

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Experimental Evidence of Behavioral Responses to Uncertainty

Permalink

<https://escholarship.org/uc/item/58f383gk>

Author

Brownback, Andrew Paul

Publication Date

2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Experimental Evidence of Behavioral Responses to Uncertainty

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

in

Economics

by

Andrew Paul Brownback

Committee in charge:

Professor James Andreoni, Chair
Professor Uri Gneezy
Professor Paul Niehaus
Professor Yuval Rottenstreich
Professor Joel Watson

2015

Copyright
Andrew Paul Brownback, 2015
All rights reserved.

The dissertation of Andrew Paul Brownback is approved,
and it is acceptable in quality and form for publication
on microfilm and electronically:

Chair

University of California, San Diego

2015

DEDICATION

To Kimly,
Your love sustains me.

EPIGRAPH

Hold on to your butts.

—Jurassic Park

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Epigraph	v
Table of Contents	vi
List of Figures	ix
List of Tables	x
Acknowledgements	xi
Vita	xii
Abstract of the Dissertation	xiii
Chapter 1 All Pay Auctions and Group Size: Grading on a Curve and other Applications	1
1.1 Introduction	2
1.2 Theoretical Model	6
1.2.1 Definitions and procedures	7
1.2.2 Bidder's Utility	8
1.2.3 Equilibrium Bidding	8
1.3 Experimental Model	11
1.3.1 Bidding Functions	11
1.4 Experimental Hypotheses	14
1.4.1 Qualitative Predictions: The Designer's Ob- jectives	15
1.5 Experimental Procedures	17
1.5.1 Recruitment and Participation	17
1.5.2 Instructions	18
1.5.3 Auction Design	18
1.6 Results	20
1.6.1 Hypotheses	21
1.7 Discussion	29
1.8 Conclusions	31
1.9 Acknowledgement	32
1.10 Chapter 1 Appendix	33

	1.10.1 Proofs	33
	1.10.2 Individual Bids and Individual Differences	35
	1.10.3 Alternative Assumptions on Utility	45
	1.10.4 Subjects' Instructions	47
Chapter 2	A Field Experiment on Effort Allocation under Relative Grading	56
	2.1 Introduction	57
	2.2 Literature	65
	2.3 A Model of Academic Effort	68
	2.3.1 Student's Utility	70
	2.3.2 Predictions From the Model	74
	2.4 Experimental Design	77
	2.4.1 Recruitment and Participation	77
	2.4.2 Quiz Design, Scoring, and Randomization	77
	2.4.3 Effort	79
	2.4.4 Ability	80
	2.5 Results	80
	2.5.1 Data and Descriptive Statistics	81
	2.5.2 Dependent Variable	82
	2.5.3 Endogenous Selection and Controls	83
	2.5.4 Heterogeneity in the Treatment Effect	85
	2.5.5 Alternatives to the Neo-Classical Assumptions of the Model	89
	2.6 Discussion	98
	2.6.1 Patterns of Effort Allocation	100
	2.6.2 Misallocation of Effort	101
	2.6.3 Policy Prescriptions	103
	2.7 Conclusion	105
	2.8 Acknowledgement	106
	2.9 Chapter 2 Appendix	107
	2.9.1 Experimental Procedures	107
	2.9.2 Theory	108
	2.9.3 Empirical Results	110
Chapter 3	On the Elicitation of Willingness to Pay for Stigmatized Goods	112
	3.1 Introduction	113
	3.2 Literature	118
	3.3 Model	123
	3.4 Experimental Design	133

3.4.1	Item Selection	135
3.4.2	WTP Elicitation	139
3.5	Conclusion	141
3.6	Acknowledgement	142
3.7	Chapter 3 Appendix	143
3.7.1	Proof of Lemma 1	143
3.7.2	Proof of Lemma 2	144
3.7.3	Proof of Proposition 2	144
3.7.4	Full Survey Results	145
3.7.5	Relationship between Stigma and Stated WTP	145
3.7.6	Auto-Filling MPL Mechanism	147

LIST OF FIGURES

Figure 1.1:	Optimal Bid Functions	12
Figure 1.2:	Aggregate Bidding	21
Figure 1.3:	Bidding Differences	22
Figure 1.4:	Predicted Difference in Bids	25
Figure 1.5:	2-Person Auction Bids, Session 1	36
Figure 1.6:	20-Person Auction Bids, Session 1	37
Figure 1.7:	Difference in Bids, Session 1	38
Figure 1.8:	2-Person Auction Bids, Session 2	39
Figure 1.9:	20-Person Auction Bids, Session 2	40
Figure 1.10:	Difference in Bids, Session 2	41
Figure 1.11:	2-Person Auction Bids, Session 3	42
Figure 1.12:	20-Person Auction Bids, Session 3	43
Figure 1.13:	Difference in Bids, Session 3	44
Figure 1.14:	Risk Averse Bidding Equilibrium	45
Figure 1.15:	Joy of Winning Equilibrium	47
Figure 2.1:	Probability a Given Ability is in the Top 70% of Abilities in a Cohort of Size N	71
Figure 2.2:	Scores at Equilibrium in the 10- and 100-Student Quizzes	73
Figure 2.3:	Difference in Score between 100- and 10-Student Quizzes at Equilibrium	75
Figure 2.4:	Cumulative Distribution Function for GPAs of All Students in the Experiment	81
Figure 2.5:	Local Polynomial Fit of the Difference in Effort between 100- and 10-Student Quizzes (Times Reported in Seconds)	87
Figure 2.6:	Perceived and Actual Percentile Ranks of Students	92
Figure 2.7:	Predicted Difference in Effort Based on Perceived Rank	93
Figure 2.8:	Seconds spent on each quiz and aggregate seconds spent on quizzes by GPA	95
Figure 2.9:	Duration of each quiz separated by risk preference (in seconds)	97
Figure 2.10:	Local Polynomial Fit of Difference in Effort between 100- and 10-Student Quizzes by Gender (Times Reported in Sec- onds)	98
Figure 3.1:	Example of Auto-Filling MPL Mechanism (Unclicked)	147
Figure 3.2:	Example of Auto-Filling MPL Mechanism (Clicked)	148

LIST OF TABLES

Table 1.1:	The difference in bids across different valuations	26
Table 1.2:	Measuring the Well-Ordering of Bids	27
Table 1.3:	The Well-Ordering of Bids for Expected Winners and Losers	28
Table 2.1:	Submitted Quizzes from Each Week and Each Treatment .	78
Table 2.2:	Descriptive Statistics on the Duration of the Quizzes	82
Table 2.3:	Testing the Relevance and Validity of the Instrument . . .	84
Table 2.4:	Mean Difference between 100- and 10-Student Quiz Duration	85
Table 2.5:	Difference between 100- and 10-Student Quiz Duration by Ability	86
Table 2.6:	100-Student Quiz Duration Minus 10-Student Quiz Duration	110
Table 2.7:	100-Student Quiz Duration Minus 10-Student Quiz Duration	110
Table 2.8:	Stated Preference for the 10-Student Quiz	111
Table 2.9:	Scores on 100-St Quiz minus Scores on 10-St Quiz	111
Table 3.1:	Summary Statistics for Stigmatized and Neutral Goods . .	136
Table 3.2:	Table of Stigmatized and Neutral Goods Selected for Exper- iment	137
Table 3.3:	Summary Statistics of Sentiment towards Candidate Goods	146
Table 3.4:	Regression of WTP on Reported Stigma	147

ACKNOWLEDGEMENTS

There are too many people to thank in this process. Here is a cursory attempt that will no doubt leave out many deserving people: Vinayak Alladi, Mitch Downey, Simone Galperti, Laura Gee, Matthew Gibson, Sarojini Hirshleifer, Vincent Leah-Martin, Marta Serra-Garcia, Joel Sobel, Charles Sprenger, and Isabel Trevino all offered wonderful advice on my first and second chapters. Michael Kuhn responded to countless frantic phone calls, emails, and text messages with guidance and (importantly) comfort. Tristan Gagnon-Bartsch and Shengwu Li deserve so much credit for making the third chapter of this dissertation happen. Tristan is due particular thanks for working under tight time pressure. A special thanks to my dissertation committee members: Joel Watson, Paul Niehaus, Uri Gneezy, and Yuval Rottenstreich. Jim Andreoni taught me, guided me, and comforted me throughout this process. I owe him so much. Finally, my wonderful fiancé, Kimly, and my parents, Mary and Sam, each took turns holding my hand, literally and figuratively.

Chapter 1, in part is currently being prepared for submission for publication of the material. Brownback, Andrew; Andreoni, James. The dissertation author was the primary investigator and author of this material.

Chapter 2, in part is currently being prepared for submission for publication of the material. Brownback, Andrew. The dissertation author was the primary investigator and author of this material.

Chapter 3, in part is currently being prepared for submission for publication of the material. Brownback, Andrew; Gagnon-Bartsch, Tristan; Li, Shengwu. The dissertation author was the primary investigator and author of this material.

VITA

- | | |
|------|---|
| 2010 | B.A. in Mathematics and Economics <i>summa cum laude</i> ,
Kansas State University |
| 2015 | Ph.D. in Economics, University of California, San Diego |

ABSTRACT OF THE DISSERTATION

Experimental Evidence of Behavioral Responses to Uncertainty

by

Andrew Paul Brownback

Doctor of Philosophy in Economics

University of California, San Diego, 2015

Professor James Andreoni, Chair

This dissertation looks at how decisions are affected by uncertainty. The first chapter examines the effect of uncertainty on costly effort in a strategic environment. The second chapter asks a similar question in a field setting, looking at student responses to the uncertainty that is brought about by smaller classes. The third chapter explores the role of uncertainty on self-signaling by looking at difference in demand for stigmatized goods under weak or strong self-signals.

Chapter 1

All Pay Auctions and Group Size: Grading on a Curve and other Applications

1.1 Introduction

Consider a student deciding how much effort to devote to a course graded on the curve. Numerous factors come into play, among them are her innate ability, the abilities of her classmates, and the percentage of students she believes will receive each grade. That is, a portion of the student's effort choice is strategic. This paper explores one element of that strategic choice: the enrollment of the course. Under asymmetric information, the enrollment of the course affects the information the student has about the distribution of her classmates. Specifically, the law of large numbers implies that the larger is the sample of students drawn from a population into a given class, the closer that class will come to reflecting the population's distributional characteristics, affecting her optimal choice of effort.

In order to capture the nature of this environment with costly effort and probabilistically awarded prizes, we will model this environment as an all-pay auction. Here, the number of bidders in the auction represents the enrollment of the course, bids represent the amount of effort each student chooses to exert in studying, and the number of prizes in the auction represents the number of "A" grades. As the enrollment of the course shrinks, the uncertainty about the ability distribution of abilities among its students increases. This uncertainty negatively affects high-skilled students who are more likely to face a draw of students containing several outliers with high abilities, while benefiting low-skilled students who are more likely to receive a draw of classmates containing several outliers with low abilities.

Tversky and Kahneman (1971) and Kahneman and Tversky (1973)

have shown that this type of inference from sample size is difficult for experimental subjects, meaning that observing adjustments in bidding to the size of the contest are not certain. We seek, therefore, to understand the degree to which subjects are able to draw inference about their competitors based on sample size and the degree to which their bids reflect this changing inference.

Specifically, we will attempt to uncover the consequences of changes in the size of a contest on effort exerted by participants. We will show here, both theoretically and through an experimental test, that there are significant effects of a contest's size on aggregate effort, individual effort, and the ranking of effort levels even when the proportion of awards to participants is fixed. Moreover, we will show that changes in the contest size heterogeneously affect individuals of different ability levels. These results will be of particular value to mechanism designers by informing them of the unintended consequences of implementing relative awarding schemes when the size of the contest may change.

To fix ideas, we will discuss the implications of this study in the context of grading on a curve, but these results apply to any setting where performance is evaluated relative to a peer group. For example, promotion contests where firm size may change, research and development contests where markets may grow or shrink, or rent-seeking when the size or divisibility of rents may shift.

Our contribution to the literature is not in being the first to explore the effect of changes in contest size on behavior, but in being the first to look at its heterogeneous impacts across bidders of different abilities. Our paper relies on a theoretical framework similar to the existing works of Moldovanu and Sela (2001), Moldovanu and Sela (2006), and Olszewski and Siegel (2013), who each

provide a more complete theoretical treatment of the impact of contest size on performance by contestants. In the experimental literature, Harbring and Irlenbusch (2005) and Orrison, Schotter, and Weigelt (2004) provide the most similar studies to ours, both experimentally test the effect of small changes in the size of rank-order tournaments on behavior of contestants while holding constant the proportion of winners. What distinguishes our study from these is that we focus on the heterogeneity of impacts across a continuum of types, we employ an all-pay auction format, our analysis is within-subjects, and we study changes in size of a much larger scale. Harbring and Irlenbusch (2005) employ homogeneous types, and Orrison et. al (2004) only address heterogeneity among two types of players, one type advantaged by the addition of a constant to their output.

Our investigation of the heterogeneous impact of changes in contest design is a natural extension of the literature. Focusing on average effects will obscure the tension present between the interests of the high- and low-skilled contestants. We identify this tension theoretically, and then confirm its existence in our experimental results. In addition to studying the heterogeneity of responses to changes in contest structure, we offer the first test of large-scale changes in the size of the contest and a powerful design that cleanly identifies changes in behavior between contests of different sizes.

There is an extensive theoretical literature on tournaments and contests dating back to Tullock (1967) and Krueger (1974). Lazear and Rosen (1981) explore rank-order tournaments as labor contracts, while Becker and Rosen (1992) use the same setting to model grading on the curve. Hillman and Riley (1989) and Baye, Kovenock, and de Vries (1993) model political rent-seeking

as a form of all-pay auction. Baye, Kovenock, and de Vries (1996) provide the full characterization of the equilibria in the complete information setting. Amann and Leininger (1996) and Krishna and Morgan (1997) expand this model to incorporate incomplete and possibly correlated information.

Other experimental papers have studied similar changes in the structure of contests, nearly all of which can be found in a comprehensive review by Dechenaux, Kovenock, and Sheremeta (2012). Potters, de Vries, and van Winden (1998) and Davis and Reilly (1998) perform early laboratory tests of different contests. Potters et al (1998) finds that behavior fits equilibrium predictions well when allowing for mistakes. On the other hand, Davis and Reilly (1998) finds over-dissipation of rents—that is, over-bidding—but tests this in an environment where prizes can be “defended.”

In addition to the previously mentioned Harbring and Irlenbusch (2005) and Orrison et. al (2004) papers, Barut, Kovenock, and Noussair (2002), Gneezy and Smorodinsky (2006), and Lim, Matros, and Turocy (2014) all look at the effect of changes in the proportion of prizes to contestants, the first varies the number of prizes awarded, while the second and third vary the number of contestants. Müller and Schotter (2010) and Noussair and Silver (2006) focus on heterogeneous behavior in contests with a fixed size. Both papers identify and explore the stark “bifurcation” of behavior between low- and high-ability types, a term coined in the former paper. Minor (2012) evaluates this “coarse” behavior in greater depth.

We hope that by addressing the effect of contest size on behavior, we can inform policy makers about the heterogeneous impact that relative evaluation policies will have in different environments. In particular, we hope to shed

light in the effect that grading on the curve will have on student effort in classes with different enrollment numbers.

1.2 Theoretical Model

In our all-pay auction framework, each bidder is given an independent private value of winning a prize drawn from a known distribution in the style of Vickrey (1961). Bidders then simultaneously place irreversible bids for one of a limited number of prizes. Bidders must pay these bids regardless of the outcome of the auction. Since prizes are awarded to the highest bidders, an increase in the bid weakly increases the probability of receiving a prize.

These environmental characteristics nicely match not only the original motivating examples of competitive rent-seeking and research and development contests, but other competitive environments as well such as a course graded on the curve, job promotions, the allocation of bonuses within a firm, or the allocation of grants to applicants. Indeed, this framework of analysis can be applied to any environment with costly, deterministic effort and probabilistically awarded prizes.

Since bidders face asymmetric information about the valuations that their competitors have for winning prizes, the number of competitors has a meaningful effect on the inference a bidder can draw about the distribution of valuations among his competitors. Specifically, the law of large numbers implies that as his draw of competitors grows, their distribution of valuations draws closer to the population distribution. This gives the bidder less uncertainty about the relative ranking of his valuation among his peers. Import-

tantly, increasing levels of uncertainty affect bidders heterogeneously based on their own valuations. Uncertainty benefits low-valuation bidders as it increases the likelihood that their distribution of competitors has a high percentage of outliers with low valuations. Similarly, uncertainty is detrimental to high-valuation bidders as it increases the likelihood that they face a draw with a high percentage of high-valuation outliers.

This sensitivity to the size of the auction gives rise to several interesting questions: What are the independent effects of the number of bidders on the effort chosen by each bidder? Is the effect heterogeneous depending on the bidder's valuation? How well does the prize allocation reflect the true relative ranking of bidders' valuations when only bids are observable? Translated, does the course enrollment affect a student's effort in studying? Does it affect high-skilled students differently from low-skilled students? Is there an enrollment level where effort choices are more or less reflective of students' abilities? For each of these questions we will provide a theoretical prediction and a result from a controlled laboratory study.

1.2.1 Definitions and procedures

Suppose there are N bidders competing for M prizes, with $M < N$. For simplicity, call the number of participants the size of the contest. Each bidder receives an independent private value of winning a prize, v_i , drawn from a known distribution, $F(v_i)$. These values represent the surplus received upon winning a prize. Thus, the framework is equivalent to imposing independent private costs of effort with higher valuation representing lower costs of effort. Effort is measured by the bid value, b_i , and is costly regardless of the outcome

of the auction. Auctions are one-shot, and bids are cast simultaneously, so bidders have no ability to condition their effort on other bids. Let $P_{N,M}(b_i)$ be the probability that bid b_i will win a prize in an auction with N participants and M prizes.

1.2.2 Bidder's Utility

For simplicity, we assume risk neutrality. We show in the appendix that the qualitative results generalize to risk aversion.¹ A bidder's utility is:

$$\begin{aligned} U(b_i; v_i, N, M) &= P_{N,M}(b_i)(v_i - b_i) + (1 - P_{N,M}(b_i))(-b_i), \\ &= v_i \times P_{N,M}(b_i) - b_i \end{aligned} \tag{1.1}$$

In this auction, the M bidders with the highest bids will each receive one of the prizes. Therefore the probability of receiving a prize is weakly increasing in the bid value.

1.2.3 Equilibrium Bidding

With incomplete information about the valuations of other bidders, we can generate a symmetric Bayesian Nash Equilibrium in pure strategies. Since the information varies with the size of the auction, the equilibrium bidding will also vary with the size of the auction. Suppose that there exists a monotonic equilibrium bidding function, $B(v_i; N, M)$, that maps from the valuation space onto the bidding space: $B : v_i \rightarrow b_i$, where $v_i \in [0, 1]$ and $b_i \in [0, 1]$.²

¹The generalization, however, does require common knowledge of risk aversion and shared risk aversion parameters.

²Monotonicity is a standard assumption for symmetric bidding functions (see Krishna and Morgan, 1997). It also satisfies intuition, as subjects likely understand that higher-

Under monotonicity, there exists a probability function that maps valuations to the equilibrium probabilities of winning an auction. Define this function $Z_{N,M} : v_i \rightarrow [0, 1]$. This function will represent the probability that a bidder's valuation is higher than the valuations of at least $N - M$ of the opposing bidders. Mathematically, this is expressed in the form of an order-statistic:

$$Z_{N,M}(v_i) = \sum_{k=N-M}^{N-1} \left(\frac{(N-1)!}{(N-1-k)!k!} \right) F(v_i)^k (1 - F(v_i))^{N-1-k}, \quad (1.2)$$

where $F(v_i)$ is the cumulative distribution function of the valuations.

So far, we have assumed that a symmetric bidding function, $B : v_i \rightarrow \mathbb{R}^+$, exists, is well-defined, and monotonic. Continuity of the bidding function follows from monotonicity by observing that any discontinuities could not represent equilibrium bidding as profitable deviations exist by jumping to just above the lower bound of the discontinuity. Continuous, monotone functions are invertible, implying that there exists a function, $B^{-1}(b_i)$, that maps bids cast onto the valuations implied by those bids. Denote this inverse function $V(b_i)$.

To demonstrate the optimality of the candidate equilibrium bidding function, the bidding function must solve the first order condition of the bidder's utility,

valuation bidders will cast higher bids.

$$U(b_i; v_i) = v_i \times \sum_{k=N-M}^{N-1} \left(\frac{(N-1)!}{(N-1-k)!k!} \right) F(V(b_i))^k (1 - F(V(b_i)))^{N-1-k} - b_i, \quad (1.3)$$

where incorporating $Z_{N,M}(V(b_i))$ captures the bidder's incentive to deviate from equilibrium by bidding an amount higher or lower than what the equilibrium bidding function would prescribe for his valuation. Differentiating (1.3) with respect to b_i and rearranging yields

$$\begin{aligned} \frac{1}{V'(b_i)} = & v_i \times \sum_{k=N-M}^{N-1} \left\{ \left(\frac{(N-1)!}{(N-1-k)!k!} \right) \times \right. \\ & \left[k \times f(V(b_i)) F(V(b_i))^{k-1} (1 - F(V(b_i)))^{N-1-k} - \right. \\ & \left. \left. (N-1-k) \times f(V(b_i)) F(V(b_i))^k (1 - F(V(b_i)))^{N-2-k} \right] \right\}. \end{aligned}$$

Since $V(b_i) \equiv B^{-1}(b_i)$, it follows that, $\frac{1}{V'(b_i)} = B'(v_i)$. Additionally, at equilibrium, $Z_{N,M}(V(b_i)) = Z_{N,M}(v_i)$. We thus arrive at a first order differential equation from which we can solve the general form of the equilibrium bidding function,

$$\begin{aligned} B'(v_i) = & v_i \times \sum_{k=N-M}^{N-1} \left\{ \left(\frac{(N-1)!}{(N-1-k)!k!} \right) \times \right. \quad (1.4) \\ & \left[k \times f(v_i) F(v_i)^{k-1} (1 - F(v_i))^{N-1-k} \right. \\ & \left. \left. - (N-1-k) \times f(v_i) F(v_i)^k (1 - F(v_i))^{N-2-k} \right] \right\}. \end{aligned}$$

Prior to solving this differential equation, we must impose an initial

condition where $B(0) = 0$. This is natural, since negative bids are not allowed, and any positive bid would guarantee a negative surplus from the auction.

1.3 Experimental Model

Equation (1.4) provides general results and implications for any values of N and M and cumulative distribution $F(v_i)$. In our experiment, we will be testing the following two pairs: $(N, M) = (2, 1)$ and $(20, 10)$ with valuations uniformly distributed, $v_i \sim U[0, 1]$. We will now proceed to derive closed-form solutions for the bidding functions given these assumptions.

1.3.1 Bidding Functions

Proceeding from Equation (1.4) and substituting in $N = 2$, $M = 1$, $F(v_i) = v_i$ gives us

$$B(v_i) = \frac{v_i^2}{2}. \quad (1.5)$$

Now evaluating the optimal bidding function for $N = 20$ and $M = 10$, we find

$$\begin{aligned} B(v_i) = & 83980v_i^{11} - 692835v_i^{12} + 2558160v_i^{13} - 5542680v_i^{14} \\ & + 7759752v_i^{15} - \frac{14549535}{2}v_i^{16} + 4564560v_i^{17} \\ & - 1847560v_i^{18} + 437580v_i^{19} - 46189v_i^{20}. \end{aligned}$$

Rather than using $f(v_i) = U[0, 1]$, as above, our subjects will draw valuations uniformly from the set $\{\$0.01, \$0.02, \dots, \$20.00\}$. We assert that subjects view this setting as continuous, so we will use a rescaled version of

the continuous bidding functions to generate experimental predictions. These rescaled optimal bidding functions can be seen in Figure 1.1.

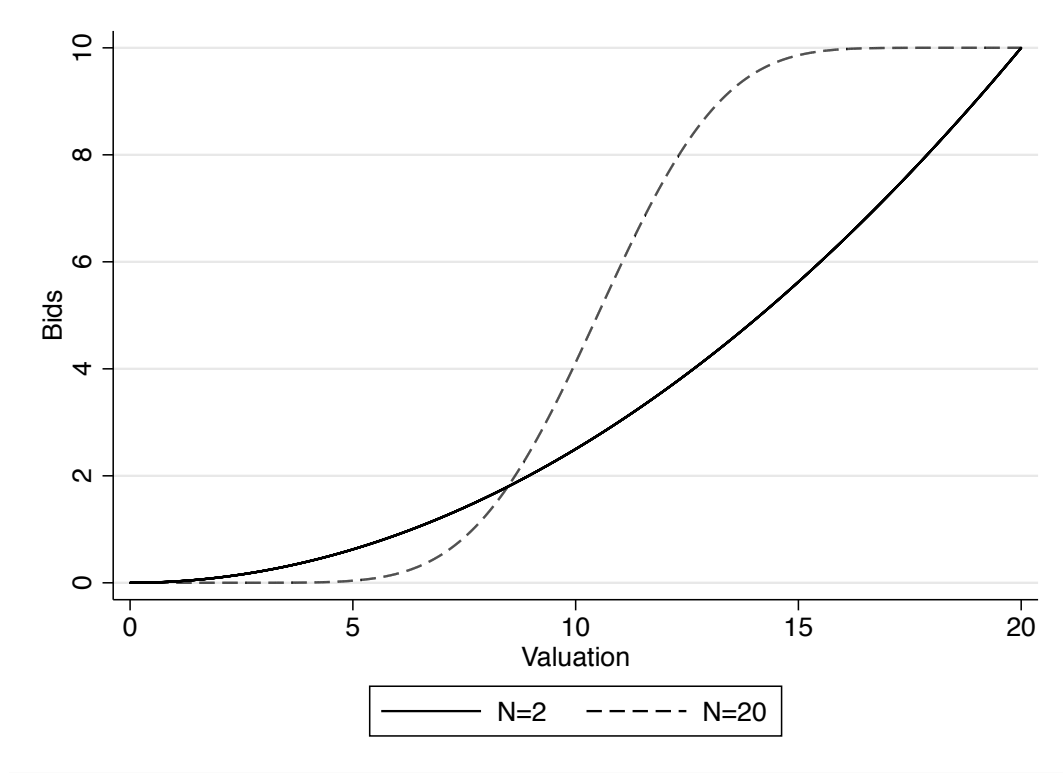


Figure 1.1: Optimal Bid Functions

To understand the intuition behind the shapes of the bidding functions, consider first the limit case where we maintain the proportion of prizes to participants ($M = \frac{N}{2}$), but let $N \rightarrow \infty$. With a large enough sample, uncertainty about the distribution of opponents' valuations is eliminated, and the bidding function approaches a step function. All bidders with $v_i < \$10$ bid $b_i = \$0$, and all bidders with $v_i > \$10$ bid $b_i = \$10$. Since only the top half of bidders receive prizes, bidders below the median know to drop out, while bidders above the median bid just enough that a bidder just below the median cannot

match their bid with positive surplus.³ This equilibrium bid for bidders with $v_i \geq \$10$ is $b_i = \$10$.

Since the large auction in our experiment is ten times the size of the small auction, its equilibrium bids will more closely reflect the limit case. This results from the distribution of valuations in the larger auction drawing closer to the underlying probability distribution than the distribution of valuations in the smaller auction. This convergence causes the median of the realized distribution in the larger auction to be closer to the median of the probability distribution, resulting in less uncertainty over the minimum bid required to win a prize.

Now consider a marginal analysis of lowering a bid from the limit case of $b_i = \$10$ for a bidder with $v_i \geq \$10$ in each of the two auctions. In either auction, the marginal *benefit* of lowering a bid is constant, since bids are paid with certainty, so foregone bids are recovered with certainty. On the other hand, the marginal *cost* of lowering a bid is paid stochastically by lowering the probability of winning a prize. For high-valuation bidders, this decrease in probability is greater in the large auction than in the small auction, so bids cast by high-valuation bidders are larger in the large auction.

Conducting the same marginal analysis for increasing a bid from the limit case of $b_i = \$0$ for a bidder with $v_i < \$10$. Here the marginal cost of increasing bids is constant, while the benefit of increasing bids is probabilistic. At low enough valuations, bidders see greater increases in the probability of

³In the limit case equilibrium all bidders have probability measure zero, and the rules must be changed to reflect the nature of the new contest. Now, prizes will be awarded to bidders in descending order of their bids until half of the bidders have received prizes. This way, there are no profitable deviations from equilibrium.

winning a prize in the small auction, so bids in the small auction rise above those of the large auction.⁴⁵

1.4 Experimental Hypotheses

Principally, a test of our model is a specific test that bidders behave according to the risk-neutral Nash Equilibrium displayed in Figure 1.1, but we also consider several more general predictions of the model. We focus on general predictions that highlight tensions between different potential objectives of the designer. We also address predictions about the ability of the auctions to generate accurate rankings of subjects' valuations based on their bids.

In the event that the equilibrium prediction holds exactly, the other predictions will hold as well. But, even if the precise equilibrium prediction is rejected by statistical tests, the relative predictions about the effects of contest size may still be informative for the policy debate about the merits of different contest sizes. Thus, we will test the qualitative predictions independently of the equilibrium test.

⁴Surprisingly, the equilibrium is not strongly affected by adding in risk-aversion to the bidders' utilities, as shown in the appendix. This results mainly from the fact that bidding is driven by changes in probability, which is unaffected by risk aversion, not changes in surplus, which is affected by risk aversion. See the appendix for a graph of the equilibrium bidding functions under common risk aversion parameters.

⁵In the appendix, we also include a graph of the equilibrium under a common joy-of-winning value. Trivially, this improves the fit of our model to the data. This is mechanical, since the joy-of-winning specification is simply the Nash Equilibrium specification with an added degree of freedom. We do not include this specification of the model in the discussion, because it provides no additional falsifiable predictions for us to compare with the Nash Equilibrium model.

Hypothesis: *Subjects bid according to the Nash equilibrium predictions*

Given the complexity of the equilibrium bidding functions, strict hypotheses about equilibrium bidding seem overly restrictive. Even with experience, the complexity of the bidding functions makes a tight fit between the data and the theory unlikely as there appears to be no obvious heuristic that subjects might adopt. Nonetheless, we will consider equilibrium as a starting point and move from there to the more general predictions identified below.

1.4.1 Qualitative Predictions: The Designer's Objectives

Objective 1: *Maximize aggregate bidding.*

In our education example, this means maximizing the total effort of students. Our equilibrium bidding function predicts that aggregate bidding will be higher in the larger auction. Increasing the size of the auction gives bidders less uncertainty about the valuation of the minimum winning bidder. With tighter predictions about the minimum winning bidder's valuation, bidders face higher probabilistic costs when decreasing their bids, which raises bids for the majority of bidders and increases the rent dissipation.

Objective 2: *Maximize bidding from high-valuation bidders.*

This corresponds to maximizing effort from high-skilled students. Our model predicts that this occurs in the larger auction for the same reasons that they maximize aggregate bidding.

Objective 3: *Maximize bidding from low-valuation bidders.*

Translated into the education context, this implies that the instructor wishes to maximize effort exerted by low-ability students. Among bidders with low valuations, the decrease in uncertainty drives the bids towards zero in the large auction. Therefore, we expect to see higher bids in the small auction for low-valuation bidders. The equilibrium bidding functions in Figure 1.1 intersect at $v_i = \$8.48$, so we will use that value to test the tension between high- and low-valuation bidding.

Objective 4: *Accurately order bidders based on their bids.*

An instructor may wish to employ the mechanism that generates the most accurate ordering of his students' unobservable abilities based only on their observable effort. In our setting, that would mean that the auction designer wants to be able to infer the true ranking of valuations from only the bids cast.

We generate our equilibrium predictions by assuming a one-to-one monotonic matching of bids to valuations. Relative bidding, thus, reflects relative valuations perfectly. In practice, deviations from the equilibrium bid will almost surely make the inferred ranking imprecise. Therefore, one objective of an instructor might be to employ the mechanism that minimizes the impact of these deviations on the accuracy of the ordering of bidders.

Figure 1.1 shows that in the large auction, low- and high-valuation bidders are expected to pool near zero and near \$10, respectively. This provides an advantage to the larger auction as it creates a greater distinction between bids by low- and high-valuation bidders. On the other hand, predicted bids

in the smaller auction increase more gradually, meaning any two bids are less likely to be close to each other. This provides an advantage to the small auction with respect to maintaining a well ordered sorting of bidders in the presence of errors.

In venturing down the path of off-equilibrium dynamic responses to bidding, it becomes clear that our model can not provide strong predictions, and the intuition gleaned above represents the fullest extent to which we are comfortable speculating.

1.5 Experimental Procedures

To test the sensitivity of effort to a change in the size of a contest, we employ an independent private value all-pay auction with multiple prizes. We use a paired bidding design similar to Kagel and Levin (1993) and Andreoni, Che, and Kim (2007) in which we elicit two bids from each bidder each round, one for the large auction and one for the small auction. This is a powerful design that allows us to pair data perfectly and test our hypotheses using the within-subject difference in bidding at a given valuation.

1.5.1 Recruitment and Participation

60 undergraduate students from the University of California, San Diego participated in our experiment. We recruited all of our subjects by means of online advertisements on the EconLab website. The experiment was conducted in 3 sessions in February of 2013 in the EconLab at UCSD. Each session required exactly 20 subjects. All subjects received a \$20 participation payment

in addition to the money gained or lost in the experiment. Our sessions lasted approximately 90 minutes, and subjects earned between \$15 and \$45.

1.5.2 Instructions

We first read the instructions aloud to all subjects and then administered a quiz to test their comprehension of the auction formats and the payoffs from different outcomes.⁶

We designed instructions to clarify the means by which we awarded prizes and the payoffs conditional on receiving or not receiving a prize. We carried out the entire experiment on computers using the Z-Tree Economics Software (Fischbacher, 1999). This allowed us to post key instructions at the top of every decision screen. We used this feature to remind subjects that bids would be deducted from their winnings regardless of the outcome of the auction, that their final payment for the experiment would be determined by one round selected at random, and that they would see two auctions for each valuation they drew, first the small auction, then the large auction. We also reinforced the number of opponents they would face and the number of prizes that would be available each round.

1.5.3 Auction Design

Every round, subjects received valuations drawn uniformly in \$0.01 increments between \$.01 and \$20.00. That is, $v_i \sim U[.01, 20.00]$. Using this valuation, subjects participated in both a “2-Person Auction,” for which they

⁶The instructions and quiz can be found in the appendix.

were randomly and anonymously paired, and a “20-Person Auction,” which consisted of all participants in the session.⁷

Subjects cast bids for both auctions in every round. Upon submission of both bids, we revealed the outcomes of all auctions to the subjects. This included the subject’s own bids and payoffs from the two auctions he participated in as well as all winning and losing bids cast by all bidders in all auctions that round. All bids were cast without knowledge of opponents’ bids or valuations, so we consider them simultaneous bids under incomplete information. No communication was allowed, and we never revealed the valuations associated with any bids.

Subjects then drew new valuations, new partners for the 2-Person Auction, and repeated the auction round, casting two more bids for each new valuation. Subjects in session 1 repeated the auction rounds 20 times, while subjects in sessions 2 and 3 repeated them 15 times. For the sake of uniformity, we analyze only the first 15 rounds of each session, but our results only strengthen with the inclusion of the final 5 rounds of the first session.⁸

After completing the auction rounds, subjects made 3 incentivized choices between gambles in order to elicit their preference for risk. We employed a discretization of the method detailed in Andreoni and Harbaugh (2009). These measures of risk aversion will be used as a control in the analysis along with demographic data we collected.

⁷This structure gave us the largest change from one size of contest to the other. Since the 20-Person Auction features a stable cohort, issues with collusion could arise. We will address this concern in the following section.

⁸We planned each session to fit 20 rounds into 90 minutes. During the first session we discovered that 20 rounds would not work due to our server speed. The second and third session were shortened to accommodate this.

1.6 Results

With 15 rounds in which subjects drew valuations and cast bids in two different auctions, we collected a total of 900 different valuations and 1800 different bids. The following tables report results using only bids that do not exceed the valuation of the bidder. These bids are strictly dominated as they guaranteed a negative payoff to the bidder.⁹ There are 14 such bids evenly distributed across periods and auction sizes. For all results, we will indicate when our analysis is substantively affected by removing these bids.

Figure 1.2 compares the aggregate patterns of bidding across different valuations to their equilibrium predictions. Here we split the data by valuation into 10 equally-sized bins. Adjacent dots show the mean and 95% confidence intervals of the bids in each auction size. Figure 1.2 shows that all bids are near-zero for low valuations, regardless of auction size. The larger auction begins to dominate over middle-valuations, as our model predicted. For high valuations, bids are close to the equilibrium predictions with little difference between the two auction sizes.¹⁰

In order to leverage the statistical power gained by our paired design, we evaluate the experimental hypotheses in terms of their predictions about the difference between the 20- and 2-Person Auction bids at a given valuation.

Figure 2.4 compares the observed differences in bids to our model's prediction

⁹Overbidding in this way has been documented in several studies using common-value all-pay auctions (Gneezy and Smorodinsky, 2006, for example), but does not represent a large fraction of bids in independent private value all-pay auctions (Noussair and Silver, 2006).

¹⁰For a finer resolution display of subject bidding, we separate individual bids by auction size and session and plot them against the bidder's valuation. We include these graphs in the appendix rather than here, since the volume of bids makes the graphs difficult to interpret parsimoniously.

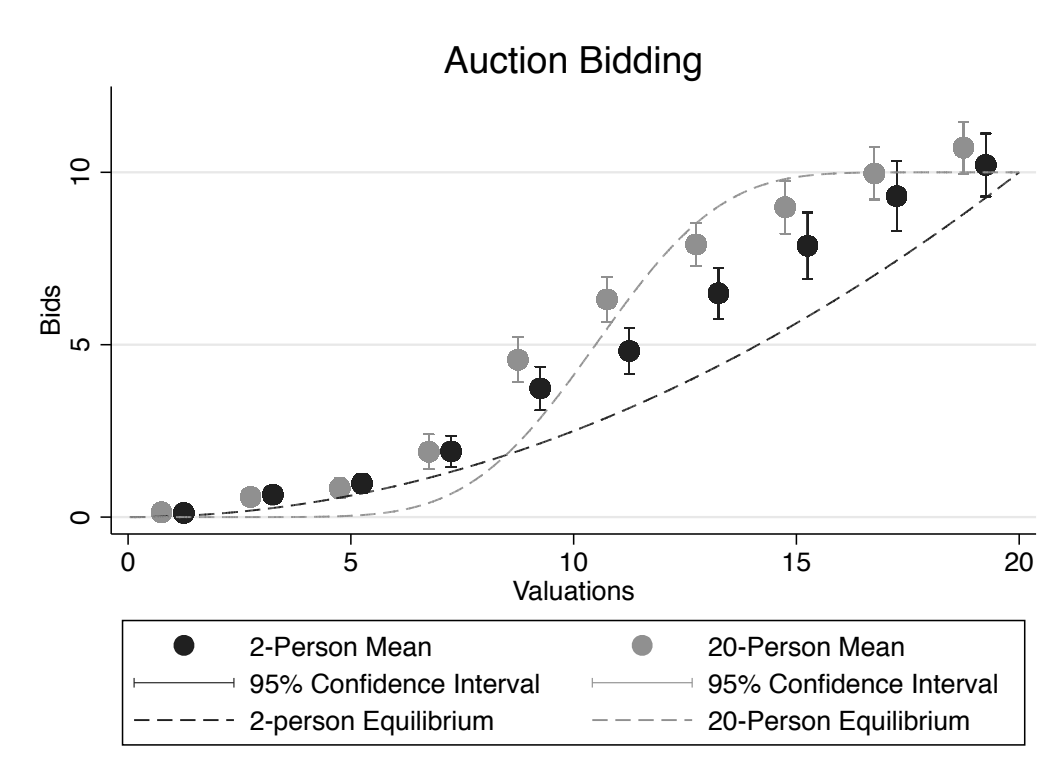


Figure 1.2: Aggregate Bidding

about this difference. The observed differences are averaged within the same 10 equally-sized bins as above.

1.6.1 Hypotheses

Hypothesis: *Subjects bid according to the Nash equilibrium predictions*

Statistical tests reveal that there are systematic deviations between individual bids and their equilibrium predictions both in the 2-Person Auction ($t = 7.03, p < 0.001, n = 893$) and the 20-Person Auction ($t = 4.73, p < 0.001, n = 893$). Comparing differences in bidding with the equilibrium predictions

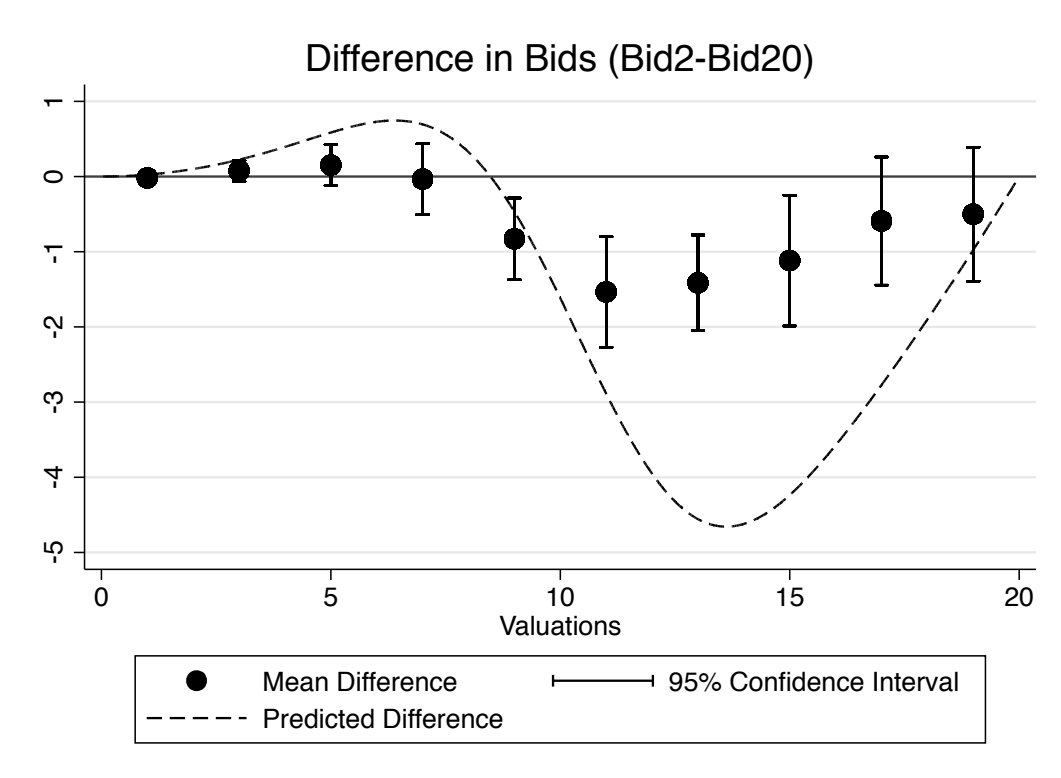


Figure 1.3: Bidding Differences

about those differences yields a similar rejection of our model's predictions ($t = 4.88$, $p < 0.001$, $n = 888$).

Of course, one should not be surprised that in this unique environment with complex and precise predictions, our model fails these statistical tests. As the literature on auction experiments reveals, the theory often provides a good benchmark but seldom predicts precise outcomes. Indeed, Figure 2.4 makes it clear that the differences between bids in the 2- and 20-Person Auctions evolve in a qualitatively similar way to the model's predictions. Therefore, the question becomes whether, despite the lack of a precise fit, the qualitative predictions of our model and the intuitions gleaned from our theory extend to the observed data. We explore this question within the context of the following

objectives.

Objective 1: *Maximize aggregate bidding.*

In 33 of the 45 rounds, revenues were greater in the larger auction, consistent with our theory. Using each round as an observation, we perform a paired t-test on the difference in total revenue between auctions and find that the larger auctions generate significantly more revenue ($t = 4.74$, $p < .001$, $n = 45$). Our paired design allows us to test this hypothesis at the individual level and perform a paired t-test matching each bid in the large auction with its counterpart in the small auction. This test rejects that bidding is equal across auction sizes, confirming that bidding is significantly larger in the large auction ($t = 5.02$, $p < .001$, $n = 888$).¹¹

Objective 2: *Maximize bidding from high-valuation bidders.*

Objective 3: *Maximize bidding from low-valuation bidders.*

Our model predicts that bids in the small auction will dominate bids in the large auction for low valuations, while bids in the large auction will dominate for high valuations. Our model specifically predicts that this crossover occurs at $v_i = \$8.48$.

Figure 2.4 shows that this qualitative pattern holds for high-valuation bidders, but not for low-valuation bidders. That is, it appears that the behavior of low-valuation bidders is largely invariant to the size of the auction, but

¹¹These results are unchanged by including all outlying bids except one particularly extreme outlier, where the bidder received a valuation of \$0.01 and bid \$99.00. While this bid did risk real money, we do not believe it to be reflective of the types of bidder behavior we are attempting to capture.

that high-valuation bidders show sensitivity to the auction size.

We test this prediction more formally in two ways. First, we estimate the point at which bids in the large auction begin to dominate bids in the small auction. Second, we directly test the sign of the bidding difference predicted by our model at the theoretical and fitted crossover points.

To estimate the valuation at which bids in the large auction cross over those of the small auction, we use a locally-linear polynomial smoothing function of the difference in bids. We plot this function in Figure 1.4, which shows the fitted crossover point as $v_i \approx \$6.65$. While this diverges from the precise theoretical prediction about the location of the single-crossing point, it confirms the single-crossing nature of the bidding functions with bids in the small auction dominating at first, before bids in the large auction begin to dominate.

We use the theoretical crossover point of $v_i = \$8.48$ to split our data into two bins, *LowerBin* and *UpperBin*, and test our theory’s predictions about relative bidding in each bin. To best exploit our paired data, our dependent variable will be the bid in the small auction minus the bid in the large auction. Our theory predicts that the difference will be positive until $v_i = \$8.48$ and negative afterwards. We regress this difference onto *LowerBin* and *UpperBin* with standard errors clustered at the subject level. We repeat the regression including controls for the risk aversion parameters elicited from each subject.¹² The results can be seen in the left-hand side of Table 1.1.

We repeat this analysis with the fitted crossover point, generating new dummy variables, *LowerFitBin* and *UpperFitBin*, for valuations below the

¹²Risk aversion does not have a clear prediction, as it affects high- and low-valuation bidders differently. Thus, we interact these controls with the bins for different valuations.

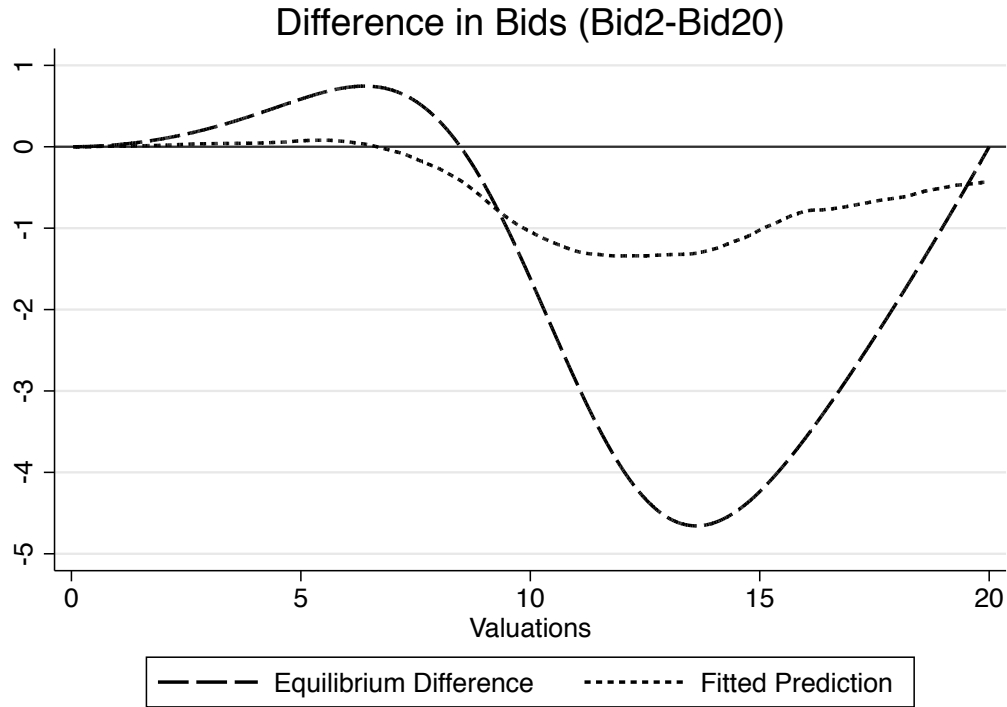


Figure 1.4: Predicted Difference in Bids

fitted crossover point, $v_i = \$6.65$. The results are found on the right-hand side of Table 1.1.

The difference in bids in the upper bin is large in magnitude and highly statistically significant. This difference in bids is also statistically distinguishable from the same difference in the lower bin. However, the difference in bids in the lower bin is not statistically distinguishable from zero.¹³ Recall that the predicted difference between the bidding functions is small over lower valuations, so it is not surprising that the deviation from zero is not statistically significant. For valuations just above the median, however, we predict large differences between bids cast in the large and small auction. This separation

¹³These results are also unchanged with the inclusion of all outlying bids except for the particularly extreme outlier mentioned in objective 1.

Table 1.1: The difference in bids across different valuations

Periods:	Difference in Bids (Small-Large)							
	Predicted Cutoff: \$8.48				Fitted Cutoff: \$6.65			
	1-15	1-15	1-15	9-15	1-15	1-15	1-15	9-15
LowerBin=1								
if $V_i \leq \$8.48$	0.0657 (0.10)	0.0714 (0.09)	-0.224 (0.32)	-0.0996 (0.14)				
UpperBin=1								
if $V_i > \$8.48$	-1.000*** (0.31)	-0.958*** (0.29)	-0.957*** (0.29)	-1.151*** (0.40)				
LowerFitBin=1								
if $V_i \leq \$6.65$					0.0296 (0.08)	0.0282 (0.08)	-0.340 (0.37)	-0.0666 (0.09)
UpperFitBin=1								
if $V_i > \$6.65$					-0.847*** (0.28)	-0.834*** (0.26)	-0.836*** (0.26)	-1.087*** (0.37)
R.A. Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Dominated Bids	No	No	Yes	No	No	No	Yes	No
N	888	888	900	417	888	900	888	417

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Standard errors clustered by subject.

is clearly displayed in statistical tests and graphical analysis.

Objective 4: *Accurately order bidders based on their bids.*

We test which auction best accomplishes Objective 4 with Kendall's (1938) tau rank correlation coefficient (Hereafter, K-T). The K-T coefficient measures the relationship between the number of concordant pairs, C , and discordant pairs, D , in a sample with N observations. Mathematically, the K-T coefficient is

$$K = \frac{C - D}{\left(\frac{N \times (N-1)}{2}\right)}. \quad (1.6)$$

A pair of bids is concordant if the ranking of the two bids is consistent with the ranking of the valuations of the bidders ($b_i < b_j \iff v_i < v_j$), while a pair of bids is discordant if it is inconsistent with the rankings of the bidders'

valuations $(b_i > b_j \iff v_i < v_j)^{1415}$

The K-T coefficients and their significance tests for the pooled data are reported in the first and second row of Table 1.2. We then split the data and generate a K-T coefficient for each auction in each period before running a pairwise analysis of the two auctions.¹⁶

Table 1.2: Measuring the Well-Ordering of Bids

Kendall-Tau Coefficient of Well-Ordering				
	Obs	Small Auction	Large Auction	Difference
Pooled [†]	900	0.546 (0.014)	0.586 (0.012)	-0.040**
Pooled*	888	0.549 (0.014)	0.591 (0.012)	-0.043***
Per Period	45	0.565 (0.014)	0.608 (0.014)	-0.042**

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[†] *Standard errors clustered at the period level.*

* *Outliers omitted, and standard errors clustered at the period level.*

Our results show that neither auction is able to generate a perfect ordering of valuations based on bids. Observing a properly ordered pair of bids in the small auction is 55% more likely than observing an improperly

¹⁴We can also make a correction found in Kendall (1975) that accommodates ties in bids or in valuations. The signs and significance of our results are maintained or increased under this specification.

¹⁵A possibly more recognizable test of rank-correlation is the Spearman (1904) coefficient. We have chosen Kendall's tau for two reasons: First, Spearman's coefficient is more heavily influenced by outliers. This will bias our results in the direction of the large auction, since there are many more outlying bids in the small auction. Second, Kendall's tau has a more straightforward interpretation. Repeating the analysis using Spearman's coefficient returns the same signs, but with larger magnitudes.

¹⁶Given that we have 45 pairs of K-T statistics, we invoke the Central Limit Theorem and perform a standard t-test on the difference between the pairs.

ordered pair of bids. In the large auction that ratio is nearly 59%. The difference in relative likelihood is statistically significant, and economically meaningful, as it leads to 365 more improperly ordered pairs of bids.

Digging deeper into the K-T coefficients reveals some of the mechanisms that drive bids out of their well-ordered ranking. While the small auction equilibrium bidding function showed a more gradual increase in Figure 1.1, the larger auction showed starker contrasts between bids of high- and low-valuation bidders. These competing mechanisms are clearly visible in the data.

Table 1.3: The Well-Ordering of Bids for Expected Winners and Losers

	Kendall-Tau Coefficient of Well-Ordering			
	Obs	Small Auction	Large Auction	Difference
$Valuation_i < 10$	414	0.375 (0.028)	0.295 (0.029)	0.079***
$5 < Valuation_i < 15$	474	0.404 (0.029)	0.482 (0.021)	-0.078***
$10 < Valuation_i$	474	0.334 (0.031)	0.345 (0.028)	-0.011

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Outliers omitted, and standard errors clustered at the period level.

We consider two groups of bidders, those above the median of the distribution and those below it. Rows 1 and 3 of Table 1.3 report the K-T coefficients for each of these two groups. The coefficients within a group differ meaningfully from the pooled coefficients from Table 1.2, and are relatively more favorable towards the small auction. The second row of Table 1.3 maintains the interval length, but compares sorting between bidders above and below the median. Over this interval, sorting statistics strongly favor the larger auction, indicating that bidders identified the starker probabilistic distinction

between expected outcomes in the large auction. This starker distinction of probable outcomes arises from the law of large numbers and appears to lead to starker separation of bids on either side of the median valuation.

The two observed sorting mechanisms closely follow the intuition of our model, which demonstrates that the larger auction should see pooling among high- and low-valuation bidders but strong separation between the two groups, while the smaller auction should see better separation of bidders with similar expected outcomes. Interpreting Table 1.2 in light of the results from Table 1.3 sheds light on the relative strengths of the two sorting mechanisms. It appears that as the auction size grows, the superior sorting across the median dominates the diminished sorting between bidders with similar expected outcomes, generating a better-ordered set of bids in the larger auction.

1.7 Discussion

Our results show that subjects are sensitive to changes in the size of their competitive environment even when the proportion of winners remains constant. Moreover, our model provided useful benchmarks for predicting the way in which subjects would adjust their behavior to the contest size.

Taking our results from the abstract environment and returning them to the environment of an undergraduate course will provide context and useful policy recommendations. It is important to note that our analysis explores the effect of contest size while holding all other inputs fixed. That is, we abstract away from systematic correlations between the quality of inputs to the students' production function and the size of the classroom. Of course, in

any discussion of policy these correlations will need to be weighed against the effects of classroom size on strategic effort that we have uncovered. We do not feel that the effects we uncovered trump other effects of classroom size, but we hope to bring them to light in order that policy-makers may factor them into the discussion of optimal classroom size.

For classroom designers concerned with maximizing aggregate effort, the larger contests dominate. Thus, students in classes graded on the curve are expected to increase effort as the class size increases. In smaller classes, the greater uncertainty surrounding the minimum effort required for a given grade may cause students to strategically lower their effort in an attempt to exploit the randomness of the environment. In larger classes, however, there is less uncertainty about the minimum effort required for a given grade, and students will likely respond by increasing their effort to match this minimum required effort for their desired grade.

A designer may also have preferences over the distribution of effort among students with different abilities. We demonstrated that this distribution of effort was strongly affected by the size of the contest. From our results, we can infer that high ability students will likely exert greater effort in larger classes, while students with lower-middle abilities may exert slightly more effort in smaller classes. The lowest ability appear to be insensitive to changes in the size of the strategic environment.

Another advantage of the larger auction was its ability to sort bidders by their valuation when only bids were observable. In the classroom, student abilities are unobservable and so classroom evaluations must serve as a sorting mechanism, attempting to distinguish students' abilities by students' outputs.

What we discovered suggests that the ability of a relative grading scheme to sort students could be systematically undermined in courses with low enrollment. That is, a student's ability seems to be less correlated to her relative performance in courses with fewer students.

With multiple grade levels (A,B,C,D,F, for example), each student will face greater uncertainty with regards to their relative position in low-enrollment courses. Our results indicate that this uncertainty may cause effort to diminish and further weaken the correlation between ability and performance. On the other hand, with large enough enrollment, we still expect students to match the minimum required effort for a given grade.¹⁷ With a finer grading scheme, the ordering of bids will be more dependent on sorting across expected outcomes, making the sorting of students in courses with high enrollment gain an additional advantage over sorting in courses with lower enrollment. This speculation, however, needs to be verified by further studies.

1.8 Conclusions

We designed and executed an experiment to examine how the size of a contest will affect the effort choices by players of different abilities. We leveraged a paired-auction design to focus on the difference in bids between all-pay auctions of different sizes and analyzed how that difference evolved across bidders with different abilities. Using this approach we were able to abstract away from the exacting question of on- or off-equilibrium bidding and focus on more general predictions about how differences in bidding between two

¹⁷For a deeper discussion of optimally selecting the fineness of a grading scheme, see Dubey and Geanakoplos (2010).

auction sizes evolve with ability. We discovered economically and statistically significant results that hold implications for a broad category of contest design.

Our model correctly predicted that larger contests elicit greater aggregate bidding and larger bids from high-ability bidders. Our model also drew attention to some undesirable effects of smaller contests. Specifically, the uncertainty that arises as the size of a contest decreases. This uncertainty causes high-ability bidders to shade down their bids and limits the ability of the designer to accurately order bidders' valuations by observing their bids.

As policy makers, administrators, and other mechanism designers continue to employ contests, this research sheds light on the effect of a contest's size on the effort exerted by its participants. While this study considered only two sizes of a stylized contest, we believe it serves as a strong foundation for further research on the independent effect of a contest's size on the performance of its participants.

1.9 Acknowledgement

Chapter 1, in part is currently being prepared for submission for publication of the material. Brownback, Andy; Andreoni, James. The dissertation author was the primary investigator and author of this material.

1.10 Chapter 1 Appendix

1.10.1 Proofs

Proposition 1: *Optimal Bidding Function is Weakly Monotonic in Valuation.*

Proof: Suppose not. Then v_i and v_j exist such that: $v_i < v_j$ but $b_i > b_j$.

Incentive compatibility for i dictates that

$$U(b_i, v_i, N) \geq U(b_j, v_i, N).$$

Substituting in the bidder's utility yields

$$v_i * P_N(b_i) - b_i \geq v_i * P_N(b_j) - b_j,$$

and rearranging we find

$$b_i - b_j \leq v_i (P_N(b_i) - P_N(b_j)).$$

A similar derivation for j reveals

$$b_i - b_j \geq v_j (P_N(b_i) - P_N(b_j)).$$

Therefore, since $b_i > b_j$, by construction this implies $P_N(b_i) \geq P_N(b_j)$. Subsequently, we have

$$v_i (P_N(b_i) - P_N(b_j)) \geq v_j (P_N(b_i) - P_N(b_j)).$$

Thus, $v_i \geq v_j$ or $P_N(b_i) = P_N(b_j)$. In the case of the former we have a

contradiction. In the case of the latter, our contradiction is satisfied by a violation of incentive compatibility. Hence, $v_i < v_j \implies b_i \leq b_j$. QED.

Corollary 1: *Monotonicity holds under Risk Aversion*

Proof: Again we start with:

$$U(b_i, v_i, N) \geq U(b_j, v_i, N)$$

and

$$U(b_j, v_j, N) \geq U(b_i, v_j, N).$$

Expand to

$$\begin{aligned} U(v_i - b_i) * P_N(b_i) + U(-b_i) (1 - P_N(b_i)) &\geq \\ U(v_i - b_j) * P_N(b_j) + U(-b_j) (1 - P_N(b_j)) \end{aligned}$$

and

$$\begin{aligned} U(v_j - b_j) * P_N(b_j) + U(-b_j) (1 - P_N(b_j)) &\geq \\ U(v_j - b_i) * P_N(b_i) + U(-b_i) (1 - P_N(b_i)) . \end{aligned}$$

Solving for common terms and combining yields

$$\begin{aligned} U(v_i - b_i) * P_N(b_i) + U(-b_i) (1 - P_N(b_i)) - U(v_i - b_j) * P_N(b_j) &\geq \\ U(v_j - b_i) * P_N(b_i) + U(-b_i) (1 - P_N(b_i)) - U(v_j - b_j) * P_N(b_j) . \end{aligned}$$

Canceling terms and grouping simplifies this to

$$P_N(b_i) [U(v_i - b_i) - U(v_j - b_i)] \geq P_N(b_j) [U(v_i - b_j) - U(v_j - b_j)].$$

We have assumed $v_j > v_i$ and $b_i > b_j$, so we can define $v_i - b_i \equiv A$, $v_j - b_i \equiv A'$, $v_i - b_j \equiv B$, and $v_j - b_j \equiv B'$ with $A' > A$, $B' > B$, and $B > A$:

$$P_N(b_i) [U(A) - U(A')] \geq P_N(b_j) [U(B) - U(B')]$$

But, since $A' - A = B' - B$ and $U'' < 0$ we know

$$[U(A) - U(A')] < [U(B) - U(B')] < 0.$$

Our assumptions also tell us that $P_N(b_j) < P_N(b_i)$, meaning

$$P_N(b_i) [U(A) - U(A')] < P_N(b_j) [U(B) - U(B')],$$

and we arrive at a contradiction. QED.

1.10.2 Individual Bids and Individual Differences

Below we present the individual bids for the first 15 periods of each session. In addition, we plot the differences in pairs of bids for each of these periods. In generating these scatter plots we have included all outlying bids except for one. This outlier was a bid of \$99 when the bidder was assigned a valuation of \$0.01. We interpret this bid as a form of protest against the valuation assigned.

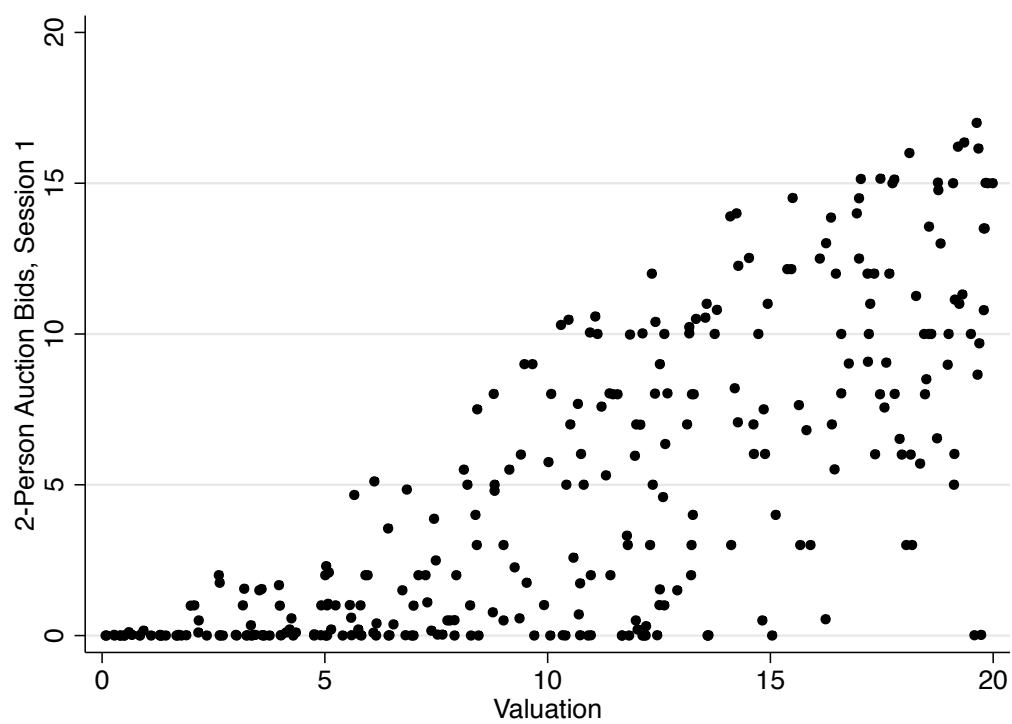


Figure 1.5: 2-Person Auction Bids, Session 1

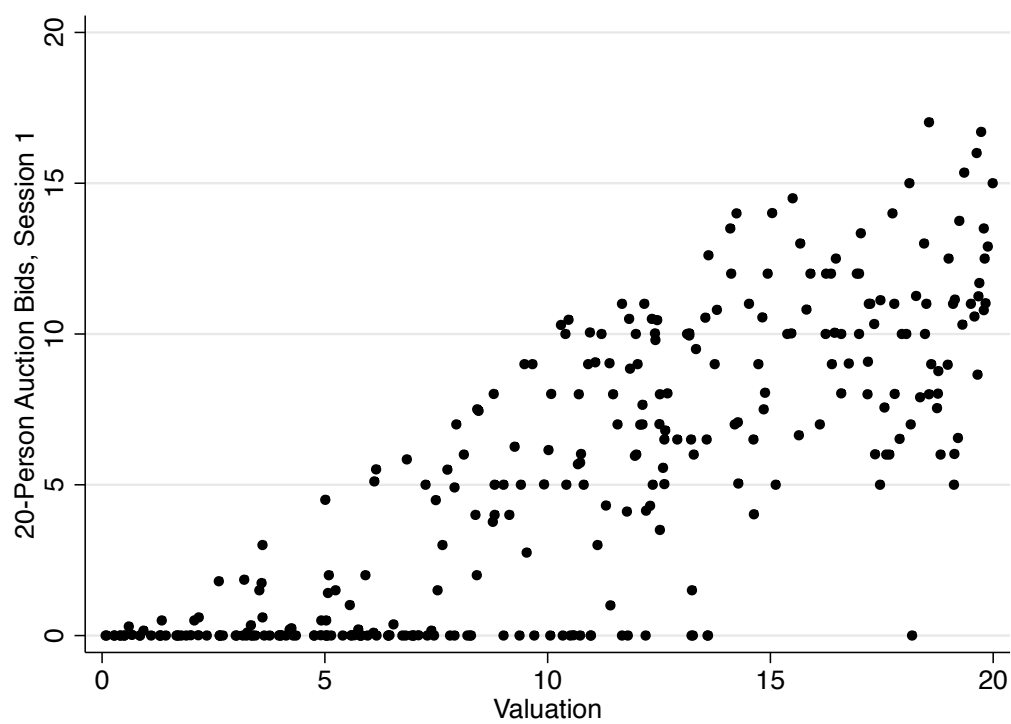


Figure 1.6: 20-Person Auction Bids, Session 1

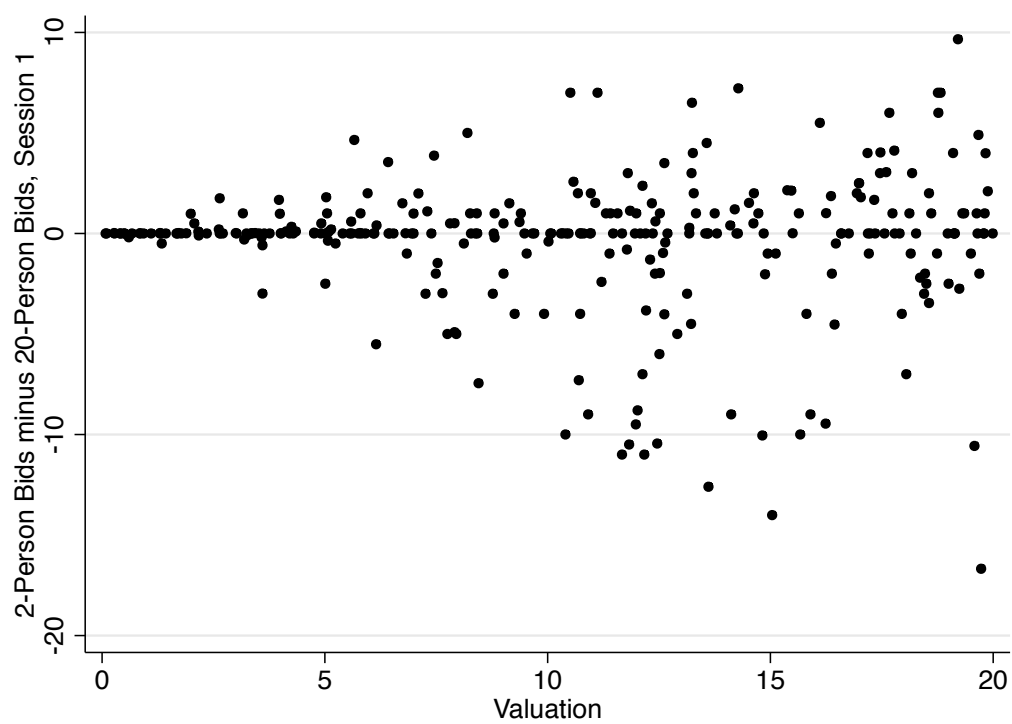


Figure 1.7: Difference in Bids, Session 1

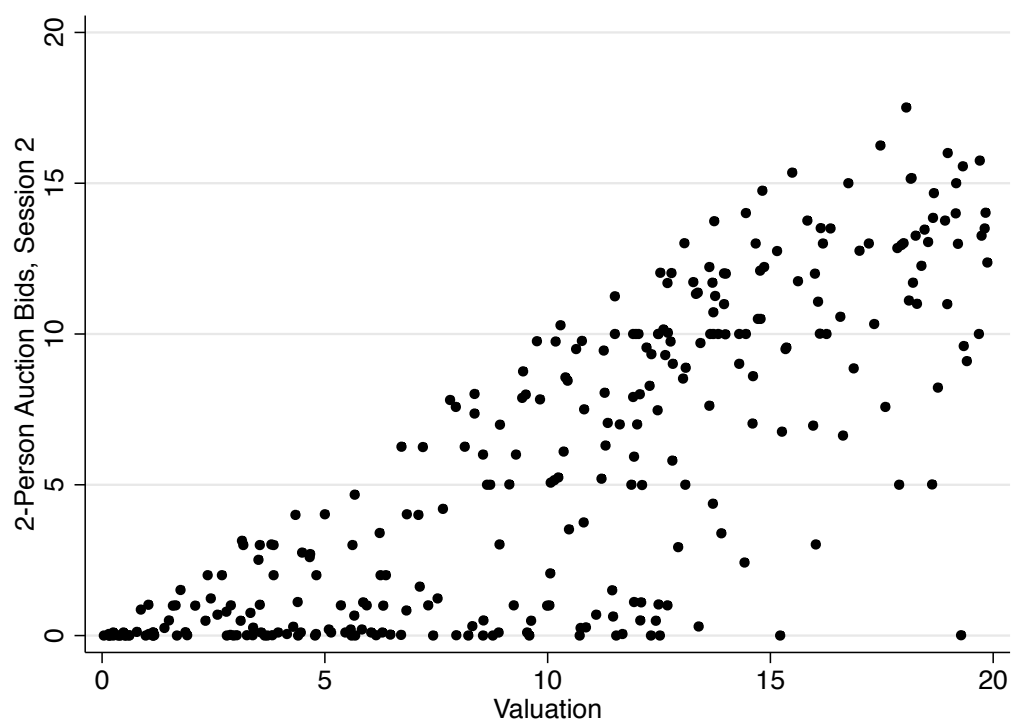


Figure 1.8: 2-Person Auction Bids, Session 2

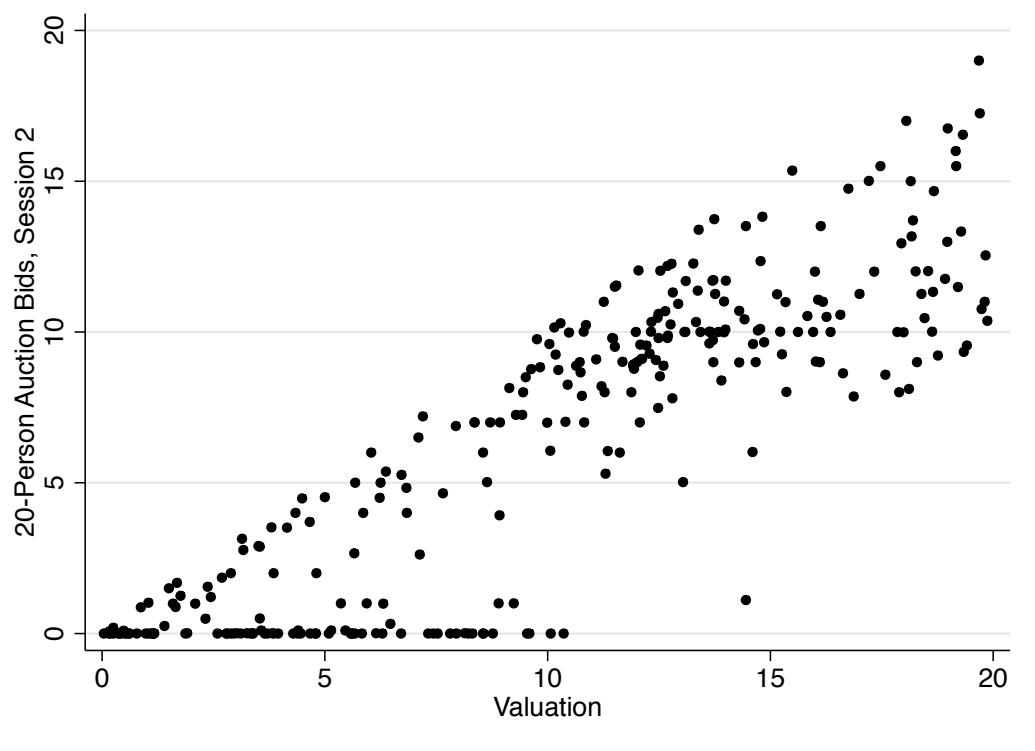


Figure 1.9: 20-Person Auction Bids, Session 2

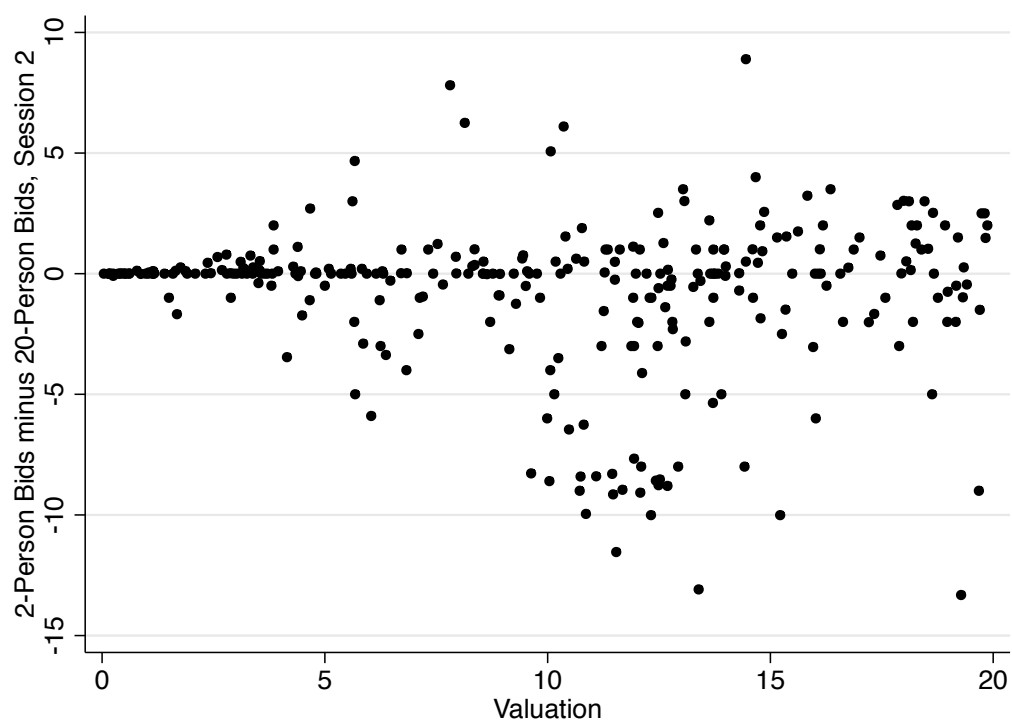


Figure 1.10: Difference in Bids, Session 2

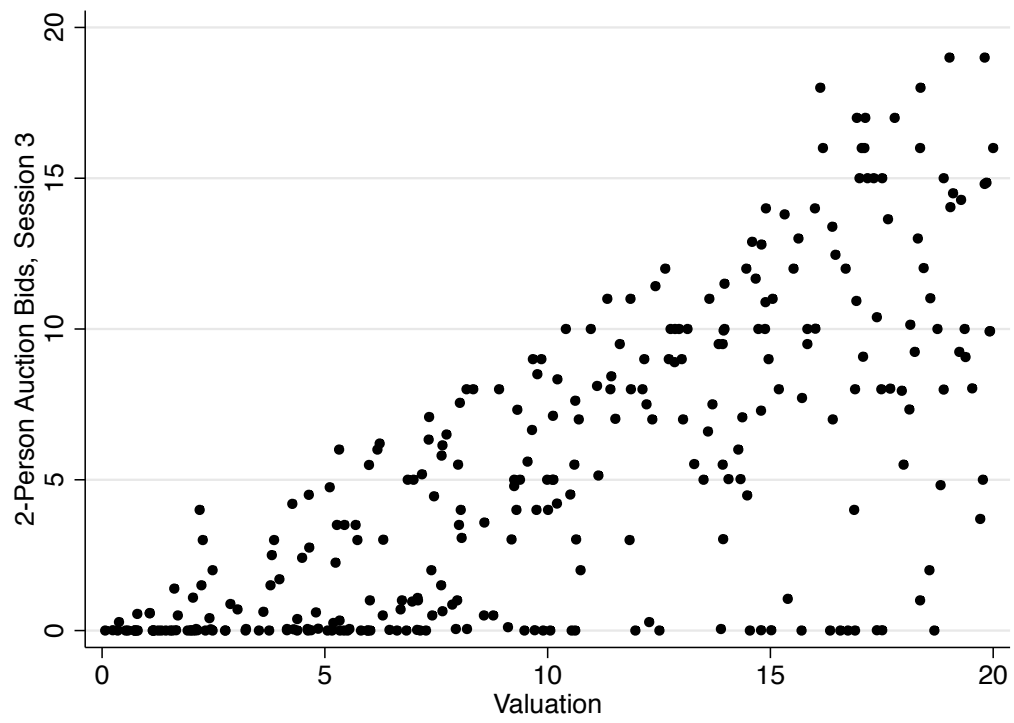


Figure 1.11: 2-Person Auction Bids, Session 3

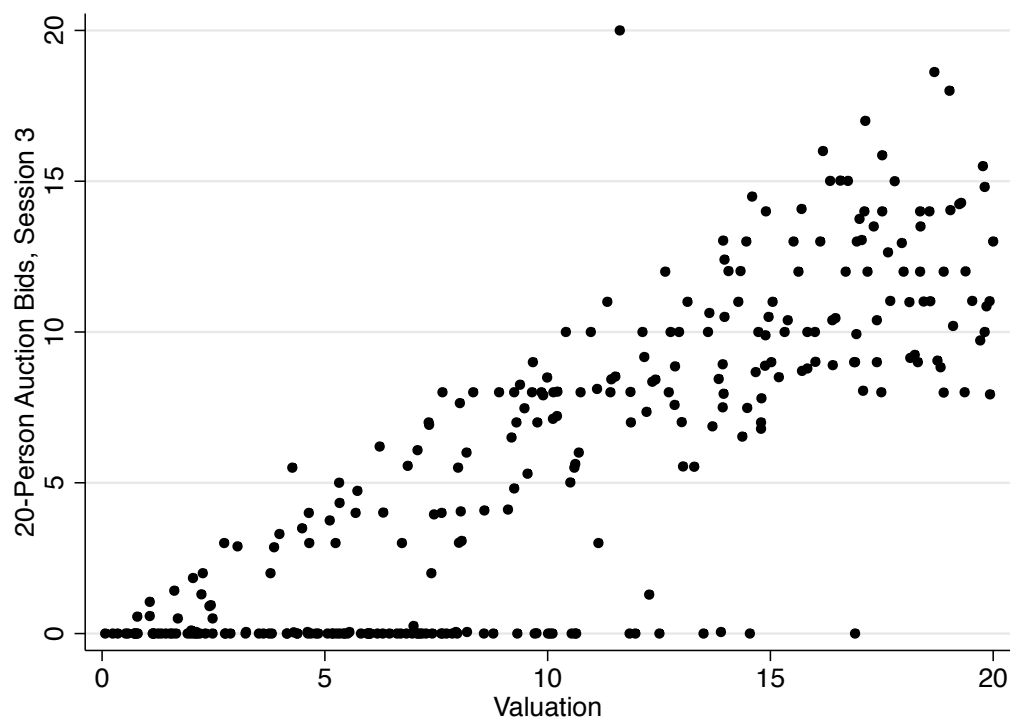


Figure 1.12: 20-Person Auction Bids, Session 3

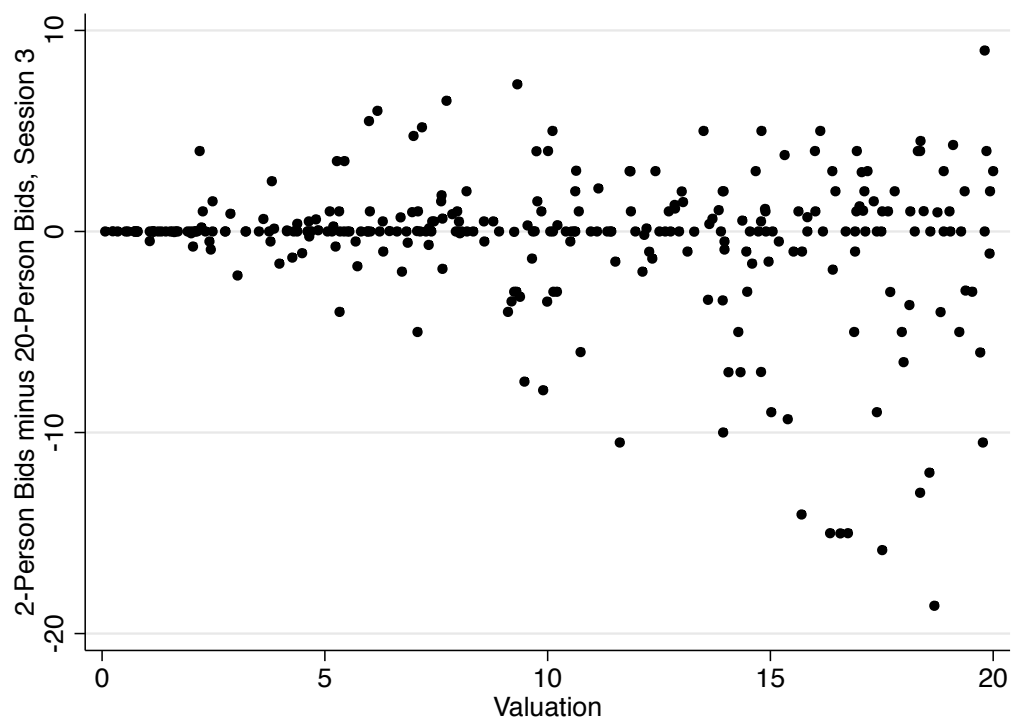


Figure 1.13: Difference in Bids, Session 3

1.10.3 Alternative Assumptions on Utility

Here we reconsider our model under assumptions of risk averse bidders, and bidders with a joy-of-winning.

Risk Averse Bidding Functions

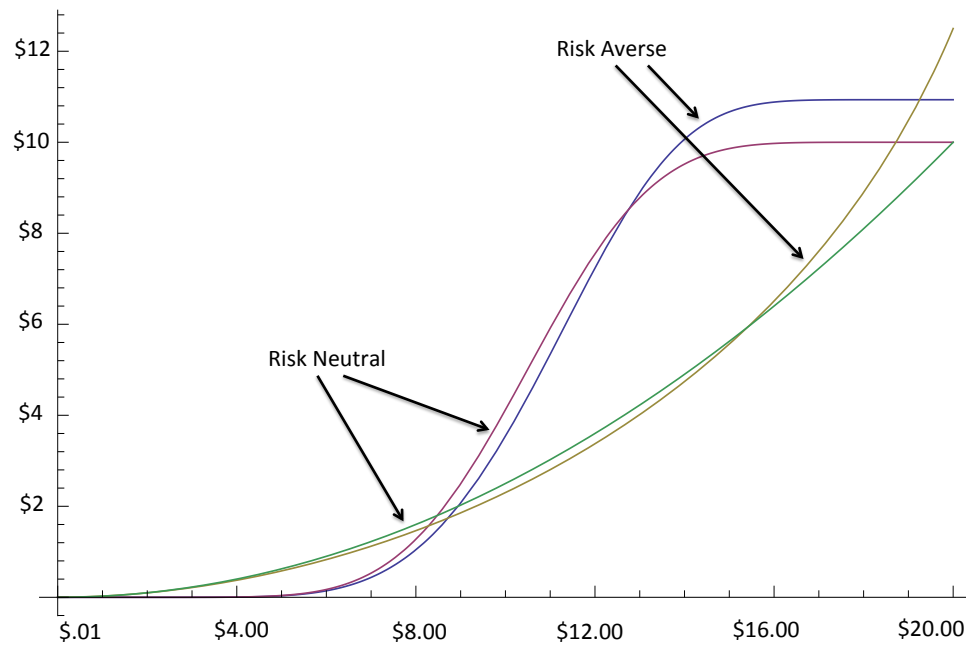


Figure 1.14: Risk Averse Bidding Equilibrium

This graph details the bidding function under CRRA utility of $U(x) = \frac{x^{1-\sigma}}{1-\sigma}$ with $\sigma = 3$. Here we can see that the risk aversion does not strongly affect the shape or magnitude of the bidding function in a symmetric equilibrium. Moreover, the risk averse Nash Equilibrium deviates from the data even more than does the risk neutral Nash Equilibrium over middle valuations.

Joy of Winning

We can modify our optimal bidding framework to include a constant utility gain simply from winning an auction. This will result in the following utility function:

$$U(\bullet) = (v_i + c) * P_N(b_i) - b_i$$

Resulting in the following bidding functions:

$$B(v_i, 2) = \frac{v_i^2}{2} + cv_i$$

$$\begin{aligned} B(v_i, 20) = & -\left(\frac{1}{2}\right)x^{10} \{2c(48620x^9 - 461890x^8 + 1956240x^7 - 4849845x^6 + \\ & + 7759752x^5 - 8314020x^4 + 5969040x^3 - 2771340x^2 + 755820x - 92378) \\ & + x(92378x^9 - 875160x^8 + 3695120x^7 - 9129120x^6 + 14549535x^5 \\ & - 15519504x^4 + 11085360x^3 - 5116320x^2 + 1385670x - 167960)\} \end{aligned}$$

Figure 2 shows the equilibrium that arises under a modest value for the Joy of Winning parameter, $c = \$1.00$.

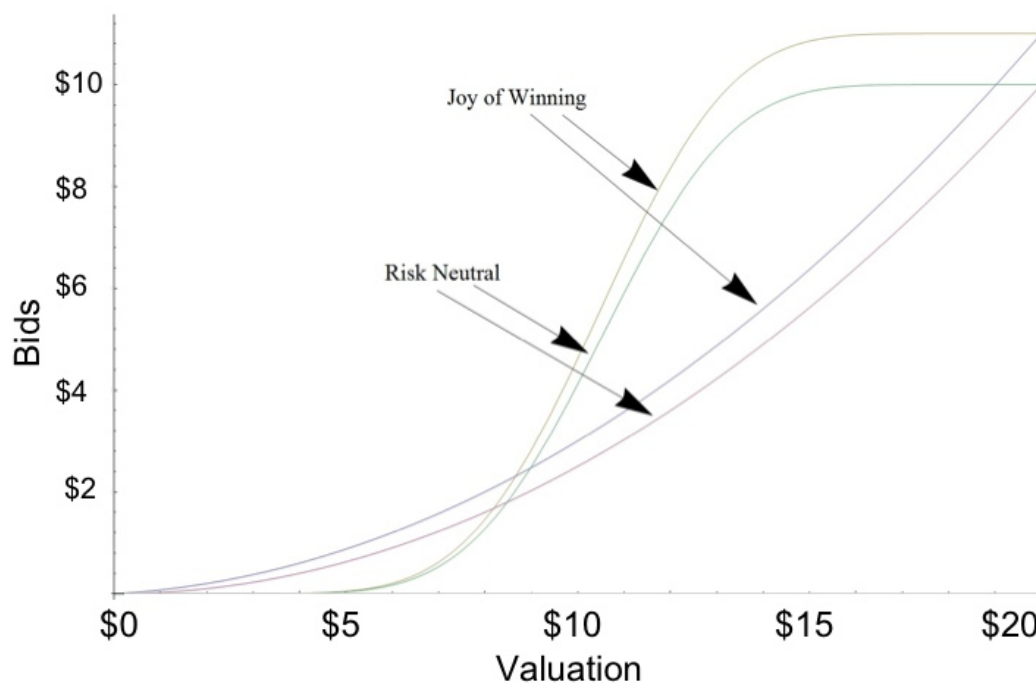


Figure 1.15: Joy of Winning Equilibrium

1.10.4 Subjects' Instructions

Before the experiment, subjects were given two sheets of paper to facilitate understanding of the experiment. The first page defined a list of terms that we would refer to throughout the experiment. The second page outlined logistics of the experiment with respect to how payments were determined. The two pages are found in the two following subsections.

Definitions

These terms were used throughout the experiment. We gave all subjects a hard copy reference sheet that they could use for clarification of the instructions.

Prize Value: In each auction you will be assigned a prize value this is the amount of money that you will win if you receive a prize

in that auction. So, if your prize value is \$6.76 then you will receive \$6.76 if you win that auction. If your prize value is \$19.95 then you will receive \$19.95 if you win that auction, and so on.

Bid: In each auction you must bid in order to win a prize. Prizes will be awarded to the participants with the highest bids. If there is one prize in the auction then the prize will go to the highest bidder. If there are ten prizes in the auction then they will go to the ten highest bidders, and so on.

NOTE: You must pay your bid regardless of the outcome of the auction. Whether you win or you lose you will be charged your bid. If you bid \$5.34 in an auction and you lose then you will be charged \$5.34. Likewise, if you bid \$5.34 in an auction and you win then you will be charged \$5.34.

Payout: The amount of money you gain from an auction will depend on your bid and your prize value. If you win then you will receive your prize value and you will be charged your bid. If you lose then you will receive nothing and you will still be charged your bid.

EXAMPLE: If you have a prize value of \$15.00 and you bid \$10.00 and you *win* the auction then you will receive your prize value, \$15.00, and you will be charged your bid, \$10.00, meaning that your payout will be (positive) \$5.00.

EXAMPLE: If you have a prize value of \$15.00 and you bid \$10.00 and you *lose* the auction then you will receive nothing and you will be charged your bid, \$10.00, meaning that your payout will be (negative) -\$10.00.

Auction That Counts: After you have completed the experiment, one round will be selected at random to be the Auction That Counts. The Auction That Counts will determine part of your Take-Home Pay and your Payout from the Auction That Counts could be positive or negative. We select one round at random as the Auction That Counts so that no participant accumulates too many losses and has a negative Take-Home Pay. Also, we select the Auction That Counts randomly so that you take every round seriously because it could be the Auction That Counts.

Take-Home Pay: You will receive a \$20.00 show-up fee for participating today. In addition to this show-up fee, we will combine

your Payout from the Money Auction to your \$20 show-up fee to determine your Take-Home Pay.

Instructions:

The page shown below explained the means by which we would calculate payments for each subject. All subjects were given a hard copy of this sheet for their reference.

Experimental Instructions

Thank you for volunteering for our experiment! Today you will participate in several different auctions in order to win money. The auctions will not be standard auctions, however, so there are some things that you need to keep in mind:

- **Participation Payment:** The payment for participating today is \$20.
- **Auctions:** Today's auctions will be different from normal auctions. Like all auctions, you will bid money in order to win a prize and the prizes will go to the highest bidders. However, the number of bidders and the number of prizes will change. The first type of auction will have 2 people bidding for 1 prize. In these auctions the pairs will be randomly reassigned each round. The second type of auction will have 20 people bidding for 10 prizes.
- **Bid:** In each auction you will specify a bid in order to win a prize. You must pay your bid regardless of the outcome of the auction so choose carefully. Whether you win or you lose you will be charged your bid.
 - **EXAMPLE:** If you bid \$5.34 in an auction and you lose, you will be charged \$5.34. Likewise, if you bid \$5.34 in an auction and you win, you will be charged \$5.34.
- **Winning Bid:** If there are more bids below your bid than above it, your bid is declared a Winning Bid and you will receive a prize. That is, if your bid is greater than half of the bids in the auction then you have a Winning Bid.
 - **EXAMPLE:** In a 2-Person auction your bid must be higher than the other bid to be a Winning Bid. In a 20-Person auction your bid must be among the 10 highest bids to be a Winning Bid.
 - **Ties:** The computer will randomly break ties between equal bids.
- **Prize Value:** The "prizes" in these auctions will be worth a certain amount of money, and this amount will be different for each bidder. Every round you will be assigned a new "Prize Value." This Prize Value is the amount of money that you will win if you have a Winning Bid in the auction.
 - **Possible Values:** Prize Values are between \$0.01 and \$20.00 and every amount of money between them is equally likely. On average, half of the Prize Values will be above \$10.00 and half of the Prize Values will be below \$10.00 but the actual numbers will vary.
 - **Hidden Values:** You will see your Prize Value but will not see anyone else's Prize Value. Likewise, no one will see your Prize Value but you. All prize values will be drawn in the same way as yours.

Pre-Experiment Quiz:

Pre-Experiment Quiz

1. Suppose I have a prize value of \$6.25 and I bid \$5.00
 - a. If I **win** my payout will be: _____
 - b. If I **lose** my payout will be: _____
2. Suppose I have a prize value of \$12.12 and I bid \$15.00
 - a. If I **win** my payout will be: _____
 - b. If I **lose** my payout will be: _____
3. Suppose I am in an auction with 2 players and 1 prize. If I have a prize value of \$10.50 and I bid \$9.00 and my opponent bids \$10.00.
 - a. My payout will be: _____
4. Suppose I am in an auction with 2 players and 1 prize. If I have a prize value of \$7.25 and I bid \$6.00 and my opponent bids \$5.00.
 - a. My payout will be: _____
5. Suppose I am in an auction with 20 players and 10 prizes. If I have a prize value of \$14.75 and I bid \$8.00 and my opponents' bids are (in increasing order): \$0 \$1.53 \$2.01 \$2.75 \$3.00 \$4.40 \$4.83 \$5.51 \$7.08 \$8.08 \$10.00 \$10.05 \$12.24 \$13.76 \$15.90 \$16.85 \$17.00 \$18.84 \$18.99.
 - a. My payout will be: _____
6. Suppose I am in an auction with 20 players and 10 prizes. If I have a prize value of \$14.75 and I bid \$8.00 and my opponents' bids are (in increasing order): \$0.54 \$1.03 \$2.91 \$3.75 \$3.80 \$4.40 \$4.83 \$5.51 \$6.08 \$7.10 \$7.75 \$10.35 \$13.34 \$13.89 \$14.90 \$17.85 \$17.90 \$18.84 \$19.99.
 - a. My payout will be: _____
7. Suppose I have a prize value of \$3.00 and I bid \$6.00.
 - a. The highest my payout can be is: _____

8. Finally, consider a more complicated situation: Here I have bid in 2 different auctions:

- a. The first auction has 20 players and 10 prizes. If I have a prize value of *\$14.75* and I bid *\$8.00* and my opponents' bids are (in increasing order): *\$0.54 \$1.03 \$2.91 \$3.75 \$3.80 \$4.40 \$4.83 \$5.51 \$6.08 \$7.10 \$7.75 \$10.35 \$13.34 \$13.89 \$14.90 \$17.85 \$17.90 \$18.84 \$19.99*.
- b. The second auction has 2 players and 1 prize. If I have a prize value of *\$14.75* and I bid *\$10.00* and my opponent bids *\$5.00*.
- c. If you are offered the payout from the first auction plus *\$1.50* your final payout will be: _____
- d. If you are offered the payout from the second auction your final payout will be: _____
- e. Lastly, if I offer you the payout from the outcome of the first auction plus *\$1.50* or the payout from the second auction you are deciding between what two numbers? _____ & _____.

Auction Round Introduction

Prior to the auction round, we posted an introduction with instructions:

Welcome to the auction round! In this round you will participate in 2 different auctions. The size of your prize has been randomly determined. We call it your "Prize Value" and you can see it below. This is the amount of money that you will receive if you win one of the prizes.

In each auction, you will compete with other participants for prizes by bidding a certain amount of money. You will be required to pay this bid regardless of the outcome of the auction. Even if you fail to win one of the prizes, you must pay your bid.

If there are more bids below your bid than above it, then you have a Winning Bid and will receive one of the prizes. That is, you will receive your Prize Value given below.

If your bid is not a Winning Bid, then you will simply lose the amount that you bid in the auction.

On the next 2 pages you will enter bids for both auctions. After all of the participants submit bids for both of the auctions, you will be told the outcomes of the auctions. You will not know the outcomes of the auctions until you have finished both auctions.

Your Prize Value \$XX.XX

Auction Round

Each screen of the auction round was identical, reminding the subject of instructions and important valuations:

This is the N -Person Auction

Remember:

- You must pay your bid regardless of the outcome.
- You will win your Prize Value if your bid is higher than $\frac{N}{2}$ opponents' bids.
- One auction will be selected randomly to be the Auction That Counts, so treat each auction as if it is the Auction That Counts.

On this page you have been randomly assigned to an auction with $N - 1$ other participants and yourself. You are competing to win one of $\frac{N}{2}$ prizes.

- Number of participants: N
- Number of prizes: $\frac{N}{2}$

Your Prize Value \$XX.XX

Please enter your bid here:

Auction Results

After the subjects all bid in both auctions, we posted all results on one page:

- Your Prize Value was: \$XX.XX
- Your bid in the 2-Person Auction was: \$XX.XX
- In the 2-Person Auction you were in group G and you (Won/Lost)
- Your Payout from the 2-Person Auction: \$XX.XX

[List all 20 bids in all 10 auctions with winning bids paired with losing bids. The subject's bid is bold.]

- Your bid in the 20-Person Auction was: \$XX.XX
- In the 20-Person Auction you (Won/Lost)
- Your Payout from the 20-Person Auction: \$XX.XX

[List all 20 bids in the auction with winning bids in the top row and losing bids in the bottom row. The subject's bid is bold.]

Risk Aversion Elicitation

There are three different risk preference elicitation tasks in the style of Andreoni and Harbaugh (2010).

Remember:

- The computer will choose 1 of the 3 pages of lotteries to determine your payment.
- You will be paid according to the outcome of the lottery you selected.

Please decide which option you prefer the most. Indicate your preference by filling in the one button next to your most preferred option:

- Win $\$X_1$ with chance C_1 in 100
- ...
- Win $\$X_{10}$ with chance C_{10} in 100

Chapter 2

A Field Experiment on Effort Allocation under Relative Grading

2.1 Introduction

Relative evaluation mechanisms are frequently used to mitigate the effects of asymmetric information between a mechanism designer and economic agents. Designers with incomplete information about agents' costs and valuations may prefer to employ relative evaluation mechanisms in which agents compete away much of each others' surpluses, since a poorly calibrated absolute mechanism offers little incentive for effort.

Consider, for example, a teacher who wishes to maximize effort from his students. Setting an absolute threshold for each grade that is too high or too low could cause students to exert minimal effort. This teacher may remedy his incomplete information about his students' abilities by grading on the curve. Under a relative or "curved" grading mechanism, the evaluation of a student's performance is based on her percentile rank within her comparison group independent of any absolute measure of her performance. These mechanisms have become a fixture in many university classrooms and law schools.¹ Indeed, mechanism designers across many areas of education employ relative evaluation in competitions for scholarships, college admissions, and even teacher pay.²³⁴

Under relative evaluation, the composition of an agent's comparison

¹Mroch (2005) estimates that 79% of law schools standardize scores according to a grading curve.

²Missouri's Bright Flight scholarship program awards scholarships to the top 3% of high school seniors based on ACT or SAT scores.

³In California, the top 9% of graduating seniors are guaranteed admission to one of the University of California campuses. In Kansas, the top 33% are guaranteed admission to the state college of their choice. In Texas, the top 10% are offered similar incentives.

⁴North Carolina Senate Bill 402 section 9.6(g) grants favorable contracts to the top 25% of teachers at each school as evaluated by the school's administration.

group is critical in determining the agent’s outcome at a given level of effort. I refer to a sample drawn from the population into a comparison group as a “cohort.” The law of large numbers implies that the larger this cohort becomes, the more it comes to resemble the distribution from which it is drawn. Since the composition of a cohort determines the incentives for effort by an agent, the size of a cohort affects those incentives by bringing its expected composition closer to the population distribution. Therefore, in theory, the incentives for effort under any relative evaluation mechanism can be modeled and understood as a function of cohort size. A mechanism designer concerned about the allocation of incentives to agents could equalize incentives across any two cohort sizes by manipulating the number or value of prizes awarded. In practice, however, mechanism designers typically operate under constraints restricting them to a given number or percentage of prizes awarded regardless of cohort size. For example, a professor may be obligated to award A’s to 20 percent of his students. This paper evaluates the theoretical and empirical effect of cohort size on incentives for effort in a relative evaluation mechanism under such a constraint.

Consider an example, Texas HB 588 grants automatic admission to any Texas state university—including the University of Texas at Austin—to all Texas high school seniors who graduate in the top 10 percent of their high school class.⁵ Each graduating class is therefore subject to a distinct relative evaluation mechanism under the restriction that each mechanism must award 10 percent of the students automatic admission, regardless of the class

⁵The bill was modified in 2009 to stipulate that the University of Texas at Austin may cap the number of students admitted under this measure to 75% of in-state freshman students.

size. Since Texas high schools vary in size by orders of magnitude,⁶ the law of large numbers implies that the incentives for effort under this policy will vary dramatically by school. Smaller high schools are more likely to draw graduating classes full of outliers, adding uncertainty around the returns to effort, while cohorts of students in larger high schools are more likely to reflect the characteristics of the population, reducing the uncertainty around the returns to effort.

While the theory clearly indicates that incentives change with the size of a cohort, whether students identify and react to these incentives is uncertain. Prior research casts doubt on the ability of economic agents to draw accurate inference about information that depends critically on sample size (Tversky and Kahneman, 1971; Kahneman and Tversky, 1973; Rabin, 2002; Benjamin, Rabin, and Raymond, 2014). Therefore, research is necessary to uncover the reactions of agents to these changing incentives.

This paper presents the results of a field experiment on relative grading in a large, upper-division economics course at the University of California, San Diego (UCSD) testing the sensitivity of student effort to cohort size. I use an experimental intervention in the size of a cohort to test its causal effect on effort by students who are graded on the curve. In order to provide the most powerful test of this effect, I present students with a pair of quizzes subject to an identical grading curve where the top 70 percent of scores in a cohort receive high grades. Each week, for each student, I randomly determine which quiz in the pair will be graded relative to a 10 student cohort and which will be

⁶Plano East High School has an enrollment of 6,015 students, while Valentine High School has an enrollment of 9 students.

graded relative to a 100 student cohort. Since cohort size is randomly assigned to quizzes, I will refer to these as the “10-Student Quiz” and the “100-Student Quiz.” I measure effort as the time a student spends on a given quiz.

By comparing a pair of quizzes taken in quick succession but with different cohort sizes, I can use the difference in the amount of time spent on each quiz to measure the causal impact of cohort size on effort within a given student, within a time period. This cleanly identifies student responses to changes in their strategic setting independent of any potential classroom-specific, student-specific, or time-specific confounds that can plague classroom studies.

In order to generate predictions about the ways in which a student subject to curved grading would respond to changes in the cohort size, I develop a theoretical model of strategic effort exertion in the classroom. Strategic concerns are just one of countless motivations a student may have to exert effort in the classroom. My results uncover the relevance of these other motivations per se and confirm that my experimental design has eliminated all of these potentially confounding factors. I therefore abstract away from any other motivation for effort and develop predictions based on expected-grade maximizing responses to the underlying strategic incentives of the curved grading environment. In this simplified model, I make several assumptions about students’ utility functions, their beliefs about their ability relative to their classmates, their responses to those beliefs, and the separability of their cost functions. As such, the value of this model is in generating qualitative predictions about the allocation of effort by strategic students responding to a change in the size of their grading cohorts. Indeed, from a policy perspective, it is the empirical

effect of the cohort size on effort allocation that is paramount. Confirming or rejecting a given model of strategic effort will always be a secondary concern.

My model predicts that mean effort will increase with the size of the cohort. As the cohort size grows and the characteristics of the cohort draw closer to the population characteristics, the uncertainty surrounding the returns to any given level of effort decrease, drawing up the mean effort.

My model also provides structure for how incentives differ by ability level. In particular, incentives diverge on either side of the the 30th percentile of scores, where high and low grades are separated. I refer to the quantile that distinguishes the two outcomes as the “cutoff.” The 10-student cohorts have higher variance, increasing the marginal benefit of effort for students with abilities below the cutoff. With lower variance, the 100-Student Quiz increases the marginal benefit of effort for students with abilities above the cutoff. Thus, my model predicts that “low-ability” students—those with GPAs below the 30th percentile—will exert more effort on the 10-Student Quiz, and “high-ability” students—those with GPAs above the 30th percentile—will exert more effort on the 100-Student Quiz. These heterogeneous effects highlight the tradeoffs between mean effort and the distribution of effort, a key tension for optimizing classroom design.

My experimental results confirm the prediction that the mean effort will increase in the cohort size. The 100-Student Quizzes elicit 5 percent more effort than the 10-Student Quizzes, and this difference is significant at the 1 percent level. Next, I use GPA as a proxy for ability in order to test my model’s predictions about the heterogeneous impact of cohort size on students with different abilities. I first confirm that high-ability students exert significantly

greater effort on the 100-Student Quiz. My experimental results then deviate from the model's predictions in an important way. While the *lowest* ability students exert more effort on the 10-Student Quizzes, this effect is not found for all students below the cutoff. In fact, *mean* effort by students below the cutoff is higher on 100-Student Quizzes. For these students, allocating greater effort to the 100-Student Quiz fails to take advantage of the higher variance of the smaller cohorts. These deviations from the theoretical predictions highlight the limits to the predictive ability of a purely strategic model of classroom effort. Specifically, a strategic model under classical assumptions about beliefs and best responses will overstate the distributional costs of increasing mean effort. That is, effort from low-ability students is less negatively affected by increasing cohort sizes than theory predicts.

I attempt to uncover the origin of the misallocation of effort by low-ability students by exploring the critical assumption that students hold accurate beliefs about their own ability relative to the distribution of abilities among their classmates. A violation of this assumption will cause the perceived incentives to deviate from the incentives the model predicts. For this assumption to hold, students must forecast their distribution of classmates taking into account the fact that higher ability students are more likely to enroll in upper-division economics courses. “Cursed” students (Eyster and Rabin, 2005), on the other hand, may wrongly perceive that the distribution of students in the class is simply an “average” draw from the undergraduate population similar to the draws they faced in previous lower-division courses. In order to examine the predictions of my model under cursed beliefs, I construct a distribution of classmates using public data on grades from from

lower-division UCSD courses. My results show that the observed behavior of students is closer to the expected behavior for students who best respond to cursed beliefs that fail to account for selection into the course.

My results suggest that cursed beliefs are not the only potential complication in evaluating strategic effort exertion. Allowing for “fully cursed” beliefs⁷ shifts the predicted single crossing point towards lower abilities and increases the number of students predicted to exert greater effort on the 100-Student Quiz but still cannot explain the behavior of many low-ability students. The misallocation of effort by these students is consistent with several possible behavioral biases, among them, overconfidence, updating failures, and reference dependence. My experiment cannot independently identify these possible behavioral responses, but I discuss the impacts of each and outline experiments that could identify the extent to which each is responsible for deviations from the theoretical predictions.

My data can cleanly reject several alternative explanations for student behavior. A model where high-ability students have greater intrinsic motivation provides some similar predictions about how incentives for effort evolve with student ability and may be present as an incentive for effort. Importantly, my experiment controls for this with its within-student analysis. Therefore, without allowing for sensitivity to the strategic incentives present, this model cannot generate the patterns of effort observed in my experiment. In a similar way, my data can reject the notion that risk preference or demographic characteristics can explain the results entirely. I can also reject the hypothesis

⁷Under cursedness, there are degrees to which agents infer information from other agents’ actions. A fully-cursed agent draws no inference from the actions—in this case, enrollment decisions—of any other agent.

that effort allocation is driven by a correlation between GPA and the ability to intuit the incentives of the environment. Indeed, my paper provides clean evidence that students respond strategically and systematically to changes in the size of their grading cohorts, despite many other observable and unobservable motivations for effort.

To fix ideas, I refer to grading mechanisms throughout this paper, but this should not distract from the generality of the results. Relative awarding mechanisms are found throughout the modern economy in job promotion contests, performance bonuses, and lobbying contests, among others.⁸ Since the costs of effort, the means of exerting it, and the ways in which heterogeneous abilities manifest themselves are all similar in academic and professional settings, my results provide a framework for predicting how certain mechanisms will affect the allocation of effort across employees.

The outline of this paper is as follows. In Section 2, I provide a survey of the relevant literature. In Section 3, I outline a simple model of incentives under relative grading. This model yields qualitative predictions that I test experimentally. The experiment itself is formally introduced in Section 4. Section 5 presents the results of the experiment and tests the predictions of the model, addressing several alternatives to my model including cursed beliefs. Section 6 discusses the results. Section 7 concludes the paper.

⁸For example, in his book *Straight from the Gut*, former GE CEO Jack Welch recommends that managers rank employees according to a 20-70-10 model of employee vitality where 20% of employees are labeled “A” players, 70% “B” players, and 10% “C” players. “A” players are rewarded, and “C” players are eliminated.

2.2 Literature

In addressing the strategic incentives of classroom grading mechanisms, this paper spans three distinct literatures: experimental economics, microeconomic theory, and the economics of education. In the realm of experimental economics, it owes a debt to many prior experimental tests of contests and auctions. My model has roots in a long theoretical literature on contests. Notably, Becker and Rosen (1992), who modify the tournament structure of Lazear and Rosen (1981) to generate predictions for student effort under relative or absolute grading mechanisms. By modeling and collecting data in a classroom setting, my paper contributes to a literature on classroom performance that has been explored in the economics of education.

Experiments testing effort exertion in different laboratory settings date back to Bull, Schotter and Weigelt (1987), who test bidding in laboratory rank-order tournaments. They find that bidders approach equilibrium after several rounds of learning. Equilibrium behavior in laboratory all-pay auctions is more elusive with the majority of studies demonstrating overbidding (Potters, de Vries, and van Winden, 1998; Davis and Reilly, 1998; Gneezy and Smorodinsky, 2006; Barut, Kovenock, and Noussair, 2002). Müller and Schotter (2010) and Noussair and Silver (2006) confirm the overbidding result, but also uncover heterogeneous effects for different types of bidders. For an exhaustive survey of the experimental literature on contests and auctions, refer to Dechenaux, Kovenock, and Sheremeta (2012).

Andreoni and Brownback (2014) provide a framework for evaluating the effects of contest size on bids in a laboratory all-pay auction along with

the first directed test of the independent effect of contest size on effort. Larger contests in this setting are found to generate greater aggregate bidding, greater bidding by high types, and lower bidding by low types. Other studies that find effects of contest size on effort restrict their focus either to small changes in the size of contest (Harbring and Irlenbusch, 2005) or changes that also affect the proportion of winners (Gneezy and Smorodinsky, 2006; Müller and Schotter, 2010; Barut et al., 2002, List et al., 2014).

This paper takes the framework of Andreoni and Brownback (2014) out of the laboratory and into a field setting and is, to my knowledge, the only study that directly measures effort as a function of the classroom size. Classroom experiments have been conducted to answer other questions. The state of Tennessee experimented with classroom sizes for kindergarten students (Mosteller, 1995), but student outcomes, not inputs, were the focus of the study and the setting was non-strategic. Studies have also explored the responsiveness of effort to mandatory attendance policies (Chen and Lin, 2008; Dobkin, Gil, and Marion, 2010) or different grading policies (Czibor et al., 2014), finding mixed results. I explicitly control for factors related to classroom instruction or procedures in order to uncover the direct effect of changes in the strategic incentives for effort.

In the microeconomic theory literature, the study of contests was originally motivated by the study of rent-seeking (Tullock, 1967; Krueger, 1974), but has since evolved into a more general branch of research that considers various environments with costly effort and uncertain payoffs. The three models most often employed are the all-pay auction (Hirshleifer and Riley, 1978; Hillman and Samet, 1987; Hillman and Riley, 1989), the Tullock contest (Tullock,

1980), and the rank-order tournament (Lazear and Rosen, 1981).

Hillman and Riley (1989) and Baye, Kovenock, and de Vries (1993) use the all-pay auction model to explore the incentives for rent-seeking in politics. Amann and Leininger (1996) introduce incomplete information about opponents' types to generate a pure-strategy equilibrium bidding function. Baye, Kovenock, and de Vries (1996) fully characterize the equilibrium of the all-pay auction and demonstrate that a continuum of equilibria exist. Moldovanu and Sela (2001) develop a model of optimal contest architecture for designers with different objectives. For a comprehensive theoretical characterization of all-pay contests that incorporates many of the existing models into one framework, see Siegel (2009).

This paper is motivated by the way in which the size of a contest changes the incentives for participants. Moldovanu and Sela (2006) capture this intuition more generally and demonstrate the single-crossing property of symmetric equilibria in differently sized contests. Olszewski and Siegel (2013) provide similar results about equilibria in a general class of contests with a large but finite number of participants.

My paper considers strategic interactions *between* students, unlike much of the previous work on grading mechanisms. The contest-like relative grading mechanisms as well as absolute grading mechanisms are often studied in the economics of education. Costrell (1994) explores the endogenous selection of grading standards by policy makers seeking to maximize social welfare, subject to students who best respond to those standards. Betts (1998) expands this framework to include heterogeneous students. Betts and Grogger (2003) then look at the impact of grading standards on the distribution of students.

Both Paredes (2012) and Dubey and Geanakoplos (2010) compare incentive across different methods of awarding grades. The former considers a switch from an absolute to a relative grading mechanism, while the latter finds the optimal coarseness of the grades reported when students gain utility from their relative rank in the class.

The education literature traditionally studies class size like an input to the production function. The aforementioned Mosteller (1995) paper operates in this vein, finding that the decrease in class size caused by the Tennessee class size project had lasting impacts on the outcomes of students. Kokkelenberg, Dillon, and Christy (2008) find a negative effect of class size on the grades awarded to individual students in a non-strategic environment. The independent effect of class size on strategically interacting students, however, remains unstudied. Contest size is often taken as given or assumed to be determined exogenously. In this paper, I demonstrate that cohort size plays a significant role in a student's selection of effort when grading on the curve. Thus, a classroom designer optimizing student outcomes needs to take into consideration the heterogeneous effects of cohort size on students with different abilities.

2.3 A Model of Academic Effort

In this section I develop a simple framework outlining the incentives for effort present when grades are awarded on a relative basis. This model will provide generic predictions about the direction a strategic student should shift effort as the cohort size changes. Providing a formalization of the way student incentives are tied to cohort size will be instructive for building intuition and

developing ways to test whether students view the classroom as a strategic environment and if they predictably respond to shifts in strategic incentives.

Strategic incentives will be present amidst myriad other incentives for effort. As such, the contribution of this model is in the structure that it gives to the *relative* incentives for effort, not the point estimates it provides for the amount of effort each student chooses. Focusing on the relative incentives for effort provides predictions independent of other motivations for effort, better matching my experiment, which explicitly controls for alternative motivations.

The heterogeneity this model predicts relies on several assumptions about the beliefs and information that each student has. In section 6, I discuss the model's dependence on these assumptions and the way predictions change when they are relaxed.

In solving this model, I borrow heavily from the independent private value auction (Vickrey, 1961) and the all-pay auction literatures (Baye, Kovenock, and de Vries, 1993). Suppose there are N students exerting costly effort in order to increase their chances of winning one of $M \equiv P \times N$ prizes in the form of high grades. Effort appears as scores on quizzes, and high grades are awarded to the students with the highest M scores.

Suppose each student has an ability, a_i , distributed uniformly from 0 to 1. That is $a_i \sim U[0, 1]$, meaning $F(a_i) = a_i$. Students are evaluated at each period, t , based on their academic output, or “score,” $s_{i,t}$ from that period. Suppose that scores have a constant marginal cost that is inversely related to

ability,⁹

$$C(s_{i,t}; a_i) = \frac{s_{i,t}}{a_i}. \quad (2.1)$$

2.3.1 Student's Utility

The expected utility of a student is determined by both the likelihood of receiving a high grade at a given score and the cost of that score. Normalize the value of receiving a high grade to one. Heterogeneity across students with different ability levels is now captured by the cost of generating a given score. Thus, a student's utility is given by

$$U(s_{i,t}; a_i) = \Pr(s_{i,t} \geq \bar{S}) - \frac{s_{i,t}}{a_i}, \quad (2.2)$$

where $\bar{S}$ represents the minimum score required to receive a high grade.

I restrict my attention to the set of functions, $S : (a_i; N, P) \mapsto s_{i,t}$, that take parameters, N and P , and map abilities to scores in such a way as to constitute a symmetric equilibrium of the model. In the appendix, I prove that any such function must be monotonic in ability. In addition to monotonicity, it is straightforward to show that scores must also be continuous in ability.¹⁰ All continuous, monotonic functions are invertible, so there exists a function that maps a given score back onto the ability implied by that score. Given that the equilibrium scores depend on the parameters, N and P , this inverse function, too, depends on these parameters. I define the function

⁹The actual value of the marginal cost will not be critical as my within-subjects experimental design controls for student-specific costs of effort.

¹⁰Suppose not. With discontinuities within the support of $S(a_i; N, P)$, some students would be failing to best respond. A student scoring just above the discontinuity would be able to increase his expected utility by lowering his score, which would lower his costs, up until the discontinuity in scores has vanished.

$$A(s_{i,t}; N, P) \equiv S^{-1}(s_{i,t}; N, P).$$

With invertibility established, the probability of receiving a high grade is equivalent to the probability that the ability level implied by a student's score is higher than the ability levels implied by the scores of $N - M$ other students. At equilibrium, this probability is represented as an order statistic giving the probability that a student's ability is among the top M abilities of her cohort. Substituting in the experimental parameters, $N = \{10, 100\}$ and $P = 0.7$, yields the order statistics presented in Figure 2.1.

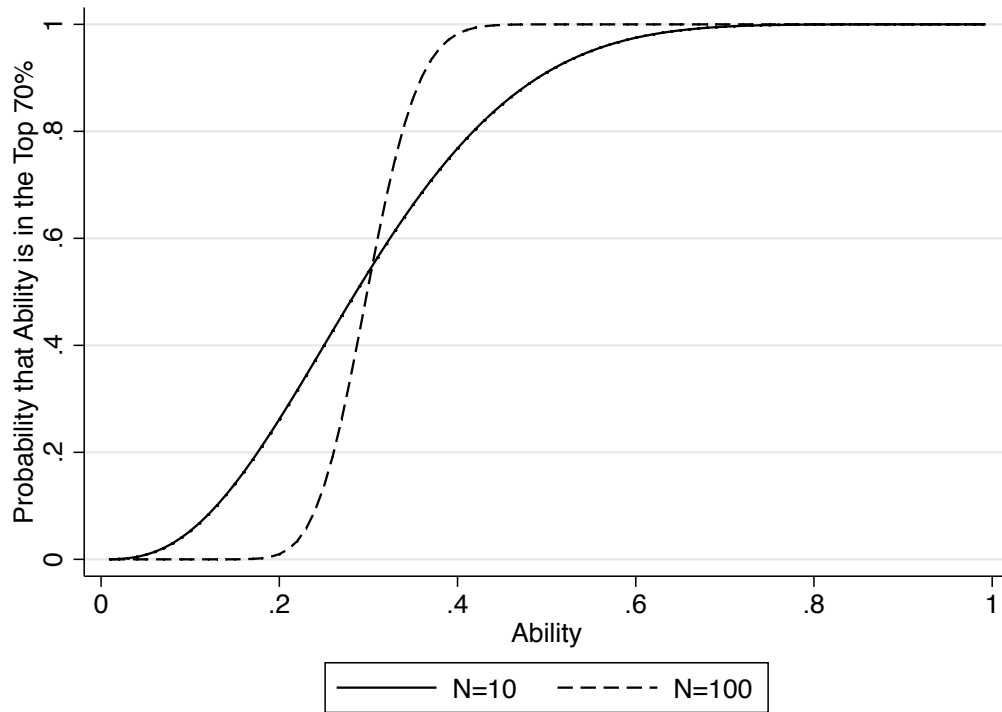


Figure 2.1: Probability a Given Ability is in the Top 70% of Abilities in a Cohort of Size N

Figure 2.1 reveals the key intuition of this paper. For low-ability students, the uncertainty of the 10-Student Quiz increases their likelihood of

encountering cohorts in which they are among the top 70 percent. For high-ability students, that same uncertainty decreases the likelihood that they are among the top 70 percent of their cohort. These probabilities heterogeneously vary the expected returns to student effort.

Plugging the order statistics into (2.2) completes the student's utility function.

$$U(s_{i,t}; a_i, N, P) = \sum_{j=N-NP}^{N-1} \left(\frac{(N-1)!}{j!(N-1-j)!} \right) A(s_{i,t}; N, P)^j \times (1 - A(s_{i,t}; N, P))^{N-1-j} - \frac{s_{i,t}}{a_i}. \quad (2.3)$$

In the appendix, I solve for the equilibrium scores as a function of ability. Figure 2.2 plots the equilibrium score functions at the experimental parameter values, $N = \{10, 100\}$. It is worth noting that the point estimates represented are not valuable per se, as my predictions relate to relative, not absolute scores.

While the function mapping the ability of a student to his or her score at equilibrium is clearly quite complicated, the intuition behind it is rather simple. low-ability students benefit from the introduction of randomness into the draw of their cohort, and put forth greater effort. Conversely, randomness is detrimental to high-ability students, so their effort decreases.

A thought experiment can reveal the intuition behind the equilibrium score functions in more depth. Consider the symmetric best response function in an environment where the proportion of winners remains constant at $P = 0.7$, but the number of students in a cohort approaches infinity. The law of large numbers ensures that the distribution of students in the cohort

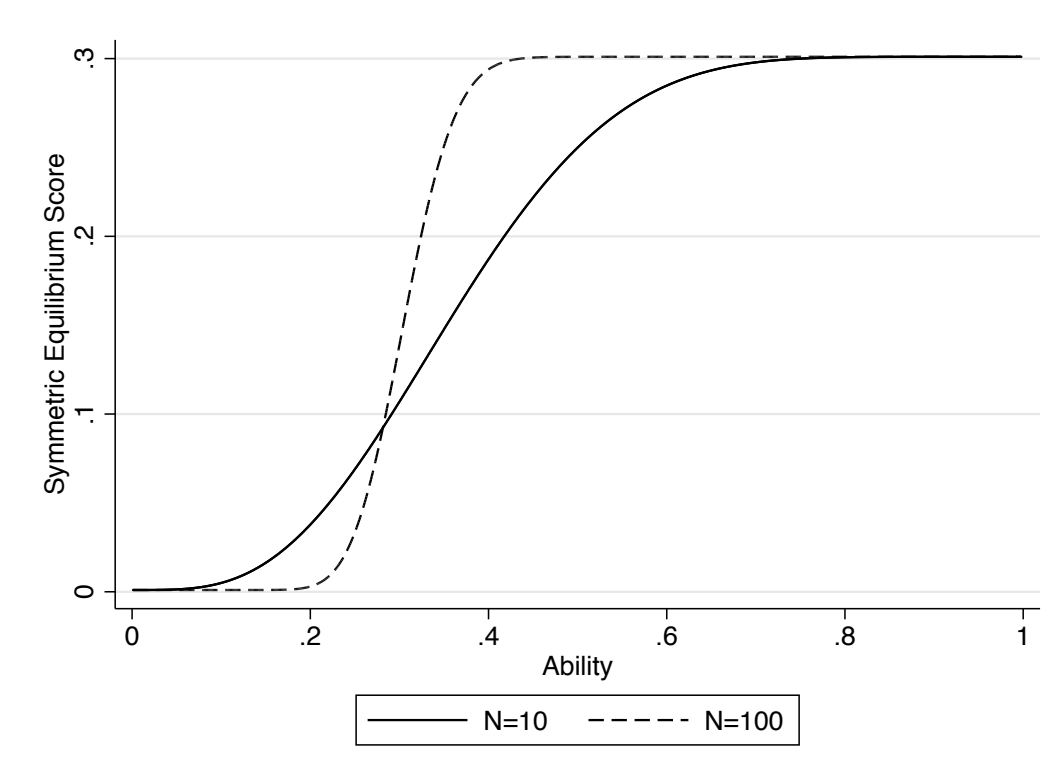


Figure 2.2: Scores at Equilibrium in the 10- and 100-Student Quizzes

approaches a perfect reflection of the probability distribution from which they are drawn. Thus, common knowledge of the probability distribution is sufficient for a student's belief about her relative position in her cohort to approach certainty.

In this infinitely large cohort, a student whose ability is greater than the 30th percentile in the probability distribution will best respond by choosing a score that no student below the 30th percentile can match and receive non-negative expected surplus. That score, given the assumed cost function, is approximately $s_{i,t} = 0.3$.¹¹ Students below the 30th percentile best respond with a score of $s_{i,t} = 0$. Thus, the equilibrium score function in this setting

¹¹This value itself means little except as an ordinal measure of academic output.

approaches a step function.¹²

Keeping in mind the limiting case, consider the equilibrium scores in Figure 2.2. Deviations from the infinite case can be explained through marginal benefits and marginal costs. For students who choose $s_{i,t} > 0$ in the limiting case, the marginal benefit of lowering a score is constant and identical across treatments—scores have a constant marginal cost, so foregone scores have a constant marginal benefit. The marginal cost of lowering a score is paid through reductions in the probability of receiving a high grade. In the 10-Student Quiz, that probability changes more gradually, making marginal reductions of scores less costly, causing the 10-Student Quiz scores to drop below the 100-Student Quiz scores.

Now consider students with $s_{i,t} = 0$ in the infinitely sized cohort. Increasing scores bears a constant marginal cost for both the 10- and 100-Student Quizzes, but holds a higher marginal benefit in the 10-Student Quiz because the randomness increases the likelihood of states of the world in which low scores receive high grades. So, the 10-Student Quiz scores rise above the 100-Student Quiz scores.

2.3.2 Predictions From the Model

My experiment pairs 10- and 100-Student Quizzes each week, and my analysis takes the difference in effort between the two quizzes—specifically,

¹²To be precise, the rules of this environment need to change to accommodate the fact that students have a probability measure of 0 in an infinitely large contest. One option is to begin by awarding a prize to the student with the highest score and continue awarding prizes to students in decreasing order of score until 70 percent of students have been awarded prizes. Thus, there is no profitable deviation where a student earns a prize with a score of $s_{i,t} = \epsilon$

the 100-Student Quiz duration minus the 10-Student Quiz duration—as its dependent variable. I refer to this difference as the “treatment effect.” I use the model’s predictions for the difference in scores as a proxy for its predictions about the difference in effort. This allows me to remain agnostic about the production function for scores, only assuming that higher predicted scores imply higher predicted effort. My within-subjects analysis of this difference in effort controls for student-specific heterogeneity in costs of production and will provide a cleaner test of the treatment effect. To see the model’s predictions for how the treatment effect will evolve with ability consider Figure 2.3, which plots the equilibrium score in the 100-Student Quiz minus the equilibrium score in the 10-Student Quiz as a function of ability.

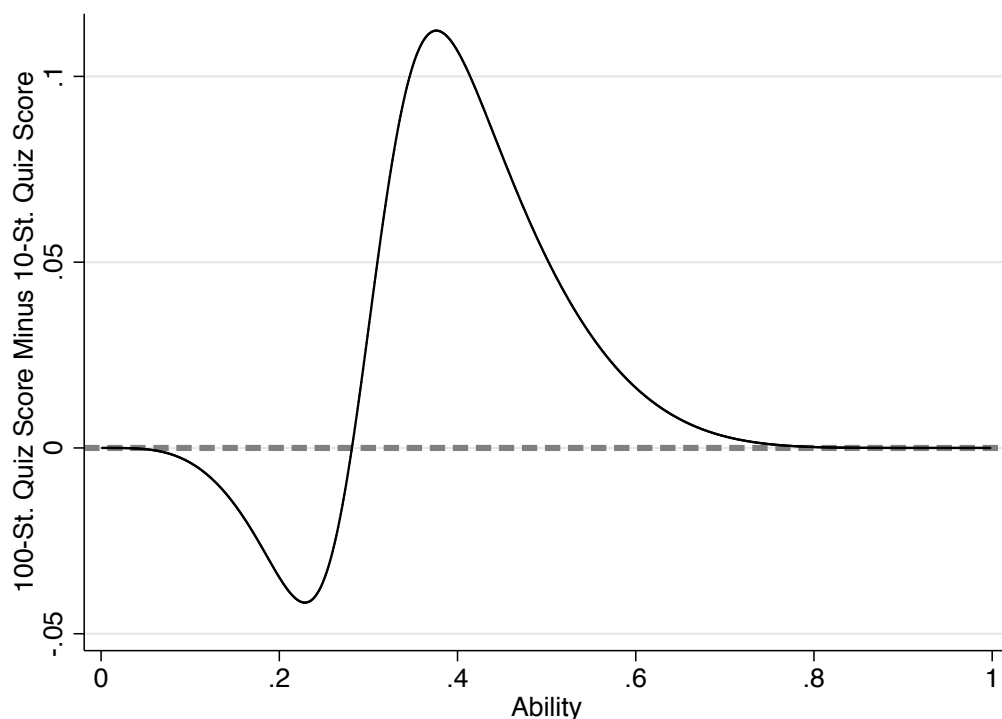


Figure 2.3: Difference in Score between 100- and 10-Student Quizzes at Equilibrium

The model provides three primary predictions about the treatment effect displayed in Figure 2.3. While the point estimates of the model are based on specific assumptions, the following predictions represent generic qualities that provide clarity about the ways in which an expected-grade maximizing student may react to changes in the grading environment.¹³

Hypothesis 1: Mean effort is increasing in cohort size.

My model predicts that the greater effort exerted by high-ability students on the 100-Student Quiz outweighs the greater effort that low-ability students exert on the 10-Student Quiz, causing average effort to increase in cohort size.

Hypothesis 2: The treatment is negative for low-ability students.

Hypothesis 3: The treatment is positive for high-ability students.

My model predicts that the treatment effect begins negative and moves positive, crossing the horizontal axis exactly once. Call this single-crossing point a^* . Based on Figure 2.1, it is natural to think of a^* as corresponding with the cutoff. This is approximately true, but the specific location of a^* will depend on the distribution of abilities, and the cost function for scores. My predictions focus on the cutoff of $a_i = 0.3$ as a^* .¹⁴

¹³Moldovanu and Sela (2006) prove these properties for a general class of cost functions. These predictions derive from a single-crossing property they prove for symmetric equilibria in contests with different N but fixed P .

¹⁴While the single-crossing point in Figure 2.3 is not precisely 0.3, the salience of the 30th percentile in my experiment and the proximity of the single-crossing point to this value make it a natural candidate.

2.4 Experimental Design

My experiment takes the paired-auction design used in Andreoni and Brownback (2014) and adapts it for a classroom context. My design simultaneously presents students with a pair of quizzes, a 10-Student Quiz and a 100-Student Quiz, and records student behavior on each. This design is inspired by the paired auction design of Kagel and Levin (1993) and Andreoni, Che, and Kim (2007).

I analyze the difference in behavior between the two simultaneous quizzes in order to control for student-specific and week-specific effects. This paired design will provide a powerful test of responses to cohort size that occur in the classroom, an environment riddled with alternative motivations for effort.

2.4.1 Recruitment and Participation

The experiment was conducted in the winter quarter of 2014 in an intermediate microeconomics course at UCSD. Enrollment in the course started at 592 students, and ended at 563 after some students withdrew from the course. All enrolled students agreed to participate in the experiment. The experiment was announced both verbally and via web announcement at the beginning of the course. The announcement can be found in the appendix.

2.4.2 Quiz Design, Scoring, and Randomization

There were five Quiz Weeks in the quarter. At noon on Thursday of each Quiz Week, two different quizzes covering material from the previous week were posted to the course website. Both quizzes were due by 5pm the following

day. Each quiz consisted of four questions and had a time limit of 30 minutes. The 30-minute limit ensures that the quiz was given focused attention with little time spent idle, meaning that the time recorded for students is reflective of their effort on the quizzes. Students could take the quizzes in any order.

I refer to the content of the two quizzes in a pair as “Quiz A” and “Quiz B.” The two quizzes were presented to students in the same order, but I randomly assigned grading treatments to each. One of the quizzes received the 10-Student Quiz treatment and one received the 100-Student Quiz treatment. Therefore, while every student was assigned both Quizzes A and B, approximately half of them had Quiz A graded as the 10-Student Quiz and half had it graded as the 100-Student Quiz. The opposite treatment was assigned to Quiz B in each case. The questions on Quizzes A and B were designed to have as little overlap as possible to eliminate order effects in effort or scores. Before beginning the quiz, students only observed the grading treatment, and not the quiz content. No student was informed of the treatments received other students. Table 2.1 shows the balance across treatments.

Table 2.1: Submitted Quizzes from Each Week and Each Treatment

Quiz Version	Treatment	Week 1	Week 2	Week 3	Week 4	Week 5
A	10-Student	201	282	258	253	249
	100-Student	262	282	259	253	243
B	10-Student	267	280	262	251	243
	100-Student	200	282	259	256	252

Note: Asymmetries across treatments may arise out of chance, failed submissions, or withdrawals. Asymmetries in completion rates will not affect the analysis, since only completed pairs will be analyzed.

The number of questions correct determined the score for each student. The top 7 scores received high grades in each 10-Student Quiz cohort, and the

top 70 scores received high grades in each 100-Student Quiz cohort. Students were informed that they would be anonymously re-randomized into cohorts each week. Cohorts consisted of students who had taken the same quiz under the same grading treatment. Students receiving high grades were awarded 3 points, while students receiving low grades were awarded 1. Non-participants received 0 points. The quizzes counted for approximately 13 percent of the total grade in the class, providing strong incentives for effort. Students whose scores were tied at the 70th percentile of a cohort were all awarded 3 points unless the tied students all failed to participate, in which case the students all received 0 points.

2.4.3 Effort

The time at which every quiz was started and completed was recorded to the millisecond. My analysis will take the amount of time that a student spent taking a quiz to be the measure of effort that the student exerted on that quiz. This measure of effort will reveal which quiz the student believed to hold the greater returns to her effort.

Both quizzes were posted simultaneously, meaning that the amount of time a student could spend studying prior to starting either quiz was roughly constant between the two quizzes. Student behavior supports this assertion. 86 percent of students waited less than an hour between the two quizzes, with a median interval of 32 minutes.

2.4.4 Ability

At the beginning of the course, students consented to the use of their grade point average (GPA) in this study.¹⁵ I use this as the measure of academic ability in the analysis. I elected not to use exam performance in the course because of the endogeneity between the allocation of effort to exams and quizzes. I contend that GPA is a more valid instrument for ability because, unlike exams, there is no sense in which quiz effort and GPA are substitutable. While there may be a correlation between the level of effort and ability as measured by GPA, my analysis will eliminate these level effects by only considering differences in effort between pairs of quizzes.

Figure 2.4 shows the cumulative distribution function of all student GPAs in the class. I use the value of the cumulative distribution function at a given GPA to represent the ability level of that student in my predictions. Importantly, the GPA at the 30th percentile is 2.72. The median and mean GPA are 3 and 2.99, respectively. Seven students have a GPA of 4.0, while only one student has the minimum GPA of 1.0.

2.5 Results

I begin this section by describing my data and their basic statistics. Then, I specify the dependent variable I will use in the analysis and test its aggregate characteristics. Next, I demonstrate heterogeneous treatment effects

¹⁵Due to administrative delays, I was not able to get a student's GPA until after the quarter. Thus, the response to the treatment will have some impact on ability. Since the quizzes only amounted to approximately 13 percent of the students grade in one of dozens of classes they have taken, I do not see this as a major problem.

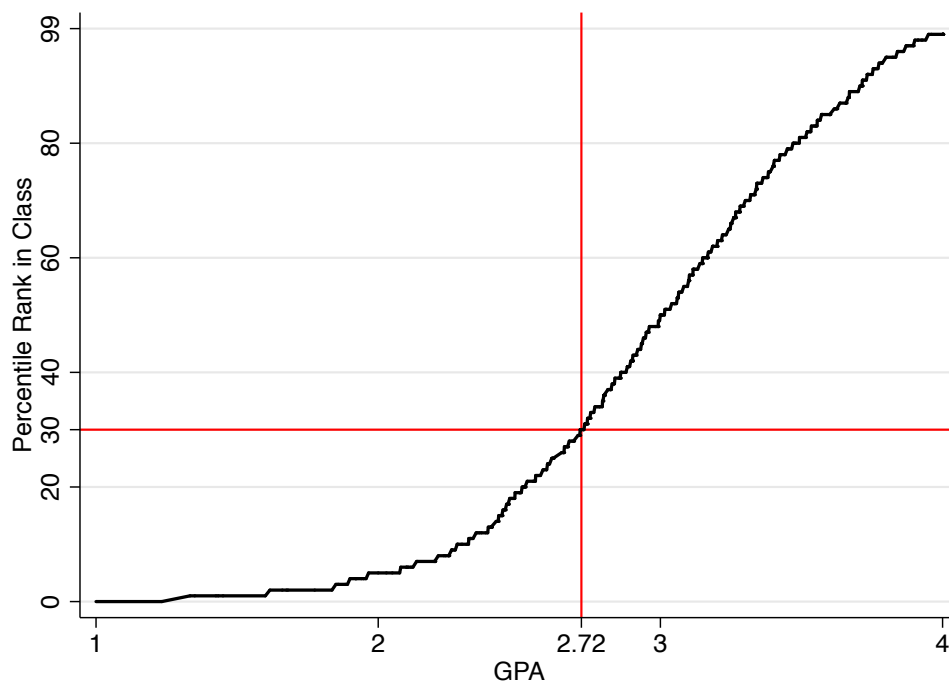


Figure 2.4: Cumulative Distribution Function for GPAs of All Students in the Experiment

across students of different ability levels and test where the model does and does not hold predictive power. Finally, I investigate possible explanations for the model's failures.

2.5.1 Data and Descriptive Statistics

In total, 579 students submitted 5,094 online quizzes in this experiment. Of those, 2,546 had been assigned to a 10-Student Quiz, and 2,548 had been assigned to a 100-Student Quiz. The duration of each quiz was recorded, and my analysis will include every recorded time appearing in a completed pair of quizzes. Table 2.2 reports the means and standard deviations of the quiz

durations for both cohort sizes.¹⁶

Table 2.2: Descriptive Statistics on the Duration of the Quizzes

	100-Std. Quiz	10-Std. Quiz
Mean Duration (<i>in minutes</i>)	14.97	14.58
Standard Deviation/Error	(9.21)	(9.00)
Note: In this table, the quiz means are unpaired, so will provide a much weaker test of significant differences.		

2.5.2 Dependent Variable

My analysis uses the difference between the time allocated to the 100- and 10-Student Quizzes as the dependent variable. Recall that I refer to this difference as the treatment effect. This dependent variable is appealing because it reveals a student's beliefs about which quiz will yield higher returns to her effort. Since random assignment leaves effort costs and quiz difficulty independent of the cohort size, if a student shows a general trend toward spending more time on the 10- or 100-Student Quiz, then the student must believe that her marginal product is higher on that quiz.¹⁷ Using within-student differences also offers the best control for individual-specific and week-specific noise in the data.

¹⁶Since I did not receive GPA data until the end of the quarter, I was not able to observe the GPAs of the students who dropped during the quarter. There were only 20 submitted pairs of quizzes from students who dropped the course. The mean treatment effect for these quizzes is approximately -36.7 seconds with a standard deviation of 432 seconds. I include these data in the analysis of the mean treatment effect, but exclude them from the analysis of heterogeneity in treatment effects, since I have no measure of ability. Each of these decisions biases my results away from the model's predictions. For the means, it lowers the average effect, diminishing my results. With respect to heterogeneity, their inclusion only strengthens my results, because they are more likely to be low GPA students, and their average treatment effect is negative.

¹⁷Since I only include times recorded in a completed pair of quizzes, my analysis reveals the perceived relative returns to effort conditional on participation in both quizzes. The results do not substantively change by replacing all skipped quizzes with values of 0, but the standard errors expand, as 0 is a much shorter duration than any observed in the data.

2.5.3 Endogenous Selection and Controls

One potential complication in these results is that the order of quiz completion could not be controlled and thus is endogenous.¹⁸ Fortunately, even though quizzes are presented simultaneously, one quiz is positioned above the other vertically. The presentation order of treatments is randomly assigned and provides a relevant instrument for the order of completion that is mechanically designed to be valid. The effect of this endogenous selection is not large—51.4 percent of 100-Student Quizzes were presented first online, while 55.3 percent were completed first—but is statistically significant.

Column 1 of Table 2.3 demonstrates that the instrument is extremely relevant. Column 2 shows that the instrument has a problematic correlation with student GPA. Despite being randomly assigned before each Quiz Week, the order in which the quizzes were presented happened to correlate to the GPAs of the students. This is unfortunate but was unavoidable, since I did not have access to student GPAs until the end of the quarter. Additionally, the explanatory power is minimal, with an R^2 value below 0.002. Column 3 demonstrates that the residual effect of GPA on the order in which a student completes the quizzes has negligible explanatory power and is not statistically significant after controlling for presentation order. To demonstrate that endogenous selection does not drive any results, all tables will feature results with and without the instrument for quiz completion order.

¹⁸This is a limitation of the online environment. In order to force students to take quizzes in a specific order the second quiz must be hidden from view until the completion of the first quiz. I posted both quizzes simultaneously in order to ensure that students knew they were assigned two quizzes.

Table 2.3: Testing the Relevance and Validity of the Instrument

	(1) Pr{100-St. Quiz completed first}	(2) Pr{100-St. Quiz presented first}	(3) Pr{100-St. Quiz completed first}
<i>100-St. Quiz presented first</i>	0.759*** (0.05)		0.757*** (0.05)
GPA		0.105** (0.05)	0.060 (0.05)
Constant	-0.243*** (0.04)	-0.290** (0.14)	-0.422*** (0.15)
R^2	0.064	0.002	0.065
N	2,486	2,486	2,486

* p<0.10, ** p<0.05, *** p<0.01

Note: The first column demonstrates that the order in which quizzes are displayed is a highly relevant instrument for the order in which the quizzes are completed. The second column shows the troubling correlation that the mechanically random instrument has with the GPAs of students, though the explanatory power is negligible. The third column shows that the relevance of the instrument is not a result of its correlation with GPA.

Hypothesis 1: Mean effort is increasing in cohort size.

Table 2.4 tests my model's straightforward prediction that mean effort is increasing in the cohort size. Column 1 shows that, on average, students spend approximately 27 more seconds on the 100-Student Quiz than on the 10-Student Quiz, an increase of 3 percent over the mean. With endogenous selection of ordering, however, this number is confounded by the effect of quiz order on quiz duration. Instrumenting for the order of completion removes this endogeneity and provides a clearer picture of the treatment effect, showing that the 100-Student Quiz elicits 46.4 seconds more effort from students, an increase of more than 5 percent over the mean.¹⁹

Result 1: *Mean effort is increasing in the cohort size.*

¹⁹Inserting direct controls for the order of completion is not a valid measure, since the endogenous order of completion is collinear with the treatment effect.

Table 2.4: Mean Difference between 100- and 10-Student Quiz Duration

	OLS	IV Regression
<i>100-St. Quiz</i>		-6.085***
<i>Taken First</i>		(1.25)
Constant	0.450**	0.773***
	(0.19)	(0.20)
Instrumented	No	Yes
N	2,507	2,507

* p $\leq$ 0.10, ** p $\leq$ 0.05, *** p $\leq$ 0.01

All values are reported in minutes.

All standard errors clustered at the student level.

2.5.4 Heterogeneity in the Treatment Effect

My model predicts heterogeneity in the treatment effect across ability levels. I test this in two ways. I first address it semi-parametrically, testing the mean treatment effect within low- and high-ability students. Second, I impose continuity on the treatment effect across students with different abilities in order to understand how it evolves.

It is important to note that the complicating influence of other incentives for effort will be most apparent when considering the heterogeneity of the treatment effect. Deviations from the theory may manifest themselves through biased beliefs, correlations between intrinsic motivation and GPA, heterogeneous comprehension of the strategic incentives, or correlations between demographic characteristics and GPA. After testing the specific predictions of the model, I address each of these in turn.

Hypothesis 2: The treatment is negative for low-ability students.

Hypothesis 3: The treatment is positive for high-ability students.

To semi-parametrically test the heterogeneity of treatment effect, I regress the treatment effect on indicator variables for low- and high-ability students. The results are presented in Table 2.5. The first column shows significant and positive treatment effects for low-ability students and positive but insignificant treatment effects for high-ability students. Including the instrument for the order of quiz completion more accurately characterizes the treatment effect. In the second column, the treatment effect remains positive for low-and high-ability students, but is now significant for both. These results reject Hypothesis 2 but confirm Hypothesis 3.

Table 2.5: Difference between 100- and 10-Student Quiz Duration by Ability

	OLS	IV Regression
<i>100-St. Quiz</i>		-6.043***
<i>Taken First</i>		(1.26)
<i>Low-Ability</i>	0.706**	0.954***
<i>GPA < 2.72</i>	(0.31)	(0.32)
<i>High-Ability</i>	0.348	0.699***
<i>GPA ≥ 2.72</i>	(0.23)	(0.25)
Instrumented	No	Yes
N	2,507	2,507

* p<0.10, ** p<0.05, *** p<0.01

All values are reported in minutes.

All standard errors clustered at the student level.

This continuous evolution of the treatment effect will allow us to explore these heterogeneous effects in more detail. Figure 2.5 plots a locally linear polynomial smoothing function estimate of the treatment effect along with 95

percent confidence intervals, providing an empirical test of the predictions of Figure 2.3.²⁰

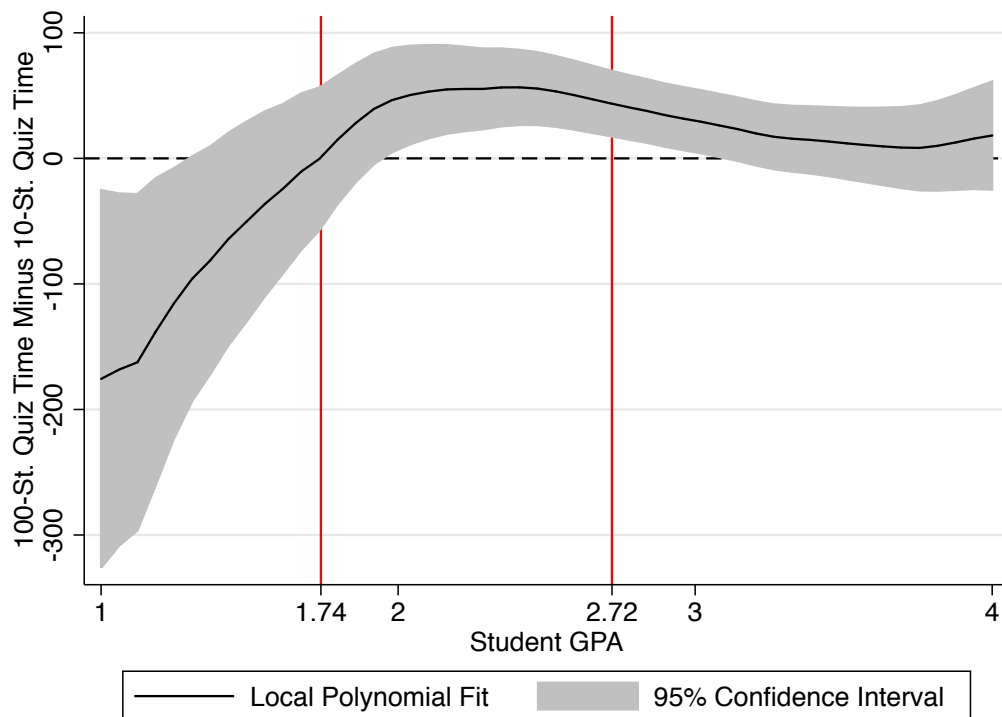


Figure 2.5: Local Polynomial Fit of the Difference in Effort between 100- and 10-Student Quizzes (Times Reported in Seconds)

Consider the evolution of the treatment effect beginning with the lowest-ability students. The model predicts that these particularly weak students understand the futility of their efforts and show no sensitivity to the treatment effects. In contrast to these predictions, the left tail of the distribution of GPAs shows a significant negative treatment effect with the minimum treatment effect occurring at the boundary, where the predicted change in effort is -175.7 seconds. This tail represents just over 2 percent of the population, but

²⁰I include a less parametric test of evolution of the treatment effect in the appendix, showing that the results are not driven by the continuity restriction.

is economically significant, as policy makers regularly design policies around the interests of the most at-risk students. For these students, changing from a 10- to 100-student cohort results in a 19.3 percent reduction in effort. This pattern is not consistent with the specific predictions of the model but is consistent with the general intuition that weaker students benefit more from the randomness of the 10-Student Quiz.

The treatment effect then crosses the axis and becomes significant and positive for the remaining low-ability students. For the high-ability students, the treatment effect is positive, but decreasing in GPA, returning to near-zero for the highest-ability students. Qualitatively, this pattern is similar to what my model predicted—that the treatment effect would cross the axis once and from below.

While these tests support qualitative predictions of my model, they also suggest that the model misses the mark with respect to specific locations of phenomena. Specifically, my model makes a focused prediction that the treatment effect crosses from negative to positive near the cutoff at the 30th percentile, or a GPA of approximately 2.72. Table 2.5 easily rejects this location for the single-crossing point. Statistical testing fails to reject the null hypothesis that the coefficient of the treatment effect is identical for low- and high-ability students ($F = 0.43$ $P = 0.512$). In fact, the raw mean difference between the 100- and 10-Student Quizzes below the cutoff is 44.4 seconds, a greater value than the mean difference above the cutoff, 20.0 seconds.

Figure 2.5 provides a more specific test of the single-crossing point of the treatment effect. The point at which the smoothing estimate crosses the horizontal axis is approximately a GPA of 1.74, or the second percentile

of student GPAs. For a more non-parametric estimate of the single-crossing point, I first note that all differences in effort above the single-crossing point should be positive, while all differences in effort below it should be negative. I then use two measures to assess the fit of different possible single-crossing points. First, I find the point that maximizes the absolute number of positive differences in effort above it and negative differences in effort below it. My second measure finds the point that maximizes the sum of the differences with the predicted sign minus the sum of the differences without the predicted sign. Using both of these measures, the single-crossing point that best fits the data occurs at a GPA of 1.81 or approximately the 3rd percentile.²¹

Result 2: *Contrary to the predictions of the model, the mean treatment effect is positive for low-ability students. The lowest-ability students, however, show significant negative treatment effects.*

Result 3: *As the model predicted, the treatment effect is positive for all high-ability students.*

2.5.5 Alternatives to the Neo-Classical Assumptions of the Model

Using my model as the backbone for understanding the patterns of effort allocation when students are graded on the curve provides several qualitatively accurate predictions, but understandably fails with regards to more focused predictions. Interactions between the treatment effect and alternative motivations for effort could cause shifts in the locations of specific phenom-

²¹In both cases, there is a flat maximum. The former case is maximized at GPAs of 1.81, 1.57, 1.54, 1.51, and 1.45. The latter is maximized at GPAs of 1.81 and 1.78.

ena but will nonetheless preserve the shape and single-crossing nature of the treatment effect that the data show.

To the extent that my experimental results provide evidence on the effects of these alternative motivations for effort, I address them below. Additional behavioral phenomena may be present but are not identified by my data. I explore the implications of several of these possibilities in the following section.

“Cursed” Beliefs

My model depends critically on the specification of students’ beliefs. Characterizing how the predictions of the model change when students are no longer required to make accurate inferences about their relative ability will allow me to test if deviations from the model can be explained by beliefs alone. Classical assumptions require that students make accurate inferences based on their past academic experiences and the selection of students expected to enroll in each class. Prior research suggests, however, that students may be “cursed” to believe that enrollment decisions are independent of ability.²² Without accounting for the selection of classmates, students will appear as if they are best responding to a distribution of classmates similar to that of previous lower-division courses.

Since my experiment takes place in what is typically the first upper-division course for economics students, the distribution of classmates differs largely from the distribution of classmates students have faced in lower-division

²²Eyster and Rabin (2005) provide an extensive review of empirical and experimental phenomena that can be attributed to cursedness.

courses. Data on the distribution of grades from all lower-division UCSD courses over the last 5 years can provide suggestive evidence that student behavior is consistent with cursedness. To construct the distribution of classmates that a cursed student would perceive based on experiences in lower-division courses, I aggregate the grade distributions from all lower-division courses in all departments. From this composite grade distribution, I find the likely GPAs associated with students at each percentile rank in lower-division courses at UCSD.²³ Figure 2.6 plots the cumulative distribution function of GPAs in this perceived grade distribution alongside the GPAs from Figure 2.4. Figure 2.6 shows how election effects draw the realized percentile ranks of each GPA in the experiment below their perceived percentile rank for low GPAs and above it for high GPAs. Cursed students fail to account for the fact the low GPA students select out of intermediate economics courses and believe their percentile rank is based on this relatively weaker distribution. Thus, a cursed student with a GPA of 2.24 perceives his percentile rank to be 30, while his actual percentile rank is approximately 10.

Figure 2.7 imposes fully-cursed beliefs about selection into the class and adjusts the equilibrium predictions to fit the perceived percentile rank.²⁴

Referring back to Figure 2.5, it is clear that the model can now more accurately

²³Specifically I suppose there are students at the 1st percentile, 2nd percentile, $\dots$, 99th percentile. The 99th percentile students receive the grades associated with the top 1 percent of grades awarded in each class, the 98th percentile students receive the grades associated with the top 2 percent, and so on. This generates a function mapping a student's percentile to their lower-division GPA. The measure is imperfect, but does capture the selection effects of upper-division economics courses.

²⁴These predictions assume that students are aware of neither their own cursed beliefs, nor the potential for cursed beliefs in others. Additionally, I assume that cursedness is limited to selection effects and does not affect equilibrium effort allocation. This allows for a simple mapping from the actual GPA to the equilibrium effort at the perceived GPA percentile. A richer model could predict cursed effort allocation at equilibrium as well.

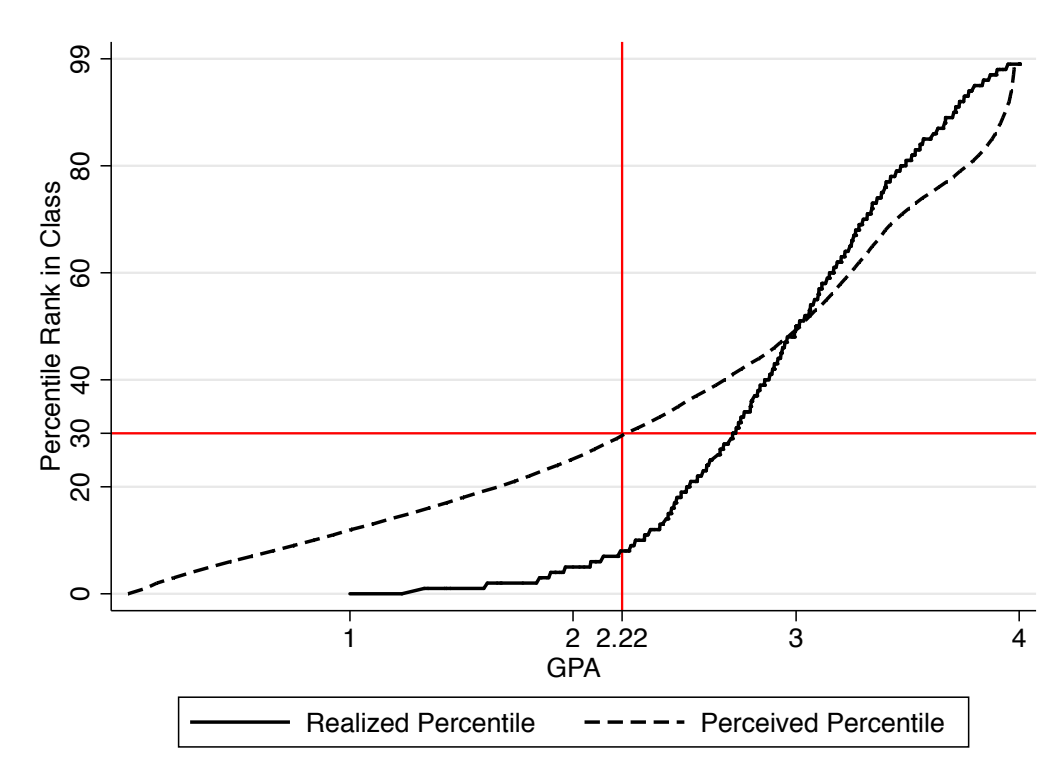


Figure 2.6: Perceived and Actual Percentile Ranks of Students

predict the location of the single-crossing point of the treatment effect.

Under cursed beliefs, my model predicts that the single-crossing point occurs at a GPA near 2.14, a shift that places it much closer to the estimated single-crossing point of 1.81. Less than 10 percent of students in my experiment possess GPAs below this adjusted single-crossing point, so the vast majority are now predicted to have positive treatment effects, a phenomenon confirmed by the data.

Cursedness adjusts the location of the single-crossing point closer to its location in the data, but even fully cursed beliefs were not sufficient for it to coincide exactly with the data. This implies that, while cursed beliefs may have a dramatic impact on the fit of the model, there are still residual

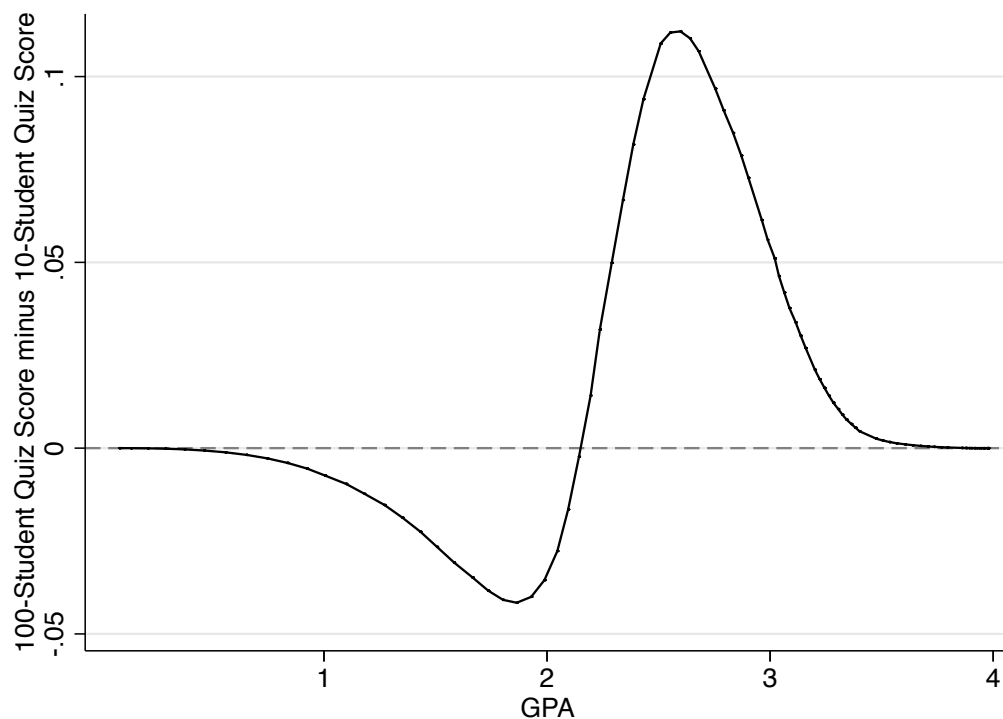


Figure 2.7: Predicted Difference in Effort Based on Perceived Rank

deviations that need to be accounted for. It is also important to note that this should not be taken as proof of cursedness among students in my experiment, rather, it is suggestive that with a better understanding of the belief structures of students, models of strategic interaction in the classroom can generate useful predictions for the allocation of effort by students. Further experimentation will be required to understand the complete process of belief formation and updating that students employ in a classroom setting. In particular, careful experimentation will be needed to distinguish the effects of cursedness from the effects of general overconfidence.

Comprehension of Strategic Incentives

The complexity of the equilibrium effort prediction raises the concern that lower ability students will be less able to intuit the benefits to randomness, while higher ability students will understand the benefits to reducing randomness. In fact, the weakest students appear to be the most sensitive to the treatments and respond as predicted to the increases in randomness by exerting more effort. Additionally, students below the cutoff are significantly more likely to state a preference for the 10-Student Quiz on post-experiment questionnaires ($z = 2.41$, $P < 0.02$).²⁵ This shows that comprehension of incentives cannot be globally increasing in GPA and cannot drive the result.

Intrinsic Motivation

Using GPA as a proxy for ability may raise the concern that the students labeled high-ability may be more intrinsically motivated to exert effort on quizzes than the students labeled low-ability. These level effects of effort will be removed by differencing the time on the 100- and 10-Student Quizzes. Thus, level effects only threaten my identification of the treatment effect if they drive changes in raw differences in effort while holding constant the ratios of effort. My data can address this hypothesis directly.

Figure 2.8 uses a locally linear polynomial smoothing function to plot the amount of time allocated to each quiz on the left-hand vertical axis and the aggregate amount of time allocated to quizzes on the right-hand vertical axis. The graph clearly shows that aggregate quiz duration is generally higher

²⁵The post-experiment questionnaires as well as the full specification of this regression can be found in the appendix.

for students with higher GPAs, but reaches a maximum at a fairly low GPA of 2.81, approximately the 37th percentile.

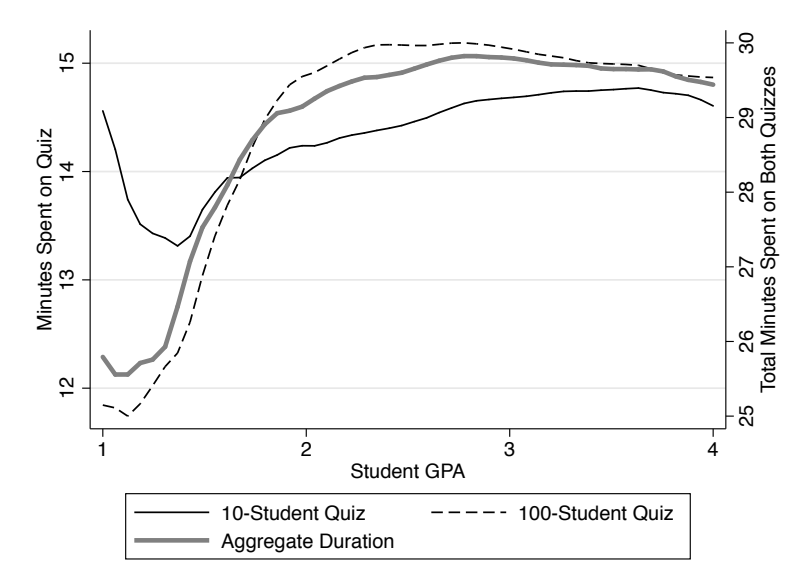


Figure 2.8: Seconds spent on each quiz and aggregate seconds spent on quizzes by GPA

Figure 2.8 clearly shows that aggregate quiz duration is generally higher for students with higher GPAs, but reaches a maximum at a fairly low GPA of 2.81, approximately the 37th percentile. The data reject the notion that level effects drive the results. While differences are large for students with the highest aggregate duration, they are also large for students with the lowest aggregate duration. Indeed, the two points where differences are smallest have substantially different levels of aggregate duration. This shows that, while level effects in effort that trend with GPA are a reality in my experiment, they cannot drive the observed pattern of differential effort.

Risk Aversion

Risk aversion could provide an alternative explanation to the observed behavior, since the two quizzes inherently bear different levels of uncertainty. In general, as risk aversion increases, low-ability students will decrease their effort, and high-ability students will increase their effort. To control for risk preferences, at the end of the course students answered a survey question about their likelihood of taking risks. Dohmen et al. (2011) show that this survey question is a strong predictor of revealed risk preferences.²⁶

In general, the patterns of effort predicted under different risk preferences will simply increase the standard errors of the estimates of the treatment effect unless risk aversion is correlated with GPA. The data refute this correlation ($t = -0.71$, $P = 0.478$), showing that a positive average treatment effect is common across different risk preferences.²⁷ Figure 2.9 splits students into risk averse and risk loving based on their responses to the survey question and plots the absolute duration of each quiz treatment to test if the impact of risk aversion is different across quizzes. The locally linear polynomial fit of the data show that the level effect of risk aversion is more pronounced than the differential effect across the two treatments. Low-ability, risk-loving students exert more effort on all quizzes than their risk-averse counterparts, but this level effect attenuates as ability increases. Importantly, the estimates of the treatment effect are similar for students with different risk preferences but equal abilities.

Figure 2.9 confirms the prediction that low-ability, risk-averse students

²⁶The question specifically asks, “How likely are you to take risks, in general.”

²⁷The full specification of this regression can be found in the appendix.

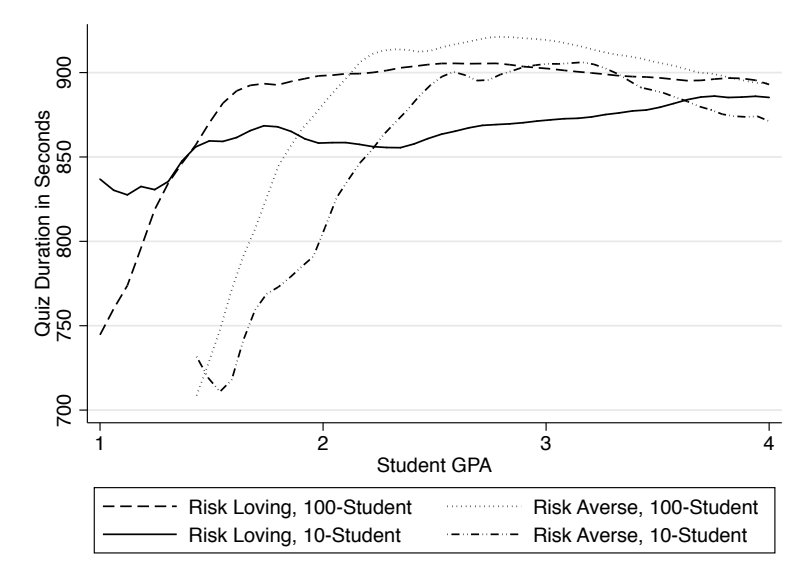


Figure 2.9: Duration of each quiz separated by risk preference (in seconds)

diminish their effort, but this impulse does not differentially affect the 10- and 100-Student Quizzes, even though one is inherently riskier than the other. Additionally, as the ability increases, there are no meaningful differences between students of equal ability but different risk preference. Thus, the general patterns of the treatment effect cannot be driven by risk preferences alone.

Gender and Competitiveness

The role of gender in determining risk and competitive preferences has been widely studied, and deserves attention in this paper, as my model supposes tournament-style competition between students.²⁸ The effect of gender does not present a likely confound in the identification of the heterogeneity of the treatment effect across abilities, as GPA does not significantly correlate

²⁸See Niederle and Vesterlund (2011) for a review of the literature on gender and competitive preferences.

with gender ($t = -1.44$, $P = 0.150$).²⁹ Nonetheless, understanding the ways that students of different genders react to the treatments will be valuable as policy makers design grading mechanisms. Figure 2.10 plots the treatment effects by gender and shows that men appear to be slightly more sensitive to the treatments, but the qualitative features of the treatment effect are similar across genders, making it an unlikely driver of the observed patterns of effort allocation between quizzes.

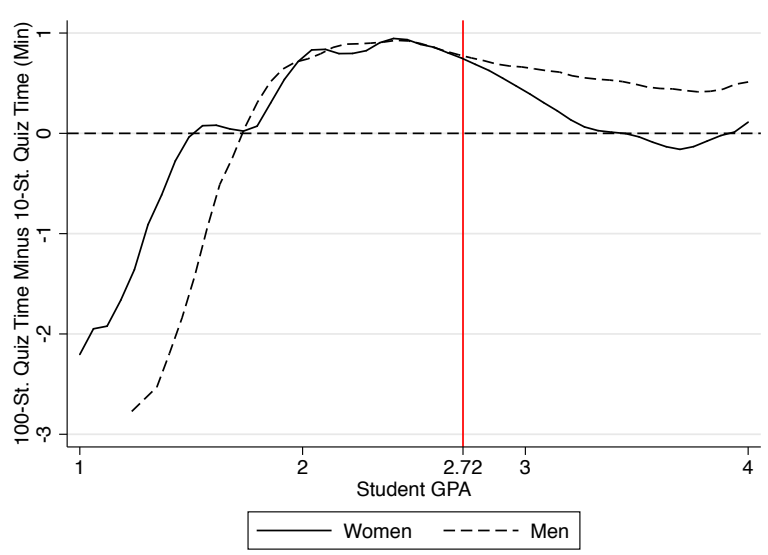


Figure 2.10: Local Polynomial Fit of Difference in Effort between 100- and 10-Student Quizzes by Gender (Times Reported in Seconds)

2.6 Discussion

A discussion of the policy implications of my experiment must first address the social benefit of inducing greater effort on classroom quizzes. One approach this is to consider the effect of the treatments on the scores of the

²⁹This regression can be found in the appendix.

students. This analysis shows no significant effect of the treatments ($t = -0.37$, $P = 0.710$).³⁰ There are three primary explanations for why time spent on the quizzes is a superior measure to scores on quizzes for my purposes. First, time is the costly resource that students exert in order to achieve higher scores. The decision of how much time to allocate to academic tasks, both in this experiment and in other classrooms, is based on an ex-ante perception of the relative returns to effort. Thus, measuring the impulse to allocate more time to a quiz based on its grading treatment is a better measure of how students will make studying decisions outside of the experiment than looking at the ex-post realization of their production of scores.

Second, quiz scores are a much coarser outcome measure, with all quizzes being graded based on four questions. The greater variation in effort than scores gives me more power in testing the effects of each treatment. Third, differences in the difficulty of the quizzes assigned to each treatment are amplified in the analysis of quiz scores, driving relatively much more variation in scores than in durations. A sufficiently large sample of quizzes could overcome this problem, but each subject saw only five pairs of quizzes, so the effect of changes in effort is lost amidst the variation in quiz difficulty.

The connection between greater effort in studying or attendance and greater academic performance is well-documented (Romer, 1993; Stinebrickner and Stinebrickner, 2008; De Fraja, Oliveira, and Zanchi, 2010; Arulampalam, Naylor, and Smith, 2012), implying that classroom motivation generates positive returns. Understanding the effect of subtle changes in the grading environment on the strategic allocation of effort by students therefore serves to

³⁰The full specification of this regression is found in the appendix.

further the social goal of increasing academic output.

Towards that end, my experiment uncovers systematic trends in the ways students of different abilities react to changes in the randomness of their grading environment that arise from changes in their cohort size. My experiment isolates the strategic uncertainty of small classes by manipulating a grading cohort size while holding constant all other characteristics of the classroom, such as teaching quality, student observability, or access to resources. With these clean controls, I can contribute to the discussion of optimal classroom size by demonstrating that even small changes in a student's strategic setting causally affect student effort exertion.

2.6.1 Patterns of Effort Allocation

Three prominent stylized facts manifest themselves in my data. First, my model is correct in predicting that effort is increasing in the cohort size. That is, on average, students increase their effort as the uncertainty of their grading environment decreases.

Second, low-ability students show positive average treatment effects, contradicting the model. This misallocation of effort may reflect biased beliefs about relative ability. The lowest ability students, however, show negative treatment effects, taking advantage of the increased randomness of the 10-Student Quiz. Stated preference also confirms that students below the cutoff are more likely to prefer the higher variance environment of the 10-Student Quiz than students above it, though both groups state a general preference for the 100-Student Quiz.

Third, the treatment effect is positive and significant for high-ability

students. This confirms the model’s prediction that these students exert more effort on the quiz with higher returns to their effort.

2.6.2 Misallocation of Effort

Despite the number of competing explanations for the results addressed in the previous section, none were able to independently generate the patterns of the treatment effect demonstrated in Figure 2.5. Indeed, the effects of alternative motivations for effort are often clear, but never consistent with the observed behavior and thus not a viable alternative to a strategic effort model. The data indicate that the strategic incentives present in the environment do cause shifts in effort that trend systematically with student ability in ways qualitatively similar to the predictions of the model. Nonetheless, the accuracy of the predictions of the model can still be improved by incorporating some of these effects, most importantly, biased beliefs. The data provide suggestive evidence that beliefs are partially responsible for effort allocations that are ex-ante suboptimal.

The positive treatment effect for students with abilities below the cutoff represents a robust misallocation of effort by students. While the mechanisms for this misallocation of effort are not clear from the data, the fit of the model is greatly improved by supposing that students may be “cursed” to believe that their classmates’ enrollment decisions for the upper-division economics course were unrelated to their abilities. To experimentally identify the effect of cursedness, one could randomly assign students to different distributions of classmates with a identical top quartiles. Then, after engaging in a task selecting the top quartile into a pool, see if the beliefs about the distribution

of students in the pool are biased by the first-stage distribution of students.

Adjusting for cursed beliefs, however, cannot fully explain the deviations from the model’s predictions. The following three behavioral biases are potential causes of this residual misallocation of effort:

First, students may possess a general overconfidence about their abilities. If students respond to their strategic grading environment according to an inflated estimate of their relative ability, then many students below the cutoff may favor the 100-Student Quiz, but students far enough below the cutoff will still favor the 10-Student Quiz. Second, students may fail to properly update their beliefs about their relative abilities. Upon receiving their results from the quizzes, students should reallocate effort based on their posterior beliefs about their relative ability, but updating failures allow students to ignore the informational content of quiz results and cling to their potentially biased prior beliefs. Both of these biases are consistent with existing results showing biased updating about self-perceived intelligence (Eil and Rao, 2011; Mobius et al., 2011). These two effects have been identified using standard belief elicitation before and after noisy information is given about relative ability. Biases at the first stage can be attributed to over-confidence, and differential biases at the second stage can be attributed to updating failures.

Third, students may possess reference dependent utility, specifically myopic loss aversion (Benartzi and Thaler, 1995).³¹ If students consider each quiz independently, they will overexert on quizzes with lower returns to their effort. Recall that increasing effort only increases the probability of high grades

³¹Here, “myopic” does not refer to time horizons, rather it refers to students evaluating outcomes of each quiz individually instead of evaluating the impact of each quiz on the total course grade.

and does not alter the grades themselves, since grades are fixed at 0, 1, or 3 points. As Sprenger (2010) points out, loss aversion over probabilities can only arise under expected value based reference points as in disappointment aversion (Bell, 1985; Loomes and Sugden, 1986; Gul, 1991), because under stochastic reference points (Kőszegi and Rabin, 2006, 2007), students are risk neutral over changes in probability within the support of the original gamble. A student who is risk neutral over changes in probability would exert more effort on the quiz that led to the greatest increases in her probability of high grades. Under loss aversion, a student would exert additional effort on the quiz that was considered to be “losing” relative to her reference probability, and in the case of students below the cutoff, this would be the 100-Student Quiz. To identify this effect, an experimenter could assign students a pair of quizzes with questions of roughly equal difficulty where one is graded such that the top 25 percent receive A’s and the other graded such that the top 75 percent receive A’s. Expected grade maximizing students between the 25th and 75th percentiles will exert greater effort on the former quiz, but reference dependent students with similar reference points across the two quizzes will exert greater effort on the latter.

2.6.3 Policy Prescriptions

My results show that, in general, mechanism designers with preferences over aggregate effort should implement a grading environment with the lowest possible variance in order to maximize the total effort exerted. This result could be accomplished through combining multiple classes into one grading unit and compensating for classroom-level differences.

While decreasing the variance increases aggregate effort, it incurs certain costs. For the lowest ability students, higher variance environments induce more effort. The intuition for this result relates back to Figure 2.1, which shows how increases in the size of a cohort make it increasingly unlikely that a low-ability student receives a high grade. From a policy-perspective, this result suggests that low-ability students can become discouraged by relative grading when the cohort size becomes large enough. If motivating low-ability students is part of an educator’s objective, then smaller cohorts can accomplish this. This could be achieved by splitting large classes up into smaller sections while grading each individually.

My results show that all students above the lowest abilities increase effort under the lower variance grading treatment. Under this allocation pattern, decreasing the variance of the grading environment bears lower costs—as measured by lost effort from low-ability students—than theory would predict. This result allows mechanism designers to re-optimize with respect to the positive and negative changes in effort that result from changes in the variance of the grading environment.

While my model captured many qualitative features of the data, it failed to capture the locations of the relevant phenomena. This failure may be attributable to students holding systematically inaccurate beliefs about their ability relative to their classmates. These beliefs could create long-term damage by causing students to misallocate effort to tasks or choose courses sub-optimally. In these cases, feedback about relative ability could increase student utility, but may reduce aggregate effort, since this misallocation generated increases in the aggregate effort on the 100-Student Quiz. The incentives of

students and classroom designers in this setting are clearly misaligned with respect to feedback, possibly causing the designer to withhold information in order to enable the biases of the students.

2.7 Conclusion

In this paper, I theoretically and empirically uncover heterogeneity in the way in which students of different abilities respond to changes in class size when the class is graded on the curve. Understanding that students identify strategic incentives in the classroom will greatly benefit educators and administrators as they seek to design classroom environments and grading mechanisms to achieve their objectives with respect to student effort.

In order to ground the intuition for why class size may affect student effort choices, I first develop a theoretical model of the situation. My experiment tests the qualitative predictions of this model and measures the causal impact of cohort size on effort.

My results highlight an important tension between mean effort and the distribution of effort implying that designers should use caution when attempting to maximize aggregate effort. This tension presents itself in the theory, and to a lesser degree, in the data. The mean effort exerted increased significantly with the cohort size, and effort among the lowest-ability students decreased with the cohort size. This confirms that using manipulations of the class size to encourage greater mean effort comes at the cost of effort by the lowest-ability students.

Several students who would benefit from a more random environment

misallocated effort to the less random environment. This misallocation may not be completely atheoretical, as it is consistent with well-documented behavioral biases, such as cursedness, overconfidence, non-Bayesian updating, and reference dependence. Further experimentation is needed to confirm or reject these theories, however.

My results make it clear that relative grading mechanisms currently in place generate unintended consequences as class size changes. This information can serve to identify the different demographics who are put at risk by different grading mechanisms. It additionally provides the basis for exploring grading mechanisms that find the desired balance between increasing mean effort and promoting a more desirable distribution of effort among students.

2.8 Acknowledgement

Chapter 2, in part is currently being prepared for submission for publication of the material. Brownback, Andy. The dissertation author was the primary investigator and author of this material.

2.9 Chapter 2 Appendix

2.9.1 Experimental Procedures

Syllabus Instructions for Quizzes

Economics 100A Quizzes

This quarter, we are studying how students respond to different grading formats by implementing two different grading methods on quizzes. Here are some reminders about the methods.

Overview:

- There will be 5 Quiz Weeks this quarter.
- Each Quiz Week, you will have to complete 2 quizzes for a total of 10 quizzes.
- All quizzes will appear on TED at Noon on Thursday of a Quiz Week and will be due no later than 5pm on Friday. That is, you will have 29 hours in which to complete the quiz.
- Each quiz will have its own 30-minute time limit.

Quiz Grading (**Grading schemes are listed in the title of the quiz**):

- **Points:**
 - All quizzes are out of 3 points for a total of 30 possible points this quarter.
 - 1 point will be awarded to any student who participates in a quiz*.
 - The remaining 2 points will be awarded in one of two different possible ways based on your student ID. We do this randomly so that all students can see both quizzes and types of grading without one being tied to the other.
 - **100-Student Quizzes:** We will select groups of 100 students randomly. The top 70 of 100 student scores will receive 2 additional points (giving them 3 of 3 points). The bottom 30 of 100 scores will receive 0 additional points (giving them 1 of 3 if they participated and 0 of 3 if they did not).
 - **10-Student Quizzes:** We will select groups of 10 students randomly. The top 7 of 10 student scores will receive 2 additional points (giving them 3 of 3 points). The bottom 3 of 10 scores will receive 0 additional points (giving them 1 of 3 if they participated and 0 of 3 if they did not).
- **Ties:**
 - All students who do not participate will get 0 points regardless of ties.
 - Any student who participates will be given 2 points if they are a part of a tie that crosses the 70% cutoff.
 - Example: Suppose we are in a **10-Student Quiz** and we have the scores: 4,4,4,4,3,3,3,3,1,1. The 70% cutoff will be 3, and all students with a score of 3 or more will receive full credit.
 - Example: Suppose we are in a **10-Student Quiz** and we have the scores: 4,4,4,3,3,1,NP,NP,NP,NP. Where “NP” means “No Participation*” The 70% cutoff will be at “NP”, but all students with a score of NP will receive 0, because they failed to participate.
 - Example: Suppose we are in a **10-Student Quiz** and we have the scores: 4,4,4,3,3,1,0,0,NP,NP. Where “NP” means “No Participation.” The 70% cutoff will be at 0, so students with a 0 who participated will receive full credit, but all students with a 0 who did not participate will receive no credit, because they failed to participate.

*Note: Participation will be judged based on accessing the quiz and attempting at least one question.

Online Instructions for Quizzes

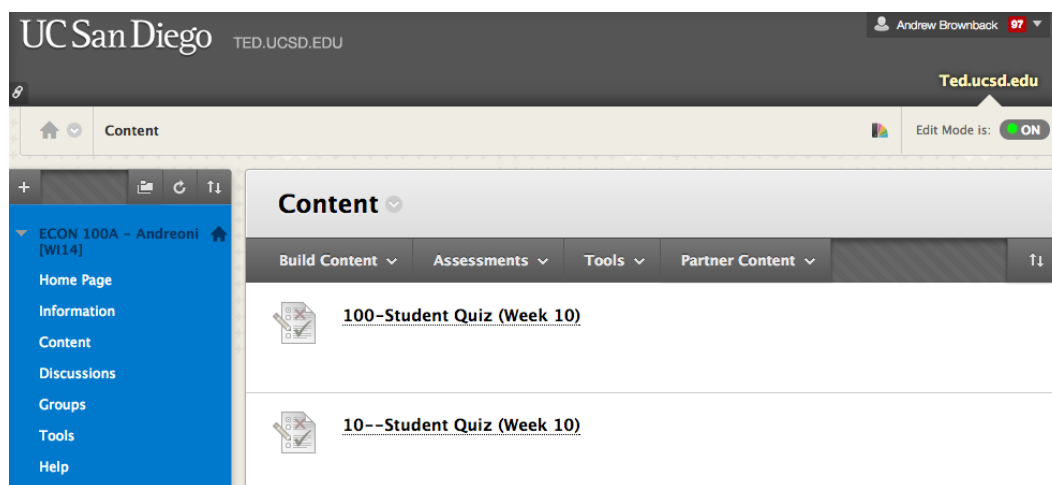
“On this quiz, there are 4 questions. Each question will be graded for every student who takes the test, giving all students a ”Score”. This Score is not your Grade, but it will help determine your Grade. Your Score will be compared to the Scores of 9 of your classmates. If your Score is among the top 7, you will receive a Grade of 3/3 for this quiz. If your score is among the bottom 3, you will receive a Grade of 1/3 simply for participating.

Your Grade on this quiz will appear in the gradebook after we have calculated it. Your Score will not appear in the gradebook.

You will have 30 minutes to complete this quiz. You are only allowed to take the quiz ONE TIME. If your application crashes, please email Andy at abrownba@ucsd.edu to work out a solution.

All answers will be in WHOLE NUMBERS.”

Online Environment



2.9.2 Theory

Score is Monotonic in Ability

Proof: Suppose not. Then a_i and a_j exist such that: $a_i < a_j$ but $s_{i,t} > s_{j,t}$.

Incentive compatibility dictates that for i and j , respectively,

$$U(s_{i,t}, a_i, N, P) \geq U(s_{j,t}, a_i, N, P)$$

$$U(s_{j,t}, a_j, N, P) \geq U(s_{i,t}, a_j, N, P).$$

Expanding these equations yields

$$P_{N,P}(s_{i,t}) - \frac{C(s_{i,t})}{a_i} \geq P_{N,P}(s_{j,t}) - \frac{C(s_{j,t})}{a_i} \quad (2.4)$$

$$P_{N,P}(s_{j,t}) - \frac{C(s_{j,t})}{a_j} \geq P_{N,P}(s_{i,t}) - \frac{C(s_{i,t})}{a_j}, \quad (2.5)$$

where $P_{N,P}(s_{i,t})$ represents the probability of receiving a high grade with score, $s_{i,t}$, parameters, N and P , and ability, a_i .

Solve for common terms and combine (2.4) and (2.5) to get

$$\begin{aligned} P_{N,P}(s_{i,t}) - \frac{C(s_{i,t})}{a_i} + \frac{C(s_{j,t})}{a_i} \\ \geq P_{N,P}(s_{j,t}) \\ \geq P_{N,P}(s_{i,t}) - \frac{C(s_{i,t})}{a_j} + \frac{C(s_{j,t})}{a_j}. \end{aligned}$$

Eliminate the middle term, and cancel the remaining terms to arrive at

$$\frac{C(s_{j,t}) - C(s_{i,t})}{a_i} \geq \frac{C(s_{j,t}) - C(s_{i,t})}{a_j}.$$

Dividing both sides by $C(s_{j,t}) - C(s_{i,t})$ reverses the inequality and yields

$$\frac{1}{a_i} \leq \frac{1}{a_j} \iff a_i \geq a_j,$$

which is a contradiction. QED.

2.9.3 Empirical Results

Means on either side of the cutoff

Table 2.6: 100-Student Quiz Duration Minus 10-Student Quiz Duration

	OLS	IV
$GPA \leq 2.72$	0.358 (0.39)	0.185 (0.39)
<i>100-St. Quiz Taken First</i>		-6.043*** (1.26)
Constant	0.348 (0.23)	0.397* (0.21)
N	2,507	2,507

* p_i0.10, ** p_i0.05, *** p_i0.01

Pairwise tests

Table 2.7: 100-Student Quiz Duration Minus 10-Student Quiz Duration

	OLS	IV	OLS	IV
<i>LowestBin</i>	-0.558 (0.47)	-0.680 (0.48)		
<i>LowBin</i>			1.180** (0.49)	1.130** (0.48)
<i>100-St. Quiz Taken First</i>		-6.160*** (1.26)		-6.029*** (1.25)
Constant	0.527** (0.21)	0.871*** (0.22)	0.275 (0.21)	0.602*** (0.22)
N	2,507	2,507	2,507	2,507

* p_i0.10, ** p_i0.05, *** p_i0.01

Stated Preference

After the course, students were asked to take a survey on their experience in the experiment. One of the questions asked:

“If we were to offer quizzes in your next econ class, but graded all quizzes in one way, which would you prefer?”

Table 2.8: Stated Preference for the 10-Student Quiz

Pr(Prefer to be graded based on 10-St. Quiz)	
$GPA \leq 2.72$	0.360** (0.15)
Constant	-1.196*** (0.09)
N	493

* p|0.10, ** p|0.05, *** p|0.01

Treatment Effect on Scores

All quizzes were scored 0-4. Column 2 makes it clear that there is no endogeneity concern with respect to scores, as the order of completion has no effect on the relative scores.

Table 2.9: Scores on 100-St Quiz minus Scores on 10-St Quiz

OLS		
<i>100 St Quiz</i>	-0.033	
<i>Taken First</i>	(0.05)	
Constant	-0.010	0.011
	(0.03)	(0.04)
N	2,491	2,419

* p|0.10, ** p|0.05, *** p|0.01

Chapter 3

On the Elicitation of Willingness to Pay for Stigmatized Goods

3.1 Introduction

Economists often treat incentivized measures of willingness-to-pay (hereafter, WTP) as accurate and stable estimates of a person's valuation for a good. However, stigmas against certain goods may impose psychological costs on people who express demand for those goods, depressing their stated WTP. Many of these psychological costs are inconstant, as research has established stigma as a "social and cultural construct" (Becker and Arnold, 1986), subject to variation across cultural and geographic boundaries. For example, the stigma one would face from expressing demand for modern medicine varies dramatically across different regions of the developed- and developing-world. These unstable stigma costs present confounds when using WTP elicitation to determine the social value of different medical treatments.

Consider, for example, a researcher eliciting people's WTP for contraceptives in an environment where birth control is socially or religiously taboo. Anonymity may help a person avoid the *social* stigma associated with making her preference transparent to the researcher. However, even with anonymity, the subject will still face *self* stigma associated with revealing certain preferences. Other goods bearing similar self-stigma may include government assistance programs, addiction treatment, unhealthy foods, drugs, contributions to unpopular causes, or anti-social behaviors like selfishness or prejudice.

If a subject's beliefs about her own character directly affect her utility, then stating a WTP for a good bearing self-stigma becomes a game of self-signaling similar to models of self-confidence and willpower by Benabou and Tirole (2002, 2004) and models of self-deception by Bodner and Prelec (2003)

and Mijović-Prelec and Prelec (2010): The subject may suppress her stated WTP in order to strategically manipulate her self-image. In evaluating the social value of these items or behaviors, policy-makers must consider the value to the individual in addition to the externalities imposed on others. However, because of these downwardly-biased valuations, policy makers may be prone to underestimate the social value of stigmatized goods. We refer to an anonymous but transparent statement of WTP as a “transparent WTP.”

This paper develops and experimentally tests a set of theoretical predictions for a WTP elicitation method that obscures the subject’s WTP for an individual good. Obscuring a subject’s WTP weakens her self-signal and allows her to reduce the psychological costs associated with expressing demand for that good. This “shrouded WTP” we elicit provides an alternative measure of the subject’s valuation that is less distorted by stigma costs than the valuation elicited using standard transparent mechanisms. Additionally, this shrouded WTP can help identify the magnitude of the impact of self-stigma on reported valuations for goods.

Our mechanism operates as follows: we present subjects with two goods, one stigmatized and one neutral. Eliciting WTP for each good individually provides two transparent WTPs that incorporate the psychological costs of expressing demand for each good. However, eliciting WTP for a *bundle* of the two goods provides a shrouded WTP that presents a weaker signal of the subject’s demand for the stigmatized good. Purchasing a sandwich and a milkshake separately provides a strong signal of a customer’s demand for each item, while purchasing them as a meal allows the customer to “blame” a portion of their demand for the (more stigmatized) milkshake on their demand

for the (more neutral) sandwich. We predict that bundling suppresses the psychological costs of expressing demand for the stigmatized good by allowing the subject to blame her WTP on demand for the neutral good. Therefore, we predict that WTP for the bundle of goods to be super-additive. That is, the WTP for the bundle of goods exceeds the sum of the two transparent WTPs.

We cannot, however, simply test for stigma-based self-deception by testing for super-additivity. Previous literature documents that the WTP for bundles of goods is systematically *less* than the sum of the WTPs for its components irrespective of the stigma associated with the goods in the bundle (Bateman, Munro, Rhodes, Starmer, and Sugden, 1997). That is, people generally display *sub*-additivity in their willingness to pay for bundles. To control for the possibility that preferences are generally not additive, we compare the additivity of WTP for a bundle containing a stigmatized good with the additivity of WTP for a “placebo” bundle that does not contain any stigmatized goods. This difference-in-differences design cleanly identifies the effect of bundling on stigmatized goods, a process that we claim reduces the psychological costs associated with expressing demand for the stigmatized good. By identifying the differential additivity of bundles whose components have varying levels of stigma, we claim to be able to provide a lower-bound estimate of the stigma costs associated with demanding a given good. Moreover, these stigma cost estimates provide a baseline for the welfare loss from naïvely assuming that transparent WTP reflects the subject’s valuation.

Since the stigma cost of a good may depend on how it is provided, shrouded WTP is particularly relevant for policy interventions that manipulate the provision mechanism for certain goods. For instance, stigma costs are

presumably lower for a good provided by default or without charge than one actively chosen or paid for. Relative to standard WTP measures, shrouded WTP more accurately represents the individual benefit of goods in these lower stigma contexts, and is thus a more appropriate measure to use when calculating the benefit of such policies.

Our bundling mechanism stands in contrast to the traditional economics view of bundling. Classical economic models show that bundles can increase a monopolist's profits depending on the covariance between the WTP for each item (Adams and Yellen, 1976; McAfee, McMillan, and Whinston, 1989). We are suggesting that bundles may affect consumer demand in an entirely different way, through the strength of the signal they send about the consumer's character. Bundles that weaken negative self-signals or strengthen positive ones augment consumer demand, increasing profits.

Clearly, self-stigma and self-signaling are subtle effects and will be challenging to observe in the laboratory. An obvious approach to this problem is to prime subjects with stigma towards certain goods. We avoid this approach, because we have no evidence to suggest that experimentally-induced stigmas operate under the same psychological processes as naturally occurring stigmas, nor it is clear that they are identifiable independent of experimenter demand effects. Moreover, if primed stigma were a compelling solution to our experimental design problem, then it would represent the most compelling solution to the problem of confounded WTP measures. All an experimenter would need to do to elicit WTP for a stigmatized good would be to prime subjects to think of stigmatized goods as neutral before measuring WTP. We believe these results would be dubious, and thus conclude that priming is a dubious

solution to our experimental design problem.

This paper presents the design of an experiment designed to test our theoretical results on the impact of bundling on consumer WTP for stigmatized goods. We hope to show there to be greater additivity in bundles that contain stigmatized goods than in those that do not, indicating that stigma confounds can be obviated when the self-signal associated with expressing demand for the stigmatized good is weakened through bundling.

In addition to a detailed design of our experiment, we discuss future directions for research in this field. Indeed, while our bundling mechanism is specifically designed to deal with circumstances where stigma presents a direct confound in the elicitation of WTP for certain goods, we consider it one part of a more general literature about the importance of self-signaling in economic decisions. Our results address the role of bundles in modifying the signaling content of a consumer's WTP, but this effect is not limited to WTP elicitation. Purchasing decisions, consumer sentiment, advertisement, and marketing may all be affected by changes in the signal content of goods that result from bundling. In this paper, we also restrict our focus to negative self-signals, but the self-signaling value of consumption is not similarly limited. Consumption of and stated WTP for “virtuous” goods should be affected just as strongly. Indeed, we hope that this paper is just the first of many to explore the self-signal value of preferences—stated or revealed—over goods of different valence.

The next section of this paper presents an overview of the literature. Section 3 introduces the self-signaling model of revealing WTP for stigmatized goods. Section 4 outlines the design of the experiment. Section 5 presents the

results of the experiment. In Section 6, we conduct power calculations for the sample size required to reliably find a significant effect of our mechanism. Section 7 concludes by discussing both how our mechanism may help inform policy decisions and directions for future research.

3.2 Literature

In order to understand the impact that self-stigma may have on economic decisions, it is important to look at the effects of social-stigma observed by psychologists and economists across many domains. For example, a “social desirability bias” has been shown to cause survey respondents to lie about characteristics that they believe are undesirable (Maccoby and Maccoby, 1954; Edwards, 1957; Crowne and Marlowe, 1960). Examples can be seen in self-reported personality traits (Fisher and Katz, 2000), self-reported dietary intake (Hebert, Clemow, Pbert, Ockene, and Ockene, 1995), and even demand for a CD player others may dislike (Fisher, 1993). Paulhus (1984) and Paulhus and Reid (1991) decompose this bias into self-deceptive positivity and impression management, showing that the impact of each can be independently identified depending on the types of questions asked of subjects.

Recently, economists have studied how the stigmas attached to undesirable traits influence economic decisions. For instance, Bharadwaj, Pai, and Suziedelyte (2015) find that stigmas towards reporting different mental health conditions can cause patients to conceal that condition on medical surveys. In a different setting, Bursztyn, Callen, Ferman, Hasanain, and Yuchtman (2014) find that Pakistani subjects were willing to forgo substantial payments in or-

der to avoid expressing gratitude towards an unpopular cause (in this case, the United States government). Coffman, Coffman, and Ericson (2013) uncovered significant stigma costs associated with either (1) identifying with the LGBT community and (2) expressing disapproval of the LGBT community. Importantly, they show that these social-stigma costs can be mitigated with sufficient protections of anonymity.

In order for social-stigma towards a given item or characteristic to be observable at the population level, it must exist at the individual level for a measurable proportion of the population. For this sub-population, anonymity may mitigate the role of social-stigma, but cannot affect the influence of self-stigma, itself a powerful motivator of decision-making. Akerlof and Kranton (2000) discuss the role of “identity” in economic decisions. In this framework, utility from a “virtuous” identity (such as, disciplined in saving for retirement or dieting) leads to negative (or stigmatized) attitudes towards certain goods or behaviors (such as, over-spending or indulging in unhealthy foods).

The concept of identity as discussed in Akerlof and Kranton (2000) relates closely to the psychology literature on positive self image. Psychologists have long held that people seek a positive self image, and a large literature explores how people foster this image through motivated reasoning and self-deception.¹ Recently, economists have modeled this process as a game of self-signaling. Bodner and Prelec (2003) argue that if one cannot perfectly introspect the motivation underlying her actions, she may purposefully distort her behavior in order to manage her impression of herself. Along these same

¹For instance, Bem’s (1972) work on self-perception studies the notion that people adopt the perspective of an outside observer when interpreting their own actions.

lines, Benabou and Tirole (2002, 2004) justify self-signaling as an attempt to influence the beliefs of a future self who cannot recall the original motivation for past behavior. To the extent that one is unaware of how she alters her behavior for the sake of self-image, this framework models self-deception. Quattrone and Tversky (1984) provides a seminal example. First, subjects completed a “cold-pressor test” (holding their arm in ice-water for as long as they could tolerate). Subsequently, they were told that tolerance is correlated with heart disease. Some subjects were told that higher tolerance predicts heart disease, while others were told the opposite. Finally, subjects repeated the cold-pressor test, and a vast majority showed changes in tolerance in the direction of “good news”—those told that low tolerance predicts heart disease increased their tolerance, and vice versa. Hence, people were willing to bear painful consequences for the sake of a positive diagnosis.

Self-deception does, however, have limits. Gneezy, Saccardo, Serra-Garcia, and van Veldhuizen (2015) task subjects with judging the merits of two alternatives and show that monetary incentives (“bribes”) can bias judgement when presented before the alternatives have been evaluated, but not when presented after the alternatives, implying that the psychological costs of self-deception are increasing in its salience.

Other evidence from economics and psychology demonstrates that people are willing to give up monetary payments to avoid privately confronting their own stigmatized behavior. For example, Dana, Weber, and Kuang (2007) find that subjects avoid information on how their actions affect others’ payoffs, providing them with “moral wiggle room” to take selfish actions. In a similar vein, Exley (2014) finds that subjects reduce charitable donations when they

can “blame” their selfishness on uncertainty over a charity’s outcomes. In this setting, risk over a donation’s outcome affords the subjects ambiguity over the origin of the desire to donate less. Preference towards risk shrouds an individuals selfish intentions, decreasing their charitable contributions. Andreoni, Rao, and Trachtman (2011) find a similar avoidance of charitable giving, showing that grocery store patrons go out of their way to avoid being asked to give to a charity. This avoidance can be interpreted as people preferring the weak signal associated with avoiding the solicitation over the strong signal associated with turning down the solicitor directly.

Settings where people express demand for this type of wiggle-room grant experimenters a chance to measure the WTP for this moral flexibility and potentially provide a measure of stigma costs. There are, however, two reasons that we propose a richer measurement methodology for self-stigma. First, few settings offer agents the ability to avoid information about the negative consequences of their actions. Second, the act of avoiding information itself contains unavoidable self-signaling information, potentially confounding the results.

Eliciting WTP for goods that the subject may deem virtuous or vicious has consistently proven to be a challenge for economists. Contingent valuation (hereafter, CV) employs stated-preference survey responses to estimate WTP and has long been the workhorse methodology for evaluating WTP for non-marketable goods. CV is, however, regularly criticized for its biased reporting of economic value of certain goods (see Carson (2012) and Carson, Flores, and Meade (2001) for a more extensive discussion of CV, its biases, and its role in the the Exxon Valdez oil spill). Bernheim, Bjorkegren, Naecker, and

Rangel (2013) demonstrate that non-choice data is often biased but can be useful in predicting choice. Their paper points out that the biases inherent in non-choice methods (of which, CV is one) can result from the unincentivized nature of the questions and the potential for other psychological factors to provide confounding incentives for certain responses. For example, in stating a CV for a clean ocean, a subject will not face economic incentives that promote truthfulness, but may hold psychological incentives that bias him to representing himself as one who values a pristine, natural coastline, causing him to inflate his CV for a clean ocean. Moving from CV to an incentivized WTP mechanism will provide economic incentives to discipline his choice, but will not eliminate his psychological incentives. Thus, incentives will weaken but not eliminate a biased report of WTP for goods that the subject may consider virtuous or vicious.

In exploring the biases of WTP mechanisms when used on goods that bear self-stigma, we borrow from two different, well-developed WTP elicitation techniques, the Multiple Price List (hereafter, MPL) framework developed by Binswanger (1980) and the Becker-DeGroot-Marschak mechanism (hereafter, BDM) developed by Becker, DeGroot, and Marschak (1964). Our primary aim is to understand the impact of stigma on the super- or sub-additivity of a bundle of goods. Bateman et. al. (1997) explore the question of additive preferences for bundles of non-stigmatized consumables, finding that consumers possess generally sub-additive preferences. Gaeth, Levin, Chakraborty and Levin (1991) find the opposite result, super-additive WTP for bundles. With the specific focus on stigmatized goods, Blair and Roese (2013) find situations in which the psychological costs associated with stigmatized goods may be

increased or decreased with bundling, depending on the composition of the bundle. We differentiate ourselves from this paper by 1) looking at purchasing decisions, rather than survey-elicited feelings and 2) avoiding bundling goods that could affect the framing of the purchasing decision, only relying on bundling goods that allow subjects flexibility in their valuation of the bundle.

Our paper makes three novel contributions to this literature. First, our mechanism identifies the impact of stigma on WTP measures and evaluates the role played by self-signaling. Second, our mechanism is easily portable across contexts, not limiting us to identifying the impact of stigma in one particular setting. Third, our method uses a difference-in-differences design that, unlike binary-choice, can precisely identify the effect of stigma on WTP independently from general effects of bundling and regardless of the size and sign of the effect.

3.3 Model

This section presents a basic self-signaling model to demonstrate how bundling a stigmatized good with a neutral good may increase WTP for the stigmatized good. There are many such models that could deliver this result; the one we describe here, which is an adaptation of Bodner and Prelec (1995, 2003) to a WTP-elicitation setting, is chosen primarily for its simplicity and accordance with evidence from psychology.

Consider a person who bids in a BDM mechanism for two goods S, B . We will consider two cases: (1) the person bids separately for each good, and (2) she bids for a single bundle containing both goods. Denote a person's

value for good $j \in \{S, B\}$ by v_j . We assume v_j is the sum of two unobserved components: $v_j = \theta_j + \tau_j$. The person's constant "disposition" for good j is represented by θ_j , while τ_j reflects a momentary feeling of temptation. We assume that $\tau_j \sim N(0, \sigma_j^2)$, so θ_j is the person's average utility from consuming j . For tractability, we assume the person's prior belief is $\theta_j \sim N(\bar{\theta}_j, \gamma_j^2)$. In addition to earning utility from consuming the good, we assume the person earns utility directly from her beliefs about her disposition. For instance, if the good in question is a cigarette, the person earns consumption utility from smoking it and "diagnostic" utility from reflecting on her beliefs about her average disposition towards cigarettes, which reveal if she is a social or heavy smoker. If the person observes only v_i , she may try to "blame" her smoking on high momentary temptation, τ_j , in order to maintain the belief that she has a low disposition toward smoking, θ_j .

Imagine the person is bidding for a single good j through the BDM mechanism. The mechanism draws a random price $p \in [0, 1]$, and the person simultaneously announces a bid (or WTP) $W(j) \in [0, 1]$.² The agent receives good j at price p if and only if $p \leq W(j)$. For sake of describing the self-signaling game, we assume the person has two selves. The first self chooses bid $W(j)$, and the second forms beliefs about θ_j given that bid. We also assume imperfect recall: the first self knows v_j when choosing $W(j)$, but the second self recalls only the bid $W(j)$. Conditional on observing $W(j)$, the second self forms a belief distribution $\mu(\cdot | W(j))$ over θ_j ; as we will see, μ is determined endogenously in the signaling equilibrium.

If the person receives the good, she feels two sources of utility. The first

²Normalizing the space of prices and bids to the unit interval is without loss of generality.

is standard consumption utility, $v_j - p$. The second source, which Bodner and Prelec (1995, 2003) call “diagnostic utility”, derives directly from her beliefs about θ_j . Diagnostic utility is defined as

$$D(\mu, W) = k_j \int \theta \mu(\theta|W) d\theta,$$

where the parameter k_j summarizes the weight attached to diagnostic utility relative to consumption utility. $k_j < 0$ models a stigmatized good—the person wishes to believe she has a low disposition for the good, whereas $k_j > 0$ corresponds to a virtuous good. $k_j = 0$ implies good j is neutral: the person’s beliefs about her disposition are irrelevant to her utility. In this exercise, we assume $k_S < 0$ and $k_B = 0$. The first self chooses bid W to maximize the expected value of her consumption and diagnostic utilities, and hence chooses $W(j)$ to solve

$$\max_W U(W, v_j, \mu)$$

where

$$U(W, v_j, \mu) = \Pr(p \leq W) \mathbb{E}[v_j - p | p \leq W] + k_j \int \theta \mu(\theta|W) d\theta. \quad (3.1)$$

To close the model, we must specify how the second self updates her beliefs about θ given bid W . Bodner and Prelec (2003) discuss two ways to model the inference function $\mu(\theta|W)$. The first approach, which we follow, is to assume the second self naively infers θ from W at face value. That is, the second self neglects the signaling motivation of the first self, and assumes the first self chooses $W(j)$ to maximize only the expected consumption utility

$\widehat{U}(W, v_j, \mu) = \Pr(p \leq W) \mathbb{E}[v_j - p | p \leq W]$ rather than the full utility function specified in Equation 3.1. Alternatively, we could assume that the second self is fully aware of the first self’s motive to distort beliefs about θ , causing the second self to discount announcements by the first self. In this case, $\mu(\theta|W)$ would constitute the receiver’s belief in a Perfect Bayesian Equilibrium of the signaling game between selves.

Note that, while the second approach is standard in economics,³ the simple “face value” model still captures the behavior we seek to illustrate. Although it is much simpler than the rational signaling model, the results of Bodner and Prelec (1995) confirm that our basic intuitions under face-value inference—that the person will bid below her true value for a stigmatized good—still hold under rational inference. Additionally, several psychology experiments report evidence in support of face-value inference. In Quattrone and Tversky (1984), for instance, only a small minority of subjects confessed to conscious efforts to influence the cold-pressor test, and those who did were already pessimistic about their chances of having a good heart. In a study specifically designed to test for rational versus face-value self-signaling, Mijovic-Prelec and Prelec (2010) find that a majority of subjects engage in self-deception. We therefore decide against a rational signaling model, as its empirical justification is tenuous and its additional complication is considerable.

We now derive the optimal bids, beginning with the case where the person bids for each good separately. Under face-value inference, the second

³For instance, Bernheim (1994) studies rational signaling when people have a taste for conformity, and assumes a sender’s objective function that is nearly identical to Equation 3.1. The primary difference between his model and ours is interpretation: the diagnostic utility in his model derives from how *other* people judge the agent’s behavior, whereas in our model the agent judges her own actions.

self assumes $W(j) = v_j$ conditional on $W(j) \in (0, 1)$. This follows trivially from the fact that under the BDM mechanism—absent signaling concerns—it is optimal to bid one's true value. Thus, because we have assumed normally distributed valuations, the second self believes the mean of θ_j conditional on $W(j)$ is

$$\mathbb{E}[\theta_j | v_j = W(j)] = \alpha(j)v_j + [1 - \alpha(j)]\bar{\theta}_j,$$

where

$$\alpha(j) = \gamma_j^2 / (\gamma_j^2 + \sigma_j^2). \quad (3.2)$$

Since the second self naively treats $W(j)$ and v_j as identical, she believes her expected disposition is $\alpha(j)W(j) + [1 - \alpha(j)]\bar{\theta}_j$ following bid $W(j) \in (0, 1)$. If the second self observes a bid $W(j) = 0$, she assumes that it could have derived from any $v_j \leq 0$, and hence infers

$$\mathbb{E}[\theta_j | v_j \leq 0] = \bar{\theta}_j - \frac{\gamma_j^2}{\sigma_j} \left(\frac{\phi\left(\frac{-\bar{\theta}_j}{\gamma_j}\right)}{\Phi\left(\frac{-\bar{\theta}_j}{\gamma_j}\right)} \right), \quad (3.3)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution functions, respectively. Finally, if $W(j) = 1$, then the second self infers

$$\mathbb{E}[\theta_j | v_j \geq 1] = \bar{\theta}_j + \frac{\gamma_j^2}{\sigma_j} \left(\frac{\phi\left(\frac{1-\bar{\theta}_j}{\gamma_j}\right)}{1 - \Phi\left(\frac{1-\bar{\theta}_j}{\gamma_j}\right)} \right). \quad (3.4)$$

Letting $\bar{\theta}_j^L$ and $\bar{\theta}_j^H$ denote the right-hand sides of equations 3.3 and 3.4, respectively, the second self's naive expectation of θ_j following bid $W(j)$ is sum-

marized as follows:

$$\widehat{\mathbb{E}}[\theta|W(j)] = \begin{cases} \bar{\theta}_j^L & \text{if } W(j) = 0 \\ \alpha(j)v_j + [1 - \alpha(j)]\bar{\theta}_j & \text{if } W(j) \in (0, 1) \\ \bar{\theta}_j^H & \text{if } W(j) = 1. \end{cases} \quad (3.5)$$

From Equation 3.1, the first self chooses $W(j) \in [0, 1]$ to maximize

$$U(W, v_j) = W \left(v_j - \frac{1}{2}W \right) + k_j \widehat{\mathbb{E}}[\theta|W],$$

where $\widehat{\mathbb{E}}[\theta|W]$ is as specified in Equation 3.5.

For a moment, let us ignore the bounded bid space in order to understand the impact of stigma on bids. If the person were to be able to choose any bid $W \in \mathbb{R}$, she would maximize the unconstrained utility function $\widetilde{U}(W, v_j) = W \left(v_j - \frac{1}{2}W \right) + k_j (\alpha(j)W + [1 - \alpha(j)]\bar{\theta}_j)$. From the first-order condition, the optimal unconstrained bid is $\widetilde{W}(j) = v_j + k_j \alpha(j)$. Hence, when $k < 0$, the person shades her bid downward by an amount $|k_j \alpha(j)|$ in order to signal a lower-than-true disposition. Notice that the extent to which the person underbids is directly proportional to both k_j —the weight she places on her belief about θ_j —and $\alpha(j)$ —the informativeness of her bid regarding θ_j .

To derive the optimal bidding strategy when bids are confined to $[0, 1]$, we must simply check whether interior unconstrained-optimal bids are preferred over either $W = 0$ or $W = 1$. Note that there is a discontinuity in perceptions when moving from an interior bid to either 0 or 1. For instance, following bid $\epsilon > 0$, the second self believes the first self had value ϵ . But

following $W = 0$, she believes her valuation was the average v conditional on $v < 0$, which is bounded below 0. Hence, a person with an unconstrained optimal bid $W = \epsilon$ may strictly prefer deviating to bid $W = 0$ because of this discontinuous reduction in beliefs about her disposition that results. Likewise, a person will never bid $W = 1$ when $k_j < 0$: reducing her bid infinitesimally will reduce her perceived disposition by a finite amount. As such, we will slightly abuse language in what follows, and denote a bid infinitesimally below 1 by $W = 1$.

Whether some valuations (or “types”) will choose interior bids will depend on the weight, k , placed on diagnostic utility. Formally, for each good j there exists a value $\hat{k}_j < 0$ such that $k_j < \hat{k}_j$ implies that all types pool at either $W(j) = 0$ or $W(j) = 1$. For small stigma costs, $\hat{k}_j < k_j < 0$, there will be a range of types who optimally select interior bids. The following lemma formally presents the optimal bidding strategy.

Lemma 1. *Consider a person bidding for the single good j with valuation v_j . Suppose $k_j < 0$, and let $\hat{k}_j \equiv \frac{1}{2(\theta^L - [1 - \alpha(j)]\theta)} < 0$. If $k_j < \hat{k}_j$, then there exists a value $v_j^* > 1 - k_j\alpha(j)$ such that the optimal bidding strategy is*

$$W^*(j|v_j) = \begin{cases} 0 & \text{if } v_j \leq v_j^* \\ 1 & \text{if } v_j > v_j^*. \end{cases} \quad (3.6)$$

If $k_j \in (\hat{k}_j, 0)$, then there exist a value $\hat{v}_j \in [-k_j\alpha(j), 1 - k_j\alpha(j)]$ such that the

optimal bidding strategy is

$$W^*(j|v_j) = \begin{cases} 0 & \text{if } v_j < \hat{v} \\ v_j + k_j\alpha(j) & \text{if } v_j \in [\hat{v}, 1 - k_j\alpha(j)] \\ 1 & \text{if } v_j > 1 - k_j\alpha(j). \end{cases} \quad (3.7)$$

Proof. See Appendix. □

The bidding strategy in Lemma 1 implies that people generally under bid for stigmatized goods. Summarizing:

Proposition 1. *Suppose the person is bidding for a single good j for which she has valuation $v_j > 0$. If $k_j < 0$, then $W(j) < v_j$. If $k_j = 0$, then $W(j) = v_j$. Hence, the person bids below her value for the stigmatized good S , and bids her value for the neutral good B .*

Now suppose that the person is bidding for the bundle containing both goods S and B . The setup is the same as the single-good case except the person's total valuation for the bundle is equal to $v_T = v_S + v_B$, where v_S and v_B have the same distributional properties as above. Following Equation 3.1, the first self chooses $W(S, B)$ to maximize

$$\Pr(p \leq W) \mathbb{E}[v_T - p | p \leq W] + k_S \hat{\mathbb{E}}[\theta_S | W(S, B)] + k_B \hat{\mathbb{E}}[\theta_B | W(S, B)].$$

Since we assume $k_B = 0$, this case differs from the single-good case only in the inference about θ_S given $W(S, B)$, $\hat{\mathbb{E}}[\theta_S | W(S, B)]$. Importantly, $W(S, B)$ provides a weaker signal of θ_S than did $W(S)$ in the single-good case. The second self is able to “blame” a high bid for the bundle on a strong disposition

for the neutral good B , whereas a high bid for the single good S in isolation sends a strong signal that she has a strong disposition for the stigmatized good S . Formally, as above, the person treats bid $W(S, B) \in (0, 1)$ as equal to v_T . Note that $v_T = \theta_S + \theta_B + \tau_S + \tau_B = \theta_S + \eta$ where $\eta \sim N(\bar{\theta}_B, \sigma_S^2 + \sigma_B^2 + \gamma_B^2)$. Thus, the second self forms expectation $\hat{\mathbb{E}}[\theta_S|W(S, B)]$ equal to

$$\mathbb{E}[\theta_S|v_T = W(S, B)] = \alpha(S, B)(W(S, B) - \bar{\theta}_B) + [1 - \alpha(S, B)]\bar{\theta}_S,$$

where

$$\alpha(S, B) = \frac{\text{var}(\theta_S)}{\text{var}(\theta_S) + \text{var}(\eta)} = \frac{\gamma_S^2}{\gamma_S^2 + \gamma_B^2 + \sigma_S^2 + \sigma_B^2}. \quad (3.8)$$

Notice that Equation 3.8 is a similar expression to Equation 3.2, but the weaker signal forces the second-self to place greater weight on prior expectations of θ . Given that this weaker signal about θ_S is the only difference between the bundled and single-good cases, the optimal bidding strategy takes an identical form with different parameter values:

Lemma 2. *Consider a person bidding for both goods S and B with respective valuations v_S and v_B . Let $\tilde{k}_S \equiv < 0$. If $k_S < \tilde{k}_S$, then there exists a value $v_T^* > 1 - k_S\alpha(S, B)$ such that the optimal bidding strategy as a function is*

$$W^*(S, B|v_T) = \begin{cases} 0 & \text{if } v_T \leq v_T^* \\ 1 & \text{if } v_T > v_T^*. \end{cases} \quad (3.9)$$

If $k_S \in (\tilde{k}_S, 0)$, then there exist a value $\hat{v}_T \in [-k_S\alpha(S, B), 1 - k_S\alpha(S, B)]$ such

that the optimal bidding strategy is

$$W^*(S, B|v_T) = \begin{cases} 0 & \text{if } v_T < \hat{v}_T \\ v_T + k_S\alpha(S, B) & \text{if } v_T \in [\hat{v}_T, 1 - k_S\alpha(S, B)] \\ 1 & \text{if } v_T > 1 - k_S\alpha(S, B). \end{cases} \quad (3.10)$$

Proof. See Appendix. □

We now analyze how the sum of the bids under separate elicitation compare to the bid for the bundle. Our main result is that bids are systematically higher for the two goods when elicited in a bundle. To see this, note the following lemma which follows immediately from Equations 3.2 and 3.8.

Lemma 3. $\hat{\mathbb{E}}[\theta_S|W]$ places less weight on W in the bundled case than in the single-good case. That is, $\alpha(S, B) < \alpha(S)$.

The lemma implies that, so long as bids are not at the boundary, the total bid for the two goods is reduced by less in the bundle case than in the separate-elicitation case. In the bundle case, $W(S, B) = v_T + k_S\alpha(S, B)$ falls short of the person's total valuation by $|k_S\alpha(S, B)|$. In the separate-elicitation case, the total bid for the two goods is $W(S) + W(B) = v_S + k_S\alpha(S) + v_B$, which falls short of her total valuation by $|k_S\alpha(S)| > |k_S\alpha(S, B)|$. The next proposition shows that the average bid across all types is strictly greater under bundled elicitation relative to separate.

Proposition 2. Suppose $k_S < 0$ and $k_B = 0$. Then for all types $(v_S, v_B) \in \mathbb{R}^2$, the optimal bids in bundled and separate elicitation satisfy $W^*(S, B|v_S + v_B) \geq W(S|v_S) + W(B|v_B)$, and the inequality is strict for a positive measure of types.

Proof. See Appendix. □

Proposition 2 presents the main testable prediction of our self-signaling model: with a sufficiently large sample of valuations, if one of the goods in the bundle is stigmatized, then we predict a significantly greater average bid under bundled elicitations relative to separate elicitations. If both goods in the bundle are neutral, then we expect the average bid to be identical across elicitation methods. The next section outlines an experiment to test this prediction.

3.4 Experimental Design

Our theoretical predictions are based on the differential strength of self-signals from expressing demand for bundles and individual goods. Since agents in our model gain utility from their self-signals, our results would not hold if multiple signals were given to the agent at the same time. Therefore, we must rely on a between-subjects design—where each subject will only gain one self-signal from expressing demand for any one good—to test the effectiveness of a bundling mechanism intended to obscure the self-signal value of expressing demand for stigmatized goods. Under this mechanism, a bundling good, B_i , shrouds the subject from the negative self-signal associated with stating a WTP for a stigmatized good, S_i . This bundling good is chosen to have ambiguous value to the subject so that any given WTP for the bundle can be more easily “blamed” on the subject’s WTP for B_i .⁴

⁴To better understand why this is important, consider the likelihood of encountering super-additivity when goods are bundled with \$5 bills.

In order to test the effectiveness of this mechanism, we will use an MPL to elicit the following WTP measures:

- WTP for B_i , denoted $W(B_i)$,
- WTP for S_i conditional on receipt of B_i , denoted $W(S_i|B_i)$,
- WTP for the bundle of S_i and B_i , denoted $W(S_i, B_i)$.

We refer to B_i as an “individual good,” $\{S_i|B_i\}$ as a “conditional good,” and $\{S_i, B_i\}$ as “bundled goods.” Classical assumptions imply that the sum of the WTPs for the individual and conditional goods should exactly equal the WTP for the bundled goods:

$$W(B_i) + W(S_i|B_i) = W(S_i, B_i) . \quad (3.11)$$

Incorporating psychological costs associated with expressing demand for S_i may break this equality. Specifically, within a self-signaling model, $W(S_i|B_i)$ provides a more precise signal of the person’s value of S_i than does $W(S_i, B_i)$. In this case, we would expect the following inequality to hold:

$$W(B_i) + W(S_i|B_i) < W(S_i, B_i) . \quad (3.12)$$

That is, we expect to see super-additivity in bundles that incorporate stigmatized goods. Super-additivity, however, may or may not occur as a result of other factors such as wealth effects or relative thinking about prices (Kahneman and Tversky, 1984). Previous studies have found mixed results with respect to super- or sub-additivity of bundles (Gaeth et al., 1991; Bateman et

al., 1997). To eliminate any confounds from general super- or sub-additivity, we will compare the additivity of bundles containing stigmatized goods with the additivity of bundles containing only neutral goods. Thus, the effectiveness of our bundling mechanism requires that the super-additivity be higher for a stigmatized good than for a neutral good, N_i . That is:

$$W(S_i, B_i) - [W(B_i) + W(S_i|B_i)] > W(N_i, B_i) - [W(B_i) + W(N_i|B_i)] \quad . \quad (3.13)$$

3.4.1 Item Selection

Central to this mechanism is the definition of stigmatized goods. In order to justify our choices for stigmatized and non-stigmatized goods, we use out-of-sample surveys from the same population on which we will test our mechanism.

The first round of surveys took place in April of 2015 at the University of California, San Diego (hereafter, UCSD). We asked 36 subjects to report their emotions associated with several different goods. These surveys provided a manipulation check ensuring that the classes of goods we considered to be stigmatized were, in fact, stigmatized in the population from which we drew our subject pool.

Table 3.1 presents the 7 most relevant goods along with summary statistics of responses to the two primary questions of the survey. The two questions we focus on are, “Would you be willing to pay any money for this item?” and “Imagine that you purchased this item, how would you feel about yourself

after having done this?”⁵

Table 3.1: Summary Statistics for Stigmatized and Neutral Goods

SUMMARY STATISTICS FOR ALL GOODS			
	DESCRIPTION	PERCENTAGE WILLING TO PAY POSITIVE AMOUNT FOR ITEM	RESPONDENTS' REPORT OF SELF- STIGMA (1-7 SCALE)
ARC	<i>“Reverse donation” to the American Red Cross</i>	0.22	2.61 (1.29)
NRA	<i>Donation to the NRA</i>	0.08	2.72 (1.73)
MAXIM	<i>Maxim Magazine</i>	0.19	3.36 (1.36)
PEOPLE	<i>People Magazine</i>	0.31	3.31 (1.51)
MUG	<i>UCSD Coffee Mug</i>	0.61	4.47 (1.30)
PEN	<i>UCSD EconLab Pen</i>	0.44	4.08 (1.00)
COIN	<i>Gamble where subjects win based on flips of coin</i>	0.42	3.53 (1.44)

Note: $N = 36$. Standard deviations in parentheses.

Note: Respondents answered the question: “How would you feel about yourself if you had purchased the item.”

From Table 3.1 it is clear that the stigmatized goods generate a more negative self-image than the neutral goods. Moreover, subjects report a lower willingness to pay for these goods. A more rigorous analysis shows that not only does the negative correlation between WTP and self-stigma exist across goods, but within goods as well. That is, conditional on the good, subjects with more self-stigma associated with the good also have lower WTP for the good. This analysis can be found in the appendix.

From the reports on this battery of candidate goods, we immediately identified 2 neutral goods $\{N_1 = \text{MUG}, N_2 = \text{PEN}\}$, a neutral bundling good, $\{B_1 = \text{COIN}\}$, and 2 stigmatized goods $\{S_1 = \text{ARC}, S_2 = \text{NRA}\}$. The

⁵The full list of candidate goods and the survey questions used to winnow them down can be found in the appendix.

remaining goods were selected based on their similarity to goods that tested highly in this survey. The final list of experimentally tested goods is presented in Table 3.2.

Table 3.2: Table of Stigmatized and Neutral Goods Selected for Experiment

ALL GOODS AND BRIEF DESCRIPTIONS	
ITEM	DESCRIPTION
S_1 =NRA	<i>\$1 Donation to the National Rifle Association.</i>
S_2 =ARC	<i>A \$1 reduction of the EconLab's donation to the American Red Cross.</i>
SET ₁ N_1 =MUG	<i>A white, porcelain coffee mug with the UCSD logo.</i>
N_2 =PEN	<i>A stylus pen with the UCSD EconLab logo.</i>
$\{B_{1a}, B_{1b}\} = \{\text{COIN-2}, \text{COIN-5}\}$	<i>Gamble where subjects win $\approx$ \$20 if 2 (5) heads are flipped.</i>
$\{B_{2a}, B_{2b}\} = \{\text{COIN-3}, \text{COIN-4}\}$	<i>Gamble where subjects win $\approx$ \$10 if 3 (4) heads are flipped.</i>
S_{3a} =BOOKM	<i>"Inside Her Mind," a book on seducing women.</i>
SET ₂ S_{3b} =BOOKF	<i>"50 Shades of Grey," an erotic romance novel</i>
N_3 =BOOKN	<i>"The Crack Up," a novel by F. Scott Fitzgerald.</i>
B_{3a}, B_{3b} =NOVEL	<i>"Ready Player One" & "The Fault in Our Stars," popular, recent novels.</i>
S_4 =TVS	<i>"Masters of Sex," a TV show about sex researcher, William Masters.</i>
SET ₃ N_4 =TVN	<i>"Elementary," a CBS drama about Sherlock Holmes.</i>
B_{4a}, B_{4b} =COMEDY	<i>"The Goldbergs" & "Trophy Wife," both popular, recent TV comedies.</i>

Note: S_3 will be evaluated as either BOOKM or BOOKF depending on the sex of the subject.

Table 3.2 is divided into 3 “sets,” each consisting of stigmatized goods, neutral goods, and the bundling goods associated with them. The sets were designed to present the subjects with the most natural bundles possible while still providing sufficient subjectivity that would allow subjects to “blame” high WTPs on the bundling goods.

The stigma goods in the first set, ARC and NRA, represent social-preference goods that subjects found to bear negative emotion. The first is a reduction of the EconLab’s pre-scheduled donation to the American Red Cross. This generated the strongest level of self-stigma from our survey. The second strongest level of self-stigma came from NRA, which is a donation to the National Rifle Association (hereafter, NRA). While this level of self-stigma clearly would not hold across all sub-samples of the population, we are

primarily interested in the goods that are stigmatized among the sub-sample of the population from which we will recruit our experimental subjects.

The neutral goods in the first set, MUG and PEN, generated near-perfect neutral responses on our self-stigma measures (a score of 4 is perfectly neutral) and had low standard errors, making them excellent control goods for the experiment. Goods that were considered virtuous could confound the results as subjects may seek out the self-signal of stating a high WTP for such goods.

Bundled with these goods are four different iterations of COIN, a gamble in which subjects receive payment based on the number of heads flipped out of 7 coins. This good proved fairly neutral in the pre-survey and affords subjects considerable wiggle-room in the WTPs they reveal.

The stigmatized goods in the second set are gender-specific E-books, labeled BOOKM and BOOKF. These were not included in our first survey, but were carefully selected for their similarity to the gender-specific magazines, which showed strong self-stigma measures in the survey, but were infeasible to distribute anonymously. We believe that the gendered E-books excite the same emotions to an even greater degree.

The neutral good associated with these gender-specific E-Books is a neutral book, BOOKN, an F. Scott Fitzgerald E-book. We selected it because we believed that few subjects would have experience with or strong feelings about it, but many subjects would be familiar with its author. It also serves as a control for the gender-specific E-books, so any unique elements of eliciting WTP for E-books will be captured in both the treatment and control.

We chose other E-Books, NOVEL, to bundle with the previous E-Books,

because we felt that they provided natural bundling goods. The books chosen, “The Fault in Our Stars” and “Ready Player One,” were both highly regarded, recent novels. The descriptions of these books were neutral, and each book was identified as a New York Times bestseller.

The final set is our most experimental, using digital versions of one episode of different television shows as our goods. It is important to first note that our conjectures about the amount of stigma surrounding any one of these goods is not founded in survey data as with the previous two sets. The stigmatized show, TVS, is a Showtime series titled “Masters of Sex,” that deals with potentially stigmatized themes. Additionally, its title is open to an indecent interpretation.

The neutral good from the final set, TVN, was a network drama titled “Elementary” that possesses none of the controversial properties of the stigmatized good. It and TVS were bundled with two versions of COMEDY, each a neutral, network comedy that was identified as one of TV.com’s top comedies from 2014. We believe network comedies to be strong bundling goods as they do not tend to elicit strong sentiment from potential viewers.

At the end of the experiment, we asked all subjects which, if any, of the E-Books they owned. We also asked which, if any, of the TV shows they had previously watched. We will use this to control for previous experience or ownership that would generate a strong preference shock.

3.4.2 WTP Elicitation

In order to conduct our test of interest, we will need to elicit the following 22 WTPs:

- for $i \in \{1a, 1b, 3a, 3b, 4a, 4b\}$: WTP for B_i , denoted $W(B_i)$,⁶
- for $i \in \{1, 2, 3, 4\}$: WTP for S_i conditional on receipt of B_i , denoted $W(S_i|B_i)$,
- for $i \in \{1, 2, 3, 4\}$: WTP for N_i conditional on receipt of B_i , denoted $W(N_i|B_i)$,
- for $i \in \{1, 2, 3, 4\}$: WTP for S_i and B_i , denoted $W(S_i, B_i)$.
- for $i \in \{1, 2, 3, 4\}$: WTP for N_i and B_i , denoted $W(N_i, B_i)$.

Our WTP elicitation will employ a simple randomized MPL mechanism. The briefer instructions associated with this elicitation allow us to shorten the length of the experiment, reducing subject fatigue and experimental costs. This randomized MPL is designed to elicit WTP at \$0.50 intervals. Each row of the MPL asks subjects to choose between the item and an amount of money. Our MPL begins at -\$4 and ends at \$8. We believe these limits will capture switching behavior for most respondents. Since observing \$0.50 intervals at this range means that we will be asking subjects to answer 25 questions, we will use an auto-fill program, where once the subject switches from the item to the monetary value, the remaining questions automatically record that the subject would prefer the monetary value.

Auto-filling answers prevents subjects from switching from one side of the price list to the other multiple times. Multiple switching points is an observed phenomenon that is typically associated with confusion about the mechanism, though it can also indicate indifference (see Andersen, Harrison,

⁶Since the gambles were very similar, we only elicited one WTP per subject, either for B_{1a} or B_{1b} .

Lau, and Rutstrom (2006) for a discussion of this behavior). To ensure that the mechanism itself is not confusing subjects, we will test subjects' comprehension of the mechanism prior to running the study. Since our analysis relies on a difference-in-differences design, we believe that any pressure that this restriction places on the switching point of an individual is independent of the stigma associated with the item in question.

We will split the 22 WTP elicitation into two groups of 11 and randomly assign each subject to the two groups. Subjects will first see 8 randomly-ordered WTP elicitation, 4 for bundled goods and 4 for conditional goods. Then each will see 3 randomly-ordered WTP elicitation for the bundling goods they did not evaluate in any of their bundles. Thus, no subject will face both a bundle and any individual component from that bundle, nor will any subject face both $\{W(G_i, B_I)\}$ and $\{W(G_i|B_i)\}$.

3.5 Conclusion

Our mechanism investigates the context-dependence of WTP elicitation, specifically focusing on distortions brought about by self-stigma. Uncovering the sensitivity of these measures to context gives policy-makers better predictions about the consumption rates of various goods and a deeper understanding of their welfare implications. For instance, healthcare policy is sensitive to both individuals' valuations for cigarettes and their valuations for programs to quit smoking. Our bundling mechanism can more accurately measure preferences over these potentially stigmatized goods like these, in particular finding that the self-signal cost of expressing demand for stigmatized

goods systematically depresses the stated WTP for them.

Since our mechanism is explicitly designed to weaken the self-signal associated with a good or action, it raises an interesting question about the response to bundling “virtuous” goods with neutral goods. Our model predicts that a weaker self-signal associated with prosocial behavior would make people less likely to express demand for such behavior. Our model generates more ambiguous predictions about bundling virtuous goods with stigmatized goods, as the WTP elicited would send a signal about both desirable and undesirable characteristics. We leave these questions for future research.

Finally, our method suggests a new way for policy-makers to distribute stigmatized goods, especially in less-developed countries. These goods include STI tests, contraceptives, addiction treatment, female education, and savings-commitment devices. Bundling these goods with neutral goods could boost uptake. For instance, bundling contraceptives with other basic medicines could make people more willing to acquire contraceptives by fulfilling demand that standard measures of WTP overlook.

3.6 Acknowledgement

Chapter 3, in part is currently being prepared for submission for publication of the material. Brownback, Andy; Gagnon-Bartsch, Tristan; Li, Shengwu. The dissertation author was the primary investigator and author of this material.

3.7 Chapter 3 Appendix

3.7.1 Proof of Lemma 1

Proof. First note that the utility from announcing $W = 0$ is $u_0 = k_j \bar{\theta}^L$. The utility from announcing $W = 1$ is $u_1(v_j) = v_j - \frac{1}{2} + k_j(\alpha(j) + [1 - \alpha(j)]\bar{\theta}_j)$. Finally, the utility from the optimal interior bid for type v_j is $\tilde{u}(v_j) = k_j[1 - \alpha(j)]\bar{\theta}_j + \frac{1}{2}(v_j + k_j\alpha(j))^2$. The only types to consider an interior bid are those for whom $\widetilde{W}(v_j) = v_j + k_j\alpha(j)$ falls in $(0, 1)$. These types comprise the open interval $V_I \equiv (\underline{v}, \bar{v})$, where $\underline{v} \equiv -k_j\alpha(j)$, and $\bar{v} \equiv 1 - k_j\alpha(j)$. Since types $v_j < \underline{v}$ optimally announce $W = 0$, we must determine the threshold value $\hat{v} > \underline{v}$ such that types $v > \hat{v}$ prefer to announce $W > 0$.

First consider when a type with $\widetilde{W}(v_j) \in (0, 1)$ prefers $W = 0$ over $\widetilde{W}(v_j)$. This happens whenever

$$u_0 > \tilde{u}(v_j) \Leftrightarrow v_j < \hat{v} \equiv -k_j\alpha(j) + \sqrt{2k_j\{\bar{\theta}^L - [1 - \alpha(j)]\bar{\theta}_j\}}$$

conditional on $\hat{v} < \bar{v}$. Such a $\hat{v}$ exists iff $k_j > \frac{1}{2(\bar{\theta}^L - [1 - \alpha(j)]\bar{\theta}_j)} \equiv \hat{k}_j$. If $0 \geq k_j > \hat{k}_j$, then the optimal bidding strategy is as follows: $v_j < \hat{v}$ chooses $W = 0$, $v_j \in [\hat{v}, \bar{v}]$ chooses $\widetilde{W}(v_j)$, and $v_j > \bar{v}$ chooses $W = 1$. If $k_j < \hat{k}_j$, then all $v_j \in V_I$ choose $W = 0$. The type indifferent between announcing $W = 0$ and $W = 1$ is v^* such that $u_1(v^*) = u_0$, hence $v^* = \frac{1}{2} + k_j\{\bar{\theta}^L - \alpha(j) - [1 - \alpha(j)]\bar{\theta}_j\}$. With $k_j < \hat{k}_j$, it is straightforward to verify $v^* > \bar{v}$, and hence the optimal bidding strategy is: $W = 0$ if $v_j \leq v^*$, and $W = 1$ if $v > v^*$. $\square$

3.7.2 Proof of Lemma 2

Proof. The logic of this proof is identical to that of Lemma 1. The key difference is that in the bundled case the signal $W - \bar{\theta}_B$ about θ_S has variance $\sigma_T^2 \equiv \sigma_S^2 + \sigma_B^2 + \gamma_B^2$, whereas in the single-good case it has variance σ_S^2 . Hence, replacing all instances of σ_S^2 in the proof of Lemma 1 yields the result. Specifically, the new parameters of interest are (1) $\bar{v}_T = 1 - k_S \alpha(S, B)$, (2) $\hat{v}_T = -k_S \alpha(S, B) + \sqrt{2k_S \{\bar{\theta}_T^L - [1 - \alpha(S, B)]\bar{\theta}\}}$, where

$$\bar{\theta}_T^L \equiv \bar{\theta}_S - \frac{\gamma_S^2}{\sigma_T} \left(\frac{\phi\left(\frac{-\bar{\theta}_S}{\gamma_S}\right)}{\Phi\left(\frac{-\bar{\theta}_S}{\gamma_S}\right)} \right),$$

$$(3) \tilde{k}_S = \frac{1}{2(\bar{\theta}_T^L - [1 - \alpha(S, B)]\bar{\theta}_S)}, \text{ and } v_T^* = \frac{1}{2} + k_S \{\bar{\theta}_T^L - \alpha(S, B) - [1 - \alpha(S, B)]\bar{\theta}_S\}.$$

□

3.7.3 Proof of Proposition 2

Proof. We show that moving from single-good elicitation to bundled elicitation increases the average bid. First note that $\tilde{k}_S < \hat{k}_S$. This implies three cases.

Case 1:

$k_S < \tilde{k}_S$. No type chooses an interior bid in either elicitation scheme.

We need only show that the threshold v^* at which a type switches from $W = 0$ to $W = 1$ is strictly lower under bundled elicitation. Note *Case 2:* $\tilde{k}_S < k_S < \hat{k}_S$, and (3) $k_S > \hat{k}_S$. □

3.7.4 Full Survey Results

Table 3.3 shows all goods evaluated in the first survey. Many are goods that we knew to not be potential candidates for the experiment (for example, cigarettes) but used to both understand the nature of stigmas in the population and to provide baseline estimates of stigma.

Each column of Table 3.3 reports average responses to our 7-point sentiment questions. The questions, which correspond to the respective column, are: (1) feelings about one's self conditional on paying for the item, (2) feelings about one's self conditional on accepting the item for free, (3) feelings about another person who paid for the item, (4) perception of how others would judge one's self for purchasing the item, (5) embarrassment felt while purchasing the item.

3.7.5 Relationship between Stigma and Stated WTP

With our survey results, we can address the connection between reported self stigma and stated WTP. This connection is central to our experiment, as our mechanism relies on the fact that psychological costs are not only paid in consumption of stigmatized goods, but in expressing demand for them as well. Table 3.4 displays the results of an OLS regression of stated WTP on reported self stigma in the first column. In the second column, we report the same regression including fixed effects for the items. This within-item analysis is an even stronger results that says that, holding constant characteristics about the item, people who report higher stigma associated with the item also state lower WTP.

Table 3.3: Summary Statistics of Sentiment towards Candidate Goods

	SELF STIGMA (IF PURCHASED)	SELF STIGMA (IF FREE)	STIGMA TOWARDS OTHERS	BELIEFS ABOUT SOCIAL STIGMA	EMBARRASSMENT
CIGARETTES	2.0833 (1.1557)	3.1944 (1.5642)	3.0278 (1.0278)	2.3611 (1.0462)	2.5278 (1.1081)
CONDOMS	4.5278 (1.5021)	4.8333 (1.4442)	4.6111 (1.1778)	4.2222 (1.4165)	3.6944 (1.6181)
REDUCTION IN RED CROSS DONATION	2.6111 (1.2935)	3.7222 (1.5966)	3.3056 (0.9202)	3.1667 (1.1588)	3.0000 (1.4142)
UCSD MUG	4.4722 (1.2980)	4.6667 (0.9258)	5.8056 (1.0642)	4.6111 (0.8711)	4.7222 (1.2097)
COIN-FLIP GAME	3.5278 (1.4439)	5.4167 (1.1052)	3.7500 (1.1802)	3.6389 (0.9305)	3.6111 (1.0496)
DONATION TO NRA	2.7222 (1.7339)	3.2778 (1.7826)	3.2500 (1.6279)	3.3056 (1.6181)	2.9167 (1.6279)
COUNSELING OR THERAPY	4.4167 (1.6966)	5.1389 (1.3342)	4.7778 (1.2215)	4.3611 (1.2684)	4.1111 (1.3686)
PRETZELS	4.2222 (1.1492)	5.3889 (1.1533)	4.3611 (0.8333)	4.1667 (0.6547)	4.1944 (0.4672)
CHOCOLATE- COVERED PRETZELS	4.0278 (1.5021)	5.3056 (1.4307)	4.2222 (1.0720)	4.1111 (0.9189)	4.1389 (1.1251)
FOIE GRAS	4.0833 (1.6453)	5.2500 (1.9328)	4.1944 (0.9804)	4.0556 (1.1198)	4.0278 (1.3199)
BOOK ON INCREASING IQ	3.4722 (1.5021)	4.3611 (1.6588)	3.6944 (1.1909)	3.2500 (1.1557)	2.8056 (1.2608)
ICE CREAM	4.6389 (1.2225)	5.6389 (1.3555)	4.5278 (0.9706)	4.4722 (0.8102)	4.3889 (1.0764)
URN-GAMBLE	3.5556 (1.4029)	5.5556 (1.1070)	3.9167 (0.8409)	3.7500 (0.8742)	3.6667 (0.9856)
UCSD STYLUS PEN	4.0833 (0.9964)	5.5000 (0.9411)	4.0000 (0.5345)	4.1111 (0.6223)	3.9167 (0.7700)
FRUIT BAR	4.7778 (1.1738)	5.5000 (1.1588)	4.4444 (0.8433)	4.3889 (0.8376)	4.2778 (1.0313)
PEOPLE MAGAZINE	3.3056 (1.5084)	4.4444 (1.4431)	3.6944 (1.2608)	3.4722 (0.8779)	3.1389 (1.1748)
MAXIM MAGAZINE	3.3611 (1.3555)	4.2222 (1.5512)	3.6111 (1.1283)	3.4444 (1.0541)	3.0833 (1.0790)

Note: $N = 36$. Standard deviations in parentheses.

Table 3.4: Regression of WTP on Reported Stigma

	WTP	WTP
Self Stigma	2.549***	2.331***
<i>(1-Negative to 7-Positive)</i>	(0.143)	(0.267)
Constant	-10.05***	-9.236***
	(0.576)	(0.998)
Item Fixed-Effect	No	Yes
<i>N</i>	612	612

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.7.6 Auto-Filling MPL Mechanism

We would like to know how you value the following item:

- One Banana

PREFER THE ITEM		PREFER THE MONEY
BANANA	OR	LOSE \$1.50
BANANA	OR	LOSE \$1.00
BANANA	OR	LOSE \$0.50
BANANA	OR	----- \$0.00 -----
BANANA	OR	GAIN \$0.50
BANANA	OR	GAIN \$1.00
BANANA	OR	GAIN \$1.50
BANANA	OR	GAIN \$2.00
BANANA	OR	GAIN \$2.50
BANANA	OR	GAIN \$3.00

Figure 3.1: Example of Auto-Filling MPL Mechanism (Unclicked)

BANANA	OR	LOSE \$1.50
BANANA	OR	LOSE \$1.00
BANANA	OR	LOSE \$0.50
BANANA	OR	----- \$0.00 -----
BANANA	OR	GAIN \$0.50
BANANA	OR	GAIN \$1.00
BANANA	OR	GAIN \$1.50
BANANA	OR	GAIN \$2.00
BANANA	OR	GAIN \$2.50
BANANA	OR	GAIN \$3.00

Figure 3.2: Example of Auto-Filling MPL Mechanism (Clicked)

Bibliography

Adams, William James and Janet L Yellen, “Commodity bundling and the burden of monopoly,” *The quarterly journal of economics*, 1976, pp. 475–498.

Akerlof, George A and Rachel E Kranton, “Economics and identity,” *Quarterly journal of Economics*, 2000, pp. 715–753.

Amann, Erwin and Wolfgang Leininger, “Asymmetric all-pay auctions with incomplete information: the two-player case,” *Games and Economic Behavior*, 1996, 14 (1), 1–18.

Andersen, Steffen, Glenn W Harrison, Morten Igel Lau, and E Elisabet Rutström, “Elicitation using multiple price list formats,” *Experimental Economics*, 2006, 9 (4), 383–405.

Andreoni, James and Andy Brownback, “Grading on a Curve, and other Effects of Group Size on All-Pay Auctions,” Technical Report, National Bureau of Economic Research 2014.

— **and William T Harbaugh**, “Unexpected Utility: Experimental Tests of Five Key Questions about Preferences over Risk,” Technical Report, University of Oregon Economics Department 2010.

— **, Justin M Rao, and Hannah Trachtman**, “Avoiding the ask: a field experiment on altruism, empathy, and charitable giving,” Technical Report, National Bureau of Economic Research 2011.

— **, Yeon-Koo Che, and Jinwoo Kim**, “Asymmetric information about rivals’ types in standard auctions: An experiment,” *Games and Economic Behavior*, 2007, 59 (2), 240–259.

Arulampalam, Wiji, Robin A Naylor, and Jeremy Smith, “Am I missing something? The effects of absence from class on student performance,” *Economics of Education Review*, 2012, 31 (4), 363–375.

- Barut, Yasar, Dan Kovenock, and Charles N Noussair**, “A Comparison of Multiple-Unit All-Pay and Winner-Pay Auctions Under Incomplete Information*,” *International Economic Review*, 2002, 43 (3), 675–708.
- Bateman, Ian, Alistair Munro, Bruce Rhodes, Chris Starmer, and Robert Sugden**, “Does part–whole bias exist? An experimental investigation,” *The Economic Journal*, 1997, 107 (441), 322–332.
- Baye, Michael R, Dan Kovenock, and Casper G de Vries**, “Rigging the Lobbying Process: An Application of the All-Pay Auction,” *The American Economic Review*, 1993, 83 (1), 289–294.
- , —, and —, “The all-pay auction with complete information,” *Economic Theory*, 1996, 8 (2), 291–305.
- Becker, Gaylene and Regina Arnold**, “Stigma as a social and cultural construct,” in “The dilemma of difference,” Springer, 1986, pp. 39–57.
- Becker, Gordon M, Morris H DeGroot, and Jacob Marschak**, “Measuring utility by a single-response sequential method,” *Behavioral science*, 1964, 9 (3), 226–232.
- Becker, William E and Sherwin Rosen**, “The learning effect of assessment and evaluation in high school,” *Economics of Education Review*, 1992, 11 (2), 107–118.
- Bell, David E**, “Disappointment in decision making under uncertainty,” *Operations research*, 1985, 33 (1), 1–27.
- Bem, Daryl J**, “Self-perception theory,” *Advances in experimental social psychology*, 1972, 6, 1–62.
- Bénabou, Roland and Jean Tirole**, “Self-confidence and personal motivation,” *Quarterly journal of economics*, 2002, pp. 871–915.
- and —, “Willpower and personal rules,” *Journal of Political Economy*, 2004, 112 (4), 848–886.
- Benartzi, Shlomo and Richard H Thaler**, “Myopic loss aversion and the equity premium puzzle,” *The Quarterly Journal of Economics*, 1995, 110 (1), 73–92.
- Benjamin, Daniel J, Matthew Rabin, and Collin Raymond**, “A Model of Non-Belief in the Law of Large Numbers,” 2014.

- Bernheim, B Douglas**, “A theory of conformity,” *Journal of political Economy*, 1994, pp. 841–877.
- , **Daniel Bjorkegren**, **Jeffrey Naecker**, and **Antonio Rangel**, “Do Hypothetical Choices and Non-Choice Ratings Reveal Preferences?,” Technical Report, National Bureau of Economic Research 2013.
- Betts, Julian R**, “The impact of educational standards on the level and distribution of earnings,” *American Economic Review*, 1998, pp. 266–275.
- and **Jeff Grogger**, “The impact of grading standards on student achievement, educational attainment, and entry-level earnings,” *Economics of Education Review*, 2003, 22 (4), 343–352.
- Bharadwaj, Prashant**, **Malles M Pai**, and **Agne Suziedelyte**, “Mental Health Stigma,” 2015.
- Binswanger, Hans P**, “Attitudes toward risk: Experimental measurement in rural India,” *American journal of agricultural economics*, 1980, 62 (3), 395–407.
- Blair, Sean** and **Neal J Roese**, “Balancing the basket: the role of shopping basket composition in embarrassment,” *Journal of Consumer Research*, 2013, 40 (4), 676–691.
- Bodner, Ronit**, “Self-knowledge and the diagnostic value of one’s actions.” PhD dissertation, Massachusetts Institute of Technology 1996.
- and **Drazen Prelec**, “Self-signaling and diagnostic utility in everyday decision making,” *The psychology of economic decisions*, 2003, 1, 105–26.
- Bull, Clive**, **Andrew Schotter**, and **Keith Weigelt**, “Tournaments and piece rates: An experimental study,” *The Journal of Political Economy*, 1987, pp. 1–33.
- Bursztyn, Leonardo**, **Michael Callen**, **Bruno Ferman**, **Ali Hasanain**, and **Noam Yuchtman**, “A Revealed Preference Approach to the Elicitation of Political Attitudes: Experimental Evidence on Anti-Americanism in Pakistan,” Technical Report, National Bureau of Economic Research 2014.
- Carson, Richard T**, “Contingent valuation: A practical alternative when prices aren’t available,” *The Journal of Economic Perspectives*, 2012, 26 (4), 27–42.

- , **Nicholas E Flores**, and **Norman F Meade**, “Contingent valuation: controversies and evidence,” *Environmental and resource economics*, 2001, 19 (2), 173–210.
- Chen, Jennjou and Tsui-Fang Lin**, “Class attendance and exam performance: A randomized experiment,” *The Journal of Economic Education*, 2008, 39 (3), 213–227.
- Coffman, Katherine B, Lucas C Coffman, and Keith M Marzilli Ericson**, “The size of the LGBT population and the magnitude of anti-gay sentiment are substantially underestimated,” Technical Report, National Bureau of Economic Research 2013.
- Costrell, Robert M**, “A simple model of educational standards,” *The American Economic Review*, 1994, pp. 956–971.
- Crowne, Douglas P and David Marlowe**, “A new scale of social desirability independent of psychopathology,” *Journal of consulting psychology*, 1960, 24 (4), 349.
- Czibor, Eszter, Sander Onderstal, Randolph Sloof, and Mirjam van Praag**, “Does Relative Grading Help Male Students? Evidence from a Field Experiment in the Classroom,” 2014.
- Dana, Jason, Roberto A Weber, and Jason Xi Kuang**, “Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness,” *Economic Theory*, 2007, 33 (1), 67–80.
- Davis, Douglas D and Robert J Reilly**, “Do too many cooks always spoil the stew? An experimental analysis of rent-seeking and the role of a strategic buyer,” *Public Choice*, 1998, 95 (1-2), 89–115.
- Dechenaux, Emmanuel, Dan Kovenock, and Roman M Sheremeta**, “A survey of experimental research on contests, all-pay auctions and tournaments,” Technical Report, Discussion Paper, Social Science Research Center Berlin (WZB), Research Area ‘Markets and Politics’, Research Professorship & Project ‘The Future of Fiscal Federalism’ 2012.
- Dobkin, Carlos, Ricard Gil, and Justin Marion**, “Skipping class in college and exam performance: Evidence from a regression discontinuity classroom experiment,” *Economics of Education Review*, 2010, 29 (4), 566–575.

- Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner**, “Individual risk attitudes: Measurement, determinants, and behavioral consequences,” *Journal of the European Economic Association*, 2011, 9 (3), 522–550.
- Dubey, Pradeep and John Geanakoplos**, “Grading exams: 100, 99, 98, or A, B, C?,” *Games and Economic Behavior*, 2010, 69 (1), 72–94.
- Edwards, Allen L**, “The social desirability variable in personality assessment and research,” 1957.
- Eil, David and Justin M Rao**, “The good news-bad news effect: asymmetric processing of objective information about yourself,” *American Economic Journal: Microeconomics*, 2011, pp. 114–138.
- Exley, Christine L**, “Excusing Selfishness in Charitable Giving: The Role of Risk,” 2014.
- Eyster, Erik and Matthew Rabin**, “Cursed equilibrium,” *Econometrica*, 2005, 73 (5), 1623–1672.
- Fischbacher, Urs**, “z-Tree: Zurich toolbox for ready-made economic experiments,” *Experimental economics*, 2007, 10 (2), 171–178.
- Fisher, Robert J**, “Social desirability bias and the validity of indirect questioning,” *Journal of consumer research*, 1993, pp. 303–315.
- and **James E Katz**, “Social-desirability bias and the validity of self-reported values,” *Psychology & Marketing*, 2000, 17 (2), 105–120.
- Fraja, Gianni De, Tania Oliveira, and Luisa Zanchi**, “Must try harder: Evaluating the role of effort in educational attainment,” *The Review of Economics and Statistics*, 2010, 92 (3), 577–597.
- Gaeth, Gary J, Irwin P Levin, Goutam Chakraborty, and Aron M Levin**, “Consumer evaluation of multi-product bundles: An information integration analysis,” *Marketing letters*, 1991, 2 (1), 47–57.
- Gneezy, Uri and Rann Smorodinsky**, “All-pay auctions: an experimental study,” *Journal of Economic Behavior & Organization*, 2006, 61 (2), 255–275.
- , **Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen**, “Motivated Self-Deception, Identity, and Unethical Behavior,” 2015.

- Grossman, Zachary**, “Self-signaling versus social-signaling in giving,” *Department of Economics, UCSB*, 2010.
- Gul, Faruk**, “A theory of disappointment aversion,” *Econometrica: Journal of the Econometric Society*, 1991, pp. 667–686.
- Harbring, Christine and Bernd Irlenbusch**, “Incentives in tournaments with endogenous prize selection,” *Journal of Institutional and Theoretical Economics JITE*, 2005, 161 (4), 636–663.
- Hebert, James R, Lynn Clemow, Lori Pbert, Ira S Ockene, and Judith K Ockene**, “Social desirability bias in dietary self-report may compromise the validity of dietary intake measures,” *International Journal of Epidemiology*, 1995, 24 (2), 389–398.
- Hillman, Arye L and Dov Samet**, “Dissipation of contestable rents by small numbers of contenders,” *Public Choice*, 1987, 54 (1), 63–82.
- **and John G Riley**, “Politically Contestable Rents and Transfers*,” *Economics & Politics*, 1989, 1 (1), 17–39.
- Hirshleifer, Jack and John G Riley**, “Elements of the Theory of Auctions and Contests,” Technical Report, UCLA Department of Economics 1978.
- Kagel, John H and Dan Levin**, “Independent private value auctions: Bidder behaviour in first-, second-and third-price auctions with varying numbers of bidders,” *The Economic Journal*, 1993, pp. 868–879.
- Kahneman, Daniel and Amos Tversky**, “On the psychology of prediction,” *Psychological review*, 1973, 80 (4), 237.
- **and —**, “Choices, values, and frames,” *American psychologist*, 1984, 39 (4), 341.
- Kendall, Maurice G**, “A new measure of rank correlation,” *Biometrika*, 1938, pp. 81–93.
- , “Rank correlation methods,” 1975.
- Kokkelenberg, Edward C, Michael Dillon, and Sean M Christy**, “The effects of class size on student grades at a public university,” *Economics of Education Review*, 2008, 27 (2), 221–233.
- Kőszegi, Botond and Matthew Rabin**, “A model of reference-dependent preferences,” *The Quarterly Journal of Economics*, 2006, pp. 1133–1165.

- and —, “Reference-dependent risk attitudes,” *The American Economic Review*, 2007, pp. 1047–1073.
- Krishna, Vijay and John Morgan**, “An analysis of the war of attrition and the all-pay auction,” *journal of economic theory*, 1997, 72 (2), 343–362.
- Krueger, Anne O**, “The political economy of the rent-seeking society,” *The American economic review*, 1974, pp. 291–303.
- Lazear, Edward P and Sherwin Rosen**, “Rank-Order Tournaments as Optimum Labor Contracts,” *The Journal of Political Economy*, 1981, 89 (5), 841–864.
- Lim, Wooyoung, Alexander Matros, and Theodore L Turocy**, “Bounded rationality and group size in Tullock contests: Experimental evidence,” *Journal of Economic Behavior & Organization*, 2014, 99, 155–167.
- List, John A, Daan Van Soest, Jan Stoop, and Haiwen Zhou**, “On the role of group size in tournaments: Theory and evidence from lab and field experiments,” Technical Report, National Bureau of Economic Research 2014.
- Loomes, Graham and Robert Sugden**, “Disappointment and dynamic consistency in choice under uncertainty,” *The Review of Economic Studies*, 1986, 53 (2), 271–282.
- Maccoby, Eleanor E and Nathan Maccoby**, “The interview: A tool of social science,” *Handbook of social psychology*, 1954, 1, 449–487.
- McAfee, R Preston, John McMillan, and Michael D Whinston**, “Multiproduct monopoly, commodity bundling, and correlation of values,” *The Quarterly Journal of Economics*, 1989, pp. 371–383.
- Mijović-Prelec, Danica and Drazen Prelec**, “Self-deception as self-signalling: a model and experimental evidence,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2010, 365 (1538), 227–240.
- Minor, Dylan**, “Coarse Strategy and Competition,” Technical Report 2012.
- Mobius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat**, “Managing self-confidence: Theory and experimental evidence,” Technical Report, National Bureau of Economic Research 2011.
- Moldovanu, Benny and Aner Sela**, “The optimal allocation of prizes in contests,” *American Economic Review*, 2001, pp. 542–558.

- **and** –, “Contest architecture,” *Journal of Economic Theory*, 2006, *126* (1), 70–96.
- Mosteller, Frederick**, “The Tennessee study of class size in the early school grades,” *Future of children*, 1995, *5*, 113–127.
- Mroch, Andy**, “Law School Grading Curves,” 2005.
- Müller, Wieland and Andrew Schotter**, “Workaholics and dropouts in organizations,” *Journal of the European Economic Association*, 2010, *8* (4), 717–743.
- Niederle, Muriel and Lise Vesterlund**, “Gender and competition,” *Annu. Rev. Econ.*, 2011, *3* (1), 601–630.
- Noussair, Charles and Jonathon Silver**, “Behavior in all-pay auctions with incomplete information,” *Games and Economic Behavior*, 2006, *55* (1), 189–206.
- Olszewski, Wojciech and Ron Siegel**, “Large Contests,” 2013.
- Orrison, Alannah, Andrew Schotter, and Keith Weigelt**, “Multiperson tournaments: An experimental examination,” *Management Science*, 2004, *50* (2), 268–279.
- Paredes, Valentina**, “Grading System and Student Effort.”
- Paulhus, Delroy L**, “Two-component models of socially desirable responding,” *Journal of personality and social psychology*, 1984, *46* (3), 598.
- **and Douglas B Reid**, “Enhancement and denial in socially desirable responding,” *Journal of Personality and Social Psychology*, 1991, *60* (2), 307.
- Potters, Jan, Casper G de Vries, and Frans Van Winden**, “An experimental examination of rational rent-seeking,” *European Journal of Political Economy*, 1998, *14* (4), 783–800.
- Quattrone, George A and Amos Tversky**, “Causal versus diagnostic contingencies: On self-deception and on the voter’s illusion,” *Journal of Personality and Social Psychology*, 1984, *46* (2), 237.
- Rabin, Matthew**, “Inference by Believers in the Law of Small Numbers,” *Quarterly Journal of Economics*, 2002, pp. 775–816.

- Romer, David**, “Do students go to class? Should they?,” *The Journal of Economic Perspectives*, 1993, pp. 167–174.
- Siegel, Ron**, “All-Pay Contests,” *Econometrica*, 2009, 77 (1), 71–92.
- Spearman, Charles**, “The proof and measurement of association between two things,” *The American journal of psychology*, 1904, 15 (1), 72–101.
- Sprenger, Charles**, “An endowment effect for risk: Experimental tests of stochastic reference points,” Technical Report, working paper 2010.
- Stinebrickner, Ralph and Todd R Stinebrickner**, “The Causal Effect of Studying on Academic Performance,” *The B.E. Journal of Economic Analysis & Policy*, 2008, 8 (1).
- Tullock, Gordon**, “The welfare costs of tariffs, monopolies, and theft,” *Economic Inquiry*, 1967, 5 (3), 224–232.
- Tversky, Amos and Daniel Kahneman**, “Belief in the law of small numbers,” *Psychological bulletin*, 1971, 76 (2), 105.
- Vickrey, William**, “Counterspeculation, auctions, and competitive sealed tenders,” *The Journal of finance*, 1961, 16 (1), 8–37.
- Welch, Jack and John A Byrne**, “Straight from the Gut,” *New York*, 2001.