redistributionary tax (either 0%, 50%, or 100%) they preferred. This design allows me to test competing models of redistributionary preferences and plausibly rule out some models as consistent with observed voting. In particular, I focus on demand for redistributionary taxation in workers with even wage rates, and consider how uncertainty over the position in the income distribution affects how different models can explain this demand.

One contribution of this paper is methodological: I expand upon the experimental

[1]The demand for redistributionary policies, such as unemployment benefits, for their role as social insurance has been widely studied (Varian, 1980; see Backus & Esteller-Moré, 2016 for recent empirical work), but typically focus on events such as job loss. While clearly important for real world demand, here I focus on a simplified case without such risks.

literature by using a real-effort task in which participants are paid a piece-rate while facing a known tax rate. While real-effort tasks are disadvantaged relative to stated-effort tasks in that the cost of effort function is not known and thus theoretical predictions are not as clear, I highlight two key advantages of piece-rate real-effort due to their particular relevance in learning about tax preferences from laboratory experiments.

Firstly, income in this task is always increasing in effort, so participants always face positive monetary incentives to work harder, similar to labor decisions when facing real-world income taxes. Stated-effort methodologies, in which participants pay a monetary cost for higher effort, by construction make it often such that after a certain point higher effort levels produce negative individual earnings. This inflection point can be calculated by participants, and is decreasing in tax rate, thus potentially depressing demand for taxation if the convexity of the cost of effort function is too exaggerated. The use of a real-effort task generates a natural source of uncertainty.

Secondly, recent research has shown that the source of inequality is relevant to preferences for redistribution, with the general finding that individuals prefer when outcomes are proportionate to effort expenditure (Starmans, Sheskin, and Bloom, 2017). As opposed to inequity aversion, which predicts that participants will try to equalize outcomes, fairness based on effort predicts that individuals will not want redistribution that reduces deserved inequalities, and so workers will not want to compress income differences that are produced by differences in real-effort. Using a real-effort task allows for testing theories that account for the source of the income inequality, and I test whether a model of preferences based on desiring effort to be proportional to income describes voting behavior.

In addition to examining preferences for redistribution, I also examine the behavioral response to tax rates. An assumption, central in the public finance literature and inherent to the MR model, is of individual optimization over labor supply and consumption resulting from the incentives in a tax schedule. This optimization can potentially be challenged, however, by a number of factors, such as the salience of a tax (Chetty, Looney, and Kroft, 2009). An open

question is what features of a tax schedule modify behavior besides the tax rates themselves.

One source of factors which may influence the impact of redistributive taxation on behavior is that such systems are often determined by voting. The political process may affect the perceived fairness of a tax schedule, or a voting outcome may signal information about the preferences and potentially the behavior of voters. For instance, a self-interested individual who expects that they will earn less than the mean in a redistributive system will vote for redistribution in the MR model. An individual who observes another person vote for high redistribution might suspect that they intend to exert lower effort, and if they have strong fairness preferences, may choose to lower their effort in response to this belief, beyond their response to the tax rate alone. Individuals may also engage in retaliatory shirking in order to punish those who implement high taxes. Focusing on the process itself, individuals who vote for their redistribution system may find whatever outcome is produced to be more or less fair than if the redistribution rate was determined by some other mechanism. To examine the effect of voting on effort provision, I additionally test the effect of exogenously determined tax rates.

I find that income is a primary determinant of demand for redistribution when position in the income distribution is clear, with a large majority of low wage participants voting for full redistribution and high wage participants voting for no redistribution. I also find that demand in even-wage workers is large, with 85% of participants voting for either a 50% or 100% tax rate.

The effort provision results are more inconclusive, hindered in part by low variation in effort levels. I do not find significant differences in effort by tax rate or wage rate, and the effects of voting were not significant either.

The rest of the paper proceeds as follows. In section 2 I present the experimental design and procedure. In section 3 I present and consider four models of redistributionary preferences. In section 4 I present the results, split into subsections on voting and effort, and in section 5 I discuss.

2.3 Experimental Design and Procedure

Between June 6th and 13th of 2017, $N = 118$ participants were recruited in groups of 8 or 10[2]. After completing consent forms, participants were then seated at individual computer stations in the Rady Incentives Lab and given brief verbal instructions to not talk to other participants or look at the screens of other participants, to silence their phones, and that all future instructions would be given through the computer.

The rest of the experiment was conducted through Z-Tree (Fischbacher, 2007) and proceeds as follows (see Appendix A.1 for screenshots of the task). First, all participants were informed that they had been randomly matched with another participant in that session that would remain entirely anonymous, and that both of them would work on a real-effort task which paid them a piece rate for every set of letters encoded into numbers (Erkal, Gangadharan, and Nikiforakis 2011).

Participants were then informed of their wage rate. In the even wage rate treatment, they were informed that both they and the person that had been matched with would earn the same \$0.10 for each word encrypted. Participants in the uneven wage treatments were first informed that one of them would be randomly determined to be "high productivity" and the other to be "low productivity," and that the high productivity individual would earn \$0.15 for every line of decryption, compared to \$0.05 for the low productivity individual. The next screen informed individuals which productivity type they were, and thus participants knew their relative wages rates before voting.

Participants in the endogenous taxation treatments were then told that they would vote on a redistributionary tax rate before beginning the task, and that one member of each group's vote would be randomly selected to count. 3 choices were presented: no tax (0%), medium (50%) and high (100%), along with explanations and examples of how the redistributionary tax

[2]Subjects were recruited in groups of 10, but occasionally a scheduled participant did not show up, and due to the necessity of running with an even number a 9th participant was randomly dismissed.

worked. After both members of a group voted, participants were informed which tax rate had been implemented.

In the exogenous taxation treatments, instead of being told that they would vote for their preferred tax rate, participants were informed that there were the three aforementioned possible tax rates, one of which would be selected randomly. On a subsequent screen they were then informed of which rate had been implemented. A breakdown of the number of participants in each treatment is presented in Table 2.1.

Table 2.1. Number of participants in each treatment.

Tax rate	**No Vote**	**No Vote**	**Vote**	**Vote**
	Even Wage	**Uneven Wage**	**Even Wage**	**Uneven Wage**
0	n=10	n=10		
50%	n=10	n=10	n=26	n=32
100%	n=10	n=10		

Participants were then given more detailed instructions and tips for completing the task, which they were informed would last for 10 minutes before automatically ending, and the task began when all participants in each session clicked to indicate that they were ready.Once 10 minutes had elapsed since the start of the task, the task period automatically ended, and participants were shown a results screen which told them their pre-tax earnings, the pre-tax earnings of the person they were matched with, and their final post-tax earnings.

Finally, participants completed a questionnaire which asked about whether they thought redistributive taxation (as a broader concept) was fair, whether they thought the tax rate impacted their effort and/or the effort of the person they were matched with and why, their political preferences, and demographic questions. After completing the questionnaire, participants were paid individually and privately, then dismissed.

2.4 Theory

In this theory section, I characterize whether votes for redistribution are consistent with various models of preferences. First I consider the MR model, then an MR with uncertainty of position in the income distribution, third a model with Fehr-Schmidt preferences, and lastly a model with preferences for income to be proportionate to effort.

2.4.1 Meltzer and Richard Model

First, consider this simplified version of the MR model, with only 2 individuals in the economy. Each individual i seeks to maximize their strictly concave utility U_i, which is determined by the wage rate w_i, their units of labor supplied (or effort) x_i , and the tax rate $t \in [0,1]$ (Equation 1). Each individual pays the tax rate from their earnings (their effort multiplied by their wage rate), and receives the average contribution of all individuals j through n in the economy. Individuals dislike exerting effort, and the cost of effort function, $c(x_i)$, is assumed to be convex and the same for all individuals. Because this experiment uses a real effort task, I do not make further assumptions about the cost of effort function.

$$U_i^{MR}(w_i, x_i, t) = (1-t) * w_i x_i - c(x_i) + \frac{1}{2} \sum_{j=1}^{2} (t * w_j x_j) \tag{2.1}$$

Optimal labor supply for each individual x_i^* is set, conditional on the tax rate, for when the marginal gain from effort is equal to the marginal cost, with $\frac{\partial U_i}{\partial x_i} = 0$ at $\frac{\partial c(x_i)}{\partial x_i} = w_i(1.5-t)$. x_{-i} is assumed to be determined solely by the implemented tax rate. Because the marginal gains from effort decrease as taxation increases but the cost of effort is not impacted, effort decreases as taxation increase, thus for any pair of tax rates t, t' where $t \leq t'$, $x_{it} \geq x_{it'}$.

In accordance with their beliefs about the disincentive effects of taxation and their relative earnings, each individual votes for the tax rate t which maximizes their own utility. In the MR model, tax policies are assumed to be enacted by the median voter, but here as there 2 voters of equal political importance, I characterize the optimal tax rate for a voter i, under the condition

that each voter has the same cost of effort function.

$$t_i^* = \begin{cases} \frac{n^2}{n^2-1} * \frac{\frac{1}{n}Z - w_i^2}{\frac{2}{n+1}Z - w_i^2} & \text{if } w_i x_i < \frac{1}{n}Z \\ 0 & \text{if } w_i x_i \geq \frac{1}{n}Z \end{cases} \tag{2.2}$$

Where Z denotes the aggregate income in the case of $t = 0$. The intuition is straightforward, as voters prefer no taxes if they would earn above the mean income in the case of no taxes, and prefer positive taxation as long as the monetary gains from taxation more than offset the decreased effort provision.

An important note is that individuals that would earn the mean income in this case should not support taxation. As by assumption each individual in this model shares the same cost of effort function, worker effort levels facing the same tax and wage rates will exert the same amount of effort, and thus have the same earnings. Taxation would only disincentivize effort equally for both parties.

Formally, this lead to the following hypothesis:

H1 MR: Medium-wage workers with will vote for no redistribution.

2.4.2 MR with Uncertainty

In the MR model, individuals have perfect knowledge, and thus do not face uncertainty of their position within the final wage distribution and thus the material effects of taxation. For an individual around the mean income that has imperfect information, uncertainty over being above or below the mean income is a reasonable concern, as individuals face a challenge both in forecasting their own future incomes and the future incomes of others, which are both likely to have some variability. Intuitively, this could induce self-interested demand for taxation in order to reduce uncertainty.

More formally, this uncertainty leads to individuals having some range of potential utilities from a given tax rate, and thus their preferred tax rate must take into account their beliefs

about whether the tax will help or hurt them.

Consider the following: conserving all other features of the previous MR model, each individual i is unsure of the relationship between their cost of effort function and that of their counterpart m. If workers are paid equally and $c_i(x) > c_m(x) \forall x \in X$, then i will exert less effort than m at any given t. The converse is true if the cost of effort relationship is flipped.

Individual i's beliefs about m's cost-of-effort function can subsequently rationalize the implementation of any tax. This leads to the following hypothesis, which contrasts with middle-wage worker voting in the basic MR model.

H2 MR + U: Medium-wage workers motivated by hedging against uncertainty may vote for no, medium, or high redistribution.

Notably, this uncertainty is unlikely to be inconsequential for workers with unequal wages, as they must hold extreme beliefs about the relative cost of effort functions. While such beliefs are not impossible and I cannot rule them out, uncertainty with self-interested preferences is an intuitively unlikely explanation for low- or -high wage workers voting other than for a 0% or 100% tax rate, respectively.

2.4.3 Fehr-Schmidt Inequity Aversion

The MR model is amenable to including various forms of social preferences, and here I include Fehr-Schmidt (FS) inequity aversion, in which individuals prefer post-tax incomes to be compressed to some degree (Equation 3). α captures the strength of distaste for the other individual earning more, and β captures the distaste for earning more than the other individual.

$$U_i^{FS}(w_i,x_i,t) = U_i(w_i,x_i,t) - \alpha \sum_{j=1}^{n} max(U_j(w_j,x_j,t) - U_i(w_i,x_i,t),0) - \beta \sum_{j=1}^{n} max(U_i(w_i,x_i,t) - U_j(w_j,x_j,t),0) \quad (2.3)$$

The standard assumption in the literature, which I follow, is that individuals more strongly disapprove of inequities that are not in their favor, thus $0 < \beta < \alpha < 1$. This condition satisfies that the ideal tax rate of an individual is weakly greater than if $\beta = \alpha = 0$.

While these preferences alone with full information allow for rational implementation of taxation by high-wage workers, they would still not lead to a tax on their own in middle-wage workers, as there would be no inequality and like in the MR model each voter would simply be worse off due to lower productivity.

H3 FS: Medium-wage workers who vote for no taxation may have FS preferences, but they do not drive behavior.

Including uncertainty however into these preferences also allows medium-wage workers to vote for a positive tax.

H4 FS + U: Medium-wage workers voting for positive taxation is consistent with uncertainty over the income distribution, and some demand may come from Fehr-Schmidt (FS + U) preferences.

2.4.4 Effort-Fairness

A second social-preferences model I refer to as Effort-Fairness (F) captures the desire for monetary outcomes to be proportionate to effort. In this model, individuals do not have an aversion to inequity per se. Let γ represent the earnings after taxation and redistribution of an individual, conditional on the tax rate t. Thereby, $\frac{\gamma}{x}$ is the ratio of earnings to units of labor supplied. I borrow the quality from FS of individuals having a self-interested ranking of F preferences, with a greater distaste for disadvantageous unfairness than advantageous unfairness. Let ι represent the distaste for an disadvantageous unfairness, and δ represent the distaste for advantageous unfairness, with $0 < \iota < \delta < 1$. The formal utility function is presented in Equation 4.

$$U_i^F(w_i,x_i,t) = U_i(w_i,x_i,t) - \iota \sum_{j=1}^{n} max(\frac{\gamma_j}{x_j} - \frac{\gamma_i}{x_i}, 0) - \delta \sum_{j=1}^{n} max(\frac{\gamma_i}{x_i} - \frac{\gamma_j}{x_j}, 0) \quad (2.4)$$

Compared to FS preferences, F preferences predict that individuals who earn the same wage rate will not support redistributionary taxation, as any redistribution would cause a deviation from post-tax earnings being proportionate to labor supply.

In the case of uncertainty, demand for taxation due to self-interested reasons would be tempered by effort-fairness concerns, as any level of redistribution would still cause a net transfer from the individual who completed more tasks to the one who completed fewer. Thus, fairness preferences cannot explain any level of support

H4 F: Effort-fairness preferences push medium-wage workers to vote for a 0% tax.

2.5 Results

A total of N = 118 individuals participated. Sessions lasted roughly 25 minutes on average, and average payment was $9.31. Participants were on average 21.6 years of age, 64% female, and 58% reported English as their first language. 17% of participants were majoring in engineering or physics, 33% in natural sciences such as chemistry, biology or pre-med, 14% in social sciences excluding economics, 34% in economics or business, and 2% in literature, art, or history.

2.5.1 Voting

Voting substantially differed depending on the wage rate of the participant (see Table 3). The overall relationship between wage rate and preferred tax rate was negative ($\beta = -0.69, p < 0.01$), in line with the prediction in all models that voter behavior will be to some degree dependent upon self-interest. Voting distributions were significantly different at the $p = 0.05$ level

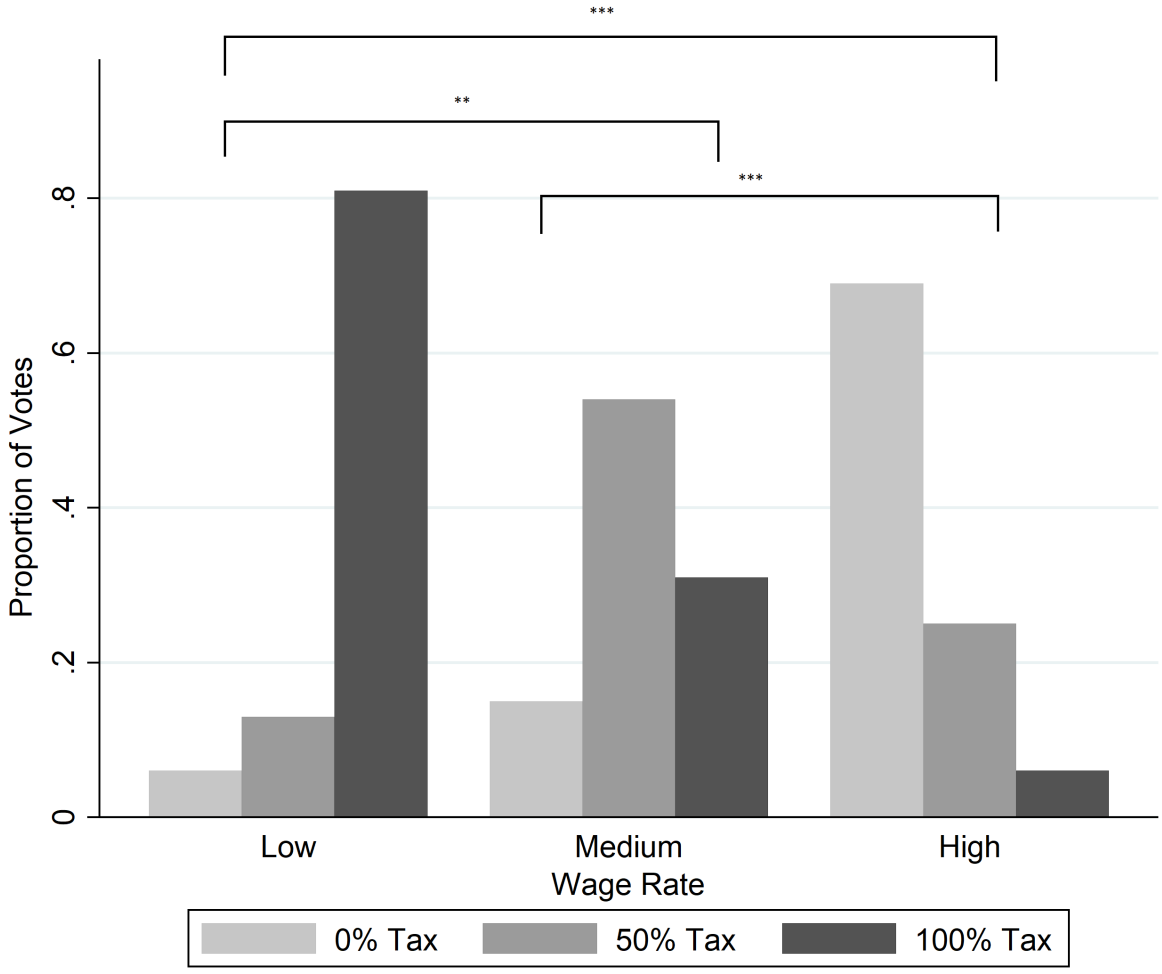

Figure 2.1. Fraction of the vote for each tax rate, separated by wage rate. Voting distributions compared using a χ^2 test of proportions. ** represents a significant difference at $p < 0.05$, *** represents a significant difference at $p < 0.01$.

Table 2.2. Votes for each tax rate by wage rate.

Tax rate	Low Wage N = 16	Medium Wage N = 26	High Wage N = 16
0	1 (6.25%)	4 (15.38%)	11 (68.75%)
50%	2 (12.50%)	14 (53.84%)	4 (25.00%)
100%	13 (81.25%)	8 (30.77%)	1 (6.25%)

or lower across wage groups by a χ^2 test of distributions (see Fig. 1), and the overall difference between groups was statistically significant (one-way ANOVA, $F(2,55) = 18.94, p < 0.01$).

The negative relationship between wage-rate and preferred tax-rate was driven by votes in the low- and high- wage rate groups. Low-wage workers voted by a large majority for the 100% tax rate, consistent with all 4 presented models of tax preferences, and similarly a large majority of high-wage workers voted for the 0% tax rate. The difference in vote distributions between high and low-wage workers was significant ($\chi^2 = 29.29, p < 0.01$).

Even-wage workers voted significantly differently from both low- and high-wage workers (Low: $\chi^2 = 10.19, p = 0.01$; High $\chi^2 = 12.60, p < 0.01$). The majority voted in favor of medium or high taxation, in line with U and FS + U models, which could indicate a desire to either hedge or compress post-tax incomes. Only 15.38% voted for the 0% tax rate, which would be in line with MR or F models.

Notably, a non-trivial fraction of low and high wage workers voted against their likely self-interest. That nearly a third (31.3%) of high wage earners did vote for some level of taxation is consistent with a substantial fraction of voters being motivated by some form of social preferences, but the 18.8% of low wage workers voting for attenuated or no taxation is not consistent with my model predictions. One possibility is that these low-wage workers simply didn't understand the task and tax, another is be that they viewed the random allocation of wage-rates as legitimate, and voted to preserve that spread.

2.5.2 Effort provision

Wage Rates

Changes in effort levels were less striking than those in voting patterns, with few significant differences. Wage rates were negatively but not significantly associated with effort levels, with an increase in wage level (e.g. from low to medium) associated with an average of $\beta = -0.44$ fewer words encoded ($p = 0.65$). Implemented tax rates also did not have a significant effect on effort ($\beta = 0.17, p = 0.92$), even when controlling for wage rate ($\beta = 0.17, p = 0.92$).

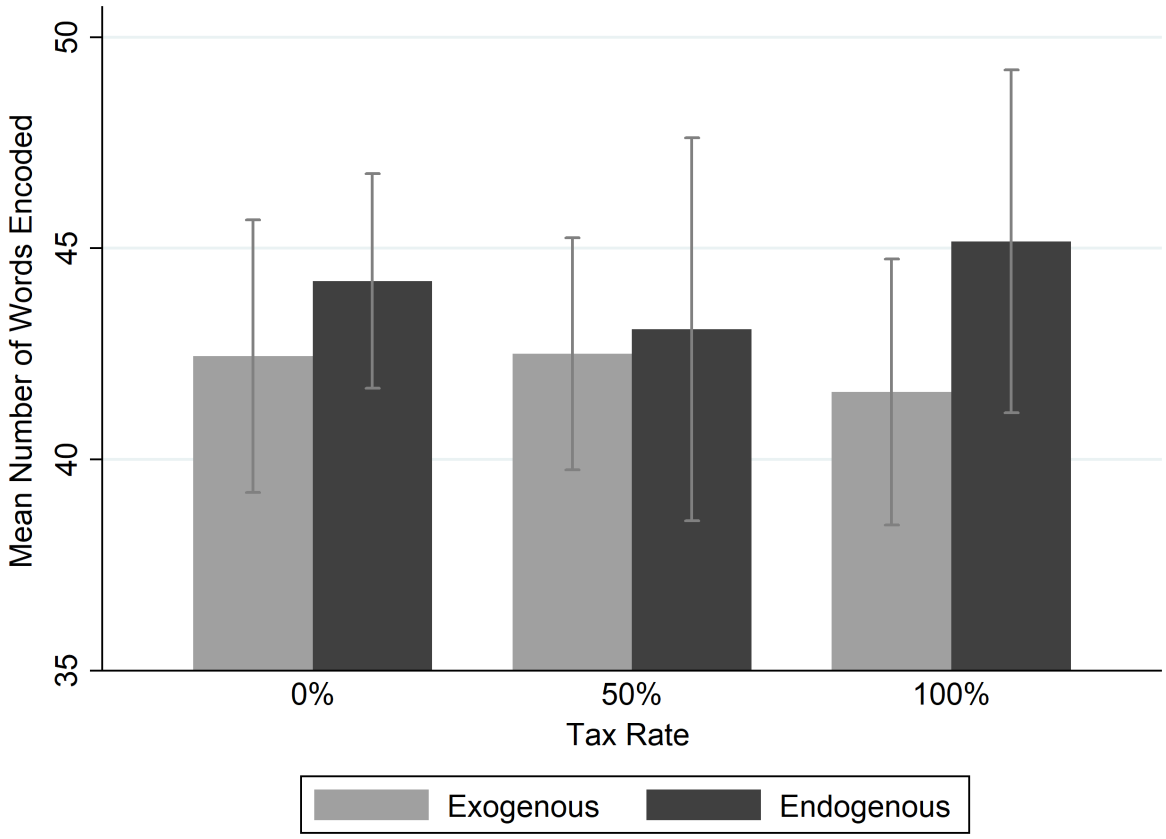

Figure 2.2. Mean number of words encoded, separated by tax rate and system of determination. 95% Confidence interval displayed.

This was surprising, as the marginal returns to effort varied substantially based on the wage and tax rates (from \$0.025 per word to \$0.15, a 6-fold increase). Participants may have been motivated to maximize their earnings regardless of these factors, corresponding theoretically to the cost of effort never exceeding the utility gains of any earnings rate in the work period.

Effects of Voting

While participants who voted tended to exert more effort, this effect was not significant ($\beta = 2.20, p = 0.12$). There was also no significant effect based on whether a participant's vote was implemented ($\beta = 0.69, p = 0.62$).

One of the principal reasons for investigating the effect of endogenous tax-rates was that voting may signal something about the intended effort level or preferences of one participant

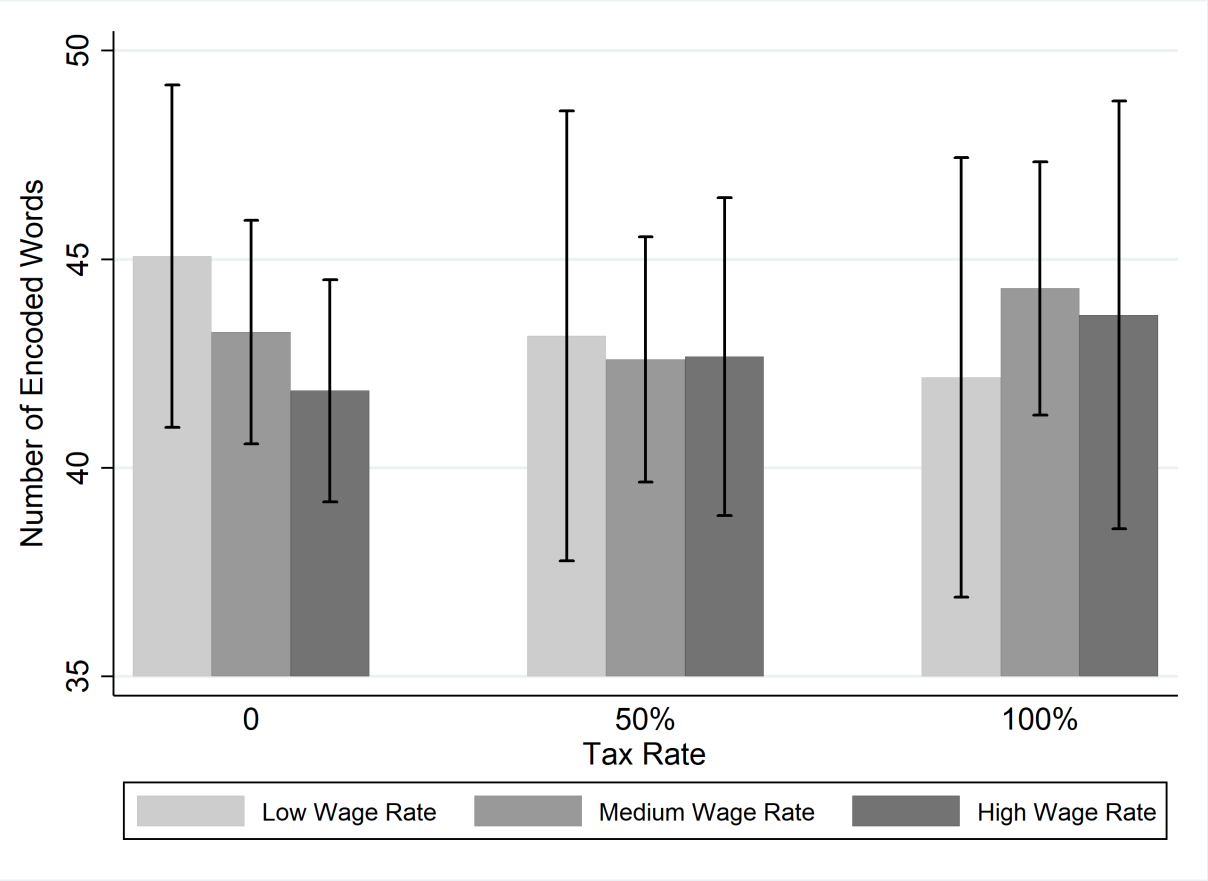

Figure 2.3. Mean number of words encoded, separated by tax rate and wage rate. 95% Confidence interval displayed.

that may affect the effort provisions of the other. Voting for higher taxes was not associated with significant differences in effort levels, even when controlling for the implemented tax rate ($\beta = 0.78, p = 0.59$). Participants who voted for a lower tax rate than was implemented (and thus know that their partner voted for a higher tax rate) did not exert significantly lower effort ($\beta = -1.00, p = 0.75$), showing that there was no evidence of retaliatory shirking.

2.6 Discussion

Voting behavior in this experiment is consistent with the hypotheses generated by multiple theories that individuals vote in their self-interest when considering taxation rates, insofar as they clearly know whether the redistribution will benefit them or not. When the relative earnings of

each partner in a pair are clear, as in the case of uneven wage rates, a large majority of low earners voted to raise taxes as much as possible. A somewhat larger proportion of high earners also voted for some degree of redistribution, consistent with having some preference for compressed post-tax earnings, but still the large majority vote for no taxation, which maximized their own earnings.

For workers with equal wages, for whom the net effect of a tax is uncertain, demand for redistribution was high. This demand for redistribution can be explained by workers having uncertainty over the pre-tax earnings of their group, which allows both self-interest and inequity aversion to theoretically drive demand.

One of the more striking findings was the lack of evidence in favor of Effort-Fairness preferences. Only a small fraction few medium-wage workers chose the 0% tax rate that an individual with strong Effort-Fairness preferences would choose. A potential reason for this discrepancy may be the ex-ante nature of redistributionary decision-making in this experiment, as opposed to ex-post in many other studies, which changes the amount of information about effort levels available to decision-makers. Another reason may be that individuals are indeed concerned with the cost-of-effort function, rather than just effort as defined by units of work completed. They may care for instance about how subjectively negative it is for others to work on the task, or even that all individuals spent an equal amount of time working.

It is more puzzling why effort levels did not differ more than was observed, given that the lowest marginal gain from a unit of work was 1/6th of the highest. Part of this could be due to the task and time limit–it may be that this task was simply not difficult enough to perform for 10 minutes that differences in motivation would yield substantial differences in provision. A potential implication of this finding is that the few voters who voted for low redistribution for fear of disincentivizing effort mis-estimated the elasticity of labor supply. Whether this mis-estimation occurs in the real world is an empirical question beyond the scope of this paper, although anecdotally a frequent argument of anti-taxation groups is that it is in the best interests of the low earners to only lightly tax high earners in order to incentivize them to work harder in

order to maximize the size of the tax base. Understanding the importance of the estimation of the disincentive effects of taxation by low income groups may be fruitful in understanding their voting preferences. Directly eliciting beliefs about the elasticity of effort would be useful for resolving these issues in a laboratory context.

2.7 References

1. Alesina, A., & Angeletos, G. M. (2005). Fairness and redistribution. American Economic Review, 95(4), 960-980.

2. Alesina, A., & Giuliano, P. (2011). Preferences for redistribution. In Handbook of social economics (Vol. 1, pp. 93-131). North-Holland.

3. Alesina, A., & La Ferrara, E. (2005). Preferences for redistribution in the land of opportunities. Journal of Public Economics, 89(5-6), 897-931.

4. Ashok, V., Kuziemko, I., & Washington, E. (2016). Support for Redistribution in an Age of Rising Inequality: New Stylized Facts and Some Tentative Explanations. Brookings Papers on Economic Activity, 2015(1), 367-433.

5. Agranov, M., & Palfrey, T. R. (2015). Equilibrium tax rates and income redistribution: A laboratory study. Journal of Public Economics, 130, 45-58.

6. Backus, P. G., & Esteller-Moré, A. (2017). Risk aversion and inequity aversion in demand for unemployment benefits. International Tax and Public Finance, 24(2), 198-220.

7. Benabou, R., & Tirole, J. (2006). Belief in a just world and redistributive politics. The Quarterly Journal of Economics, 121(2), 699-746.

8. Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. American Economic Review, 96(5), 1652-1678.

9. Cervellati, M., Esteban, J., & Kranich, L. (2010). Work values, endogenous sentiments redistribution. Journal of Public Economics, 94(9-10), 612-627.

10. Downs, A. (1957). An economic theory of political action in a democracy. Journal of Political Economy, 65(2), 135-150.

11. Durante, R., Putterman, L., & Van der Weele, J. (2014). Preferences for redistribution and perception of fairness: An experimental study. Journal of the European Economic Association, 12(4), 1059-1086.

12. Erkal, N., Gangadharan, L., & Nikiforakis, N. (2011). Relative earnings and giving in a real-effort experiment. American Economic Review, 101(7), 3330-48.

13. Esarey, J., Salmon, T., & Barrilleaux, C. (2012). Social insurance and income redistribution in a laboratory experiment. Political Research Quarterly, 65(3), 685-698.

14. Fehr, E., & Fischbacher, U. (2002). Why social preferences matter–the impact of non-selfish motives on competition, cooperation and incentives. The Economic Journal, 112(478), C1-C33.

15. Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. Trends in Cognitive Sciences, 8(4), 185-190.

16. Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics, 114(3), 817-868.

17. Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics, 10(2), 171-178.

18. Fong, C. (2001). Social preferences, self-interest, and the demand for redistribution. Journal of Public Economics, 82(2), 225-246.

19. Hansen, J. W. (2005). Uncertainty and the size of government. Economics Letters, 88(2), 236-242.

20. Kittel, B., Paetzel, F., & Traub, S. (2015). Competition, income distribution, and the middle class: An experimental study. Journal of Applied Mathematics, 2015.

21. Krawczyk, M. (2010). A glimpse through the veil of ignorance: Equality of opportunity and support for redistribution. Journal of Public Economics, 94(1-2), 131-141.

22. Lefgren, L. J., Sims, D. P., & Stoddard, O. B. (2016). Effort, luck, and voting for redistribution. Journal of Public Economics, 143, 89-97.

23. McCarty, N., Poole, K. T., & Rosenthal, H. (2016). Polarized America: The dance of ideology and unequal riches. MIT Press.

24. Moene, K. O., & Wallerstein, M. (2001). Inequality, social insurance, and redistribution. American Political Science Review, 95(4), 859-874.

25. Meltzer, A. H., & Richard, S. F. (1981). A rational theory of the size of government. Journal of Political Economy, 89(5), 914-927.

26. Meltzer, A. H., & Richard, S. F. (1983). Tests of a rational theory of the size of government. Public Choice, 41(3), 403-418.

27. Mulligan, C. B. (2014). Uncertainty, redistribution, and the labor market since 2007. IZA Journal of Labor Policy, 3(1), 8.

28. Posner, R. A., & Rasmusen, E. B. (1999). Creating and enforcing norms, with special reference to sanctions. International Review of Law and Economics, 19(3), 369-382.

29. Rawls, J.A. (1971). A Theory of Justice. Cambridge: Belknap Press of Harvard University Press.

30. Roberts, K. W. (1977). Voting over income tax schedules. Journal of Public Economics, 8(3), 329-340.

31. Romer, T. (1975). Individual welfare, majority voting, and the properties of a linear income tax. Journal of Public Economics, 4(2), 163-185.

32. Starmans, C., Sheskin, M., & Bloom, P. (2017). Why people prefer unequal societies. Nature Human Behaviour, 1(4), 1-7.

33. Sutter, M. (2006). Endogenous versus exogenous allocation of prizes in teams—Theory and experimental evidence. Labour Economics, 13(5), 519-549.

34. Tausch, F., Potters, J., & Riedl, A. (2013). Preferences for redistribution and pensions. What can we learn from experiments?. Journal of Pension Economics & Finance, 12(3), 298-325.

35. Almås, I., Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2010). Fairness and the development of inequality acceptance. Science, 328(5982), 1176-1178.

36. Tyran, J. R., & Sausgruber, R. (2006). A little fairness may induce a lot of redistribution in democracy. European Economic Review, 50(2), 469-485.

37. Varian, H. (1980). Redistributive taxation as social insurance. Journal of Public Economics, 14(1), 49-68.

Chapter 3

Assessing the Role of Gender in Choosing a Primary Care Specialty in Medical Students; A Longitudinal Study

3.1 Abstract

Background: There is a general shortage of primary care physicians in the United States, and women are disproportionately more likely than men to become primary care physicians.

Purpose: Increase understanding as to why men and women choose to enter primary care specialties.

Methods: A longitudinal annual survey from 2013-2019 was administered to students at the University of Iowa Carver College of Medicine. Using a logistic regression model, we examine the factors associated with students pursuing careers in primary care specialties.

Results: Men were significantly more likely than women to report that their choice of specialty was influenced by debt (N=278, $\chi^2(1)=10.88$, p=0.001), and students who reported that debt influenced their specialty choice were approximately 1/3rd as likely to enter a primary care specialty (N=189, 95% CI [0.11-1.06], p=0.06). For men, but not women, the subjective importance of potential salary was negatively associated with entering a primary care specialty (p=0.03).

Discussion: A key gender differentiator is that men's specialty choice is more negatively

influenced by the financial concerns of debt and future salary.

Translation to Health Education Practice: medical schools seeking to increase the number of students who match into primary care specialties should target interventions based on financial concerns.

3.2 Introduction

The American medical system faces a shortage of primary care physicians (PPs), with a projected shortage of 46,000 PPs by 2025 in the US1. The most pervasive shortages are found in rural and underserved areas2. The reduced access for patients leads to overutilization of acute care providers for routine or preventable health care3, 4. The dearth of PPs is consequential for the overall quality of healthcare in the United States, as nations with robust primary care systems demonstrate better healthcare outcomes and lower costs4. The shortage persists despite the number of residency positions in the core primary care specialties of internal medicine, pediatrics, and family medicine (the specialties defined as comprising primary care in this study) reaching an all-time high in 20181.

In response to this shortfall, many medical schools have striven to change student selection criteria and pursue curricular changes to increase the number of students choosing a primary care specialty. Strategies implemented by medical schools include mandatory primary care rotations, loan forgiveness, longitudinal programs, and rural track programs. Meta-analyses conducted by Pfarfaller et al. (2015) found that longitudinal interventions throughout medical school are the most effective6, but there is still a need for more research to identify effective and cost-effective targets for interventions at multiple stages in the medical education process.

In this study, we examine the factors associated with entering a primary care specialty in order to better guide future interventions and policy changes. We employ a theoretical framework based on work from Pfarfaller et al. (2015), in which factors affecting specialty choice are organized into conceptually linked groups which reflect individual characteristics,

preferences, and experiences. Specifically, we examine financial considerations, including debt and future salary potential; mentorship and early exposure to PPs; lifestyle and job preferences; and demographics.

In addition to these conceptual groups, we examine the role of gender, which varies dramatically across different medical specialties. Women comprise an increasing majority of domestic PP residents, accounting for 52.9% of total domestic residents in family care, internal medicine, and pediatrics in 20197. This includes female majorities in family medicine (54.9%) and pediatrics (75.3%), and a female minority in internal medicine (42.2%)7.

Our analysis uses a unique, prospective dataset from University of Iowa Carver College of Medicine (UICCM), comprised of comprehensive surveys administered annually containing questions about background, current goals, debt, and other factors. Our long-running study has a large sample size and multiple cohorts with repeated, regular surveys. This allows us to examine the importance of various factors, such as the subjective importance of time spent in patient contact, both early in medical school and after having matched. The added depth also allows for subcategorization of students by demographics and other characteristics—as well as a more integrative picture of microsystem factors, like mentorship and research experiences, with macrosystem factors including debt burden and future employment prospects.

3.3 Methods

Data was procured via the Carver College of Medicine Specialty Choice survey, a series of surveys given annually to medical students at UICCM, beginning in 2013 and still ongoing. This study was approved by the University of Iowa Institutional Review Board. The most recent data was recorded after the 2019 Match, and thus our data set includes 2013, 2014, and 2015 matriculants. The survey is given to all students, with an average completion rate of roughly 70% per cohort. Survey design was based on literature review and discussions with medical students and residents, who endorsed certain factors as being important in their specialty selection. Each

student was surveyed at the beginning of each year, as well as after the match in 4th year. We define primary care in this study as internal medicine, pediatrics, and family medicine.

For all descriptive statistics, only those students for whom there is information on their specialty match are reported. The majority of three cohorts have now completed medical school and are included in the data. Of these, N=374 students provided specialty match data by the summer of 2019. Descriptive statistics comparing men and women, and those who did or did not match into primary care specialties are presented in Tables 1-4.

3.3.1 Analysis

To examine which factors are strongly associated with matching into a primary care specialty, we compare levels of surveyed attributes using 2-sided t-tests and χ^2 tests where appropriate, as well as employ multivariate logistic (logit) regressions to calculate adjusted odds-ratios (OR) for which factors are significantly associated with selection into a primary care specialty. We present results both from pooling genders and by analyzing genders separately. We also pool together data from the three cohorts. We examine data in three time periods: from before medical school, from the first 2 years, and from after match. This was done for both theoretical and practical reasons. On a practical level, response rates for M3 and M4 were comparatively low, likely due to many students being on rotation. On a theoretical level, these 3 tranches are each relevant to different types of potential interventions, for instance pre-matriculation information being useful for admissions, and early medical school experiences relevant to interventions in this time period. All statistical analysis was done using STATA 14.0.

3.4 Results

3.4.1 Gender

Female gender is positively associated with selecting into a primary care specialty, a finding that is significant and robust to most tested regression specifications. When controlling

for student characteristics measured in the M1 survey, which includes factors determined prior to matriculation, women are 2.13 times (95% CI [1.30-3.54], p¡0.001, Appendix Table 1) more likely than men to enter a primary care specialty. Female gender was also significantly associated with entering a primary care specialty when controlling for factors measured through the M2 survey (N=264, Adjusted OR=2.06, 95%CI[1.12-3.79], p=0.02, Table 5 Column 4, and Figure 1). When controlling for all post-match factors, female gender is still positively associated, but not significantly (N=189, Adjusted OR=1.61, 95%CI[0.76-3.41], p=0.21, Table 6 Column 4, and Figure 2).

3.4.2 Mentorship and Exposure to Primary Care Physicians

Over time, students tended to find mentors in the fields they ultimately matched into. Thus, when examining the role of mentorship, we primarily consider mentorship experiences in the first 2 years of medical school, which are likely to play a greater role in determining specialty choice, rather than being determined by it. 32.54% of female medical students reported a mentor in a primary care specialty in their first 2 years, significantly more than the 14.19% of men (N=281, χ^2=13.45, p¡0.001). Of the 41 women who reported having a primary care mentor in their first 2 years, 23 would go on to match into a primary care specialty. When controlling for other variables from the first 2 years of medical school, the relationship between having a primary care mentor in the first 2 years and entering a primary care specialty is not statistically significant for women (N=116, Adjusted OR = 1.55, 95% CI[0.55-4.35], p=0.40, Table 5, Column 6). Having a mentor in a primary care specialty in the first 2 years of medical school was also not associated with a significant change in the likelihood of matching into a primary care specialty for men in any tested specification (N=150, Adjusted OR = 0.57, 95% CI [0.17-1.91], p=0.37, Table 5, Column 5).

Having a family member who practices family care is positively associated with entering a primary care specialty. The effect is significant at the 0.05 level when pooling genders, with an overall 2.87 times increased likelihood of entering a primary care specialty (N=303, 95% CI

[1.14-7.19], p=0.03). This effect is stronger in women, who are 3.72 times more likely to enter a primary care specialty (N=133, 95% CI [0.89-15.48], p=0.07), and men have a positive but statistically insignificant effect (N=120, Adjusted OR = 2.76, 95% CI [0.77-9.89], p=0.12). The proportion of students with a family member who practiced primary care did not differ between genders (N=332, χ^2=0.15, p=0.70). The importance of having a family member who practices primary care is also significant post-match when controlling for other variables as well (N=189, Adjusted OR=5.44, 95%CI[1.14-25.91], p=0.03].

3.4.3 Research

Pre-matriculation research experiences were negatively associated with the likelihood of women entering a primary care specialty for women, but not men. If a woman had pre-matriculation research experience, they had a 0.35 times likelihood of entering a primary care specialty (N=133, 95%CI[0.17-0.73], p=0.01), and even if the research was related to primary care, it had a negative but statistically insignificant association with entering a primary care specialty (N=133, Adjusted OR = 0.52, 95%CI [0.20-1.37], p=0.19). Conducting primary care research in medical school was not strongly associated with the likelihood of entering a primary care specialty, with the only significant relationship found in our tested regression being positive for men who conducted primary care research in their first 2 years of medical school (N=153, Adjusted OR=2.64, 95%CI[0.94-2.00], p=0.07, Table 5 Column 2].

3.4.4 Earnings and Debt

83.91% of students reported being in educational debt by the end of medical school. There was no significant difference in the rate of having debt between students who did or did not match into primary care (N=286, $\chi^2(1)$=1.43, p=0.23), and having debt did not have a significant effect on the likelihood of entering into a primary care specialty (N=189, Adjusted OR = 2.02, 95% CI [0.74-5.50], p=0.17). However, students who reported post-match that debt had a significant impact on their specialty choice were 67% less likely to enter a primary care

specialty (N=189, 95% CI [0.11-1.06], p=0.06). 26.79% of women reported that medical school debt had an impact on their specialty choice, compared to just 10.09% of men, a significant difference (N=278, $\chi^2(1)$=10.88, p¡0.001).

Students were also asked about the importance of potential salary in their specialty choice, and there was a significant negative relationship for men at their M2 survey. For each 1 point increase in the subjective importance of potential salary, men were 34% less likely to ultimately matching into a primary care specialty (N=148, 95%CI[0.45,0.96], p=0.03]. This relationship persisted at post-match as well (N=108, Adjusted OR=0.50, 95%CI[0.27-0.94], p=0.03]. For women, there was no significant relationship between the importance of potential earnings and likelihood of entering a primary care specialty. Men on average rated the importance of potential salary significantly higher than women (N=362, 2-sided t-test, t(360)=2.64, p=0.01).

3.4.5 Lifestyle and Job Preferences

A number of questions were asked pertaining to the importance of various job and lifestyle attributes in preferring or selecting a specialty. Academic vs. private practice opportunities and intellectual stimulation were not significantly related to choosing a primary care specialty in either M2 or post-match survey.

The amount of time in patient contact was significantly negatively related to matching into a primary care specialty at M2 when pooling genders, with each increase in importance level related to a 19% lower likelihood of matching into primary care (N=264, 95%CI[0.66-0.99], p=0.04, Table 6 Column 4). Specialty status/reputation was only significantly positively related to choosing a primary care specialty for men at M2 (N=148, Adjusted OR=1.46, 95%CI[1.01-2.12], p=0.05, Table 5 Column 5).

Quality of life was not significantly related to choosing a primary care specialty at M2, but significantly negatively associated for both men and women post-match, where each increased level of importance of quality of life related to a 35% lower likelihood of matching into a primary care specialty (N=189, 95%CI[0.45-0.93], p=0.02, Table 6 Column 4).

The importance of technical skills necessary was not related to specialty choice in M2, but post-match, was significantly negatively related for both men and women, with a pooled genders effect of a 46% lower likelihood (N=189, 95%CI[0.39-0.75], p¡0.01, Table 6 Column 4).

3.5 Discussion

3.5.1 Mentorship and Exposure to Primary Care Providers

Having a medical mentor in a given specialty is a medical career choice factor frequently examined in the literature. Some have found strong positive associations between having a medical mentor in a given specialty and matriculating into that field6, 8. There are, though, difficulties in identifying a causal relationship between mentorship and specialty match. One plausible conclusion is that exposure to a mentor itself leads to the increased likelihood of matching into their specialty. What may overstate the effect of mentorship, is that students frequently choose their own mentors in a given specialty, to increase their competitiveness. This occurs by way of letters of recommendation and the faculty member's professional network. That is, students may develop preferences for the specialty before finding a mentor in the specialty, and so there may not be much of an effect of mentorship per se. Thus, policy changes based on such findings (e.g. increasing opportunities for exposure to role models) have an uncertain effect.

In our analysis, we attempt to overcome some of these challenges by focusing on mentorship experiences in the first 2 years, a time period when a student's specialty preferences are not yet fully formed, and they are less likely to have sought a mentor in a field to bolster their match competitiveness. We find that having a primary care mentor is not significantly associated with matching into a primary care specialty, but that women are substantially more likely than men to have a primary care mentor.

Having a family member primary care practitioner was strongly associated with an increased likelihood of entering a primary care specialty, and this effect was particularly strong for women. There are several possible, non-exclusive explanations for this finding. The first is

that exposure to primary care physicians could increase interest in primary care practice–learning more about a specialty and associated lifestyle as an individual grows up could make it more appealing. The second is that having a family member who is a primary care practitioner could be correlated with a broader set of common values in the family, which favors entering primary care.

3.5.2 Lifestyle and Job Factors

Survey responses to lifestyle and personality factors were stronger predictors of matching into a primary care specialty the further along students progressed in their medical education. This is not entirely surprising, considering the limited exposure to specialties and clinical medicine during the orientation period. However, it does speak to an opportunity that increasing student awareness of some lifestyle and professional differences between medical specialties early in their medical education may influence their preferences. Relatedly, attempts to increase interest in primary care should account for the limited exposure and experience of pre-clinical medical students. Students may not be aware of desirable aspects of each specialty until their clinical clerkships. For instance, the importance of time spent in patient contact as assessed in year 2 is negatively associated with entering a primary care specialty, but this negative association is gone post-match.

3.5.3 Earnings and Debt

Medical school debt is a subject of increasing scrutiny. For the graduating class of 2018, 75% of students had student loan debt, with an average indebtedness of $200,000; with standard federal interest rates and repayment plans, this can grow to more than 400k with accrued interest through the life of the loan9. Numerous students matriculate into medical school with several hundred thousand dollars of existing student loan debt. Some argue that these costs drive students into high-paying subspecialties, in order to shed oneself of his/her debt burden as early as possible. By the time of match, 240 of 286 (83.91%) of students reported being in debt from

medical school. We do not find that having debt alone significantly explains choosing a primary care specialty, and the proportion of students in debt from medical school did not differ across primary care or non-primary care specialties, or across genders. A limitation of our study is that in initial survey years students were only asked about having debt, not levels of debt, although this has now been changed. Philips et al. (2014) find that high debt levels are associated with lower selection into primary care specialties in public schools, which can be assessed with future cohorts of our ongoing study.

There were, though, substantial differences based on whether students responded affirmatively to whether their debt impacted their specialty choice. Men were overall much more likely than women to say that their educational debts impacted their specialty choice, and men who responded that their specialty was influenced by debt were significantly less likely to enter a primary care specialty. This trend was also observed in women, but the relationship was not significant. Working in the same direction, for men but not women, the importance of future salary was negatively and significantly associated with likelihood of entering a primary care specialty.

Further research is necessary to elucidate why the financial considerations of debt and future salary are more influential in pushing men out of primary care than women. It may be related to broader motivations for entering the medical profession, perceived societal expectations, or differences in opportunities to enter higher paying specialties. Understanding more about this gap may help create interventions which encourage more men into selecting primary care specialties and help to alleviate the expected shortage of primary care physicians in the future.

3.5.4 Limitations

There are several limitations to this study. First, specialty is not a perfect proxy for primary care physicians; studies have found that 60% of internal medicine graduates subspecialize, and many primary care physicians are switching to more lucrative hospitalist positions1. Family medicine physicians have the most years of direct primary care, with 4-5 times the number of

"primary care" career years compared to internal medicine graduates10. However, eventual residency specialty is the best estimate at the level of the medical school selection process.

Second, this study only covers three graduating classes at one medical college, limiting the scope of which questions can be answered. Collaborating across institutions would provide a more representative sample of medical students, as well as increase statistical power.

Third, this study has difficulty disentangling medical student preferences for different specialties and their ability to enter them. Integrating longitudinal data with measures of student performance would improve this.

3.6 Conclusion

Increasing the number of primary care practitioners will require many changes; one key element is medical schools taking greater initiative to increase the number of students entering into primary care. Designing interventions to accomplish this is aided by a better understanding of the characteristics, preferences, and experiences associated with specialty choice.

In this study, we find several factors associated with an increased likelihood of entering primary care specialties, including female gender, having family members who practice family care, and placing a low importance on technical skills at the end of medical school. We also find evidence that the primary care gender gap may be attributable to differences in the subjective importance placed on future earnings and educational debt, with these factors associated with men choosing other specialties.

3.7 Acknowledgments

I thank my co-authors on this study Corry McDonald, Patrick Barlow, and Jerrod Keith for their collaboration on this project. Written permission has been granted by the co-authors for the use of this chapter.

3.8 References

1. Petterson, S. M., Liaw, W. R., Tran, C., & Bazemore, A. W. (2015). Estimating the residency expansion required to avoid projected primary care physician shortages by 2035. The Annals of Family Medicine, 13(2), 107-114.

2. Tu, H. T., & O'Malley, A. S. (2007). Exodus of male physicians from primary care drives shift to specialty practice. Tracking Report, 17, 1-6.

3. Rust, G., Ye, J., Baltrus, P., Daniels, E., Adesunloye, B., & Fryer, G. E. (2008). Practical barriers to timely primary care access: impact on adult use of emergency department services. Archives of internal medicine, 168(15), 1705-1710.

4. Coster, J. E., Turner, J. K., Bradbury, D., & Cantrell, A. (2017). Why do people choose emergency and urgent care services? A rapid review utilizing a systematic literature search and narrative synthesis. Academic Emergency Medicine, 24(9), 1137-1149.

5. Ridic, G., Gleason, S., & Ridic, O. (2012). Comparisons of health care systems in the United States, Germany and Canada. Materia socio-medica, 24(2), 112.

6. Pfarrwaller, E., Sommer, J., Chung, C., Maisonneuve, H., Nendaz, M., Perron, N. J., & Haller, D. M. (2015). Impact of interventions to increase the proportion of medical students choosing a primary care career: a systematic review. Journal of general internal medicine, 30(9), 1349-1358.

7. Association of American Medical Colleges. Report on Residents, 2018.

8. Osborn, E. H. (1993). Factors influencing students' choices of primary care or other specialties. Academic medicine: journal of the Association of American Medical Colleges, 68(7), 572-574.

9. Education Debt Manager. AAMC: FIRST Program, 2018.

10. Bowman, R. C. (2008). Measuring primary care: the standard primary care year. Rural & Remote Health, 8(3).

11. Phillips, J. P., Petterson, S. M., Bazemore, A. W., & Phillips, R. L. (2014). A retrospective analysis of the relationship between medical student debt and primary care practice in the United States. The Annals of Family Medicine, 12(6), 542-549.

3.9 Appendix

Table 3.1. Descriptive statistics of those who did/did not match into a primary care specialty Comparison of descriptive statistics of students who matched into internal medicine, pediatrics, or family medicine specialties. Likert scores taken from M1 survey. Pooled Genders.

Characteristic	Matched Into Primary Care Specialty, Pooled Genders. Mean (SD) or N (%).	Did Not Match Into Primary Care Specialty, Pooled Genders. Mean (SD) or N (%).	Test of differences
Count	137 (36.63%)	237 (63.37%)	-
Avg. age at matriculation	23.3 (1.97), N=110	23.54 (2.41), N=197	N=307, 2-sided t-test, t(305) = 0.87, p=0.38
Gender is female	77 (56.20%)	87 (36.71%)	N=374, $\chi^2(1)$=13.40, p¡0.01
Ethnicity is white	106 (77.37%)	189 (79.75%)	N=374, $\chi^2(1)$=0.29, p=0.59
Physicians in family	27 (22.13%)	44 (20.95%)	N=332, $\chi^2(1)$=0.06, p=0.80
Undergrad science based major	95 (87.96%)	173 (88.72%)	N=303, $\chi^2(1)$=0.04, p=0.84
Debt from pre-med education	48 (36.36%)	89 (40.09%)	N=354, $\chi^2(1)$=0.48, p=0.49
Pre-med research	54 (44.26%)	115 (54.76%)	N=332, $\chi^2(1)$=3.40, p=0.07
Pre-med mentor	45 (36.89%)	82 (39.05%)	N=332, $\chi^2(1)$=0.15, p=0.70
Ever married	27 (19.71%)	54 (22.78%)	N=374, $\chi^2(1)$=0.48, p=0.49
Mentor in primary care specialty in 1st 2 years	29 (21.17%)	34 (14.35%)	N=374, $\chi^2(1)$=2.88, p=0.09
Research in medical school	91 (66.42%)	205 (86.50%)	N=374, $\chi^2(1)$=21.20, p¡0.01
Amount of time in patient contact (Likert scale)	2.20 (1.59), N=108	2.46 (1.44), N=195	N=303, 2-sided t-test, t(301) = 1.44, p=0.15
Quality of life (Likert scale)	2.32 (1.67) , N=108	2.53 (1.59), N=195	N=303, 2-sided t-test, t(301) = 1.08, p=0.28

Table 3.1. Descriptive statistics of those who did/did not match into a primary care specialty, continued Comparison of descriptive statistics of students who matched into internal medicine, pediatrics, or family medicine specialties. Likert scores taken from M1 survey. Pooled Genders.

Characteristic	Matched Into Primary Care Specialty, Pooled Genders. Mean (SD) or N (%).	Did Not Match Into Primary Care Specialty, Pooled Genders. Mean (SD) or N (%).	Test of differences
Technical skills necessary (Likert scale)	2.07 (1.03) , N=108	2.13 (1.02), N=195	N=303, 2-sided t-test, t(301) = 0.42, p=0.63
Med school debt at match	90 (87.38%)		150 (81.97%) N=286, $\chi^2(1)$=1.43, p=0.23
Med school debt influenced specialty choice	9 (9.09%)	44 (24.58%)	N=278, $\chi^2(1)$=9.91, p¡0.01

Table 3.2. Descriptive statistics of men Comparison of descriptive statistics of students who matched into internal medicine, pediatrics, or family medicine specialties. Likert scores taken from M1 survey.

Characteristic	Matched Into Primary Care Specialty, Men. Mean (SD) or N (%).	Did Not Match Into Primary Care Specialty, Men. Mean (SD) or N (%).	Test of differences
Count	60 (28.57%)	150 (71.43%)	-
Avg. age at matriculation	23.63 (1.95)	24.00 (2.91)	N=172, 2-sided t-test, t(170) = 0.84, p=0.20
Ethnicity is white	48 (80.00%)	120 (80.00%)	N=210, $\chi^2(1)$=0.00, p=1.00
Physicians in family	11 (21.15%)	32 (24.06%)	N=185, $\chi^2(1)$=0.18, p=0.67
Undergrad science based major	40 (85.11%)	109 (88.62%)	N=170, $\chi^2(1)$=0.39, p=0.53
Debt from pre-med education	22 (39.39%)	60 (42.55%)	N=197, $\chi^2(1)$=0.18, p=0.68
Pre-med research	28 (53.85%)	69 (51.88%)	N=185, $\chi^2(1)$=0.06, p=0.81
Pre-med mentor	18 (34.62%)	51 (38.35%)	N=185, $\chi^2(1)$=0.22, p=0.64
Ever married	16 (26.67%)	39 (26.00%)	N=210, $\chi^2(1)$¡0.01, p=0.92
Mentor in primary care in 1st 2 years	6 (10.00%)	16 (10.67%)	N=210, $\chi^2(1)$=0.02, p=0.89
Research in medical school	46 (76.67%)	134 (89.33%)	N=210, $\chi^2(1)$=5.62, p=0.02**
Amount of time in patient contact (Likert scale)	2.34 (1.54)	2.45 (1.39)	N=170, 2-sided t-test, t(168) = 0.44, p=0.66
Quality of life (Likert scale)	2.53 (1.57)	2.55 (1.57)	N=170, 2-sided t-test, t(168) = 0.07, p=0.94
Technical skills necessary (Likert scale)	2.00 (1.04)	2.24 (1.07)	N=170, 2-sided t-test, t(168) = 1.29, p=0.20
Med school debt at match	42 (87.50%)	94 (81.03%)	N=164, $\chi^2(1)$=1.00, p=0.32
Med school debt influenced specialty choice	6 (13.04%)	35 (30.97%)	N=119, $\chi^2(1)$=2.06, p=0.15

Table 3.3. Descriptive statistics of women Comparison of descriptive statistics of students who matched into internal medicine, pediatrics, or family medicine specialties. Likert scores taken from M1 survey.

Characteristic	Matched Into Primary Care Specialty, Women. Mean (SD) or N (%).	Did Not Match Into Primary Care Specialty, Women. Mean (SD) or N (%).	Test of differences
Count	77 (46.95%)	87 (53.05%)	-
Avg. age at matriculation	23.05 (1.95)	22.75 (1.29)	N=135, 2-sided t-test, t(133)=-1.05, p=0.30
Ethnicity is white	58 (75.32%)	69 (79.31%)	N=164, $\chi^2(1)$=0.37, p=0.54
Physicians in family	16 (22.86%)	12 (15.58%)	N=147, $\chi^2(1)$=1.26, p=0.26
Undergrad science based major	55 (90.16%)	64 (88.89%)	N=133 $\chi^2(1)$=0.06, p=0.81
Debt from pre-med education	26 (34.21%)	29 (35.80%)	N=157, $\chi^2(1)$=0.04, p=0.83
Pre-med research	26 (37.14%)	46 (59.74%)	N=147, $\chi^2(1)$=7.49, p¡0.01***
Pre-med mentor	27 (38.57%)	31 (40.26%)	N=147, χ^2 (1)=0.04, p=0.83
Ever married	11 (14.29%)	15 (17.24%)	N=164, $\chi^2(1)$=0.27, p=0.61
Mentor in primary care in 1st 2 years	23 (29.87%)	18 (20.69%)	N=164, $\chi^2(1)$=1.84, p=0.18
Research in medical school	45 (58.44%)	71 (81.61%)	N=164, $\chi^2(1)$=10.59, p¡0.01***
Amount of time in patient contact (Likert scale)	2.09 (1.64)	2.49 (1.53)	N=133, 2-sided t-test, t(131) = 1.41, p=0.16
Quality of life (Likert scale)	2.16 (1.73)	2.50 (1.60)	N=133, 2-sided t-test, t(131) = 1.15, p=0.25
Technical skills necessary (Likert scale)	2.13 (1.02)	1.96 (0.91)	N=133, 2-sided t-test, t(131) = -1.03, p=0.31
Med school debt at match	48 (87.27%)	56 (83.58%)	N=122, $\chi^2(1)$=0.33, p=0.57

Table 3.3. Descriptive statistics of women, continued Comparison of descriptive statistics of students who matched into internal medicine, pediatrics, or family medicine specialties. Likert scores taken from M1 survey.

Characteristic	Matched Into Primary Care Specialty, Women. Mean (SD) or N (%).	Did Not Match Into Primary Care Specialty, Women. Mean (SD) or N (%).	Test of differences
Med school debt influenced specialty choice	3 (5.66%)	9 (13.64%)	N=159, $\chi^2(1)$=5.49, p=0.02

Table 3.4. Descriptive statistics comparing women and men Comparison of descriptive statistics of men and women. Likert responses taken from M1 survey.

Characteristic	Men	Women	Test of differences
Count	210	164	-
Matched into primary care specialty	60 (28.57%)	77 (46.95%)	N=374, $\chi^2(1)$=13.40, p¡0.01***
Avg. age at matriculation	23.90 (2.68), N=172	22.89 (1.63), N=135	N=307, 2-sided t-test, t(305) = 3.86, p¡0.01***
Ethnicity is white	168 (80.00%)	127 (77.44%)	N=374, χ^2 (1)=0.36, p=0.55
Physicians in family	43 (23.24%)	28 (19.05%)	N=332, $\chi^2(1)$=0.66, p=0.35
Undergrad science based major	149 (87.65%)	119 (89.47%)	N=303, $\chi^2(1)$=0.24, p=0.62
Debt from pre-med education	82 (41.62%)	55 (35.03%)	N=354, $\chi^2(1)$=1.60, p=0.21
Pre-med research	97 (52.43%)	72 (48.98%)	N=332, $\chi^2(1)$=0.39, p=0.53
Pre-med mentor	69 (37.30%)	58 (39.46%)	N=332, $\chi^2(1)$=0.16, p=0.69
Ever married	55 (26.19%)	26 (15.85%)	N=374, $\chi^2(1)$=5.80, p=0.02**
Mentor in primary care in 1st 2 years	22 (10.48%)	41 (25.00%)	N=374, $\chi^2(1)$=13.87, p¡0.01***
Research in medical school	180 (85.71%)	116 (70.73%)	N=374, $\chi^2(1)$=12.52, p¡0.01***
Amount of time in patient contact (Likert scale)	2.42 (1.43)	2.31 (1.59)	N=303, 2-sided t-test, t(301) = 0.63, p=0.53
Quality of life (Likert scale)	2.55 (1.57)	2.35 (1.69)	N=303, 2-sided t-test, t(301) = 1.07, p=0.28
Technical skills necessary (Likert scale)	2.17 (1.07)	2.04 (0.96)	N=303, 2-sided t-test, t(301) = 0.48, p=0.63
Med school debt at match	136 (82.93%)	104 (85.25%)	N=286, $\chi^2(1)$=0.28, p=0.60
Med school debt influenced specialty choice	41 (25.79%)	12 (10.08%)	N=278, $\chi^2(1)$=10.88, $p < 0.01$

Table 3.5. M1-2 factors associated with matching into a primary care specialty Dependent variable is matching into a primary care specialty. Primary care is comprised of internal medicine, pediatrics, and family medicine specialties. Adjusted odds ratios are presented, with standard errors in parentheses. Lifestyle and debt responses taken from M2 survey. Pseudo R2 is McFadden's. $***p < 0.01, **p < 0.05, *p < 0.1$ for odds ratio different from 1.

	Both Genders	Men Only	WomenOnly	Both Genders	Men Only	Women Only
Female	2.21*** (0.62)	-	-	2.06** (0.64)	-	-
Ethnicity is White	0.58 (0.20)	0.42* (0.22)	0.63 (0.30)	0.54* (0.19)	0.54 (0.29)	0.34* (0.20)
Age at Matriculation	1.12 (0.08)	1.00 (0.10)	1.43* (0.26)	1.14 (0.09)	0.96 (0.11)	1.57** (0.30)
Family Member Practices Primary Care	3.41** (1.78)	2.13 (1.47)	6.67** (6.16)	3.53** (2.00)	2.30 (1.86)	17.16*** (18.86)
Married in 1st 2 Years	1.70 (0.65)	3.36** (1.79)	1.03 (0.60)	1.92 (0.78)	3.15** (1.83)	1.58 (1.16)
Has children in 1st 2 Years	0.09** (0.09)	0.12* (0.15)	0.05 (0.09)	0.08** (0.08)	0.13 (0.18)	0.02 (0.05)
Mentor in Primary Care in 1st 2 Years	1.11 (0.38)	0.62 (0.37)	1.48 (0.66)	1.35 (0.50)	0.81 (0.57)	1.55 (0.83)
Research in Primary Care in 1st 2 years	1.44 (0.52)	2.64* (1.40)	1.04 (0.55)	1.15 (0.44)	2.07 (1.23)	0.79 (0.47)
Academic vs private practice opportunities	-	-	-	0.95 (0.11)	0.97 (0.18)	0.98 (0.22)
Amount of time in patient contact	-	-	-	0.81** (0.09)	0.81 (0.13)	0.87 (0.16)
Intellectual stimulation	-	-	-	1.06 (0.10)	0.88 (0.14)	1.19 (0.20)

Table 3.5. M1-2 factors associated with matching into a primary care specialty, continued
Dependent variable is matching into a primary care specialty. Primary care is comprised of internal medicine, pediatrics, and family medicine specialties. Adjusted odds ratios are presented, with standard errors in parentheses. Lifestyle and debt responses taken from M2 survey. Pseudo R2 is McFadden's. $***p < 0.01, **p < 0.05, *p < 0.1$ for odds ratio different from 1.

	Both Genders	Men Only	WomenOnly	Both Genders	Men Only	Women Only
Potential Salary	-	-	-	0.78* (0.10)	0.66** (0.13)	0.95 (0.24)
Quality of Life	-	-	-	1.05 (0.11)	1.17 (0.18)	0.96 (0.18)
Responsibilities at home	-	-	-	1.15 (0.14)	1.33 (0.27)	1.03 (0.18)
Specialty status/reputation	-	-	-	1.21 (0.14)	1.46** (0.28)	1.03 (0.18)
Spouse/partner's career	-	-	-	1.33** (0.17)	1.13 (0.23)	1.99*** (0.49)
Technical skills necessary	-	-	-	1.06 (0.13)	0.87 (0.15)	1.46 (0.34)
Debt from Medical Education	1.55 (0.66)	1.64 (1.03)	1.38 (0.85)	1.67 (0.76)	2.49 (1.77)	1.87 (1.31)
Debt Influences Specialty Preference	0.99 (0.27)	0.61 (0.26)	1.49 (0.57)	0.91 (0.26)	0.58 (0.27)	1.71 (0.79)
Constant	0.03** (0.05)	0.48 (1.18)	0.00** (0.00)	0.00** (0.01)	0.62 (1.96)	0.00*** (0.00)
Observations	273	153	120	264	148	116
Pseudo R2	0.08	0.09	0.08	0.12	0.17	0.20

Table 3.6. Post-match factors associated with matching into a primary care specialty Dependent variable is matching into a primary care specialty. Primary care is comprised of internal medicine, pediatrics, and family medicine specialties. Adjusted odds ratios are presented, with standard errors in parentheses. Lifestyle and debt responses taken from match survey. Pseudo R2 is McFadden's. $***p < 0.01, **p < 0.05, *p < 0.1$ for odds ratio different from 1.

	Both Genders	Men Only	Women Only	Both Genders	Men Only	Women Only
Female	1.76* (0.59)	-	-	1.61 (0.62)	-	-
Ethnicity is White	0.66 (0.27)	0.65 (0.39)	0.59 (0.35)	0.57 (0.27)	0.57 (0.41)	0.51 (0.38)
Age at Matriculation	1.18** (0.10)	1.03 (0.12)	1.46** (0.25)	1.11 (0.11)	0.88 (0.15)	1.59** (0.36)
Family Member Practices Primary Care	4.05** (2.83)	2.62 (2.41)	7.22 (8.94)	5.45** (4.34)	5.71 (6.87)	8.18 (10.93)
Married in Medical School	1.10 (0.52)	2.08 (1.41)	0.93 (0.65)	1.26 (0.69)	4.03 (3.52)	0.85 (0.78)
Had Children in Medical School	0.22* (0.20)	0.30 (0.34)	0.26 (0.46)	0.22 (0.23)	0.15 (0.20)	1.62 (3.23)
Primary Care Mentor in 1st 2 Years	1.45 (0.59)	0.67 (0.50)	2.23 (1.17)	1.30 (0.59)	0.71 (0.64)	2.4 (1.46)
Conducted Primary Care Research in Medical School	0.83 (0.29)	1.54 (0.78)	0.54 (0.28)	0.93 (0.36)	2.61 (1.71)	0.45 (0.29)
Academic vs private practice opportunities	-	-	-	0.85 (0.15)	0.78 (0.21)	0.93 (0.29)

Table 3.6. Post-match factors associated with matching into a primary care specialty, continued Dependent variable is matching into a primary care specialty. Primary care is comprised of internal medicine, pediatrics, and family medicine specialties. Adjusted odds ratios are presented, with standard errors in parentheses. Lifestyle and debt responses taken from match survey. Pseudo R2 is McFadden's. $***p < 0.01, **p < 0.05, *p < 0.1$ for odds ratio different from 1.

	Both Genders	Men Only	Women Only	Both Genders	Men Only	Women Only
Amount of time in patient contact	-	-	-	0.99 (0.16)	0.83 (0.21)	1.09 (0.29)
Intellectual stimulation	-	-	-	1.28 (0.27)	1.67 (0.59)	1.57 (0.49)
Potential Salary	-	-	-	0.73 (0.15)	0.50** (0.16)	0.95 (0.34)
Quality of Life	-	-	-	0.65** (0.12)	0.55* (0.18)	0.57* (0.18)
Responsibilities at home	-	-	-	1.44* (0.29)	1.55 (0.55)	1.47 (0.46)
Specialty status/ reputation	-	-	-	1.37 (0.27)	1.50 (0.46)	1.50 (0.52)
Spouse/partner's career	-	-	-	1.29 (0.21)	1.75** (0.50)	1.35 (0.34)
Technical skills necessary	-	-	-	0.54*** (0.09)	0.42*** (0.12)	0.54** (0.15)
Debt from Medical Education	1.53 (0.70)	1.86 (1.20)	1.17 (0.82)	2.02 (1.03)	2.62 (2.10)	1.58 (1.30)
Debt Influences Specialty Preference	0.29** (0.16)	0.24** (0.17)	0.39 (0.34)	0.33* (0.20)	0.34(0.29)	0.32 (0.34)
Constant	0.01** (0.02)	0.15 (0.42)	0.00** (0.00)	0.18 (0.49)	97.07 (418.00)	0.00** (0.00)
Observations	197	111	86	189	108	81
Pseudo R2	0.08	0.08	0.11	0.19	0.29	0.21

Table 3.7. Pre-matriculation factors associated with matching into a primary care specialty Adjusted odds ratios, standard errors in parentheses. Primary care is comprised of internal medicine, pediatrics, and family medicine specialties. Pseudo R2 is McFadden's. $***p < 0.01, **p < 0.05, *p < 0.1$ for odds ratio different from 1.

	Pooled	Men	Women
Female	2.126*** (0.545)	-	-
Ethnicity is white	0.664 (0.200)	0.683 (0.302)	0.671 (0.286)
Age at matriculation	0.994 (0.0586)	0.925 (0.0705)	1.220 (0.150)
Science major in undergraduate education	0.842 (0.335)	0.571 (0.306)	1.581 (1.028)
Debt from undergraduate education	0.936 (0.244)	1.099 (0.399)	0.762 (0.298)
Mentor in primary care	1.637 (0.563)	1.764 (0.887)	1.810 (0.917)
Performed primary care research	0.893 (0.312)	1.350 (0.686)	0.522 (0.258)
Family member practices primary care	2.859** (1.340)	2.724 (1.777)	.719* (2.706)
Constant	0.645 (0.998)	4.313(8.576)	0.00802 (0.0246)
Observations	302	169	133
Pseudo R2	0.05	0.03	0.05

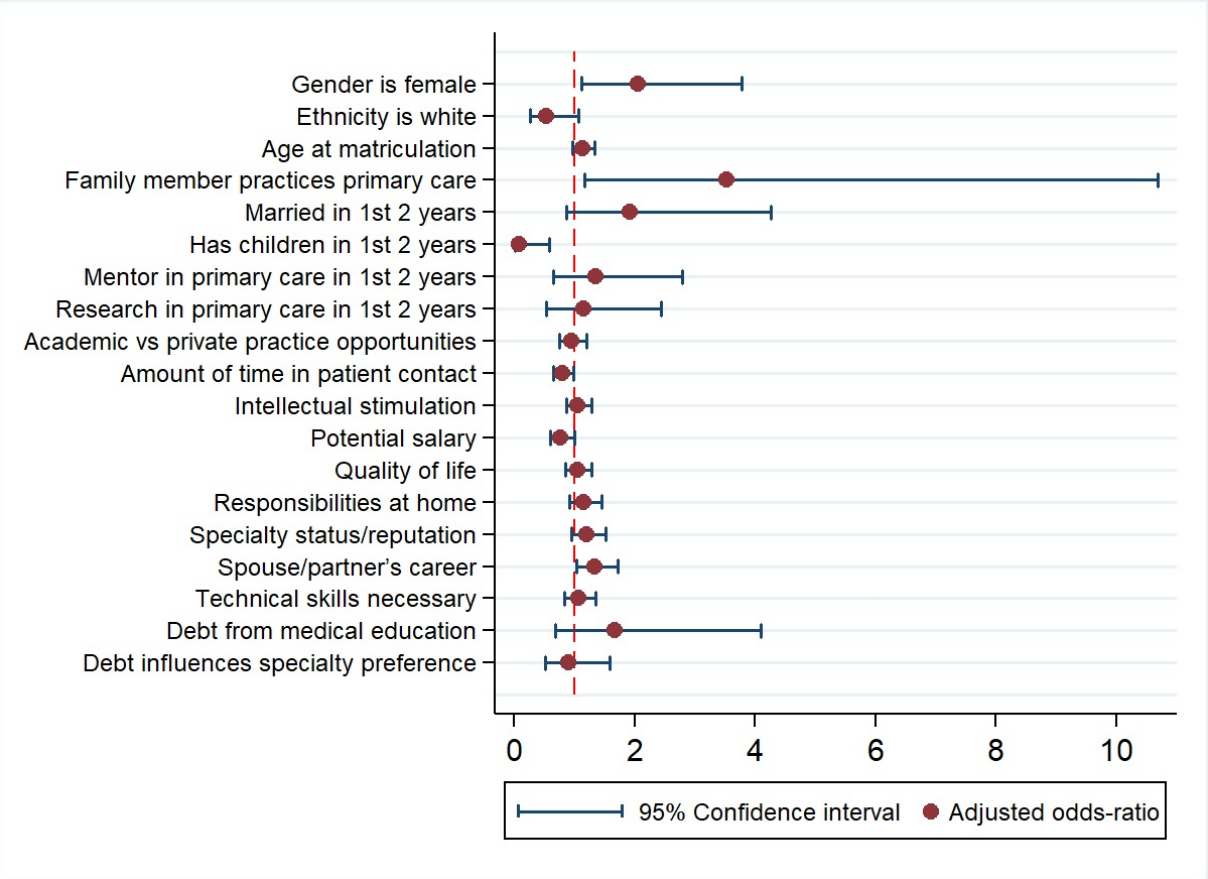

Figure 3.1. M1-2 factors associated with matching into a primary care specialty. Dependent variable is matching into a primary care specialty. Primary care is comprised of internal medicine, pediatrics, and family medicine specialties. Adjusted odds ratios are presented, bars representing 95% confidence interval. Reference line is set at 1, values higher than 1 indicated increased likelihood of entering primary care, values lower than 1 indicated reduced likelihood. Lifestyle and debt responses taken from M2 survey.

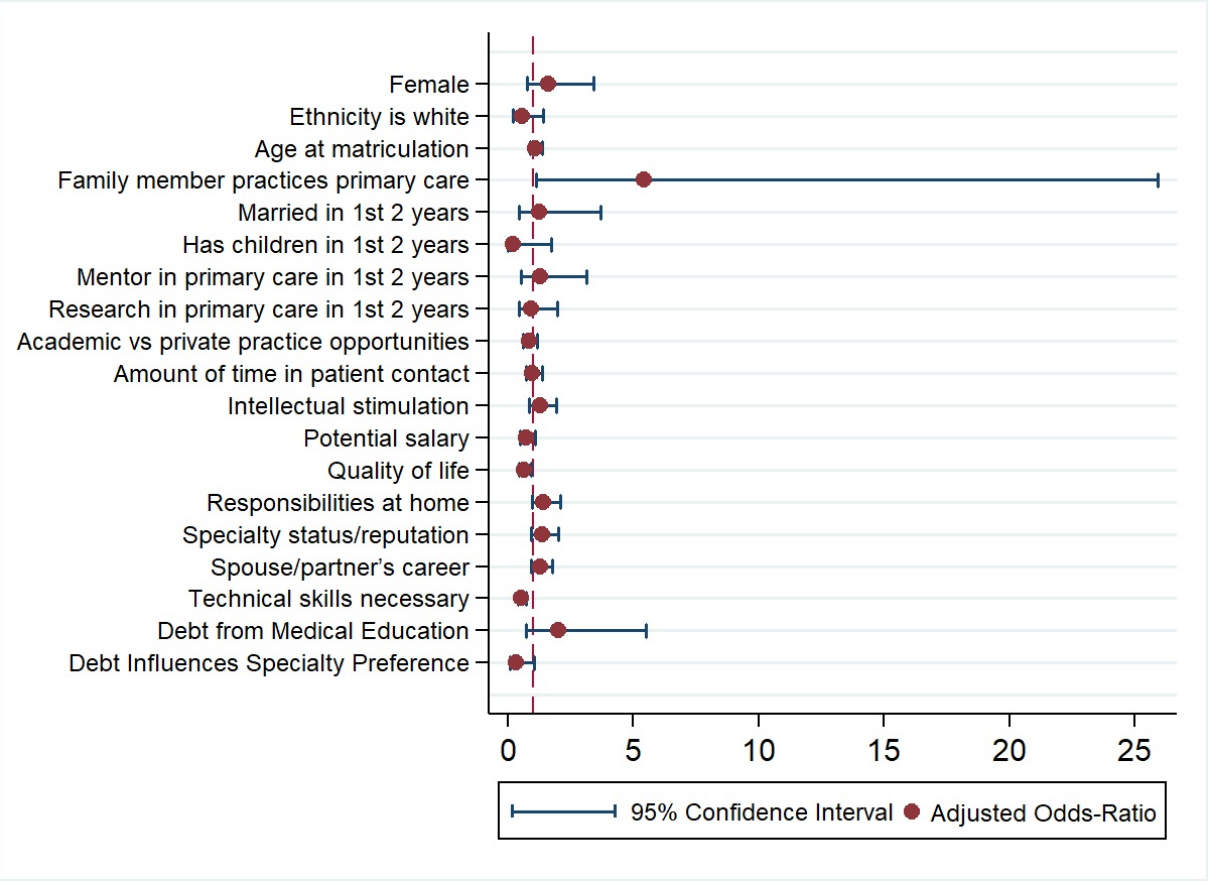

Figure 3.2. Post-match factors associated with matching into a primary care specialty. Dependent variable is matching into a primary care specialty. Primary care is comprised of internal medicine, pediatrics, and family medicine specialties. Adjusted odds ratios are presented, bars representing 95% confidence interval. Reference line is set at 1, values higher than 1 indicated increased likelihood of entering primary care, values lower than 1 indicated reduced likelihood. Lifestyle and debt responses taken from M2 survey.

Chapter 4

Testing the influence of testosterone administration on men's honesty in a large laboratory experiment

4.1 Abstract

The impact of testosterone on decision-making is a growing literature, with several reports of economically relevant outcomes. Similar to Wibral et al. (2012), we investigate the effects of exogenous testosterone administration on deception in a double-blind placebo controlled study. Participants ($N = 242$) were asked to roll a die in private and were paid according to their reported roll, which creates the opportunity to lie about the outcome to increase earnings. We find evidence for self-serving lying in both treatment and control groups and a statistically insignificant negative effect ($d = -0.17$, 95% CI[-0.42, 0.08]) indicating more honest behavior (i.e., lower reports) following testosterone administration. Although insignificant, the direction was the same as in the Wibral et al. study, and the meta-analytic effect of the two studies demonstrates lower reporting (i.e., more honesty) following testosterone (vs. placebo) administration, significant at the 0.05 level ($d = -0.27$, 95% CI[-0.49, -0.06]). We discuss how our results and methodology compare with Wibral et al. and identify potential causes for differences in findings. Finally, we consider several plausible connections between testosterone and lying that may be further investigated using alternative methodologies.

Introduction

Lying plays an important role in interpersonal relationships and many types of economic transactions, as it can create strategic advantages from informational asymmetries. Investigations of the determinants of lying have recently attracted widespread attention, and include research of the roles played by other-regarding preferences (Gneezy, 2005), social and cultural norms (Gáchter and Schulz, 2016; Mann et al., 2016), the size and nature of incentives (Fischbacher and Franziska, 2013; Kajackaite and Gneezy, 2017; Charness, Masclet, and Villeval 2014), the likelihood and costs of detection (Becker, 1968), performance in an antecedent competition (Schurr and Ritov, 2016), the opportunity for self-justification or self-signalling (Shalvi, Handgraaf, and De Dreu, 2011; Shalvi, Eldar, and Bereby-Meyer, 2012; Mazar, Amir, and Ariely, 2008), and the role of individual differences, and gender in particular (Dreber and Johannesson, 2008; Muehlheusser, Roiser, and Wallmeier, 2015).

Deception is a part of the behavioral repertoire of many animal species (Trivers, 2000; Bugnyar and Kotrschal, 2002; Whien and Byrne, 1988). The understanding of the biological foundations of deceptive behavior or lying in humans, however, is limited. Functional Magnetic Resonance Imaging (fMRI) studies suggest that deception is associated with increased activation in brain regions involved in socio-cognitive processes, such as the right tempero-parietal junction, precuneus and anterior frontal gyrus, and executive functions, such as the anterior cingulate cortex and amygdala (Volz, Vogeley, Tittgemeyer, von Cramon and Sutter, 2015; Lisofsky, Kazzer, Heekeren, and Prehn, 2014; Garrett, Lazzaro, Ariely, and Sharot, 2016). In addition, two studies reported that intranasal administration of the neuropeptide oxytocin promotes group-serving dishonesty (Shalvi and De Dreu, 2014)and decreases the ability to detect lies told by members of the opposite sex (Pfundmair, Erk, and Reinelt, 2017). However, it should be noted that several methodological reviews have recently challenged the validity of the intranasal oxytocin literature, casting uncertainty over these findings (Lane, Luminet, Nage, and Mikolajczak, 2016; Walum, Waldman, and Young, 2016; Nave, Camerer, McCullough, 2015).

The male sex steroid hormone testosterone plays a central role in physical development, and has been shown to have considerable psychological effects, such as on mood in hypogonadal men (Pope, Kouri, and Hudson, 2000; Wang et al., 2004) and cognition (Gray et al., 2005; Newman, Sellers, and Josephs, 2005; Janowsky, Oviatt, and Orwoll, 1994; Nave, Nadler, Zava, and Camerer, 2017). There are also several reports documenting the hormone's impact on decision making in a variety of economically important contexts, such as financial risk taking (Apicella, Carré, and Dreber, 2015), asset trading (Nalder, Alexander, Johnson, and Zak, forthcoming; Coates and Herbert, 2008; Cueva et al., 2015), and economic games assessing trust, reciprocity, and cooperation (Boksem et al., 2013; Van Honk, Montoya, Bos, van Vugt, and Terburg, 2012; Eisenegger, Naef, Snozzi, Heinrichs, and Fehr, 2010).

While much of testosterone behavioral research has focused on antisocial behaviors, such as aggression (Hermans, Ramsey, and van Honk, 2008; Carré, Ruddick, Moreau, and Bird, 2017), testosterone has also been shown to promote prosocial behavior in certain contexts, such as increasing fair bargaining behavior (Eisenegger, Haushofer, and Fehr, 2011). A common explanation for testosterone's promotion of prosocial behavior in some contexts and antisocial behavior in others is that testosterone may increase the desire for social status and thus promotes status seeking behavior (Eisenegger, Naef, Snozzi, Heinrichs, and Fehr, 2010; Dreher et al., 2016; Nave, Nadler, Zava, and Camerer, 2017). Along this line of argumentation, lying is a socially complex behavior that can affect social status. Hence, testosterone may impact lying in ways that increase social status, even at an economic cost. Consistent with this notion, a study of Dutch females who were administered testosterone before playing bluff poker found that the participants who received testosterone were less likely to bluff and more likely to call bluffs (Van Honk et al., 2016). The authors argued that while random bluffing was the payoff maximizing strategy in the game, exhibiting dishonesty was harmful for the player's social status.

In a study closely related to our own, Wibral et al. (2012) investigated the influence of testosterone administration on lying with a die-roll task (originally introduced in Fischbacher & Föllmi-Heusi, 2013) — an active behavioral measure of deception that has also been shown

to predict dishonest behavior in the field (Dai, Galeotti, and Villeval, 2017). In this study, German male participants ($N = 91$) were randomly administered testosterone or placebo under a double-blind exogenous administration protocol and were given monetary incentive to lie without possibility of being discovered. Wibral et al. found that testosterone, in comparison to placebo, significantly reduced deception. The authors speculated that this decrease in lying was caused by testosterone's effects on pride and self-image, two psychological constructs that are related to status concerns but do not require the actualization of status outcomes to impact behavior.

The current study aims to further test the robustness and generalizability of the findings of Wibral et al., because of the following reasons. First and foremost, recent large scale investigations have repeatedly demonstrated the importance of building a robust epistemological foundation that allows science to progress cumulatively (Open, 2015). While an encouraging fraction of laboratory economics experiments can be successfully replicated, a considerable proportion of significant effects either cannot be replicated or the replicated effect size is of a smaller magnitude (Camerer et al., 2016).

It should be noted that this experiment is not a direct replication of the Wibral et al. study. Although the overall experimental design is similar, several methodological modifications (detailed later in Differences from Wibral et al. (2012)) were purposefully made to increase our likelihood of detecting behavioral effects from testosterone administration. Such iterative methodological changes, along with employing ample sample sizes, are important for testing for the robustness and generalizability of the effect.

Second, although the effect reported by Wibral et al. is seemingly strong and with a relatively small p-value (2-sided t-test, $t(89) = 2.65$, $p < 0.001$), the die-roll task was a part of an experimental battery comprising of 11 tasks - a common research practice in behavioral endocrinological research that is aimed at maximizing the knowledge gained from each participant undergoing a pharmacological treatment (Zethraeus et al., 2009). The Bonferroni corrected p-value is $p = 0.11$, which means that the statistical evidence were not overwhelming.

Finally, while the sample size ($N = 91$) was larger than previous testosterone administration studies, the number of participants who faced an opportunity to lie was effectively smaller, due to the random nature of the task. This is because participants whose die roll outcome is high face no incentive to misreport.

To this end, we conducted a double-blind placebo-controlled investigation of exogenous testosterone's effect on the die-roll task, in a sample of $N = 242$ American participants (mostly college students, for full demographic details see Supplementary Materials S1). Our sample size is over 2.5 times larger than the Wibral et al. (2012) study - a magnitudinal difference that is in line with the "small telescope" heuristic which provides the statistical power to test whether the original study was underpowered to detect the reported effect size (Simonsohn, 2015). As in the Wibral et al. study, participants were seated privately in cubicles without the possibility of observation by researchers or other participants. They were each given a 6-sided die, a pen, and a slip of paper, and instructed via computer to roll the die and report the outcome both on the slip of paper and into the computer, thus earning them the dollar amount of what they reported. As in the original study, the task was a part of an experimental battery. Given the Wibral et al. report, we hypothesized that participants who received testosterone would be more honest (i.e., report lower outcomes compared to placebo).

Methods

Participants

Males over the age of 18 (mean = 23.65, SD = 7.24), mostly college students, were recruited via e-mail and posters to participate in an experiment on testosterone and economic decision making at the Center for Neuroeconomics Studies, Claremont Graduate University. 125 participants were administered testosterone gel and 118 were administered a placebo gel. One participant who was administered the placebo gel left the experiment before participating in the die-roll task and was therefore excluded from all analyses, bringing the total number

of participants used in analysis to $N = 242$. The institutional review boards of Caltech and Claremont Graduate University approved this study, all participants gave informed consent, and no adverse events occurred. The study was performed in accordance with the guidelines set forth by both IRBs. Descriptive statistics of the participants are presented in Supplementary Materials Table S1.

Procedure

In each experimental day there were two sessions, with one in the morning and one in the afternoon. The morning session lasted from 9:00am to 9:45am, and the afternoon session lasted from 2:00pm until roughly 4:15pm. Participants provided 4 saliva samples, one in the morning, and three in the afternoon. Participants completed the die roll task immediately before the 4th saliva sampling, which took place on average at 4:17pm (SD = 12.2 minutes), and were dismissed shortly thereafter.

Chronologically, participants arrived at the laboratory in the morning in groups of 12 or 16, whereupon they were given an informed consent form and signed it upon assent. They then proceeded to a separate room where their hands were scanned (digit ratio is a purported measure of pre-natal testosterone exposure (Lutchmaya et al., 2004) and facial photographs were taken (facial characteristics are associated with testosterone levels (Penton-Voak and Chen, 2004)). Next, they went to another room where they completed brief demographic and mood surveys in randomly assigned private cubicles. The private cubicles had a desk, computer, keyboard, monitor, and mouse, and all activity on the computer or desk was out of sight of any other participant or researcher. A saliva sample was taken at the cubicles to assess baseline testosterone levels, the first of a total of 4 samples taken for each participant (the three others were taken during the afternoon session).

Participants then proceeded to another separate room in groups of 2-6 where they were given a small paper cup containing either 10g of topical testosterone 1% (2 x 50 mg packets Vogelxo® by Upsher-Smith) or volume equivalent of an inert placebo of similar texture and

viscosity (80% alcogel, 20% Versagel®) under a double-blind protocol (the paper cups were filled by the lab manager, who did not interact with the participants or reveals its contents to the research assistants). Participants were instructed to remove their shirts and self-apply the entirety of the cup's contents to their shoulders, upper arms, and chest, as demonstrated by a research assistant. Participants were also instructed to not put their shirts back on until the gel had fully dried. Following application of the gel, all participants were asked to avoid touching any part of their body before washing their hands, and then brought into an adjacent restroom in order to thoroughly wash their hands with warm water and soap.

Participants were then given a strict set of instructions (which were also in the informed consent and recruitment materials), both verbally to the group and on a printed hand-out given to each participant, of what to do preceding the afternoon session and for the next 23 hours. Participants were told to refrain from bathing or any activities that might cause excessive perspiration, not to eat after 1:00pm (in order to produce high quality saliva samples), and to return to the lab by 1:55pm. Participants were then dismissed from the laboratory until the afternoon session. They were also told to abstain from any skin-to-skin contact with females, as per the recommendations of the testosterone gel manufacturer. A researcher contacted each participant via text message shortly before 1pm to remind them to not eat any more and that they were only allowed to consume water before the afternoon session. Upon return for the afternoon session, a researcher verbally confirmed with each participant whether they had adhered to the guidelines, and no participants admitted noncompliance. Participants were not allowed to drink water for the 10 minutes preceding a saliva sample, which was enforced via observation by researchers. Saliva samples were also inspected for abnormalities, e.g. whether it was dark from smoking or oral bleeding, and any such were marked in an experimental log for monitoring and potential exclusion.

For the afternoon session, participants returned to the same private cubicle they had used in the morning session. They then provided a second saliva sample at 2:05pm. In each cubicle there were also a standard 6-sided die, slip of paper, and pen, which were used in the

die-roll task. Upon arrivals (with no incidents of lateness) participants took part in a battery of seven behavioral tasks that included math-based competitions, risk preference questionnaires, the cognitive reflection test, and others as part of another experiment. The third saliva sample was taken at 3:15pm, and the fourth sample was taken at 4:15pm, with the die task immediately preceding the fourth sample. Once the researcher had collected the reported rolls from all participants, the participants were paid in cash for their earnings in the study, and then provided their final saliva samples.

Die-roll task

Through Qualtrics, an online survey platform, participants were given instructions to roll the die on their desk, and both record the result of the die roll into the survey and onto the slip of paper (see Supplementary Materials S2 for full instructions). The instructions informed participants that they would receive the dollar value of their reported roll - a report of 3 would earn $3, a report of 5 would earn $5, *et cetera*. The instructions stated that participants could roll the die more than once, but that only their first roll would count. Once participants had recorded their roll, they brought their slips to the research assistant, who was standing at the far end of the room, on the other side of the cubicle walls that surrounded each participant. This ensured that the roll and recording of the roll outcome were both done privately.

Differences from Wibral et al. (2012)

In this section, we consider the *a priori* differences between our study design and that of Wibral et al., and how these differences may impact results. These differences are summarized in Table 1.

First, our research differs from that of Wibral et al. in the loading period of testosterone. Our choice of testing schedule was based on the recommendations of a report by Eisenegger et al. (2013) which studies the pharmacokinetics of testosterone in healthy young men. The study documented clear elevation in testosterone levels between 3 and 7 hours after topical

Table 4.1. Methodological differences between this study and Wibral et al. (2012)

	Wibral et al. (2012)	**This study**
1. Testosterone Loading Time	21-24 hours	Approximately 7 hours
2. Administration method & Dosage	Topical gel application: 1 packet of Testogel®, which contains 5 grams of total gel and 50mg testosterone (1% concentration).	Topical gel application: 2 packets of Vogelxo® Gel 1%, each of which contains 5 grams of total gel and 50mg testosterone (1% concentration), for a total of 100mg of testosterone. The two packets were opened and precombined by the lab manager into a disposable cup before experimental sessions in order to preserve the double-blind for participants and researchers who interacted with participants.
3. System of Payoffs	Subjects receive Euro amount for reporting 1-5, but 0 for reporting a 6.	Subjects receive dollar amount for reporting 1-6.
4. Antecedent task (see SOM S6 for full list of tasks in battery for each experiment)	"Devil's Task," a risk preference task in which participants make a series of increasingly risky choices.	A risk task where participants were ranked by their performance.
5. Method of Measurement	Blood serum	Saliva
6. Subject Pool	German	American

administration. The Eisenegger et al. report explicitly recommended testing for behavioral effects 7 hours after administration, and noted that peak testosterone levels were at 3 hours after administration. The findings of the Eisenegger et al. report are qualitatively similar to those of Chik et al. (2006), who also find that a transdermal application of testosterone (of lower dose than Eisenegger et al. or our study) in healthy young men led to peak serum testosterone levels roughly four hours after administration. In the Wibral et al. study, the die roll task took place about 21-24 hours after administration (thus between 18 and 21 hours after peak testosterone levels), whereas in our study it took place roughly 7 hours after testosterone administration, as

suggested by Eisenegger et al. (2013).

One reason to be concerned that this methodological difference might cause the attenuation of the behavioral effect is due to lower treatment potency. Because our study used saliva sampling and Wibral et al. used blood sampling, we cannot directly compare measurements of testosterone levels. However, we confirmed a significant elevation in testosterone levels in our experiment, and this elevation did lead to behaviorally significant impacts in other tasks (Nave, Nadler, Zava, and Camerer, 2017; Nave, Nadler, Zava, and Camerer; forthcoming). Relatedly, another study found that testosterone administration significantly increased aggression in some participants after only an hour (Carré, Ruddick, Moreau, and Bird, 2017). Given the pharmacokinetics of testosterone, it is likely that this administration schedule led to higher testosterone levels in our study than in Wibral et al. (2012), and thus we would expect to see greater treatment potency (though we acknowledge that non-linear dose-dependency cannot be entirely ruled out). It should be noted, however, that the Eisenegger et al. study stopped sampling saliva after 7 hours, and more information is needed on the pharmacokinetics of testosterone over longer time periods as in Wibral et al. (2012).

Second, we differ in the amount of testosterone administered to participants. Whereas Wibral et al. (2012) used 50mg, and the Eisenegger et al. (2013) study used 150mg of topical testosterone gel, we decided to use 100mg of testosterone gel. Our reasoning for using a larger dosage than Wibral et al. is that we wanted to increase the potency of our treatment in order to increase our probability of detecting the behavioral effects of testosterone. However, we did not increase the dose up to 150mg, as in Eisenegger et al. (2013), in order to maintain ecological validity: 50mg and 100mg, but not 150mg, are typical dosages indicated by prescription guidelines provided by the manufacturer of Vogelxo®. (the maximum recommended dose is 100mg, with the advice to begin all patients at 50mg daily for 14 days and adjust the dose upwards if serum testosterone levels are measured to still be below the normal range). The Eisenegger et al. report also notes that the pharmacokinetic data found in their study are qualitatively similar to those found by Chik et al. (2006), who studied the effects of 50mg of testosterone in healthy young

men. This suggests that the pharmacokinetics of the intermediate dose of 100mg are likely to be similar as well.

Third, in the Wibral et al. (2012) study participants were paid the monetary value of their reported rolls of 1-5, but paid 0 for a reported roll of 6, where our study used a simpler payment scheme, with payoffs matching the reported roll. Although the salient decision individuals faced in either methodological design is essentially the same - whether to misreport a private die roll in order to increase earnings - this change does modify, to some degree, the stakes of the game - as in our study the worst a participant can do is earn $1, as opposed to nothing, and therefore so his incentive to lie may be reduced. However, as a meta-analysis of the die-roll task found no differences in reporting even when the differences in stakes are 500 times larger (Abeler, Nosenzo, and Raymond, 2016), we find it unlikely that this difference made a substantial impact.

Fourth, in both studies the die-roll task was a part of an experimental battery (which is common practice in pharmacological experiments), but the batteries consisted of different behavioral tasks. In our experiment, the die roll was the last behavioral task, and it took place immediately following a task where participants made a series of either risky or safe bets, and then were publicly ranked and identified as "winners" and "losers" according to whether they were in the top or bottom half of earners. Previous research has shown that participants who win a competition tend to lie more afterwards in a die-roll game where the reported roll of one participant is subtracted from a shared amount to be split with another participant (Schurr and Ritov, 2016). This study differs from our own in that reported rolls in our study did not impact the earnings of other participants. We test for any effect of winning or losing in the risk task, as well as an interaction with treatment, in our Results section. The antecedent task in Wibral et al. (2012) was the Devil's Task, a risk preference measure in which participants either take or reject a series of gambles, wherein winning the gamble adds to a cumulative payoff or losing the gamble eliminates the entire payoff (Slovic, 1966).

Fifth, in our study we used saliva to measure testosterone levels, compared to blood draws in Wibral et al. The advantage of using a saliva test is that it is operationally simpler as it does

not require a blood draw. Relevant to the behavioral differences between the studies, blood draws cause some amount of pain and stress as compared to a saliva draw. This pain and stress could lead to an increase in cortisol levels, and the interaction between testosterone to cortisol might be important for deceptive behavior, as it is for aggressive behavior (Montoya, Terburg, Bos, and van Honk, 2012). However, we did not find evidence that cortisol moderated the relationship between treatment and reported die roll in our study (OLS, coefficient for interaction between treatment and log cortisol levels $\beta = 0.22$, 95% CI[-0.43, 0.86], $p = 0.51$, see Supplementary Materials S6).

Last, participants in Wibral et al. were German and our study participants were American. This difference may have non-trivial consequences, as culture may influence perceptions of social status and actions which will lead to its elevation (DiMaggio, 1982). It may be the case that money is relatively more important for social status in America than Germany, and thus the same increased drive for social status will produce relatively less honesty among Americans. Propensities for honest behavior indeed vary substantially between cultures. For instance, one study found that in a task where participants were instructed to anonymously report the result of a coin flip (with a material incentive to misreport) only 3.4% of British participants lied, compared to 70% of Chinese participants (Hugh-Jones, 2015). While we are not aware of any studies that directly compare German and American behavior using similar methodologies, it is worth investigating if the impact of testosterone on behavior may differ according to cultural context in line with broader mechanisms associated with testosterone (e.g., status seeking).

Measures

Saliva Sampling

A total of 4 saliva samples were taken throughout the experimental day, the 1st occurring before treatment administration between 9:25 and 9:34 am, the 2nd upon return to the lab for the afternoon session between 1:55 and 2:15 pm, the 3rd in the middle of the behavioral tasks battery between 3:02 and 3:38pm, and the 4th at the very end between 4:10 and 4:44 pm. Participants

were not allowed to bring food or drink into the laboratory, and the only water break allowed was immediately following the 3rd saliva sample, which occurred an hour before the 4th sample.

Hormonal Assays

Saliva samples were immediately stored on dry ice in coolers after collection and shipped to ZRT Laboratories (Beaverton, OR) for assay. Salivary steroids (estrone, estradiol, estriol, testosterone, androstenedione, DHEA, 5-alpha DHT, progesterone, 17OH-progesterone, 11-deoxycortisol, cortisol, cortisone, and corticosterone) were measured by liquid chromatography tandem mass spectrometry (LC-MS/MS) using an AB Sciex Triple Quad 5500. Further details about the assay procedure are available in the Supplementary Materials. A series of one-sample Kolmogorov-Smirnov tests for conformity to Gaussian (Supplementary Materials Table S2) indicated that all hormonal measurement distributions were better approximated by a Gaussian distribution following a log transformation, as indicated by higher p-values (i.e., the Gaussian normality hypotheses were less likely to be rejected after log-transformations). Thus, all hormonal measurements were log transformed prior to data analysis in order to make their distributions closer to Gaussian. It should be noted that these log transformations only impact our supplementary analysis, which is based primarily on OLS regression and is thus benefited from a normal distribution. Three saliva samples (two from sample 2, and one from sample 3) could not be analyzed due to insufficient fluid and thus excluded from analyses involving these hormonal samples.

After experimental session 13 (of 17) it was discovered that some of the pre-treatment baseline saliva samples from both treatment groups had testosterone measures exceeding those expected in healthy young men (i.e., greater than 400 pg/mL). Crucially, hormonal panel data show normal upstream and downstream testosterone metabolites dihydrotestosterone and androstenedione, respectively, among all participants with these abnormally high samples. Interpreting this singular hormonal abnormality, only the samples themselves were affected, but not participants' physiological levels of testosterone. Following discovery of the viral spread into

samples, a thorough experimental sterilization protocol was enacted, and the number of samples with abnormally high testosterone was drastically reduced. Ultimately, it was deduced that the testosterone gel had been transferred from common surfaces (e.g., door knobs, mouse pads) onto participants' hands, and then into the saliva sampling tubes. Full details of the issue and our response are available in the supporting materials S4. In light of the resulting unreliability of measured testosterone levels, we avoid relying on measured hormones for analysis such as a regression of die rolls on testosterone levels; instead, we use treatment groups in our analysis.

Digit Ratio and Facial Masculinity

Digit ratio was calculated by first measuring the length of the second and fourth digits from the hand scans taken during the morning session, and then dividing the length of the second digit by that of the fourth. Facial masculinity was defined as the facial width-height ratio, calculated by measuring the distance between the cheekbones (width) and the upper eyelid to the top of the upper lip (height) and dividing width by height. Both facial and hand measurements were made using a software tool which counted the number of pixels between two points selected on an image. Two trained research assistants independently made each measurement, and the mean of the two measurements was used. Any discrepancies between the two measurements greater than 5% were reviewed by a senior researcher, of which there was only one instance. Further details on these measures are available in the Supplementary Materials.

Statistical Approach

The first aim of our analysis is to provide a straightforward comparison between our results and those of Wibral et al. To that end, we perform the same set of statistical tests as those reported by Wibral et al., juxtapose their results against our own, and note differences. Our second aim is to make an assessment of the cumulative evidence on the relationship between exogenous testosterone and lying on the die roll task. To do so, we perform a joint analysis using a fixed effects model.

Results

Manipulation Check

We observed elevated levels of T and its metabolites (e.g., dihydrotestosterone) in the saliva measurements of the testosterone group but not in the placebo group following gel administration relative to baseline, and average levels of testosterone were significantly higher in the testosterone group than in the placebo group following gel administration. In order to verify that the participants who had received testosterone gel indeed experienced an elevation in their testosterone levels compared to those who received placebo, we submitted the logged testosterone levels to a repeated-measures ANOVA, that included treatment status as a between-subject factor, measurement time as a within subject factor, and the interaction between the two. The F-ratio of the interaction term was significant at the 0.01 level ($F(3,716) = 311.58$, $p < 0.001$), indicating unequal mean levels of testosterone across sampling points and treatment status. We further tested for differences in logged testosterone levels between the two treatment groups in each of the four time point of saliva sampling, using 2-sided t-tests. Comparing log testosterone levels in the morning baseline sample across treatment groups yielded a non-significant difference ($t(239) = 1.440$, $p = 0.15$). The mean (SD) non-logged testosterone levels in the morning were 480.13 (826.95)pg/mL in the treatment group, and 616.24 (1052.93)pg/mL in the placebo group. However testosterone levels were significantly higher in the treatment group in the second ($t(239) = -18.61$, $p < 0.001$), third ($t(239) = -24.70$, $p < 0.001$) and fourth ($t(239) = -25.80$, $p < 0.001$) saliva sample, providing a robust and successful manipulation check to our pharmacological testosterone treatment. Mean (SD) non-logged testosterone levels 11,342.27 (15,270.73)pg/mL in treatment and 249.00 (274.20)pg/mL in placebo at the second saliva sample, 20,609.34 (20027.17)pg/mL in treatment and 353.36 (570.76)pg/mL in placebo at the third saliva sample, and 9.16 (1.40)pg/mL in treatment and 5.19 (0.92)pg/mL in placebo at the fourth saliva sample. These changes in salivary testosterone levels are in line with other studies which also used topical testosterone and progesterone administration (Mayo, Macintyre,

Wallace and Ahmed, 2004; Du et al., 2013). There were no treatment effects on either mood, treatment expectancy, or levels of all other measured hormones, ruling out these potential indirect treatment influences on the task (see Supplementary Materials Table S4 for further details).

The Influence of Testosterone Administration on Deception

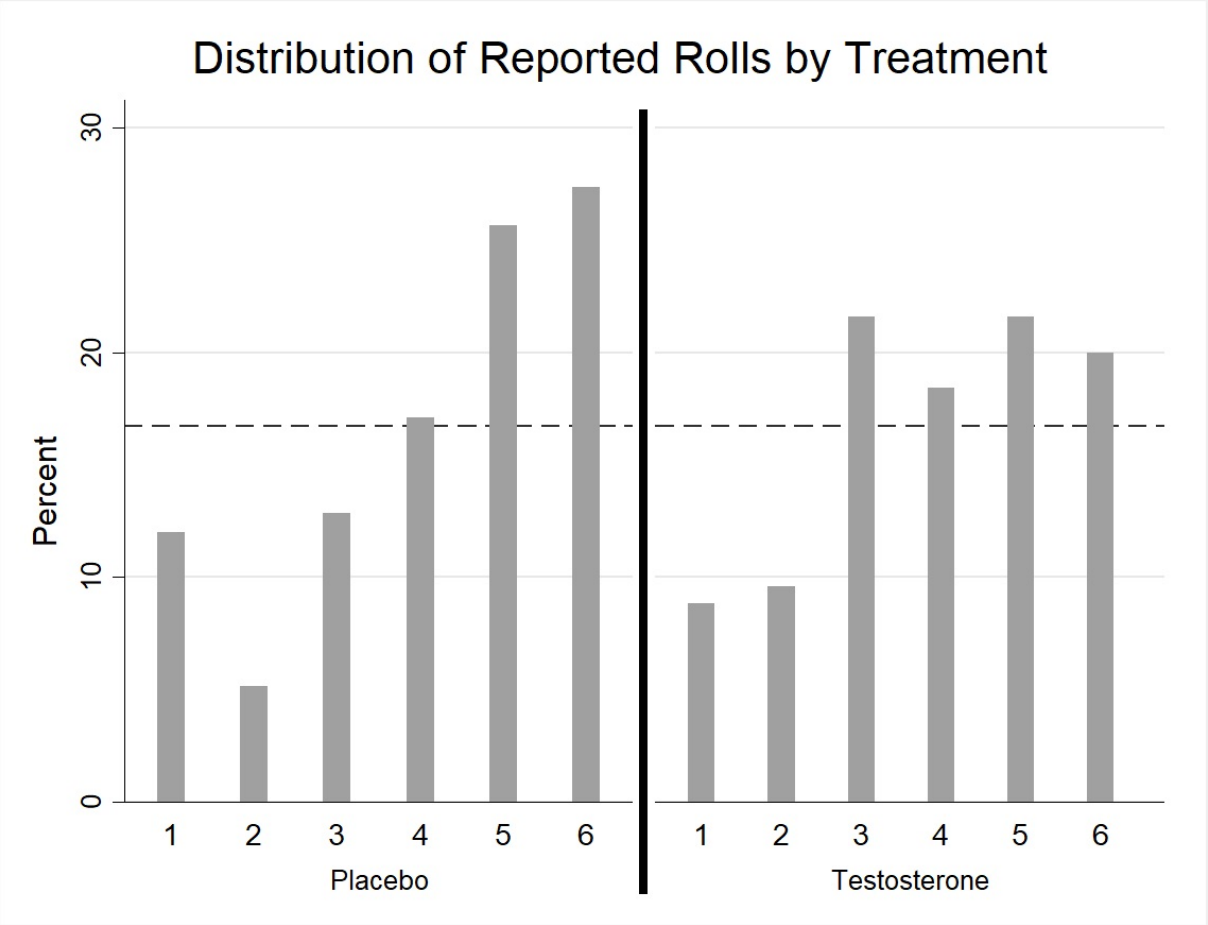

Figure 4.1. Distribution of reported rolls by treatment. Reference line is the expected frequency of each outcome with fully honest participants.

We use three non-parametric measures to compare the two treatment groups: the distribution of rolls via a χ^2-test against equal distributions, the mean reported roll via a Mann-Whitney U-test of differences, and the reported proportion of the highest possible roll via a Fisher's exact test.

A χ^2-test confirms that both treatment groups exhibited evidence of self-serving lying, as

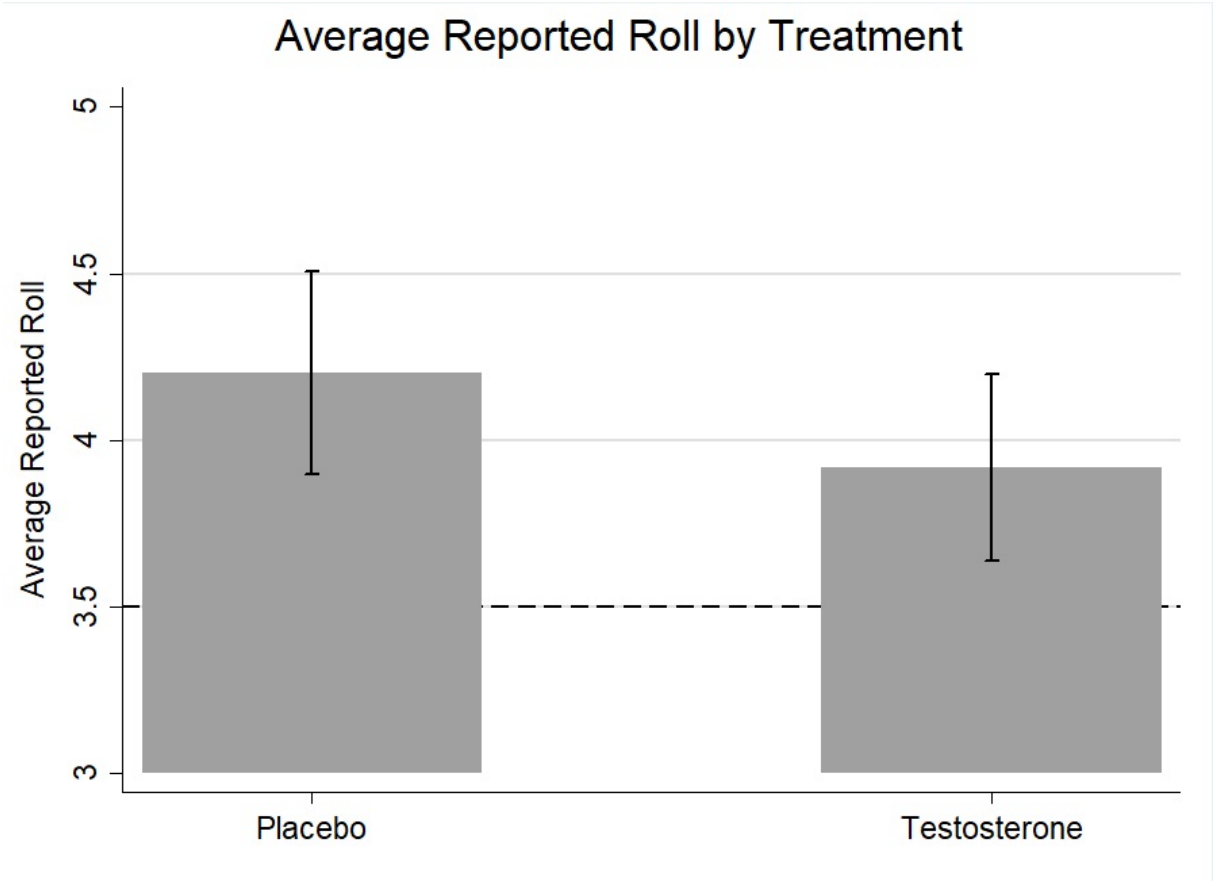

Figure 4.2. Average reported roll by treatment. Reference line is the expected average roll with fully honest participants. Bars represent 95% CI.

indicated by a right-skewed distribution (χ^2-test of even distribution, Testosterone $\chi^2(5) = 26.71$, $p = 0.02$, Placebo $\chi^2(5) = 13.10$, $p < 0.001$, see Figure 1). This common use of lying is in line with other research using the die-roll task (Shalvi, Eldar, and Bereby-Meyer, 2012; Erat and Gneezy, 2012; Fischbacher and Föllmi-Heusi, 2013; Abeler and Nosenzo, 2016) and demonstrates that participants grasp that they are able to misreport their rolls presumably in order to increase their earnings and do so. A Mann-Whitney U-test of differences in the distributions of reported roll yielded could not reject the null that the distributions are the same ($z(240) = 1.57$, $p = 0.12$). The Fisher's exact test could not reject the null that the proportion of 6's reported in each treatment group are the same (Fisher's exact, $p = 0.17$).

We also report the result of a parametric t-test, and use this summary statistic in order

to perform a joint analysis of our study together with the results of Wibral et al. Overall, our findings are similar, regardless of whether we use a parametric or non-parametric approach. The average reported die roll in our sample of the placebo group was 4.21 (95% CI[3.91, 4.52]) and treatment group was 3.94 (95% CI[3.67, 4.22]), which by a 2-sided t-test did not significantly differ ($t(240) = -1.31$, $p = 0.19$, Cohen's $d = -0.17$, 95% CI[-0.42, .08], see Figure 2).

Comparison to Wibral et al. (2012)

In Table 2 we juxtapose the major statistical results from our study and those from Wibral et al. (2012). Overall, our results are directionally the same in that testosterone is associated with a decrease in reported rolls, but we do not find statistical significance by any measure at the 10% level. The key reported measures of Wibral et al. were the Mann-Whitney U-test of different distributions between treatment groups ($z(89) = 2.78$, $p = 0.01$) and the Fisher's exact test of different frequencies of reporting the number with the highest material incentive ($p = 0.01$), which we contrast with our Mann-Whitney U-test result ($z(240) = 1.57$, $p = 0.12$), and Fisher's exact test result ($p = 0.18$).

In terms of effect size, Wibral et al. found a medium effect size of Cohen's $d = -0.56$ (95% CI[-0.97, -0.14]) of the impact of testosterone on average reported die-roll. Based on this effect size, a sample size of 81 would be sufficiently powered at $\beta = 0.80$ at the 5% level. A typical finding in the replication literature is that the replicated effect size is smaller than the original by about a half in psychological experiments (Open, 2015) and a third in experimental economics (Camerer et al., 2016). With our sample size of N = 242, we achieved $\beta = 0.88$ at the 5% level for detecting the 2/3 of the original effect size, or $\beta = 0.68$ at the 5% level for detecting one half of the original effect size. Our small effect size of Cohen's $d = -0.17$ (95% CI[-0.42, 0.08]) suggests that we would have needed a sample size of $N = 878$ to detect a significant difference in means for the point estimate, and $N = 142$ to detect the upper bound of our confidence interval at the 5% level with $\beta = 0.80$.

Table 4.2. Comparisons of major statistical findings with Wibral et al.

Statistical Test	**Description**	**Wibral et al.**	**Current Study**
Mann-Whitney U-test	Comparison of mean post-treatment testosterone levels between groups	$N = 91$, $p = 0.03$	$N = 241$, $z(239) = -12.79$, $p < 0.001$
χ^2-test against uniform distribution	Test for evidence of self serving lying, i.e. for a right-skewed distribution in the reported die rolls	Treatment $N = 46$, $\chi^2(5) = 13.22$, $p = 0.02$ Placebo $N = 45$, $\chi^2(5) = 63.47$, $p < 0.001$	Treatment $N = 125$, $\chi^2(5) = 13.10$, $p = 0.02$ Placebo $N = 117$, $\chi^2(5) = 26.71$, $p < 0.001$
Mann-Whitney U-test	Test for differences in the distributions of reported die rolls between testosterone and placebo	$N = 91$, $z(89) = 2.78$, $p = 0.01$	$N = 242$, $z(240) = 1.57$, $p = 0.12$
Fisher's exact test	Test whether the number with the highest payoff was reported more frequently in treatment as compared to control (in Wibral et al. it was the number 5, in our study it is the number 6)	Proportion of 5's in treatment $16/46 = 35\%$ Proportion of 5's in placebo $28/45 = 62\%$ $p = 0.01$, $n = 91$	Proportion of 6's in treatment $25/125 = 20\%$ Proportion of 6's in placebo $32/117 = 27\%$ $p = 0.17$, $n = 242$
2 sided T-test	Comparison of the mean reported roll between testosterone and placebo groups	Mean (SD) treatment = 3.33 (1.67) Mean (SD) placebo = 4.18 (1.37) $t(89) = 2.65$, $p < 0.001$	Mean treatment (SD) = 3.94 (1.39) Mean placebo (SD) = 4.21 (1.66) $t(240) = 1.31$, $p = 0.19$

Joint Analysis of Studies

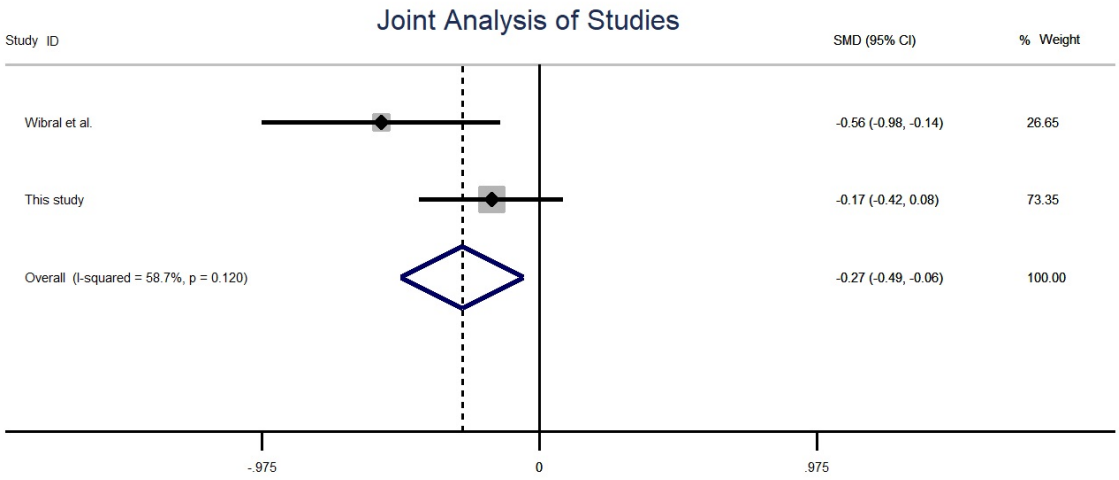

Figure 4.3. Meta-analysis of effect size using fixed effects model. Bars represent 95% CI.

To perform the joint analysis we use a fixed effects model using a weighted average of both studies, in line with previous work on replications (Open, 2015; Camerer et al., 2016). Because the system of payoffs in the Wibral et al. study was such that reporting a 6 earned nothing, we transformed their data to match our own based on the payoffs associated with each report, such that a report of 5 was coded as 6, 6 was coded as 1, 1 was coded as 2, *et cetera.* Using the fixed effects model, Cohen's d is equal to -0.27 (95% CI[-0.49, -0.06]) and a test of $d = 0$ is rejected at the 0.05 level ($z(1) = 2.46$, $p = 0.01$, see Figure 3). The achieved power is > 0.999, calculated using G-Power.

Further details, including robustness to a random effects specification and a search for comparable studies are in the Supplementary Materials.

Effect of Winning or Losing in Risk Task

As discussed previously in Differences From Wibral et al., this task was part of an experimental battery, and the preceding task was a risk task in which participants were divided into winners and losers based on their performance. In order to test an association between

the competition outcomes and a potential interaction between competition outcomes and treatment, we ran a two-way ANOVA with competition outcomes (winning/losing) and treatment (testosterone/placebo) as between-subject factors, as well as an interaction term. We found no significant effects of competition outcome on reported die roll ($F(1,234) = 0.71$, $p = 0.401$), treatment condition ($F(1,234) = 3.21$, $p = 0.074$]), or interaction ($F(1,234) = 1.41$, $p = 0.236$).

Facial Masculinity and Digit Ratio

To test for the impact of digit ratio and facial masculinity on behavior, we performed a number of ordinary least squares regressions. Regressing die roll on treatment, digit ratio, and the interaction of digit ratio and treatment did not yield any coefficients significantly different from 0 (treatment $\beta = -5.159(5.835)$, 95%CI[-16.654, 6.335], $t(235) = -0.88$, $p = 0.377$, digit ratio $\beta = -4.278(4.352)$, 95%CI[-12.851, 4.294], $t(235) = -0.98$, $p = 0.327$, interaction $\beta = 5.147(6.149)$, 95%CI[-6.966, 17.260], $t(235) = 0.84$, $p = 0.403$). Similarly, regressing die roll on treatment, facial masculinity, and the interaction of treatment and facial masculinity did not yield any coefficients significantly different from 0 (treatment $\beta = -0.018(0.644)$, 95%CI[-1.287, 1.251], $t(227) = -0.03$, $p = 0.978$, facial masculinity $\beta = 0.160(0.282)$, 95%CI[-0.396, 0.715], $t(227) = 0.57$, $p = 0.572$, interaction $\beta = -0.126(0.390)$, 95%CI[-0.895, 0.643], $t(227) = -0.32$, $p = 0.746$). The results remained insignificant when including measurements of other hormonal levels and demographic characteristics, as reported in the Supplementary Materials.

Discussion

The present study found modest evidence for testosterone reducing self-serving dishonesty. Although statistically insignificant, the direction was in same direction of the original study, with our joint analysis indicating a significant effect ($p = 0.01$) of testosterone administration on the mean die roll report. It should be taken as suggestive, but not conclusive, evidence of a relationship between testosterone and reduced lying, that should encourage further exploration.

In this section we discuss the limitations of our study, and then suggest avenues for future research elaborating on the association between testosterone and deception.

Despite the advantages of strict experimental control, there are several limitations inherent in the methodology of the current study.

First, a general limitation of laboratory studies is that at the participants are aware that they are taking part in an experiment. One cannot entirely rule out the possibility that such knowledge might bias behavior (e.g., via experimenter demand effect), in a way that could interact with the treatment.

A second limitation of the particular task at hand (where we do not directly observe the behavior of the participants), is the incapacity to measure whether or by how much each individual participant lied. This limits, to some degree, the capacity to explore which factors might moderate of the behavioral effect.

Third, while the use of college students in scientific experiments, particularly the behavioral sciences, is a widely accepted practice, it comes with specific considerations to be made when generalizing findings to other populations. The possibility of an interaction between our subject population characteristics (young males) and our experimental design is particularly relevant, as the levels of testosterone decrease with age after 20 (Harman et al., 2001) and vary significantly between sexes (Torjesen and Sandnes, 2004). Thus, our sample is not representative of the baseline physiology of the general population. Furthermore, the proposed psychological mechanism through which testosterone impacts lying is through social-status concerns, and it may be that different demographic groups would not pursue social status goals through honesty on this task. Therefore, we advise that any generalized interpretation of our findings to other populations should be made with caution.

Going forward, elucidating the relationship between testosterone and deception requires clear hypotheses of connecting mechanisms and methodologies that directly test them. The relatively complex chain of reasoning connecting testosterone and deception in the die-roll task proposed by Wibral et al. is that testosterone increases status seeking, and thus elevates the

decision maker's need for pride, which in turn promotes honest behavior. However, as deception is typically associated with material benefits that are also important for one's social status, it is not *a priori* clear whether testosterone-induced status seeking should decrease, rather than increase honesty in this task. Using deception tasks with more obvious social status interpretations would provide a stronger test of this potential connection.

Another potential mechanism by which testosterone impacts die roll reports may be through its influence on impulsivity[**?**]. Greater impulsivity may reduce the propensity of an individual to engage in processes which either increase or decrease their ultimate willingness to lie. For example, reflection could either increase lying by justifying it as harming no one (Shalvi, Eldar, and Bereby-Meyer, 2012), or decrease it by reflecting upon moral considerations. Further experimental work that aspires to explore this issue should have clear predictions about whether lying in the specific behavioral task used is more associated with either impulsive or deliberative decision-making.

A final possibility is that testosterone may increase the feelings of distrust in participants. Several studies have found a negative relationship between testosterone administration and trust, as measured by reduced offers in the trust game (Boksem et al., 2013) and facial trustworthiness evaluations by women (Bos, Terburg, and van Honk, 2010). Moreover, studies of anabolic steroid users found that they are more likely to report paranoia, even after short-term use (Perry, Anderson, and Yates, 1990; Pope and Katz, 1988; Wilson, Prange, and Lara, 1985). Even though both in our study and in Wibral et al. the researchers made efforts to provide privacy for the participants and ensure them of this fact, recent research suggests that lying in the die-roll task is partly driven by fears of detection (Kajackaite and Gneezy, 2017). Thus, increased feelings of distrust might lead participants to doubt that the researchers were truly unable to observe their actions, or to be concerned of a hidden or unstated punishment for being observed deceiving. Further research could attempt to address this issue by including survey measures to assess whether or not participants felt as if their actions were truly performed in privacy if applicable, or using methodologies where a lie is completely undetectable. An example of such a methodology

is a "mind" game in which participants think of a number and then roll a die in private, and report whether the rolled number matched the number they thought of, which was used in Kajackaite and Gneezy (2017).

In summation, we find a statistically insignificant negative effect of testosterone administration on mean reported die roll. When jointly considered along with results from a previous and similar study by Wibral et al., there is overall evidence of a negative association between testosterone and lying. There are a number of plausible mechanisms which might explain this association, but currently with data only from the die-roll task it is not possible to determine which mechanism(s) play a central role. In addition to designing future studies around straightforward tests of these mechanisms, researchers should use large sample sizes and facilitate the replications of their findings. Evidence is growing that testosterone impacts behavior in diverse ways, and practices which help build a robust knowledge base on these impacts is paramount for progress.

4.2 Acknowledgments

Chapter 4, in full, is a reprint of material as it appears in Scientific Reports 2018. Written permission has been granted by the co-authors for the use of this chapter.

4.3 References

1. Abeler, J., Nosenzo, D., & Raymond, C. (2016). Preferences for truth-telling. IZA Discussion Paper 10188.
2. Apicella, C. L., Carr, J. M., & Dreber, A. (2015). Testosterone and economic risk taking: A review. Adaptive Human Behavior and Physiology, 1, 358-385.
3. Becker, G. S. (1968). Crime and punishment: An economic approach. The Economic Dimensions of Crime., 13-68.
4. Boksem, M. A., et al. (2013). Testosterone inhibits trust but promotes reciprocity. Psychological Science, 24, 2306-2314.17

5. Bos, P. A., Terburg, D., & van Honk, J. (2010). Testosterone decreases trust insocially naive humans. Proceedings of the National Academy of Sciences, 107, 9991-9995.

6. Bugnyar, T., & Kotrschal, K. (2002). Observational learning and the raiding of food caches in ravens, corvus corax: is it tactical deception? Animal Behaviour, 64, 185-195.

7. Camerer, C. F., et al. (2016). Evaluating replicability of laboratory experiments in economics. Science, 351, 1433-1436.

8. Carr, J. M., Ruddick, E. L., Moreau, B. J., & Bird, B. M. (2017). Testosterone and human aggression.

9. Charness, G., Masclet, D., & Villeval, M. (2014). The dark side of competitionfor status.Management Science, 60, 38-55.

10. Chik, Z., Johnston, A., Tucker, A. T., Chew, S. L., Michaels, L., & Alam,C. A. S. (2006). Pharmacokinetics of a new testosterone transdermal delivery system, testosterone in healthy males.British journal of clinical pharmacology, 61, 275-279.

11. Coates, J. M., & Herbert, J. (2008). Endogenous steroids and financial risktaking on a london trading floor.Proceedings of the National Academy ofSciences, 105, 6167-6172.

12. Collaboration, O. S. (2015). Science,349.

13. Cueva, C., et al. (2015). Cortisol and testosterone increase financial risk takingand may destabilize markets. Scientific Reports, 5, 6167-6172.

14. Dai, Z., Galeotti, F., & Villeval, M. C. (2017). Cheating in the lab predictsfraud in the field: An experiment in public transportation. Management Science.

15. DiMaggio, P. (1982). Cultural capital and school success: The impact of statusculture participation on the grades of us high school students.American sociological review, 47, 189-201.

16. Dreber, A., & Johannesson, M. (2008). Gender differences in deception. Economics Letters, 99, 197-199.

17. Dreher, J., Dunne, S., Pazderska, A., Frodl, T., Nolan, J. J., & O'Doherty,J. P. (2016). Testosterone causes both prosocial and antisocial status-enhancing behaviors in human males.Proceedings of the National Academy of Sciences, 113, 11633-11638.

18. Du, J. Y., et al. (2013). Percutaneous progesterone delivery via cream or gel application in postmenopausal women: a randomized cross-over study of progesterone levels in serum, whole blood, saliva, and capillary blood. Menopause, 20, 1169-1175.

19. Eisenegger, C., Haushofer, J., & Fehr, E. (2011). The role of testosterone insocial interaction.Trends in Cognitive Sciences,15, 263-271.

20. Eisenegger, C., Naef, M., Snozzi, R., Heinrichs, M., & Fehr, E. (2010). Prejudiceand truth about the effect of testosterone on human bargaining behaviour.Nature,463, 356-361.

21. Eisenegger, C., von Eckardstein, A., Fehr, E., & von Eckardstein, S. (2013). Pharmacokinetics of testosterone and estradiol gel preparations in healthyyoung men. Psychoneuroendocrinology, 38, 171-178.

22. Erat, S., & Gneezy, U. (2012). White lies.Management Science,58, 723-733.Escasa, M. J., Casey, J. F., & Gray, P. B. (2011). Salivary testosterone levelsin men at a us sex club. Archives of sexual behavior, 40, 921-926.

23. Fischbacher, U., & Franziska, F. (2013). Lies in disguisean experimental studyon cheating. Journal of the European Economic Association, 11, 525547.

24. Garrett, N., Lazzaro, S. C., Ariely, D., & Sharot, T. (2016). The brain adaptsto dishonesty. Nature Neuroscience,19, 1727-1732.

25. Gneezy, U. (2005). Deception: The role of consequences. The American Economic Review, 95, 384-394.

26. Gray, P. e. a. (2005). Dose-dependent effects of testosterone on sexual function,mood, and visuospatial cognition in older men. The Journal of Clinical Endocrinology & Metabolism, 90, 3838-3846.

27. Gchter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. Nature, 531, 496-499.

28. Harman, S. M. e. a. (2001). Longitudinal effects of aging on serum total and free testosterone levels in healthy men. The Journal of Clinical Endocrinology & Metabolism, 86, 724-731.

29. Hermans, E. J., Ramsey, N. F., & van Honk, J. (2008). Exogenous testosterone enhances responsiveness to social threat in the neural circuitry of social aggression in humans. Biological Psychiatry, 63, 263-270.

30. Hugh-Jones, D. (215). Honesty and beliefs about honesty in 15 countries. University of East Anglia Discussion Paper.

31. Janowsky, J. S., Oviatt, S. K., & Orwoll, E. S. (1994). Testosterone influences spatial cognition in older men. Behavioral Neuroscience, 108, 325.

32. Kajackaite, A., & Gneezy, U. (2017). Incentives and cheating.Games and Economic Behavior, 102, 433-444.

33. Lane, A., Luminet, O., Nave, G., & Mikolajczak, M. (2016). Is there a publication bias in behavioural intranasal oxytocin research on humans? Opening the file drawer of one laboratory. Journal of Neuroendocrinology, 28.

34. Lisofsky, N., Kazzer, P., Heekeren, H. R., & Prehn, K. (2014). Investigating socio-cognitive processes in deception: a quantitative meta-analysis of neuroimaging studies. Neuropsychologia, 61, 113-122.

35. Lutchmaya, S., Baron-Cohen, S., Raggatt, P., Knickmeyer, R., & Manning,J. T. (2004). 2nd to 4th digit ratios, fetal testosterone and estradiol. Early Human Development, 77, 23-28.

36. Mann, H., Garcia-Rada, X., Hornuf, L., Tafurt, J., & Ariely, D. (2016). Cut fromthe same cloth: Similarly dishonest individuals across cultures. Journal of Cross-Cultural Psychology, 47, 858-874.

37. Mayo, A., Macintyre, H., Wallace, A., & Ahmed, S. (2004). Transdermal testosterone application: pharmacokinetics and effects on pubertal status, short-term growth, and bone turnover. The Journal of Clinical Endocrinology & Metabolism, 89, 681-687.

38. Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. Journal of Marketing Research, 45, 633-644.

39. Montoya, E., Terburg, D., Bos, P. A., & van Honk, J. (2012). Testosterone, cortisol, and serotonin as key regulators of social aggression: A review and theoretical perspective. Motivation and Emotion, 36, 65-73.

40. Muehlheusser, G., Roider, A., & Wallmeier, N. (2015). Gender differences in honesty: Groups versus individuals. Economics Letters, 128, 2529.

41. Nadler, A., Jiao, P., Alexander, V., Johnson, C. J., & Zak, P. J. (in press). The bull of wall street: Experimental analysis of testosterone and asset trading. Management Science.

42. Nave, G., Camerer, C. F., & McCullough, M. (2015). Does oxytocin increase trust in humans? a critical review of research.Perspectives on Psychological Science, 10, 772-789.

43. Nave, G., Nadler, A., Dubois, A., Zava, D., Camerer, C., & Plassmann, H. (inpress). Single-dose testosterone administration increases mens preferencefor status goods. Nature Communications.

44. Nave, G., Nadler, A., Zava, D., & Camerer, C. (2017). Single-dose testosterone administration impairs cognitive reflection in men. Psychological Science, 28, 1398-1407.

45. Newman, M. L., Sellers, J. G., & Josephs, R. A. (2005). Testosterone, cognition,and social status. Hormones and Behavior, 47, 205-211.

46. Penton-Voak, I. S., & Chen, J. Y. (2004). High salivary testosterone is linked to masculine male facial appearance in humans.Evolution and Human Behavior, 25, 229-241.

47. Perry, P. J., Anderson, K. H., & Yates, W. R. (1990). Illicit anabolic steroiduse in athletes. a case series analysis. Am J Sports Med, 18, 422-428.

48. Pfundmair, M., Erk, W., & Reinelt, A. (2017). Lie to me: oxytocin impairs lie detection between sexes. Psychoneuroendocrinology, 84, 135-138.

49. Pope, H. G., & Katz, D. L. (1988). Affective and psychotic symptoms associated with anabolic steroid use. Am J Psychiatry, 1, 487.

50. Pope, H. G., Kouri, E. M., & Hudson, J. I. (2000). Effects of supraphysiologicdoses of testosterone on mood and aggression in normal men: a randomized controlled trial.Archives of General Psychiatry, 57, 133-140.

51. Salameh, e. a., W. A. (2010). Validation of a total testosterone assay using high-turbulence liquid chromatography tandem mass spectrometry: total and free testosterone reference ranges.Steroids, 75, 169-175.

52. Schurr, A., & Ritov, I. (2016). Winning a competition predicts dishonest behavior. Proceedings of the National Academy of Sciences, 113, 1754-1759.

53. Shalvi, S., Dana, J., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior.Organizational Behavior and Human Decision Processes, 115, 181-190.

54. Shalvi, S., & De Dreu, C. K. W. (2014). Oxytocin promotes group-servingdishonesty. Proceedings of the National Academy of Sciences, 111, 5503-5507.20

55. Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications).Psychological Science,23, 1264-1270. Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results.Psychological Science,26, 559-569.

56. Slovic, P. (1966). Risk-taking in children: Age and sex differences. Child Development,37, 169-176.

57. Spence, S., et al. (2004). A cognitive neurobiological account of deception:evidence from functional neuroimaging. Philosophical Transactions of theR oyal Society B: Biological Sciences, 359, 1755-1762.

58. Torjesen, P. A., & Sandnes, L. (2004). Serum testosterone in women as measured by an automated immunoassay and a ria. Clinical chemistry, 50, 678-679.

59. Trivers, R. (2000). The elements of a scientific theory of self-deception. Annals of the New York Academy of Sciences, 907, 114-131.

60. Van Honk, J., Montoya, E. R., Bos, P. A., van Vugt, M., & Terburg, D. (2012). New evidence on testosterone and cooperation. Nature, 485.

61. Van Honk, J., Will, G.-J., Terburg, D., Raub, W., Eisenegger, C., & Buskens,V. (2016). Effects of testosterone administration on strategic gambling in poker play. Scientific Reports, 6, 18096.

62. Volz, K. G., Vogeley, K., Tittgemeyer, M., von Cramon, D. Y., & Sutter, M.(2015). The neural basis of deception in strategic interactions. Frontiers in Behavioral Neuroscience, 9, 27.

63. Walum, H., Waldman, I., & Young, L. J. (2016). Statistical and methodological considerations for the interpretation of intranasal oxytocin studies. Biological Psychiatry, 79, 251-257.

64. Wang, C. e. a. (2004). Long-term testosterone gel (androgel) treatment maintains beneficial effects on sexual function and mood, lean and fat mass, and bone mineral density in hypogonadal men. The Journal of Clinical Endocrinology & Metabolism, 89, 2085-2098.

65. Whiten, A., & Byrne, R. W. (1988). Tactical deception in primates.Behavioral & Brain Sciences, 11, 233-244.

66. Wibral, M., et al. (2012). Testosterone administration reduces lying in men. PloS one, 7, e46774.

67. Wilson, I., Prange, A., & Lara, P. (1985). Methyltestosterone and imipramineinmen: conversion of depression to a paranoid reaction. Am J Psychiatry, 131, 21-24.

68. Zethraeus, N., et al. (2009). A randomized trial of the effect of estrogen and testosterone on economic behavior. Proceedings of the National Academy of Sciences, 106, 6535-6538.

Chapter 5

Experimental methods: Measuring effort in economics experiments

5.1 Abstract

The study of effort provision in a controlled setting is a key research area in experimental economics. There are two major methodological paradigms in this literature: stated effort and real effort. In the stated-effort paradigm the researcher uses an "effort function" that maps choices to outcomes. In the real-effort paradigm, participants work on a task, and outcomes depend on their performance. The advantage of the stated-effort design is the control the researcher has over the cost of effort, which is particularly useful when testing theory. The advantage of the real-effort design is that it may be a better match to the field environment, particularly with respect to psychological aspects that affect behavior. An open question in the literature is the degree to which the results obtained by the two paradigms differ, and if they do, why. We present a review of methods used and discuss the results obtained from using these different approaches, and issues to consider when choosing and implementing a task.

5.2 Introduction

Understanding when and how individuals exert effort is critical to many questions in economics. While a large literature in experimental economics studies effort provision, different

approaches have been used to operationalize it experimentally. Experimental economists have primarily utilized two methodological paradigms: stated effort and real effort. There is limited theoretical and/or experimental evidence to guide researchers in deciding which task to use. Furthermore, there are many real-effort tasks and ways to implement them.

With stated effort, the choice of "effort" involves clear numerical costs and benefits. In a typical implementation, participants are presented with a menu that displays a discrete selection of effort levels (e.g., from 1 to 10) and a corresponding list of costs. These costs often influence the profits of another subject, as in a gift-exchange situation (Fehr et al., 1993, Fehr, Gächter and Kirchsteiger, 1997, Charness, 2004), or in a tournament involving effort (Müller and Schotter, 2008 and Bull, Schotter and Weigelt 1987). The advantage of the stated-effort approach is that there is no uncertainty regarding an individual's cost of effort. A potential drawback of the method is that simply choosing a number may not capture the field environment and the psychological forces involved in putting forth actual effort.

Real-effort tasks measure the behavior of participants given specific observable tasks, such as solving mazes (Gneezy, Niederle and Rustichini, 2003), solving anagrams (Charness and Villeval, 2009), adding series of two-digit numbers (Niederle and Vesterlund 2007), counting the number of zeros in a large grid (Abeler et al., 2011), transcribing meaningless "greek" letters (Augenblick, Niederle, and Sprenger, 2015), and cracking walnuts (Fahr and Irlenbusch, 2000). The effort could be physical, as in folding pieces of paper and stuffing envelopes, cognitive, as in solving a series of math equations, or creative, as in writing stories or packing quarters. The advantage of the real-effort method is that it is closer to the psychology of working. For example, the cost of effort might vary over time: solving mazes might be fun initially, but might gradually become less motivating. A potential drawback is that the researcher does not know the cost of effort (and perhaps not even the sign of the effort cost; Gross, Guo, and Charness, 2015) for participants, so that testing theories is more challenging.

A key purpose for conducting a laboratory experiment is to use the advantages of a controlled environment to learn about an economically-interesting phenomenon. We identify

several dimensions that are important when deciding about effort measurement, such as the timing of the effort decision, the existence of goal-oriented decision-making, and the particulars of decisions over effort and money. Our aim is to help organize the considerations involved in both picking the methodology best suited to the research question at hand and understanding the key limitations of that methodology.

5.3 Stated-Effort Experiments

Testing specific models is a central focus of many effort experiments, and this typically requires experimental control over the relevant components of the theory. One needs a clear mapping from the cost of effort to the resulting productivity. Models may rely upon specific characterizations of the properties of the cost of effort function. For example, the cost of function it may be linear such that each unit of effort has the same associated cost, or it could be convex, such that the cost of each additional unit of effort is increasing. Such properties may be important to the predictions of specific models.

Smith (1976) introduced and argued for induced value, which forms the logical basis for stated effort. Although many economic experiments make use of the induced-value paradigm, we focus here on papers that explicitly used it (at least in motivation) to study effort. The gift-exchange game using induced values and stated effort was first tested in Fehr, Kirchsteiger, and Riedl (1993) and has led to important insights and has had great impact on our understanding of labor relations. In a simplified version of this game, a firm chooses a wage between 0 and 100, and the firm's earnings are determined by (100-w)*e. The worker's earning is the wage less the cost of the effort level chosen. This is the cost-of-effort schedule:

e	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
c(e)	0	1	2	4	6	8	10	12	15	18

This method is useful when considering social preferences, since the relationship between the firm's payoff and the worker's payoff is completely known to the worker and the "sacrifice" of freely-chosen higher effort provides clear benefits.

Stated effort is also useful for testing models in tournament settings. For example, Müller and Schotter (2010) consider the prize structure in contests, testing the Moldovanu and Sela (2000) model that shows the optimal structure depends on whether the cost-of-effort function is convex or not. The experimental results show that low-ability workers tend to "drop out" and provide little or no effort (this is not part of the equilibrium in the theoretical model), while high-ability workers provide excessive levels of effort, so that there is a bifurcation of effort. Nevertheless, the firm overall receives the expected amount of effort. The cost of effort was implemented as either a linear function or a quadratic function of the "decision number" (effort). The 2x2 experimental design also varied whether one prize or two prizes were awarded for the group of four participants. It seems clear that one would be unable to test this model with real effort, since the cost of effort would be unknown for each individual.

We list below a number of prominent papers that use a stated-effort methodology, by their research areas, main findings, and significance. Several previous and more extensive literature reviews examined experiments that used stated effort in specific fields such as labor (Charness and Kuhn, 2011) and coordination (Devetag and Ortmann, 2007). A number of experimental public-goods games (reviewed in Chaudhuri, 2011), trust games (reviewed in Johnson and Mislin, 2011), and principal-agent games (e.g., Charness and Dufwenberg, 2006, and Brandts, Charness, and Ellman, 2016) also use the logic of stated effort, but are not explicitly about effort. Our list is neither meant to be exhaustive or an attempt to rank the most important papers, but rather to highlight how stated-effort has been used productively in a variety of research areas. For more detail, we refer people to the literature reviews mentioned above.

Table 5.1. A range of stated-effort studies.

Abbreviated citation	Research area	Experimental design	Main experimental finding and significance
Van Huyck, John, Raymond Battalio, and Richard Beil (1990)	Coordination	The article examines a class of tacit pure coordination games with multiple equilibria, which are strictly Pareto ranked. It reports experiments that provide evidence on how human subjects make decisions under conditions of strategic uncertainty	Inefficient play is typically the result in these games. This is not the result of conflicting objectives or to asymmetric information. "Instead, coordination failure results from strategic uncertainty: some subjects conclude that it is too 'risky' to choose the payoff-dominant action and most subjects focus on outcomes in earlier period games.
Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl (1993)	Labor	A 2-stage design wherein first some participants ("employers") made wage offers which other participants ("workers") could choose to accept. The sellers then made a decision of how much effort to exert	. Workers responded to higher wage offers with higher effort, providing support for the fair-wage hypothesis.
Charness, Gary (2000)	Labor	Participants were either "employers" or "employees," and the wage of the employee was either presented to them as resulting from a random process or assigned by the experimenter.	Participants only responded with very high effort levels when the wage offers were seen as being made by random processes, indicating that the perceived responsibility of a wage rate is behaviorally important.

Table 5.1. A range of stated-effort studies, continued.

Abbreviated citation	Research area	Experimental design	Main experimental finding and significance
Brown, Martin, Armin Falk, and Ernst Fehr (2004)	Labor	Participants were either assigned as "firms" or "workers," and firms offered contracts (either to individual workers or publicly) with a wage and desired effort level. Treatments varied whether there was 3rd party contract enforcement, and there were 15 rounds, with stable identities throughout.	Stable long-term relationships between firms and employees emerged even in absence of 3rd party contract enforcement. Successful relationships, common in the no 3rd party enforcement condition, had both generous rent sharing and high effort from the beginning. With 3rd party enforcement, most interactions were one-shot.
Charness, Gary, and Martin Dufwenberg (2006)	Principal-agent	The principal chooses to hire an agent or take an outside option. The agent decides whether to exert effort, with 5/6 chance of success if effort is chosen. In the communication treatments, the agent can send a free-form message to the principal.	Communication leads to more agents being hired and better performance by the agents. Statements of intent (promises) appear to have commitment power. The results are seen through the lens of guilt aversion.

Table 5.1. A range of stated-effort studies, continued.

Abbreviated citation	Research area	Experimental design	Main experimental finding and significance
Brandts, Jordi, and David J. Cooper (2007)	Labor	Participants were either "managers" or "employees" and participated in a minimum effort game, in which all payoffs depended upon the lowest effort exerted by one of the employees. Managers were able to pick bonuses for employees corresponding to increases in the minimum effort. Treatments varied the level of communication between managers and employees—no communication, one-way managerial messages to employees, or two-way communication.	Allowing communication, and the particularities of the communications, were more important for overcoming coordination failures than good monetary incentives.
Chaudhuri, Ananish, Andrew Schotter, and Barry Sopher (2009)	Coordination	Participants play in a series of minimum effort games, in which groups of 8 participants play 10 rounds of the game, then give advice to their successors who play after them. Successors are also in some treatments able to see the messages and actions of their predecessors.	Efficient outcomes were more likely when the advice from each predecessor to a round was made common knowledge, rather than given to one successor each.

Table 5.1. A range of stated-effort studies, continued.

Abbreviated citation	Research area	Experimental design	Main experimental finding and significance
Müller, Wieland, and Andrew Schotter (2010)	Labor	Participants competed in an effort tournament, in which they were first randomly assigned their type and then made an effort allocation decision. The cost of effort function was varied across treatments between linear and convex, and the payoff structure was varied across treatments to either give a prize to those who exerted the highest or the two highest effort levels.	Low-type workers exerted less than theoretically-predicted effort, and high-type workers exerted greater than predicted effort. This bifurcation of effort contradicts the hypothesis that in such an effort tournament that effort should be a continuous and increasing function of ability.
Brandts, Jordi, Gary Charness, and Matthew Ellman (2016)	Principal-agent	The principal can hire an agent, with either a rigid or flexible contract, to perform a task. If hired, the agent chooses a quality level (high, normal, or low), where non-normal quality is costly for the agent. The higher the quality, the better for the principal. There is a 50% chance of a cost shock.	Free-form communication is very effective at producing flexible contracts that achieve efficiency (high quality) and that take into account the cost shock. Everyone earns more with free-form communication, although restricted communication is ineffective.

5.4 Real-effort experiments

Researchers have used different real-effort tasks in laboratory and extra-laboratory (lab-in-the-field) settings. In Table 2, we present a partial list of real-effort tasks used in these types of settings; we then qualitatively evaluate these based on several criteria relevant to either practical research considerations or to the suitability of the task for addressing the specific research question.

On the far end of realistic effort-provision experiments are studies that directly look at performance in a specific area of interest, such as the impact of incentives on exercise (Charness and Gneezy, 2009) or the impact of incentives on effort provision to support children's education (Glewwe, Ilias, and Kremer, 2010). Research that directly deals with complex behavioral patterns such as exercise habits or support for education may well provide more convincing conclusions about those specific behaviors than analogous behavior in a laboratory, but is costly to conduct and difficult to generalize to other behavior.

The difficulty of implementation is a question of whether the method requires specific materials or preparation on the part of the researcher. In Table 2 we term a task "Low" to indicates that it can be run through a computer or with minimal materials, "Med." (Medium) indicates that there are some special materials or preparation required, but the overall burden is otherwise not high, and "High" indicates the need for a significant investment in preparing for or conducting the experiment.

The productivity column indicates whether the task requires participant to do work that has outside value, such as cracking walnuts that can later be sold (Fahr and Irlenbusch, 2000) or entering presumably-useful research data (Gneezy and List, 2006; Dutcher, Salmon, and Saral, 2015; Charness, Cobo-Reyes, and Sánchez, 2016). If the task can be modified in a straightforward manner to make the output genuinely useful, an asterisk is placed next to "No." For example, the sorting and counting of coins in Bortolotti et al. (2009) could be the sorting of loose change from a business so that it could be deposited at a bank. Although some tasks not marked with an asterisk could conceivably be made productive, we only marked cases where it would be easiest.

A skill or ability confound means that participants would be likely to have greater variance in performance at a task, meaning that larger sample sizes are needed to capture treatment effects. This may also be a theoretical confound, as in some cases we might expect higher skill at performing a task (e.g., throwing a ball at a target, as in Ariely et al, 2009) to be correlated with higher enjoyment of the task, and thus less net cost of effort. In addition, there is evidence that

tasks produce different emotional responses and effort provision may be affected by different emotions (Lezzi, Fleming, and Zizzo, 2015). Learning is also an issue for some tasks, as participants may improve at translating effort into productivity over time, again making the link between observable actions and theory tenuous.

Control over the cost-of-effort function, seen as one of the major advantages of the chosen-effort paradigm, has been addressed primarily through qualitative means, for example by juxtaposing results from "easy" and "hard" real-effort tasks, although some such as the ball-catching task (Gächter, Huang, Sefton 2015) add quantitative control as well.

Table 5.2. Some real-effort experiments. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Task	First Use	Description	Prod.	Diff.Imp	S/A	Learn	Comments
Task	First Use	Description	Prod.	Diff.Imp	S/A	Learn	Comments
Data Entry							
Library Data Entry	Gneezy and List (2006)	Participants entered information from a stack of books into a computerized database.	Yes	High	Low	Low	May be difficult to consistently find new productive data entry tasks. Task has been used many times. Counting the number of errors provides an additional measure of effort.
Classifying Reviews	Bushong and Gagnon-Bartsch (2016)	Participants listen to Amazon book reviews, classifying them as either endorsing or criticizing the group. An annoying noise can be played to increase the cost of effort.	No*	High	Low	Low	Use of annoying noise gives researcher qualitative control over the cost of effort. Straightforward implementation through Amazon mTurk.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Visual Search							
Counting Zeros	Abeler, Falk, Goette, and Huffman (2011)	Participants are given a table with 150 randomly ordered 0's and 1's, and asked to count the correct number of 0's. A typical implementation is to count as many tables as possible within a time period.	No	Low	Low	Low	Can be implemented in several ways: by requiring the correct number of 0's on a table to proceed, or by allowing errors and then not giving credit for an incorrect table.
Counting Sevens	Mohnen, Pokorny, and Sliwka (2008)	Participants are given a block of random numbers and must count the number of 7's in the block.	No	Low	Low	Low	Same as above, but potentially less difficult and thus a lower cost of effort.
Puzzles							

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Packing Quarters into Boxes	Ariely et al. (2009)	Participants must pack 9 metal quarter-circles into a wooden box, a feat that can only accomplished with a particular arrangement of the metal pieces, within some time period. Performance is measured by amount of time to solve.	No	Med.	Med.	High	Simple to implement outside of a laboratory. Some participants may enjoy the task.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Labyrinth	Ariely et al. (2009)	Participants navigate a ball through a wooden maze on a plane by tilting the plane on 2 planes, while avoiding trap holes in the maze. Success is measured by number of trap holes passed within some time period.	No	Med.	Med.	High	Simple to implement outside of a laboratory. Some participants may enjoy the task. Luck may play a role in success in small sample sizes.
Solving Mazes	Gneezy, Niederle, and Rustichini (2003)	Participants solve computerized mazes by navigating a marker through a maze using the arrow keys.	No	Low	Med.	Low	Puzzles may have unequal difficulty, and some participants may enjoy the task. Task has been used many times.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Tetris-like game	Augenblick, Niederle, and Sprenger (2015).	Participants must complete 4 rows of Tetris: blocks of various shapes descend slowly from top of screen and fall into place at the bottom. But descent rate does not increase, and there is no progression in the difficulty of the game.	No	Low	Low	Low	Participants are very likely familiar with the task, and some may enjoy it despite efforts to make it unenjoyable. By construction, participants cannot increase effort within a time period, and can only increase effort by increasing the amount of time they work.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Computerized Tower of Hanoi	Rutström and Williams (2000)	Participants on a computer play a game in which the goal is to move "disks" of various sizes onto "pegs" such that a larger disk is never placed on a smaller disk.	No	Low	Med.	High	Some participants may enjoy solving the puzzle. Can only be used once. Researcher has little control over the cost of effort.
Memory							
Simon	Ariely et al. (2009)	An electronic device flashes a sequence of colored lights and corresponding sounds that the participant must duplicate.	No	Low	Med.	Low	Simple to implement outside of a laboratory. Differences in short-term memory may confound interpretation.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Recall Last 3-digits	Ariely et al. (2009)	An experimenter reads a sequence of digits to the participant, then suddenly stops and asks the participant to recall the last 3 digits that were read.	No	Low	Med.	Low	Simple to implement outside of a laboratory. Differences in short-term memory may confound interpretation.
Physical Challenge							
Dart Ball	Ariely et al. (2009)	Participants throw a tennis ball at a target with attached Velcro patches to which the tennis ball will adhere.	No	Med.	High	Med.	Simple to implement outside of a laboratory. Some participants may enjoy the task.
Roll-up	Baumeister (1984)	Participants must maneuver a ball into a target hole by spreading apart then pushing together two metal rods	No	Med.	High	Med.	Simple to implement outside of a laboratory. Some participants may enjoy the task.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Running	Gneezy and Rustichini.(2004)	Participants run twice along a 40m track.	No	Low	High	Low	Easy to explain to children and implement. High variance in ability for adults makes it difficult to interpret results as due to variance in effort.
Hand Dynamometer	Imas (2014)	Participants squeeze a specially calibrated dynamometer that requires them to exert a steady amount of pressure over a long period of time.	No	High	Low	Low	Requires some special equipment and calibration. Calibration means that variance in strength is accounted for and thus results are more easily interpreted as effort. Researchers have a high degree of control over the level of effort required from participants.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Clicking on a Target	Houy, Nicolai, and Villeval. (2016).	Participants must click on the center of a target within 8 seconds while random perturbations move mouse pointer.	No	Low	Low	Low	Researcher has a high degree of control over amount of effort needed to succeed by changing the magnitude of the perturbations.
Repetitive Task							
Sorting and Counting Coins	Bortolotti, Devetag, and Ortmann (2009)	Participants must sort and count a number of coins worth 1, 2,5, and 10 Euro cents within a given time interval.	No*	Med.	Low	Med.	Possible some participants may have experience with task due to cashier experience.
Cracking Walnuts	Fahr and Irlenbusch. (2000)	Participants are given a pile of walnuts and nutcracker and must produce some mass of cracked walnuts in a given time.	Yes	High	Low	Low	Not much researcher control over cost of effort, unless some participants get better tools than others?

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Filling Envelopes	Konow (2000)	Participants fold letters, stuff them into envelopes, and place them through a slot in a sealed box.	Yes	High	Low	Low	Can be difficult to find an appropriate reason to need envelopes stuffed. Task has been used many times with minor variations.
Sliders	Gill and Prowse (2011).	Participants are presented with "sliders" which they must click and drag to the center of a bar.	No	Low	Low	Low	Researcher has a high degree of control over the amount of effort.
Ball Catching	Gächter, Huang, and Sefton (2015)	Participants click a "left" or "right" to move a "tray" in order to catch balls on a screen that fall at fixed time intervals. The number of clicks and balls caught are recorded.	No	Low	Low	Low	Researcher has a high degree of control over the amount of effort and the cost of effort, as the cost per click can be easily manipulated.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Dragging a Ball on a Screen	Heyman and Ariely (2004)	Participants drag a ball across a screen, at which point it disappears and new one appears. Do as many as possible.	No	Low	Low	Low	Intuitively seems very frustrating, sine there is no discernable progress.
Typing Alternative Keys	Swenson (1988)	Participants receive some amount of income per keystroke, typing alternatingly "!" and the return key.	No	Low	Low	Low	Task might require very little attention, so cost of effort could be low.
Repeatedly Typing Paragraph	Dickinson (1999)	Participants exactly type out the same paragraph over and over.	No	Low	Low	Low	Requires more attention than above. Multiple dimensions of effort, as errors can be measured.
Decoding							

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Transcribing Greek Letters	Augenblick, Niederle, and Sprenger (2015)	A row of random and blurry Greek letters appears on a screen; the participant replicates it by clicking on a list of "greek" letters.	No	Low	Low	Low
Decoding Character Strings (Computer Cards)	Chow (1983)	Participants are given a set of pre-punched computer cards and a decoding key that they use to translate the card punches to a character string.	No	Med.	Low	Low

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Encoding 3-letter words into numbers	Erkal and Nikiforakis (2009)	Participants are given a table that codes unique numbers to each letter of alphabet, then is presented with a list of words and must convert the words into their numerical codes.	No	Low	Low	Low
Decoding a Number from a Letter Grid	Lévy-Garboua, Masclet, and Montmarquette (2009)	Participants are given a grid of letters and a decoding key, and they convert the letters into numbers.	No	Low	Low	Low

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Solving CAPTCHAs	McMahon (2015)	Participants solve as many CAPTCHAs (text distorted in a way so as to be unreadable to standard computerized text scanners) as possible within a given time period.	No	Low	Low	Low	
Cognitive							
Summing Large Matrices	Corgnet, Hernan-Gonzalez, and Rassenti (2011)	Participants are given 36 numbers in a matrix and must sum them. Notably, they did so for 100 minutes.in this experiment	No	Low	Med.	Low	Some participants may enjoy solving math problems, some might have math anxiety.
IQ Test	Gneezy, and Rustichini (2000)	Participants are presented with an IQ test and must provide correct answers.	No	Low	High	Low	Some participants may be intrinsically motivated to perform well.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Adding 2-Digit Numbers	Niederle and Vesterlund. (2007)	Participants add a series of 2-digit numbers in a given time period.	No	Low	Med.	Low	Some participants may enjoy solving math problems, some might have math anxiety.
Impossible Math Problem	Heyman and Ariely (2004)	Participants are given a grid of numbers and told they must select a group of numbers that add up to 100. However, task is impossible, as no combination of numbers does so. Effort is measured as the time spent on the task before giving up.	No	Low	Med.	High	
Miscellaneous							

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Door-to-Door Fundraising	Gneezy and Rustichini (2000)	Participants go door-to-door collecting donations for charitable causes.	Yes	High	High	Low	Might be confounded by individual preferences for the task or skill at fundraising.
Numerical Optimization (multi peaked)	Montmarquette et al. (2004)	No Participants search for the highest value of a one or more peaked function displayed in a two-dimensional space by clicking a button repeatedly or continuously to uncover the space. Different buttons uncover the space at different rates.	Low	Low	Low	High	control over the cost of effort by changing the costs attached to the buttons.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Numerical Optimization (single peaked)	Van Dijk, Sonnemans, and Van Winden (2001)	Participants search for the highest value of a single-peaked function displayed in a two-dimensional space by clicking a button repeatedly or continuously to uncover the space. Different buttons uncover the space at different rates.	No	Low	Low	Low	Less effort than above.
Creativity							

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Creating New Product Ideas	Girotra, Terwiesch, and Ulrich (2010)	Participants are tasked with identifying new product concepts for a given market, e.g., sporting goods that might be sold to students.	No*	Low	Med.	Low	Easy to implement and explain. Several degrees of researcher freedom with regards to quantity vs. quantity of creative effort. Some participants may have skills such as marketing or business experience that could be confounds. Fairly realistic.
Creating Words from Letter Sets (short)	Charness and Villeval. (2007).	Participants are given a set of 7 letters and must form as many words as possible within a given time period.	No	Low	Med.	Low	Easy to implement and explain. Easy to judge output. Some participants may enjoy the task, so less control over cost of effort. Not so realistic.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Expressing words with Materials	Laske and Schröder (2017)	Participants given set of materials consisting of a string, 2 O-rings, 4 wooden sticks, and 12-colored glass pebbles, and must construct representations of words with the materials.	No	High	Med.	Low	
Writing a story	Charness and Grieco (2012)	Participants wrote a story about a future city, something they would like to invent, or using specified words.	No	Low	Med.	Low	Easy to implement. Quality is subjective.

Table 5.2. Some real-effort experiments, continued. *Asterisks indicate that the task is not necessarily productive, but could be modified in a straightforward manner to make the effort useful.

Using combinations of math operations	Charness and Grieco (2012)	Participants were given a number and designated a series of math operations to transform it to another designated number.	No	Low	Low	Low	Easy to implement. Quality is subjective.

5.5 Practical Differences between Stated effort and Real Effort

In this section we mention some dimensions on which stated effort and real effort differ in practical terms. To the extent that realism is an important characteristic (perhaps for external validity in labor settings), these considerations tend to favor real-effort designs.[1] Of course, one must consider the trade-off between the value of knowing the effort cost and the heterogeneity involved with actual task performance. We discuss the timing of decisions, planned actions versus actual behavior, and differences between time and money.

5.5.1 Timing of decisions

While stated effort is a one-time decision (even in a repeated game, it is so in each period), real effort is a dynamic process in which the participant may change their effort while performing the task. In the stated-effort paradigm, each participant typically makes one immediate decision when choosing an effort level. In real-effort experiments, which naturally occur over time, the impact of the treatment on effort exertion is not always the same throughout.

[1] There are some topics, such as creativity, in which it is precisely the behaviorally-interesting particularities around the topic that a stated-effort design would miss, and so it seems to us that using a real-effort design is necessary.

Effort levels over a period of hours may vary in a way that can drastically change the conclusions. Effects that appear consistently in a short-term setting may or may not ultimately produce changes in effort levels in a setting with more duration. In this section we discuss several studies that look at effort exertion over time and suggest reasons for why effort levels might change over time, including learning, shifting emotional states, and limitations in sustained effort expenditure.

A plausible explanation of variation in effort levels over time, particularly the existence of sometimes-temporary shifts in effort, is related to hot-versus-cold decision-making (Loewenstein and Schkade, 1999; Loewenstein, 2005). For example, participants who have just received a gift may feel a transient "rush" of gratitude that impels them to reciprocate. Once this rush fades, so does the increased effort. For example, participants in Gneezy and List (2006) were recruited to perform data entry with an advertised hourly wage rate for six hours of work, split into two 3-hour sessions separated by a lunch break. Before beginning work, participants in one treatment were surprised with a higher-than-expected wage. The immediate response was an increase in effort relative to a control group that received the advertised wage. However, the effort waned after the lunch break, eventually reaching the same level for both treatments. Gneezy and List (2006) found similar results in a door-to-door fund raising experiment.

Several questions about this interpretation remain: how long do specific emotional states continue to influence behavior? Is the emotional effect present only in the short-term, as in Gneezy and Imas (2014) who show how anger can affect strategic behavior in the short run, but that this effect vanishes after a ten-minute cooling-off period? Yet it is possible that some emotional states are strong enough to push effort levels for long periods of time. For example, negative reciprocity, which seems to produce stronger psychological effects than positive reciprocity (Offerman, 2002; Charness, 2004), produce a permanent shift in effort expenditure? Kube, Maréchal, and Puppe (2013) extend the Gneezy and List (2006) design to include a negative wage surprise, finding a persistent and significant negative reciprocity as measured by the decrease in effort. The absolute magnitude of the decrease in effort relative

to the control treatment was twice as large as the highest gap between the positive and control treatments.

Hennig-Schmidt, Sadrieh, and Rockenbach (2010) investigate the fair wage-effort hypothesis over time, as well as in a more typical short-term laboratory setting using a real-effort task. In the longer setting, participants performed a data-entry task in two discrete one-hour sessions separated by a month, with an expected show-up fee and hourly wage. When participants arrived for the second session, they were either paid the expected wage or given a pay increase of 10% or 40%. When participants believed they were providing a surplus to the employer, a wage increase significantly increased effort, and when they did not, a wage increase had no effect.

Kosfeld and Neckermann (2011) examine the impact of a non-monetary reward on effort in a data-entry setting wherein participants received a fixed wage for two hours of work. The possibility of a non-monetary reward given to the best performer, in this case a signed card from the director of the organization benefitting from the data-entry task, led to sustained levels of higher effort, a result driven by a small number of highly-productive workers who stood a reasonable shot of winning the prize.

The conclusion from this section is clear: researchers should consider whether time is relevant to the research question. Longer-term experiments can help in capturing aspects of the decision making process. If the duration of the behavioral response to a stimulus is relevant to the theoretical importance of a phenomenon, then the experimental methodology should reflect this. Emotional states are particularly likely to change over specific events, such as eating a meal or sleeping. Such events can "reset" emotional states, rapidly accelerating the pace at which the impact of an emotional state on actions decreases. Further research in this domain might look at effort levels over increasingly long durations, potentially through field experiments in real workplaces where participant behavior can be measured over time.

Similarly, further research is also needed on the persistence of stated-effort decisions over time. An experiment could perhaps test whether there is any difference in stated-effort decisions over a time span in which there has been found to be significant changes in real-effort provision.

5.5.2 Planned actions versus actual behavior

Even when individuals have a strong intention to meet goals they have decided to pursue, they may fail to do so because they do not effectively deal with self-regulatory problems – goal striving may not be enough by itself (see Gollwitzer & Sheeran, 2006 for a review). Individuals allocate their effort by planning future effort (e.g., setting a goal) and by exerting current effort. Effort planning involves scheduling future behavior, whether in informal circumstances such as planning a gym routine or a study session, or in more formal circumstances such as creating a work schedule or negotiating a contract. A plan to allocate effort in the future is purposeful; for example, a gym routine is planned because the expected effort cost from going to the gym is outweighed by the expected benefits of improved health.

Buser and Peter (2012) find that people have problems with scheduling when required to perform multiple tasks (Sudoku and Word Search). In three treatments, participants either are required to perform these tasks sequentially, required to multi-task, or they can organize and schedule the work as desired. People who were required to multi-task perform significantly worse than those who were required to work sequentially. It is interesting that participants who were allowed to create their own schedule also perform significantly worse, suggesting that scheduling is an important aspect of productivity. A final result goes against the stereotype that females are better at multi-tasking than males, since their performance is reduced by just as much as men when required to multi-task and are even *less* likely to multi-task when free to choose.[2]

When individuals consider exerting current effort, pressures outside of goal-seeking may also be in force. For example, the unpleasantness of actually exercising in a gym may discourage one from following through with his or her plan. The goal-oriented valuation that drove initial goal-setting competes with more myopic valuation systems for influence over behavior at the

[2]On the (emerging) topic of multi-tasking, Offerman and van der Veen (2015) create a dual-task environment in which one task involved making public-good contributions and the other involved keeping a randomly-moving red dot inside a box on the screen. They consider how people react to either a slow or quick increase of a subsidy for contributions to the public good. With the dual task, people seem to fail to react to a series of small changes in the decision problem.

time of action.

The economics literature in domains such as savings and healthy behavior describes the difficulties people experience with following through on plans. Individuals state a preference for saving or exercising, but often fail to follow through. An outgrowth of these observations is the development of commitment devices and other behavioral tools or strategies, including social incentives, which align an individual's future incentives with their current incentives (Thaler and Benartzi, 2004; Ashraf, Karlan, and Yin, 2006; Kast, Meier, and Pomeranz, 2012).

When designing an experiment, it is appropriate to consider the degree of difficulty individuals might have in following up on their planned behavior. Stated effort may measure the desire to attain a goal, such as winning a tournament (e.g. Müller and Schotter, 2010). However, stated effort may fail to be predictive of actual effort. When using existing research to make predictions about external phenomena, the interpretation of results from either stated-effort or real-effort should be carefully considered in the light of whether desire is likely to translate into action.

5.5.3 Differences between effort and money

Individuals may not always behave similarly when making decisions over money and effort. We consider three empirical patterns here for their relevance to selecting a methodology or interpreting results from effort experiments: individuals can exhibit a preference for donating effort rather than money in charitable giving; exhibit differently-shaped time preferences over money and effort; and money can crowd out motivation from other sources and change the nature of a social interaction. As the stated-effort task is fundamentally a decision over money, a concern is that some divergences might exist between results obtained from stated-effort and real-effort tasks in these domains. Developing an encompassing theory to explain why decisions over money and effort may not always be equivalent is beyond the scope of this paper. We limit ourselves to simply presenting these patterns along with examples from oft-studied domains. This should help researchers become aware of these behavioral differences and hopefully lead to

useful formal models.

That individuals can exhibit a preference for effort exertion over monetary donation is sometimes referred to as the "volunteering puzzle" (Handy and Katz, 2008).[3] Consider a lawyer who volunteers in a soup kitchen. If the lawyer's goal is to maximize the amount of food served at the soup kitchen, then spending an hour working at her occupation and then donating the wages to the soup kitchen is much more effective than working at the kitchen. The donated money could be used to employ several lower-skilled workers in her place.

A potential driver of donations of effort or time rather than money may be differences in the warm-glow (Andreoni 1990) attained from donating. Individuals may derive utility from effort for reasons of social signaling, self-image, or the pleasure of performing the task itself. Andreoni et al. (1996) present a model wherein individuals derive utility from donations that is separable over money and effort, which has been supported by subsequent experimental evidence. Brown et al. (2013) find that participants in a laboratory experiment are more likely to donate and donate more when they can work directly for a charity rather than work for themselves and later donate to the charity. Even when participants in this study could freely toggle between working for themselves and working for charity and wages for self were 33% higher, they still give substantially more time to charity. It seems that donations of effort are more motivated by private warm-glow than monetary donations.

Comparing time preferences over money and effort, Augenblick, Niederle, and Sprenger (2015) measure the shape of time preferences over money and consumption, operationalized as a period of time that must be spent working on a boring task. They find no evidence for present bias in money, but do find evidence for present bias in consumption in their two experiments. Additionally, participants exhibited a demand for commitment devices for effort, but not money. Money appears to be fungible between time periods, while effort does not. Bisin and Hyndman (2014) find present-bias over real effort in a field experiment in which students must complete

[3] In 2015 there were over 7.9 billion volunteer hours provided by 62.6 million volunteers in the United States for an estimated value of $184 billion. (Corporation for National & Community Service, 2015), and the estimated value of monetary charitable donations was over $358 billion in 2014. (Giving USA, 2015).

tasks by a fixed deadline, and further find that demand for a self-imposed commitment device is stronger in students who describe themselves as less conscientious, indicating that they are sophisticated about their time-preferences.

Money and effort are not always interchangeable in social interactions. Introducing monetary exchanges into a social interaction can change the character of the social interaction, potentially crowding out other incentives. Consider asking your friend to come over and help you move your sofa to your new home. Paying your friend $20 at the end would seem odd, but telling him that you will be happy to help him whenever he will need help or buying him dinner would not. Gneezy and Rustichini (2000) and Heyman and Ariely (2004) find that low levels of monetary compensation can produce less effort than no monetary compensation. If one wishes to study a social interaction that is often denominated in terms of effort, using money as the currency of exchange may crowd out key factors relevant to decision making.

When designing an experiment, researchers should consider whether in the domain they study effort and money are interchangeable. As there does not yet appear to be data that describes the degree of interchangeability in many domains, further research that facilitates such comparisons would helpful for making inferences based on laboratory experiments and deciding which methodology is most appropriate.

5.5.4 Comparative Studies

We found only few empirical investigations that directly compare results with parallel methodologies—both stated-effort and real-effort that are applied to the same treatment effect or decision-making environment. The treatment effect or environment itself is not chosen specifically as a test of comparability. Three studies (Brüggen and Strobel, 2007; Charness, Cobo-Reyes, Lacomba, Lagos, and Perez, 2016; Dutcher, Salmon, and Saral, 2015) find general equivalence between the methodologies in the environment tested, and one study (Lezzi, Fleming, and Zizzo, 2015) finds significant differences between the results obtained from the stated-effort task and several real-effort tasks.

Brüggen and Strobel (2007) used a gift-exchange game, with participants responding to a monetary transfer by either solving as many math problems as possible in five minutes or by selecting an effort level. There was evidence of positive reciprocation in both treatments, with higher average earnings and greater variance in the real-effort treatment. Charness, Cobo-Reyes, Lacomba, Lagos, and Perez (2016) investigated the role of social comparisons (both for wages and wage-decision rights) on workers' performance. The main treatments involved stated effort, but an additional treatment featured an adding-numbers task. They find qualitatively similar results from both paradigms, with quantitatively similar earnings.

Dutcher, Salmon, and Saral (2015) use a repeated public-goods setup with three treatments ("useful effort," "trivial effort," and "stated effort"). Participants were matched into groups of four for multiple periods; in each period they could either contribute money to the group fund (earning \$0.40 for the group per unit) or keep the currency in an individual fund (earning \$0.20 per unit). In the "useful effort" treatment it was made clear that the data entry contributed to a research project, whereas in the "trivial effort" treatment subjects were not given any context for the task. The maximum amount of data entry per period was capped at 10 lines, and in each period of the stated-effort treatment participants were given 10 tokens, allowing participants in all treatments to have access to a comparable number of tokens. There was no difference across treatments for either average contributions or trends in contributions.

Gächter, Huang, and Sefton (2015) used a computerized ball-catching task, in which participants move a slider at some cost to catch balls dropping randomly from the top of a screen. The argument is that performing an activity, even one that requires almost no physical or mental effort, captures the relevant aspects of a real-effort task. They use this task to study team production, gift exchange, and effort tournament, and obtain results in line with stylized findings from previous studies which use stated effort.

Lezzi, Fleming, and Zizzo (2015) directly compare the relationship between effort exertion and anxiety, risk preferences, and gender across the slider task (Gill and Prowse, 2012), adding numbers (Niederle and Vesterlund, 2007), counting zeros (Abeler et al., 2011), and the

stated-effort task. Participants in each task competed in a two-person 10-round all-pay format tournament wherein the participant who exerted the highest effort in each round won the round. Men exerted higher effort than women in the slider task but not the other tasks, anxiety decreased performance in the counting-zeros task but not others, and risk aversion was positively associated with performance on the counting zeros task but not others. The authors conclude that the task specificity of their results indicates that researchers should be careful when generalizing their experimental findings. Conducting such experiments would provide valuable knowledge on the situations in which specific laboratory effort experiments make useful predictions according to the methods employed.

5.6 Conclusion

Designing experiments that test real-effort and stated-effort on some of the dimensions identified in this paper stand to help provide a stronger empirical basis for differentiating the situations where each methodology may be appropriate. An extension of this is to study whether different treatment effects observed in a laboratory context using stated- and real-effort map to equivalent differences in more realistic settings. That is, if there is a scenario in which it is found that stated-effort and real-effort methodologies produce different results, then testing which results more closely align with the field phenomena of interest would provide careful consideration of the settings to which we can expect stated effort or real effort to generalize.

Stated effort is quite useful in a variety of situations, particularly when one is interested in a task that can be done quickly and immediately. Knowing the cost of effort is critical in many cases, particularly when one is making social comparisons (usually regarding payoffs) or testing theory. But when the field setting involves sustained effort, as in most labor environments in the field, an experimental task involving real effort seems advisable for external validity. Since initial behavior may be driven by an early rush of emotion that fades quickly or since emotional or physical fatigue may well manifest over time, one must be careful when drawing

conclusions from quick monetary choices. While to some degree having multiple periods can simulate periods of work, this would not seem to have the same psychological feel.

While it is desirable to have tasks with duration and real effort in an environment where effort must be supplied for a period of time, it may not be clear ex ante how long the duration must be with real effort in the lab. If Gneezy and List (2006) had only tested behavior in the 3-hour morning session, they would have concluded that a high-wage surprise leads to higher real effort. Having a second 3-hour session after a lunch break was crucial for the interpretation that positive feelings about this higher wage fade over time, perhaps as the sense of surprise fades and one's sense of entitlement grows.

A big question for experimenters is whether one gets different results with real effort and stated effort. Direct comparisons between results with stated and real effort are still scarce to date, but in several cases the effects are qualitatively similar. The relationship between effort or production and the other party's earnings is clear in these cases and this seems necessary for this equivalence. In fact, Hennig-Schmidt, Rockenbach, and Sadrieh (2010) find that "explicit cost and surplus information that enables an exact calculation of an employer's surplus from the work contract is a crucial prerequisite for a positive wage-effort relation." Ideally, one would like a real-effort task where there is not much variance in production across ability, so that there is a reasonably strong connection between effort and observed production.

The methodology used to measure effort in the laboratory should be appropriate to the specific research question under consideration. We have identified several considerations researchers should make to pick a methodology that best suits their needs, including the timing of the effort decision, the duration of the phenomenon, and goal orientation. Further, we provide a review of real-effort tasks along with qualitative assessments of methodological and logistical attributes.

Perhaps the main added value of the stated-effort approach is allowing the researcher to connect the experimental results to theory. A researcher who wishes to learn whether behavior in an experiment is consistent with comparative-statics predictions of an equilibrium theory must

know the function that maps the costs of effort to production. On the other hand, the main added value of the real-effort approach is the better connection to the psychology of effort, since one must be cautious in interpreting levels of behavior with stated effort.

It would be nice to have more papers with realistic real-effort tasks. At the same time, more papers that test interesting theories with the stated-effort approach would also be quite welcome. Furthermore, even unrealistic real-effort experiments would be useful to the extent that they help us identify interesting psychological mechanisms. Our goal is not to suggest that one methodology is superior, since both approaches clearly have their merits. To be clear, we have used both forms of effort elicitation techniques in our own research (and we have different views on the relative merits). Rather, our goal was to highlight some of the relevant parameters that researcher should consider when designing their method for measuring effort.

5.7 Acknowledgments

Chapter 4, in full, is a reprint of material as it appears in the Journal of Economic Behavior and Organization 2018. Written permission has been granted by the co-authors for the use of this chapter.

5.8 References

1. Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. American Economic Review 101.2: 470–492.
2. Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. The Economic Journal 100.401: 464-477.
3. Andreoni, J., et al. (1996). Charitable contributions of time and money. Working Paper, University of Wisconsin-Madison.
4. Ariely, D., et al. Large stakes and big mistakes (2009). The Review of Economic Studies 76.2: 451-469.
5. Ashraf, N., Dean, K., & Yin, W.. (2006). Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines. The Quarterly Journal of Economics 121.2: 635-672.

6. Augenblick, N., Niederle, M., & Sprenger, C. (2015). Working over time: Dynamic inconsistency in real effort tasks. The Quarterly Journal of Economics 130.3: 1067-1115.

7. Baumeister, R.F. (1984). Choking under pressure: self-consciousness and paradoxical effects of incentives on skillful performance. Journal of Personality and Social Psychology 46.3: 610.

8. Bisin, A., & Hyndman, K. (2014). Present-bias, procrastination and deadlines in a field experiment. Working paper, National Bureau of Economic Research, No.w19874.

9. Bortolotti, S., Giovanna, D., & Ortmann, A. (2009). Exploring the effects of real effort in a weak-link experiment. Working Paper, University of Trento, No. 0901.

10. B., Jordi, Charness, G. , & Ellman, M. (2016). Let's talk: How communication affects contract design. Journal of the European Economic Association 14.4: 943-974.

11. Brandts, J., & Cooper, D.J. (2007). It's what you say, not what you pay: An experimental study of manager-employee relationships in overcoming coordination failure. Journal of the European Economic Association 5.6: 1223-1268.

12. Brown, M., Falk, A. & Fehr, E. (2004). Relational contracts and the nature of market interactions. Econometrica 72.3: 747-780.

13. Brown, A.L., Meer, J., & Williams, J.F. (2013). Why do people volunteer? An experimental analysis of preferences for time donations. No. w19066. National Bureau of Economic Research.

14. Brüggen, A., & Strobel, M. (2007). Real effort versus chosen effort in experiments. Economics Letters 96.2: 232-236.

15. Bull, C., Schotter, A. & Weigelt, K. (1987). Tournaments and piece rates: An experimental study. Journal of Political Economy 95.1: 1-33.

16. Buser, T., & Peter, N. (2012). Multitasking. Experimental Economics 15.4: 641-655.

17. Bushong, B., & Gagnon-Bartsch, T. (2016). Misattribution of Reference Dependence: Evidence from Real-Effort Experiments. Working Paper, Harvard University.

18. Charness, G. (2000). Responsibility and effort in an experimental labor market. Journal of Economic Behavior & Organization 42.3: 375-384.

19. Charness, G., & Dufwenberg, M. (2006). Promises and Partnership. Econometrica 74.6: 1579-1601.

20. Charness, G., Cobo-Reyes, R., Lacomba, J., Lagos, F., & Maria Perez, J. (2016). Social comparisons in wage delegation: Experimental evidence. Experimental Economics 19.2: 433-459.

21. Charness, G., & Gneezy, U. (2009). Incentives to exercise. Econometrica 77.3: 909-931.

22. Charness, G., & Kuhn, P. (2011). Lab labor: What can labor economists learn from the lab? Handbook of Labor Economics 4: 229-330.

23. Charness, G.,Cobo-Reyes, R. & Sánchez, Á. (2016). The effect of charitable giving on workers' performance: Experimental evidence. Journal of Economic Behavior and Organization 131.PA: 61-74.

24. Charness, G., & Villeval, M.C. (2009). Cooperation and competition in intergenerational experiments in the field and the laboratory. American Economic Review 99.3: 956-978.

25. Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. Experimental Economics 14.1: 47-83.

26. Chaudhuri, A., Schotter, A., & Sopher, B. (2009). Talking ourselves to efficiency: Coordination in inter-generational minimum effort games with private, almost common and common knowledge of advice. The Economic Journal 119.534: 91-122.

27. Chow, C. W. (1983). The effects of job standard tightness and compensation scheme on performance: An exploration. Accounting Review 58.4: 667-685.

28. Corgnet, B., Hernan-Gonzalez, R., & Rassenti, S. (2011). Real effort, real leisure and real-time supervision: Incentives and peer pressure in virtual organizations. Working Paper, No.11-05.

29. Corporation for National & Community Service. Trends and Highlights Overview. https://www.nationalservice.gov/vcla/national.

30. Devetag, G., & Ortmann, A. (2007). When and why? A critical survey on coordination failure in the laboratory. Experimental economics 10.3: 331-344.

31. Dickinson, D.L. (1999). An experimental examination of labor supply and work intensities. Journal of Labor Economics 17.4: 638-670.

32. Dutcher, G., Salmon, T., & Saral, K.J. (2015). Is "Real" Effort More Real? Working Paper.

33. Erkal, N., Gangadharan, L., & Nikiforakis, N. (2011). Relative earnings and giving in a real-effort experiment. The American Economic Review 101.7: 3330-3348.

34. Fahr, R., & Irlenbusch, B. (2000). Fairness as a constraint on trust in reciprocity: earned property rights in a reciprocal exchange experiment. Economics Letters 66.3: 275-282.

35. Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. The Quarterly Journal of Economics 108.2: 437-459.

36. Fehr, E., Gächter, S., & Kirchsteiger, G. (1997). Reciprocity as a contract enforcement device: Experimental evidence. Econometrica 65.4: 833-860.

37. Gächter, S., Huang, L., & Sefton, M. (2015). Combining "real effort" with induced effort costs: The ball-catching task. Experimental Economics 19.4: 687-712.

38. Gill, D., & Prowse, V.L. (2011). A novel computerized real effort task based on sliders. IZA Discussion Paper, No.5801.

39. Girotra, K., Terwiesch, C. & Ulrich, K.T. (2010). Idea generation and the quality of the best idea. Management Science 56.4: 591-605.

40. Giving USA (2015). Giving USA 2015: The Annual Report on Philanthropy for the Year 2014.

41. Glewwe, P., Nauman, I., & Kremer, M. (2010). Teacher incentives. American Economic Journal: Applied Economics 2.3: 205-227.

42. Gneezy, U., & Rustichini, A. (2004). Gender and competition at a young age. American Economic Review 94.2: 377-381.

43. Gneezy, U., & Rustichini, A. (2000). A fine is a price. The Journal of Legal Studies 29.1: 1-17.

44. Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. The Quarterly Journal of Economics 115.3: 791-810.

45. Gneezy, U., & List, J.A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. Econometrica 74.5: 1365-1384.

46. Gneezy, U., & Imas, A. (2014). Materazzi effect and the strategic use of anger in competitive interactions. Proceedings of the National Academy of Sciences 111.4: 1334-1337.

47. Gneezy, U., Niederle, M.,& Rustichini, A. (2003). Performance in competitive environments: Gender differences. The Quarterly Journal of Economics 118.3: 1049-1074.

48. Gollwitzer, P.M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. Advances in Experimental Social Psychology 38: 69-119.

49. Gross, T., Guo, C., & Charness, G. (2015). Merit pay and wage compression with productivity differences and uncertainty. Journal of Economic Behavior & Organization 117.C: 233-247.

50. Handy, F., & Katz, E. (2008). Donating behavior: If time is money, which to give? A preliminary analysis. Journal of Economic Studies 35.4: 323-332.

51. Hennig-Schmidt, H., Sadrieh, A., & Rockenbach, B. (2010). In search of workers' real effort reciprocity—a field and a laboratory experiment. Journal of the European Economic Association 8.4: 817-837.

52. Heyman, J., & Ariely, D. (2004). Effort for payment a tale of two markets. Psychological Science 15.11: 787-793.

53. Houy, N., Nicolai, J.P., & Villeval, M.C. (2016). Doing Your Best when Stakes are High? Theory and Experimental Evidence. IZA Discussion Paper, No.9766.

54. Imas, A. (2014). Working for the "warm glow": On the benefits and limits of prosocial incentives. Journal of Public Economics 114.C: 14-18.

55. Johnson, N. D., & Mislin, A.A. (2011). Trust games: A meta-analysis. Journal of Economic Psychology 32.5: 865-889.

56. Kast, F., Meier, S., & Pomeranz, D. (2012). Under-savers anonymous: Evidence on self-help groups and peer pressure as a savings commitment device. Working paper, National Bureau of Economic Research, No.w18417.

57. Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. American Economic Review 90.4: 1072-1091.

58. Kosfeld, M., & Neckermann, S. (2011). Getting more work for nothing? Symbolic awards and worker performance. American Economic Journal: Microeconomics 3.3: 86-99.

59. Kube, S., Maréchal, M.A., & Puppe, C. (2013). Do wage cuts damage work morale? Evidence from a natural field experiment. Journal of the European Economic Association 11.4: 853-870.

60. Laske, K. & Schröder, M. (2016). Quantity, quality, and originality: The effects of incentives on creativity. Working Paper, CGS, No.7–1.

61. Lévy-Garboua, L., Masclet, D., & Montmarquette, C. (2009). A behavioral Laffer curve: Emergence of a social norm of fairness in a real effort experiment. Journal of Economic Psychology 30.2: 147-161.

62. Lezzi, E., Fleming, P. & Zizzo, D.J. (2015). Does it matter which effort task you use? A comparison of four effort tasks when agents compete for a prize. Working Paper.

63. Loewenstein, G. (2005). Hot-cold empathy gaps and medical decision making. Health Psychology 24.4S: S49.

64. Loewenstein, G., & Schkade, D. (1999). Wouldn't it be nice? Predicting future feelings. Well-being: The foundations of hedonic psychology: 85-105.

65. McMahon, M. (2015). Better lucky than good: The role of information in other-regarding preferences. Working Paper.

66. Mohnen, A., Pokorny, K., & Sliwka, D. (2008). Transparency, inequity aversion, and the dynamics of peer pressure in teams: Theory and evidence. Journal of Labor Economics 26.4: 693-720.

67. Moldovanu, B., & Sela, A. (2001). The optimal allocation of prizes in contests. American Economic Review 91.3: 542-558.

68. Montmarquette, C., et al. (2004). Redesigning teams and incentives in a merger: An experiment with managers and students. Management Science 50.10: 1379-1389.

69. Müller, W., & Schotter, A. (2010). Workaholics and dropouts in organizations. Journal of the European Economic Association 8.4: 717-743.

70. Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? The Quarterly Journal of Economics 122.3: 1067-1101.

71. Offerman, T. (2002). Hurting hurts more than helping helps. European Economic Review 46.8: 1423-1437.

72. Offerman, T., & van der Veen, A. (2015). How to subsidize contributions to public goods: Does the frog jump out of the boiling water? European Economic Review 74.C: 96-108

73. Rutström, E. E., & Williams, M.B. (2000). Entitlements and fairness: An experimental study of distributive preferences. Journal of Economic Behavior & Organization 43.1: 75-89.

74. Smith, V. L. (1976). Experimental economics: Induced value theory. American Economic Review 66.2: 274-279.

75. Swenson, C.W. (1988). Taxpayer behavior in response to taxation: An experimental analysis. Journal of Accounting and Public Policy 7.1: 1-28.

76. Thaler, R.H., & Benartzi, S. (2004). Save more tomorrow™: Using behavioral economics to increase employee saving. Journal of Political Economy 112.S1: S164-S187.

77. Van Dijk, F., Sonnemans, J. & van Winden, F. (2001). Incentive systems in a real effort experiment. European Economic Review 45.2: 187-214.

78. Van Huyck, J., Battalio, R.,& Beil, R. (1990). Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure. American Economic Review 80.1: 234-248.