

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Economics

2013

Essays on Incentives and Effort

Lau, Yan

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp011z40ks95z>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

ESSAYS ON INCENTIVES AND EFFORT

YAN LAU

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF ECONOMICS

ADVISOR: ALEXANDRE MAS

SEPTEMBER 2013

© Copyright by Yan Lau, 2013. All rights reserved.

Abstract

Incentives motivate economic agents to exert effort in producing outcomes, and the way governments, firms, and educators structure incentive schemes can influence the amount of effort exerted by these agents. This collection of essays examines various factors affecting such effort decisions. Chapter 1 investigates the role of academic competition in student effort decisions. I develop a tournament model to describe how an increase in the size of a particular schooling cohort in Hong Kong can motivate better students to “step up” effort exertion, but discourage less-able students. I use test score data from Hong Kong to empirically test these predictions. While Chapter 1 involves a natural experiment, Chapter 2 examines how changes in tournament structure (the winning cutoff and prize amount) in a laboratory experiment can influence effort decisions. Generating data in a controlled experimental setting where participants are asked to compete with one another in memorization tasks, I find that effort is positively correlated with prize amount, and that only low-performing participants are effort-responsive to movements in the winning cutoff. Moving from tournament incentive schemes to piece-rate incentive schemes, Chapter 3 re-examines an experiment conducted by Fehr and Goette (2007) in which the wage rate of Zurich bicycle messengers is increased for a short period of time. Fehr and Goette use a reference dependent preferences model to explain patterns in effort response. To explain their findings without appealing to behavioral mechanisms, I build a two-step non-time-separable neoclassical model with uncertainty resulting from limited work availability. I conduct additional analyses with their data to argue my case for the neoclassical model. The findings of these three essays suggest that the way incentive schemes are structured does matter in motivating agents to exert effort.

Acknowledgements

I am extremely grateful to Alexandre Mas, my advisor, as well as to Harvey Rosen and Cecilia Rouse, members of my committee, for their guidance, advice, and support. Without their wisdom and generosity, this thesis would not have been possible.

I would like to thank the participants of seminars and workshops at the Industrial Relations Section, and the Public Finance Working Group, for their suggestions and comments. I would especially like to acknowledge Benjamin Brooks, Gonzalo Cisternas, Henry Farber, Nikolaj Harmon, Tatiana Homonoff, Nicholas Lawson, David Lee, Jay Lu, Quynh Nguyen, and Phyllis Sun. Financial support from the Department of Economics and the Industrial Relations Section at Princeton University sustained my effort exertion. Additional acknowledgements can be found in each individual chapter.

This thesis is dedicated to my parents—Dominic Lau and Suzanne Lee. Your love and understanding form the foundation of my intellectual pursuit.

Contents

Abstract	iii
Acknowledgements	iv
1 Step Up or Give Up: Student Response to	
Increased Competition from the Dragon Cohort	1
1.1 Introduction	2
1.2 Tournament Model of Academic Competition	6
1.2.1 Theoretical Framework	6
1.2.2 Extensions	14
1.2.3 Numerical Illustration	15
1.2.4 Additional Applications	19
1.3 Background on the Dragon Cohort	20
1.4 Empirical Setup	24
1.4.1 Data	24
1.4.2 Identification Strategy	27
1.5 Empirical Results	28
1.5.1 Effort and Test Score Responses	29
1.5.2 Falsification Exercise	40
1.5.3 Teacher Reallocation	42
1.6 Conclusion	44
1.7 Proofs	47
1.7.1 Proof of Proposition 1.1	47
1.7.2 Proof of Proposition 1.2	49
1.8 Data Appendix	50
1.8.1 Effort Comparability	53
1.8.2 Test Score Normalizations	53

1.8.3	Predicted Score Percentile	54
1.9	Differences in Cohort Characteristics	55
1.10	Supplemental Results	57
1.10.1	University Intake by Program	57
1.10.2	Regression of Outcomes on Controls Only	59
1.10.3	Additional Coefficient Estimates of Main Results	61
1.11	Selection Issues	66
1.12	Teacher Reallocation Details	66
2	Tournament Structure and Effort in an Experimental Setting	71
2.1	Introduction	72
2.2	Model	75
2.3	Experimental Procedure	79
2.4	Results	86
2.4.1	Does effort affect outcomes?	88
2.4.2	Effort Response to Changes in Tournament Structure	94
2.4.3	Discussion	101
2.5	Conclusion	103
2.6	Recruitment Email	106
2.7	Instructions Provided to Participants	106
2.8	Screen-shots of Computer Interface	108
2.9	Local Polynomial Regressions	108
3	A Dynamic Model of Labor Supply for Bicycle Messengers: A Re-examination of Fehr and Goette's Experiment	114
3.1	Introduction	115
3.2	Fehr and Goette (2007)	116

3.3	Neoclassical Model of Labor Supply	120
3.3.1	Measuring Risk Aversion Versus Loss Aversion	122
3.3.2	Deterministic Framework	124
3.3.3	Two-Step Decision Framework with Uncertain Effort Caps	128
3.3.4	Predictions of Reference Dependent Preferences Model in Fehr and Goette (2007)	137
3.4	Additional Empirical Analyses	142
3.4.1	Revenue Per Shift (Prediction [1])	145
3.4.2	Number of Shifts and Total Revenue (Predictions [2] and [3])	147
3.4.3	Take-home Pay (Predictions [4] and [5])	150
3.4.4	Weather Effects (Predictions [7] and [8])	151
3.4.5	Effort of Surrounding Periods (Prediction [6])	155
3.5	Conclusion	158
3.6	$E[x_t]$ and $\text{Var}[x_t]$	160
3.7	Endogenous Reference Points	162
	References	164

List of Tables

1.1	Substitutions for Numerical Illustration	16
1.2	Enrollment of Dragon Cohort	22
1.3	Summary Statistics	26
1.4	Changes in Effort of Dragon Cohort by Quartile	32
1.5	Changes in Test Scores of Dragon Cohort by Quartile	37
1.6	Falsification Exercise Using Other Countries	42
1.7	Details for Student-level Variables	51
1.8	Details for Parent-level Variables	52
1.9	Details for School-level Variables	52
1.10	Differences in Predetermined Characteristics	56
1.11	Regressions of Effort and Test Score on Controls Only	60
1.12	Changes in Effort of Dragon Cohort by Quartile	62
1.12	Changes in Effort of Dragon Cohort by Quartile (Continued)	63
1.13	Changes in Effort of Dragon Cohort by Quartile	64
1.13	Changes in Effort of Dragon Cohort by Quartile (Continued)	65
1.14	Teacher Reallocation Parameter Estimates	70
2.1	Relative Winning Cutoffs by Session Size	83
2.2	Possible Prize-Cutoff Combinations	83
2.3	Treatments	84
2.4	Treatment Sequences and Groups	85
2.5	Summary Statistics by Treatment Group	85
2.6	Score Regressions	90
2.7	Percentile Rank Regressions	92
2.8	Pooled Regressions (Assumes Symmetry)	97
2.9	Regressions with Separate Treatment Effects	100

3.1	Reproduction of Table 6 from Fehr and Goette (2007)	120
3.2	Summary of Model Predictions	136
3.3	Summary Statistics	144
3.4	Revenue (Effort) per Shift Regressions	146
3.5	Number of Shifts per Block Regressions	147
3.6	Total Revenue per Block Regressions	149
3.7	Standard Deviation of Take-home Pay	151
3.8	Weather Summary Statistics	152
3.9	Weather Regressions	154
3.10	Prior Effort Regressions	156
3.11	Standard Deviation of Take-home Pay by Treatment	163

List of Figures

1.1	Hours Per Week Studying Math	2
1.2	Effort by Ability Percentile	17
1.3	Effort Before and After by Ability Percentile	17
1.4	Change in Effort by Ability Percentile	18
1.5	Change in Effort by Ability Percentile (Errors with Larger Spread)	19
1.6	Hong Kong Live Births by Year	21
1.7	Intake for All Programs at Local Universities by Year	23
1.8	Hong Kong Students Going Abroad Annually by Year	24
1.9	Sample Time-line	29
1.10	Effort Response Across Relative Test Score Distribution	38
1.11	Change in Math Score Across Relative Test Score Distribution	39
1.12	Change in Reading Score Across Relative Test Score Distribution	39
1.13	Change in Science Score Across Relative Test Score Distribution	40
1.14	Intake for Business and Engineering Programs by Year	57
1.15	Intake for Law and Medicine Programs by Year	58
1.16	Dragon Year Births by Month (Total)	67
1.17	Dragon Year Births by Month (Female and Male)	67
2.1	Histogram of Session Size	80
2.2	Histograms of Responses	86
2.3	Histograms of Time Spent	87
2.4	Histograms of Change in Time Spent	88
2.5	First Screen: Word-Number Pairs	109
2.6	Second Screen: Words with Text Boxes	110
2.7	Third Screen: End-of-Round Waiting	111
2.8	Non-parametric regressions	113

3.1	Wage Effects in Reference Dependent Preferences Model	139
3.2	Weather Effects in Reference Dependent Preferences Model	140

1 Step Up or Give Up: Student Response to Increased Competition from the Dragon Cohort

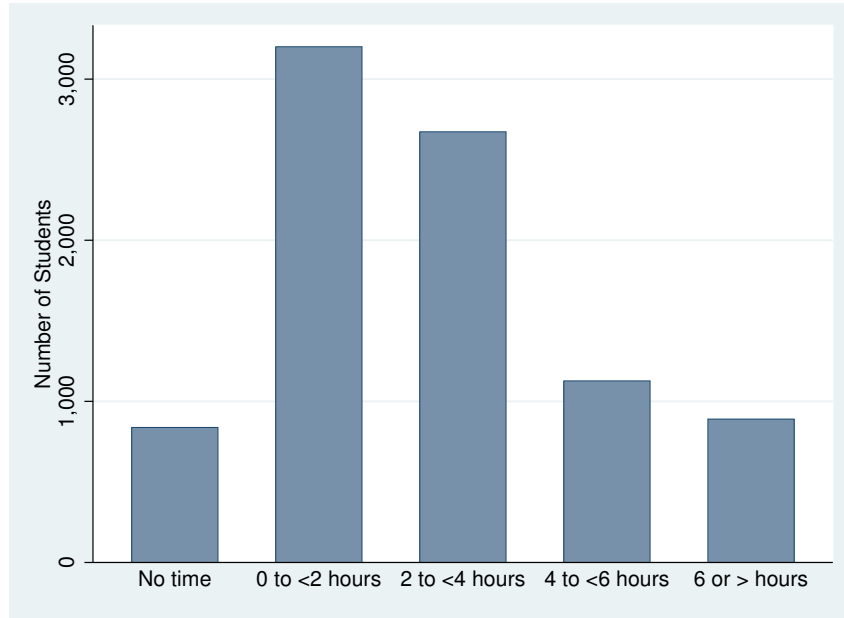
Abstract

Academic competition plays a potentially important role in any student's effort-exertion decision, which in turn affects academic achievement. A larger schooling cohort could alter competitive behavior especially if the rewards of competition (e.g., number of university places) do not increase concomitantly. I develop a tournament model of academic competition which predicts that, in response to an increase in cohort size, better students will be motivated to "step up" effort exertion, while less-able students will be discouraged and "give up."

I test these implications empirically by exploiting exogenous variation in cohort size arising from parental preference for babies born in the Chinese Zodiac year of the dragon. Using an identification strategy that avoids selection bias, I find that Hong Kong students do alter their effort in response to being part of the significantly larger "dragon cohort." Specifically, better students in the dragon cohort exert more effort (+0.5 hours/week studying math for top-quartile) relative to counterparts in other normal-sized cohorts, whereas less-able dragon cohort students only reduce effort insignificantly (-0.03 to -0.05 hours/week for bottom-quartile). Further results suggest that this higher effort translates into higher test scores. A falsification exercise fails to find such patterns in other countries. These findings are partially consistent with model predictions, and highlight the potential importance of considering cohort size and competition in policy decisions and other tournament structures.

I am very grateful to Alexandre Mas for his invaluable guidance on this project. Furthermore, I would like to thank Cecilia Rouse, Harvey Rosen, Henry Farber, Nicholas Lawson, and the seminar and workshop participants at the Industrial Relations Section and the Public Finance Working Group at Princeton University for helpful comments.

Figure 1.1: Hours Per Week Studying Math



Notes: Hong Kong students in grades 9 and 10.

Source: PISA 2003 and 2006 surveys.

1.1 Introduction

Student effort matters for educational outcomes.¹ But as Figure 1.1 shows for Hong Kong students, there is considerable variation in effort. Some students spend more than 6 hours per week studying math, while others spend no time at all. There are numerous reasons why some students are motivated to work harder than others. Engaged parents², good teachers³, financial incentives and peer competition are all potential factors.

Motivating students with financial incentives has been a main focus of the education economics literature. Some of these incentives include conditional payments linked to academic performance⁴, fulfilling certain requirements⁵, or attendance⁶. Studies typically find

¹See Bishop (2006) and OECD (2003).

²E.g., Wentzel (1997).

³E.g., Skinner and Belmont (1993) and Wentzel (1997).

⁴See Angrist et al. (2006); Angrist et al. (2009); Angrist and Lavy (2009); Fryer (2011); Kremer et al. (2009).

⁵See Bettinger (2010); Leuven et al. (2010).

⁶See Behrman et al. (2005).

that the effects of offering financial incentives on academic performance range from zero to significantly positive, though estimates vary within and across studies by subject, gender and other factors.

A less well studied and more inherent source of student motivation is effort induced by peer competition. Throughout their educational careers, students compete with one another for scarce rewards which only a fraction will receive; these include places in popular schools or universities, financial aid, scholarships, grades on a curve, and jobs. Competition between students arises naturally as they exert effort to outdo peers to obtain such rewards. Should the number of competitors for these rewards increase without a concomitant increase in the number of rewards (perhaps due to limited resources or stagnant policy), the competitive behavior among students may change as they fight for relatively fewer opportunities. Some students may be motivated to “step up” their effort in response to the augmented pool of competitors, to ensure that they receive the reward. Others may “give up” and decrease their effort exertion in the face of increased rivalry, discouraged by the fact that they are now even less likely to win the reward. This paper will focus on such responses to increased peer competition.

Academic competition of this sort can be best thought of as a rank-order tournament, a framework used to model the effort decisions of contestants in situations where prizes are awarded based on relative performance, regardless of absolute performance. Following the seminal work of Lazear and Rosen (1981), tournament models have been extended upon⁷ and applied empirically to a diverse array of situations⁸. A subset of this empirical literature investigates how the number of contestants can influence outcomes under tournament structures, in settings involving software programming contests⁹, retail sales contests¹⁰, within-firm promotions to top positions in personnel economics¹¹ and various experimental

⁷See Nalebuff and Stiglitz (1983) and Green and Stokey (1983), among others.

⁸For a survey, see Konrad (2009).

⁹See Boudreau et al. (2012).

¹⁰See Casas-Arce and Martinez-Jerez (2009).

¹¹See Main et al. (1993); Eriksson (1999).

setups¹². This study, which explores how Hong Kong students respond competitively to an increase in the schooling cohort size, contributes to this strand of empirical literature. I first construct a tournament model of academic competition to characterize the effort decisions of students with heterogeneous abilities. I then use a natural experiment from Hong Kong in which there is an exogenous increase in cohort size arising from superstitions related to birth timing, to estimate student responses in effort and changes in test scores.

Each year in the Chinese lunar calendar is represented by one of twelve animals, one of which is the dragon. The “dragon year” is considered by parents to be an especially auspicious year to give birth to a child. As such, many superstitious parents target childbirth to have a “dragon baby,” resulting in a significant spike in the number of births during this period. This increase in births translates into certain schooling cohorts being larger because these cohorts contain a majority of students who are dragon babies. I use this variation to compare outcomes of students in these enlarged “dragon cohorts” relative to their counterparts in normal-sized cohorts.

In a hypothetical experiment, experiencing the larger dragon cohort is the “treatment,” while being in a normal-sized cohort serves as the “control.” However, since birth timing is endogenously chosen by (superstitious) parents, selection bias may arise. I avoid this selection bias by applying an identification strategy which exploits the misalignment of the Chinese lunar calendar’s new year and the school cohort assignment policy boundary. As such, only students assigned into the dragon cohort who are not in fact dragon babies are used in the analysis.

The empirical findings suggest that the effort response of students to being in a larger cohort depends on their relative position in the ability distribution. Effort response is positive for the top students, with those in the highest quartile of the distribution increasing their

¹²See Orrison et al. (1997); Harbring and Irlenbusch (2003); Garcia and Tor (2009). While the motivation of Garcia and Tor (2009) stems from the social-comparison process in the psychology literature, their results are nevertheless consistent with tournament models. Their main experiment finds that participants exert more effort in finishing an easy test faster if they perceive to be competing in a pool of 10 rather than 100 people.

studying time of math per week by about 0.5 hours. This is in contrast to the negligible negative response of less-able students, with those in the lowest quartile of the distribution decreasing their studying time by about 0.03 to 0.05 hours. These results are partially consistent with the tournament model of academic competition presented, in that students at the top will step up their effort, while students at the bottom do not seem to exert less effort. Similar results are observed for changes in test scores in response to being in a larger cohort. A falsification exercise using data from other countries suggests that this pattern is not present in other countries.

These results highlight the potential importance of considering cohort size and competition when making policy decisions, in particular those which affect (or because of efficiency considerations, should be influenced by) cohort size and the relative supply of rewards. Cohort size variation arises naturally from gradual demographic shifts, and much research regarding this topic has been devoted to post-war baby boomers in the United States.¹³ Once in a while though, major policy shifts can induce shocks to cohort sizes. In 2003, Ontario, Canada’s elimination of 13th grade resulted in a “double cohort” of students entering university that year.¹⁴ A similar double cohort occurs in Hong Kong due to its transition from a 3- to 4-year university program in 2012.¹⁵ One famous historical example is the legalization (1957) and subsequent re-criminalization (1966) of abortion in Romania, which caused sudden fluctuations in the birth rate.¹⁶ These sudden cohort size changes are likely to have an impact on the competitive behavior, and hence the outcomes, of the cohort members in question. Furthermore, many policies that deal with reward allocation create tournament structures; examples include policies awarding scholarships, or Texas’ guarantee of admission to state-funded universities for the top 10% of any graduating high school class. If new policy decisions change the existing parameters in these tournament structures (e.g. relative

¹³See Welch (1979); Stapleton and Young (1984); Stapleton and Young (1988); and Bound and Turner (2007).

¹⁴See Morin (forthcoming), who estimates its effect on earnings, for a description.

¹⁵The timing of this change will not affect this paper’s analysis.

¹⁶See Berelson (1979).

reduction in number of rewards), then the competitive behavior of students will be affected.

The remainder of this chapter proceeds as follows. In Section 1.2, I construct a tournament model of academic competition to describe the student responses in effort choice and test score outcomes to an increase in cohort size. Next, in Section 1.3, I present background information on the dragon cohort's relative size and university attendance. Then, in Section 1.4, I describe the empirical setup (in terms of the data and identification strategy) which I will use to estimate the effect of being in the larger dragon cohort, the results of which are presented in Section 1.5. Section 1.6 concludes.

1.2 Tournament Model of Academic Competition

The following tournament framework models students competing for a limited number of rewards: places at universities. Students have heterogeneous ability draws, and derive utility from doing well in a test, as well as from attending university. Students choose effort to maximize utility, but effort exertion is costly. Effort and ability are inputs in the test score production function, which also contains a random error component. Admission into universities is limited to the top-performing fraction of students, whose test scores pass a certain relative score cutoff, regardless of absolute performance in the test.

After establishing the framework, I briefly consider possible extensions to the model. I then present a numerical illustration of the model using reasonable functional forms and parameter values. Additional applications to previous empirical literature will also be discussed.

1.2.1 Theoretical Framework

Let the cohort size be a measure of n students. Let θ_i denote the ability of student $i \in [0, n]$, drawn from the ability distribution $F(\theta_i)$. Let e_i denote the effort choice of student i . Suppose only $(1 - p)n$ students are admitted into university. (In other words, p is the proportion of students not going to university.) A student is admitted only if his/her test

score is above a certain cutoff score $\bar{S}_p$ at the $(100p)$ th percentile. $\bar{S}_p$ will depend on p through the aggregation of scores of all students. Suppose if admitted, a student always attends university.

The student's realized test score is described by the function

$$S(\theta_i, e_i) - \eta_i$$

which contains the production function component $S(\theta_i, e_i)$ dependent on the student's ability and effort choice as inputs, and the random error component η_i which is i.i.d. and orthogonal to effort, ability, and cohort size.¹⁷ This test score is realized only after effort decisions are made. A student derives utility from scoring well on the test and attending university.¹⁸ The student's utility function is given by

$$u(e_i) = \begin{cases} \phi [S(\theta_i, e_i) - \eta_i] + \alpha - c(e_i) & \text{if student attends university} \\ \phi [S(\theta_i, e_i) - \eta_i] - c(e_i) & \text{otherwise} \end{cases}$$

where $\alpha > 0$ is the utility derived from attending university and $\phi > 0$ is the rate at which the test score directly generates utility. Effort must be exerted at a cost of $c(e_i)$ prior to test taking and university attendance. The following assumptions are made.

Assumption A. Test score production is increasing in effort and non-decreasing in ability.

That is,

$$\frac{\partial S(\theta_i, e_i)}{\partial e_i} > 0 \text{ and } \frac{\partial S(\theta_i, e_i)}{\partial \theta_i} \geq 0$$

However, test score returns on effort are non-increasing in effort and increasing in

¹⁷Such an error term may arise due to the student being ill on test day, the student making mistakes in answering questions, the student guessing correctly for a question he/she does not know, the grader committing a grading error in favor of or against the student, etc.

¹⁸A test score component is included in the utility function to make the solutions tractable. If it is excluded, then the solution degenerates into a bang-bang solution, for which there may or may not be a separating equilibrium.

ability¹⁹. That is,

$$\frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} \leq 0 \text{ and } \frac{\partial^2 S(\theta_i, e_i)}{\partial \theta_i \partial e_i} > 0$$

Assumption B. Effort costs are increasing and strictly convex. That is,

$$c'(e_i) > 0 \text{ and } c''(e_i) > 0$$

These are reasonable assumptions, considering that common functional forms of $S(\theta_i, e_i)$ (e.g., Cobb-Douglas) and $c(e_i)$ (e.g., quadratic costs) satisfy them.

The student's expected utility conditional on ability and effort choice is

$$\begin{aligned} E[u(e_i) | e_i, \theta_i] &= \phi S(\theta_i, e_i) - \phi E[\eta_i] + \alpha \Pr(\text{Attends University}) - c(e_i) \\ &= \phi S(\theta_i, e_i) - \phi E[\eta_i] + \alpha \Pr(S(\theta_i, e_i) - \eta_i \geq \bar{S}_p) - c(e_i) \\ &= \phi S(\theta_i, e_i) - \phi E[\eta_i] + \alpha H(S(\theta_i, e_i) - \bar{S}_p) - c(e_i) \end{aligned}$$

where $H(\eta_i)$ is the c.d.f. of η_i (and $h(\cdot)$ is the corresponding p.d.f.).

Assumption η . Let the distribution of η_i be single peaked. Let this peak be (without loss of generality) at zero.

In choosing effort to maximize expected utility, the student's first order condition (FOC) is

$$\left[\phi + \alpha h(S(\theta_i, e_i) - \bar{S}_p) \right] \frac{\partial S(\theta_i, e_i)}{\partial e_i} - c'(e_i) = 0 \quad (1.1)$$

This equation defines student i 's optimal effort decision e_i^* as a function of ability θ_i .²⁰ In par-

¹⁹This would be so if higher ability students make better use of their effort exertion.

²⁰In deriving this FOC, I assume that $\frac{d\bar{S}_p}{de_i}$ is small; that is, any particular student's effort decision will not move the cutoff. The probability that any single student's effort decision is pivotal is small, given the continuum of students on the ability distribution. This arises from the uncertainty in the final realized score from the error term η . Even if a student is at

$$E[S(\theta_i, e_i) - \eta_i] = \bar{S}_p$$

this is only in expectation, so the student cannot know for sure that he/she will be pivotal.

ticular, the closer a particular student is to the cutoff score (i.e., the higher $h(S(\theta_i, e_i) - \bar{S}_p)$ is), the more effort he/she will exert in order to overcome the potential of an adverse error, get above the cutoff, and receive the additional α utility. This generates a “bump” in effort near the cutoff.

The second order condition (SOC) is given by

$$\alpha h'(S(\theta_i, e_i) - \bar{S}_p) \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 + [\phi + \alpha h(S(\theta_i, e_i) - \bar{S}_p)] \frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} - c''(e_i) < 0 \quad (1.2)$$

This is negative as I will only consider interior solutions with $e_i^* > 0$.²¹

Proposition 1.1. *Under Assumptions A and B, test score production $S(\theta_i, e_i)$ (taking into account equilibrium effort choice) is strictly increasing in ability. Hence, expected realized test scores are strictly increasing in ability. That is, the total derivatives*

$$\frac{dS(\theta_i, e_i)}{d\theta_i} > 0 \text{ and } \frac{dE[S(\theta_i, e_i) - \eta_i]}{d\theta_i} > 0$$

Proof. See Appendix 1.7. □

The intuition behind this proposition is that although high ability students may decrease their effort because their higher ability can compensate for lower effort, they will never do so to the extent that it will decrease their test score production or realized test score in expectation. Doing so would not only reduce expected utility directly (through the ϕ term), but also hurt their chances of getting into university, thereby reducing expected utility indirectly.

One important implication of Proposition 1.1 is that if a test were designed such that its underlying test score production function satisfies Assumption A, then scores from this test

²¹In the case where the only solution to the FOC has a positive SOC (i.e., the FOC solution is utility-minimizing), then the corner solution $e_i^* = 0$ is utility-maximizing.

will be on average a good measure of ability. This is true whether the incentives to performing well in the test enter the utility function directly through caring about the test score (the ϕ term), or indirectly through some relative prize (the α term), or some combination of both.²²

A second implication of Proposition 1.1 is that the equilibrium distribution of expected realized test scores will be a monotonic transformation of the ability distribution $F(\theta_i)$. To see this, note that the realized test score is given by $S(\theta_i, e_i) - \eta_i$. A student with ability θ_i (choosing equilibrium effort decision e_i) will have a realized test score which comprises a deterministic test score production component $S(\theta_i, e_i)$, and the random error component η_i . First, note that the distribution of $S(\theta_i, e_i)$ will be given by

$$F_S(S_i) = F \circ S^{-1}(S_i)$$

for some test score production choice S_i , where S^{-1} is the inverse²³ of $S(\theta_i, e_i)$ with respect to θ_i , conditional on equilibrium effort. Therefore, the distribution of realized test scores will be given by

$$F_E(x) = \int F \circ S^{-1}(x + \eta_j) dH(\eta_j)$$

for some expected realized test score x .²⁴ Thus, realized test scores can be thought of as draws

²²For the case where students have no incentive ($\alpha = \phi = 0$), the equilibrium is for all students to exert zero effort. In that case, whether the test score is a good measure of ability will depend on the test score production function at $e_i = 0$.

²³From Proposition 1.1, $S(\theta_i, e_i)$ is a monotonic bijection between test scores and ability, and hence invertible.

²⁴The equation is derived as follows. $F_E(x)$ is a convolution of two distributions: the distribution of $S(\theta_i, e_i) \sim F_S(\cdot)$, and the distribution of $\eta_i \sim H(\eta_i)$. Let S and η denote the random variables representing these distributions. For some value of realized test score x ,

$$\begin{aligned} F_E(x) &= \Pr(S - \eta \leq x) \\ &= \int \Pr(S - \eta \leq x \mid \eta = \eta_j) dH(\eta_j) \\ &= \int \Pr(S \leq x + \eta_j) dH(\eta_j) \\ &= \int F_S(x + \eta_j) dH(\eta_j) \\ &= \int F \circ S^{-1}(x + \eta_j) dH(\eta_j) \end{aligned}$$

from the distribution $F_E(.)$ in the same way that abilities are drawn from the distribution $F(\theta_i)$.²⁵

I make the following assumption with regard to the ability composition of the enlarged cohort.

Assumption C. Regardless of their size, cohorts consist of students drawn from the same ability distribution. In other words, the distribution $F(\theta_i)$ does not change with changes in cohort size (n or p).

I make this simplifying assumption so as to enable a direct (*ceteris paribus*) comparison between cohorts of different sizes in the model, abstracting away from distributional differences in ability.

Consider the effects of an increase in the cohort size n . Depending on how universities respond to this cohort size increase, the intake of students, and therefore p , may also change. This, in turn, may move the cutoff $\bar{S}_p$. In response, students will adjust their effort decisions accordingly. This process can be expressed as

$$\frac{de_i}{dn} = \frac{de_i}{dp} \times \frac{dp}{dn} \quad (1.3)$$

As cohort size n increases, universities may offer relatively fewer places ($\frac{dp}{dn} > 0$, since p is proportion *not* going to university), relatively more places ($\frac{dp}{dn} < 0$), or the same relative proportion of places ($\frac{dp}{dn} = 0$). Suppose university intake does not increase enough to fully accommodate for a larger cohort. Such a regime can be stated as the following assumption.

Assumption D. Let $\frac{dp}{dn} > 0$.

With these two additional assumptions, the following proposition can be derived.

²⁵This is also the reason why the expected test score cutoff $\bar{S}_p$ is known to all students. Since the ordering of the distributions $F_S(.)$ and $F(.)$ are maintained, and the full distribution of $F(.)$ is known, as well as the fact that there are a large number of students, all students know that it is the ability level at the $(100p)$ th percentile which will generate the test score cutoff in expectation.

Proposition 1.2. *Under Assumptions A, B, η , C, and D, a student will exert more effort in response to a cohort size increase if and only if his/her relative ability in the ability distribution $F(\cdot)$ is above a certain threshold. In particular, for student i with relative ability $a_i = F(\theta_i)$,*

$$\frac{de_i}{dn} > 0 \text{ iff } a_i > F \circ S^{-1} \left(S \circ F^{-1}(p) - E(\eta_i) \right)$$

For error distributions with mean zero, the condition becomes

$$\frac{de_i}{dn} > 0 \text{ iff } a_i > p$$

Proof. See Appendix 1.7. □

The intuition behind this result is that when cohort size increases without a concomitant increase in university intake, the cutoff for acceptance into university becomes higher. Students above the threshold exert more effort because the higher acceptance cutoff is now closer to them, increasing their risk of falling below the cutoff due to the error term; in response, they “step up” effort to compete for the relatively fewer spots. On the other hand, students below the threshold exert less effort because the higher acceptance cutoff is now further away, decreasing the probability of them ever surpassing it; in response, they “give up” competing for the relatively fewer spots as effort exertion is costly.

From the proof,

$$\frac{de_i}{dn} = - \frac{-\alpha h' \left(S(\theta_i, e_i) - \bar{S}_p \right) \frac{\partial \bar{S}_p}{\partial p} \frac{\partial S(\theta_i, e_i)}{\partial e_i} \frac{dp}{dn}}{\alpha h'(\cdot) \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 + [\phi + \alpha h(\cdot)] \frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} - c''(e_i)} \quad (1.4)$$

This equation allows us to further describe the comparative statics of a cohort size change.

Firstly, as $h' \left(S(\theta_i, e_i) - \bar{S}_p \right)$ decreases in magnitude, the response $\frac{de_i}{dn}$ becomes smaller in magnitude. This will occur as you move away from the peak of the $h(\cdot)$ distribution, where the tails flatten out. Hence, students who are progressively further away from the cutoff

(higher $(S(\theta_i, e_i) - \bar{S}_p)$) will attenuate their effort response to a cohort size increase. The intuition behind this is that though the cutoff shifts, these students are so far away from it that their chances of getting accepted into university (or getting rejected) are unaffected by the cutoff shift.

Secondly, as α increases, the response $\frac{de_i}{dn}$ becomes larger in magnitude. The intuition is that the larger the utility benefit from passing the cutoff, the more pronounced the “bump” around the cutoff will be, as students have more to lose from falling below the cutoff, and therefore, any shift in the cutoff will induce larger effort responses.

Thirdly, as $c''(e_i)$ increases, the response $\frac{de_i}{dn}$ becomes smaller in magnitude. The intuition is that if the marginal cost of effort exertion increases quickly, then all students will exert less effort and the “bump” around the cutoff will be smaller. Again, any shift in the cutoff will then induce smaller effort responses.

Proposition 1.2 can be extended to describe consequences for realized test scores (in expectation) by noting that

$$\frac{dE[S(\theta_i, e_i) - \eta_i]}{dn} = \frac{dS(\theta_i, e_i)}{dn} = \frac{\partial S(\theta_i, e_i)}{\partial e_i} \frac{de_i}{dn} \quad (1.5)$$

resulting in the following similar proposition.

Proposition 1.3. *Under Assumptions A, B, η , C, and D, a student’s realized test score (in expectation) will increase in response to a cohort size increase if and only if his/her relative ability in the ability distribution $F(\cdot)$ is above a certain threshold. In particular, for student i with relative ability $a_i = F(\theta_i)$,*

$$\frac{dE[S(\theta_i, e_i) - \eta_i]}{dn} > 0 \text{ iff } a_i > F \circ S^{-1}(S \circ F^{-1}(p) - E(\eta_i))$$

Proof. This follows from applying the result in Proposition 1.2 to equation (1.5), and noting that under Assumption A, $\frac{\partial S(\theta_i, e_i)}{\partial e_i} > 0$. □

Most importantly, these two propositions state that whether a student's effort or test score responses (to a cohort size increase) are positive or negative depend on the student's relative position in the ability distribution, and not the actual ability level.

The results stated in Propositions 1.2 and 1.3 offer two testable implications which will be empirically considered in subsequent sections. The reasonableness of Assumption D will also be considered.

1.2.2 Extensions

There are many possible substantive extensions to the model presented.

Firstly, the entire framework can be revised to describe a penalty rather than a reward. For instance, if the latent score were something undesirable, then a penalty $\alpha < 0$ may be imposed for surpassing the relative cutoff. On the other hand, in situations where the latent score is desirable, the utility function can be rewritten so as to impose the penalty for falling below the relative cutoff. An example of this is a downsizing company which must fire a specific number of its lowest-performing salespeople.

Secondly, finding the optimal reward mechanism to put in place is a natural next step of the model. This can inform policymakers as to the most efficient way to choose the size of the reward and proportion of prize winners, conditional on the number of competitors. Given a social welfare function, a policy-maker can choose reward size α (translated into util units), and the number of rewards $(1 - p)n$, through choosing p , possibly subject to a budget constraint involving α , p and other variables. This question has been examined extensively in the tournament literature.

Thirdly, one can think about incorporating multiple score cutoffs as opposed to just one, with higher cutoffs offering successively larger rewards. The two cutoff case can be expressed

cumulatively with the new utility function

$$u(e_i) = \begin{cases} \phi[S(\theta_i, e_i) - \eta_i] + \alpha_1 + \alpha_2 - c(e_i) & \text{if both cutoffs are passed} \\ \phi[S(\theta_i, e_i) - \eta_i] + \alpha_1 - c(e_i) & \text{if only lower cutoff is passed} \\ \phi[S(\theta_i, e_i) - \eta_i] - c(e_i) & \text{otherwise} \end{cases}$$

where α_1 and α_2 are strictly positive. Multiple “bumps” in effort will be generated around these multiple cutoffs.

Finally, the case where the error term is not orthogonal to ability, effort, and/or cohort size can be explored, though the algebra for such an endeavor quickly becomes much more complicated.

1.2.3 Numerical Illustration

This subsection presents a numerical illustration by substituting exact functional forms and parameter values as described in Table 1.1 into the model. Doing so allows for the comparison between two sets of calculations representing results for before and after a cohort size increase. This illustration is meant to develop intuition for the model.²⁶

The functional forms of the test score production function and cost of effort function satisfy Assumptions A and B respectively. The cohort size measure increases from 8 to 10 while the university intake is kept constant at 4, satisfying Assumption D. Computations make use of ability levels before and after this cohort size increase as if drawn from the same uniform distribution, satisfying Assumption C.

The following figures illustrate the results for this set of substitutions. To gain insight into the model and the result of Proposition 1.2, I plot effort and effort change against percentile of ability on the x-axis. Even though the x-axis is free of the error component, the outcomes on the y-axis are nonetheless chosen (computed) by the student taking the error

²⁶Moreover, in the process, I confirm that the comparative statics of taking derivatives with respect to n are similar to increasing n from one value to another, and then comparing the results at the two points.

Table 1.1: Substitutions for Numerical Illustration

Function/Parameter	Notation	Substitution
Score Prod. Function	$S(\theta_i, e_i)$	$\theta_i^{0.5} e_i^{0.5}$
Cost of Effort	$c(e_i)$	e_i^2
Reward	α	0.03
Utility from Score	ϕ	1.00
Ability Distribution	$F(\theta_i)$	$U(1, 2)$
Error Distribution	$H(\eta_i)$	$N(0, 0.08)$
Cohort Size	n	From 8 to 10
University Intake	$n(1 - p)$	4 (constant)
Prop. Below Cutoff	p	From 0.5 to 0.6

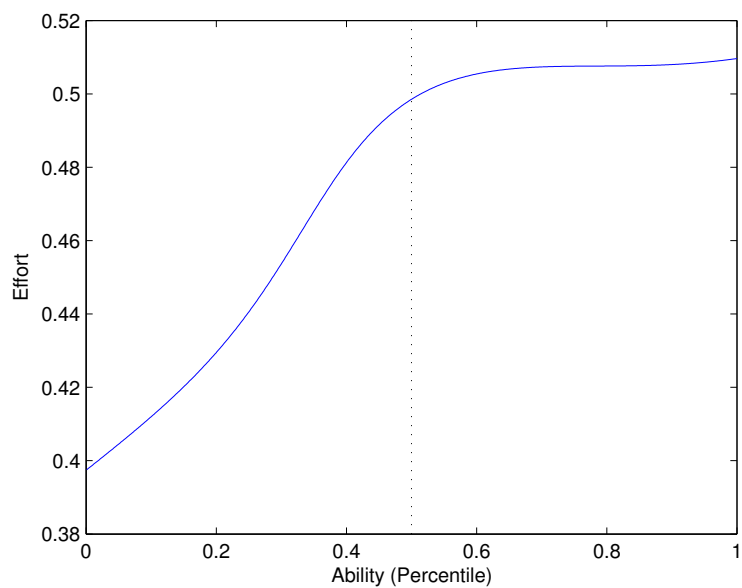
distribution into account.

Figure 1.2 plots effort choice by ability percentile prior to the cohort size change (i.e. $n = 8$ and $p = 0.5$). Overall, effort is increasing in ability percentile. This is to be expected since higher ability students exert more effort to obtain a higher test score (and hence a higher utility level), by taking advantage of the test score production function's returns to effort which are increasing in ability. However, around the score cutoff percentile $p = 0.5$ (indicated by the vertical line), there is an additional increase in effort exertion. The reason for this “bump” is that students near this cutoff have a realistic chance of affecting whether they pass the cutoff with their effort decision, thereby gaining (or potentially losing) the additional utility α for passing (or falling below) it.²⁷ Conversely, those far below or above the cutoff have no realistic chance of getting above or falling below the cutoff (respectively) through changing their effort decision. Thus, the bump in effort attenuates further away from the cutoff.

Suppose now that the cohort size increases and university intake remains unchanged (i.e. $n = 10$ and $p = 0.6$). Figure 1.3 plots the previous graph (in Figure 1.2) before the cohort size increase, and superimposes an analogous graph for after the cohort size increase. The cutoff, and hence the bump in effort, has now shifted to the right, as the proportion of

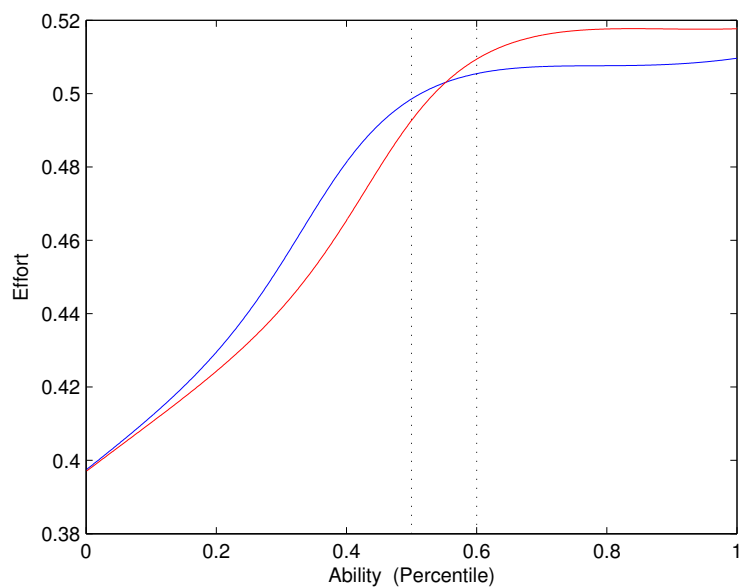
²⁷This “chance” corresponds to the term $\alpha h(.)$ in the FOC in equation (1.1).

Figure 1.2: Effort by Ability Percentile



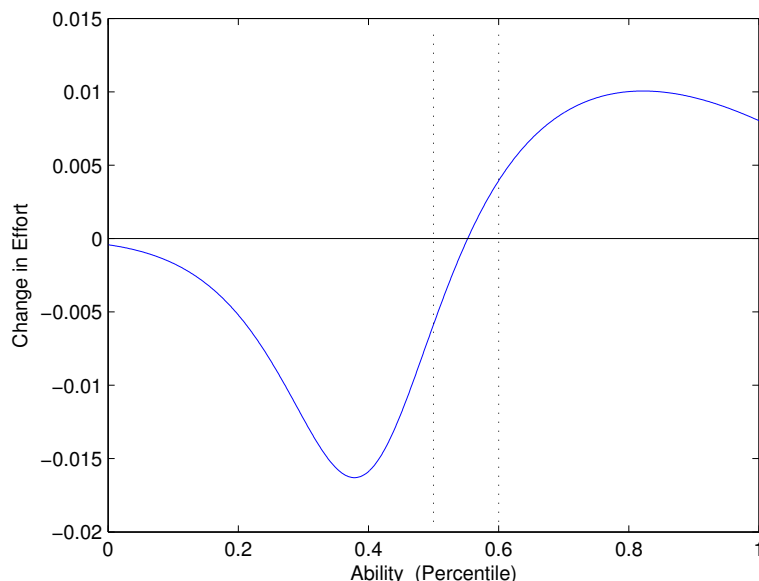
Note: The vertical line represents the cutoff. Refer to Table 1.1 for the substitutions made for this numerical illustration.

Figure 1.3: Effort Before and After by Ability Percentile



Note: The vertical lines represent the cutoffs before and after. Refer to Table 1.1 for the substitutions made for this numerical illustration.

Figure 1.4: Change in Effort by Ability Percentile



Note: The vertical lines represent the cutoffs before and after. Refer to Table 1.1 for the substitutions made for this numerical illustration.

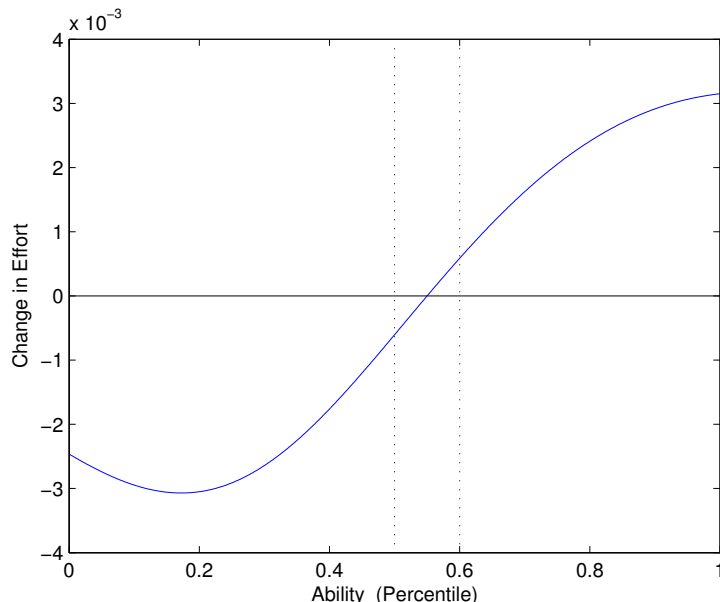
students not entering university (p) increases.

Figure 1.4 plots the change in effort before and after the cohort size increase, the net difference between the two graphs in Figure 1.3. It shows that there is a threshold in ability percentile (between the two cutoffs), above which effort responds positively, and below which effort responds negatively, to a cohort size increase (Proposition 1.2).²⁸ Students above the threshold “step up” their effort as the cutoff is now closer to them, increasing their risk of falling below it. Students below the threshold “give up” on exerting effort as the cutoff is now further away from them, decreasing their chances of ever surpassing it. For ability percentiles progressively further away from this threshold, the magnitude of the response diminishes as these students are too far from the cutoff to be affected. A similar graph for change in test score by ability percentile can be produced (Proposition 1.3).

The shape of this change in effort plot will depend on the parameters, especially the

²⁸The threshold is not exactly at p because this illustration is a discrete change from one p to another, rather than a small derivative change as in the earlier analytical results.

Figure 1.5: Change in Effort by Ability Percentile (Errors with Larger Spread)



Note: The vertical lines represent the cutoffs before and after. The error term distribution is now changed to $\eta \sim N(0, 0.16)$. Refer to Table 1.1 for other substitutions made for this numerical illustration.

spread of the error term distribution. For an error term distribution with a higher standard deviation, the bump in the effort plot will be more spread out because students further away from the cutoff could now potentially be affected by the error term. Figure 1.5 shows a similar numerical illustration of the change in effort, but with an error term distribution with double the original standard deviation ($\eta \sim N(0, 0.16)$). Notice that the plot is “stretched horizontally” because of the wider error term distribution.

1.2.4 Additional Applications

This model and its comparative statics can be applied to other tournament situations in which the number of competitors is subject to change.²⁹ Boudreau et al. (2012) and Casas-Arce and Martinez-Jerez (2009) offer two examples of empirical results consistent with equation (1.4).

²⁹For example, while many sporting tournaments give out a fixed number of prizes every year (e.g. three medals), the number of athletes or teams qualifying to compete may vary.

Furthermore, this particular setup of using a production function in effort and ability is applicable to situations in which the latent score that determines winning is a function of both effort and a heterogeneous unobservable. Many of the rewards structures in the financial incentives literature rely on test scores to allocate prizes, some of which may be of a tournament format, where (for example) only students who place in the top 10% of their grade in school receive rewards. Leuven et al. (2010) find empirical results which mirror this model’s predictions with regard to the introduction of academic performance rewards for first-year university students. A real world example of the tournament model in action is Texas’s guarantee of admission to state-funded universities to any student who is in the top 10% of his/her high school’s graduating class. Aside from education contexts, this form of the production function is also applicable in settings involving tournaments based on output production within firms.

1.3 Background on the Dragon Cohort

To test the predictions of my model, I make use of the natural experiment of the “dragon cohort.” The Chinese Zodiac consists of twelve animal signs³⁰, each assigned to a particular year in the Chinese lunar calendar within a twelve year cycle. In particular, children born in the “dragon year” are deemed to be especially auspicious, resulting in a significant spike in the number of “dragon babies” being born during this period.³¹ Figure 1.6 shows the spikes for two such dragon years in Hong Kong.³²

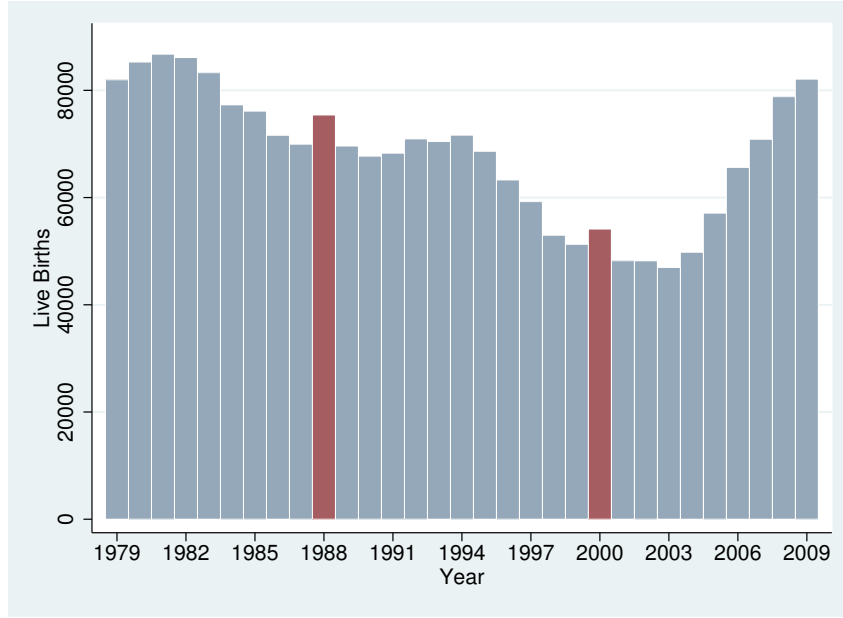
This increase in births translates into certain “dragon cohorts” being larger because they contain a majority of students who are dragon babies. Since subsequent analysis is identified off of the dragon cohort of students born in 1988, the analysis to follow will focus on this

³⁰These are (in order of the cycle) rat, ox, tiger, rabbit, dragon, snake, horse, goat, monkey, rooster, dog and pig.

³¹Other papers which have looked into this phenomenon and used it in their identification include Vere (2008) and Yip et al. (2002). Another instance in which superstitions have been used for identification is Rohlfs et al. (2010).

³²This phenomenon has also been documented in other countries with appreciable ethnic Chinese populations, including Singapore and Taiwan; see Yip et al. (2002).

Figure 1.6: Hong Kong Live Births by Year



Note: Years are Gregorian calendar years, not Chinese lunar calendar years. The highlighted years are those in which the majority of births fall in the Chinese lunar calendar dragon year.

Source: Annual Digests of Statistics, Hong Kong Census and Statistics Department

cohort.

To get a sense of how the size of the 1988 dragon cohort compares to those of other cohorts throughout their entire schooling career, I use enrollment data for all grades over all years after 1991 to run the ordinary least squares (OLS) regression

$$enrollment_{gy} = \beta_0 + \beta \text{dragoncohort}_{gy} + \mu_g + \mu_y + \varepsilon_{gy} \quad (1.6)$$

where

- $enrollment_{gy}$ is the aggregate Hong Kong enrollment for grade g in year y
- $dragoncohort_{gy}$ is an indicator variable which is 1 when the students of grade g in year y are the 1988 dragon cohort, and

Table 1.2: Enrollment of Dragon Cohort

Dep. Var.:	(1)	(2)
Enrollment	Levels	Logs
Dragon Cohort	2954*** (927.3)	0.048*** (0.017)
Constant	71033*** (2485)	11.13*** (0.049)
N	234	234
R^2	0.926	0.949
Grade FEs	Yes	Yes
Year FEs	Yes	Yes

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

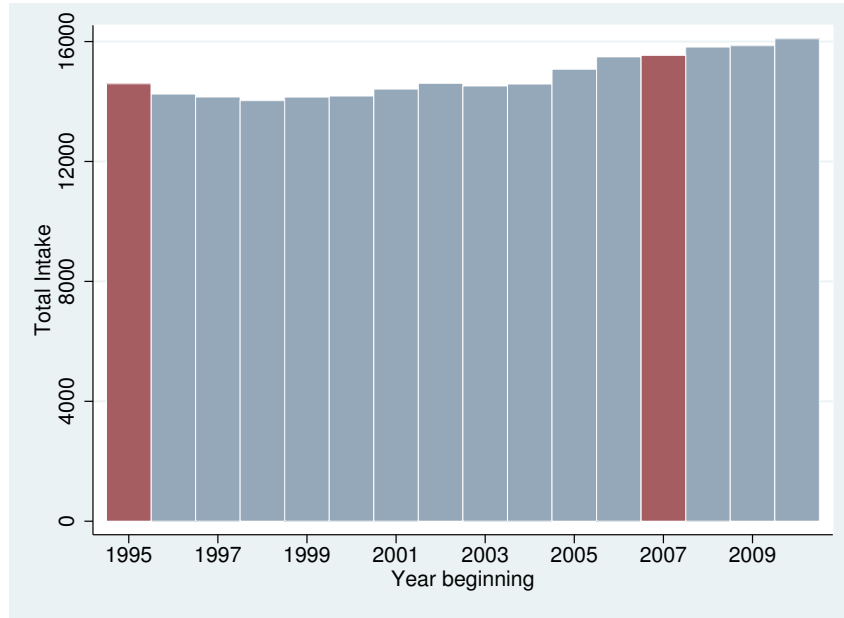
Note: Robust standard errors are in parentheses.

- μ_g and μ_y are grade and year fixed effects respectively

Table 1.2 shows the coefficient estimates for regression (1.6). Column (1) shows the results for enrollment in levels; column (2) shows the results for enrollment in logarithms. On average, the dragon cohort is 2954 students larger than other cohorts every year. In percentage terms, the dragon cohort is on average 4.8% larger every year. All coefficient estimates are significant at a 1% level. These estimates strongly suggest that students in the dragon cohort in question did in fact experience the “treatment” of being in a larger cohort throughout their schooling career.

In spite of the larger cohort size, university intake (both locally and abroad) did not seem to change in response. Figure 1.7 shows the total intake of first-year university students by year for all local universities. For the dragon cohort born in 1988, the intake year of interest is the year beginning 2007. From the figure, there seems to be a slow upward trend in local university intake over time. However, there does not appear to be any discernible difference between intake for the 1988 dragon cohort and that of surrounding years. The year-to-year changes are very small relative to the enlarged size of the dragon cohort. There also does not seem to be much substitution across years, suggesting that students are not delaying or

Figure 1.7: Intake for All Programs at Local Universities by Year



Note: Intake figures are for all undergraduate programs across all local universities. Shaded bars represent dragon years.

bringing forward university matriculation.³³

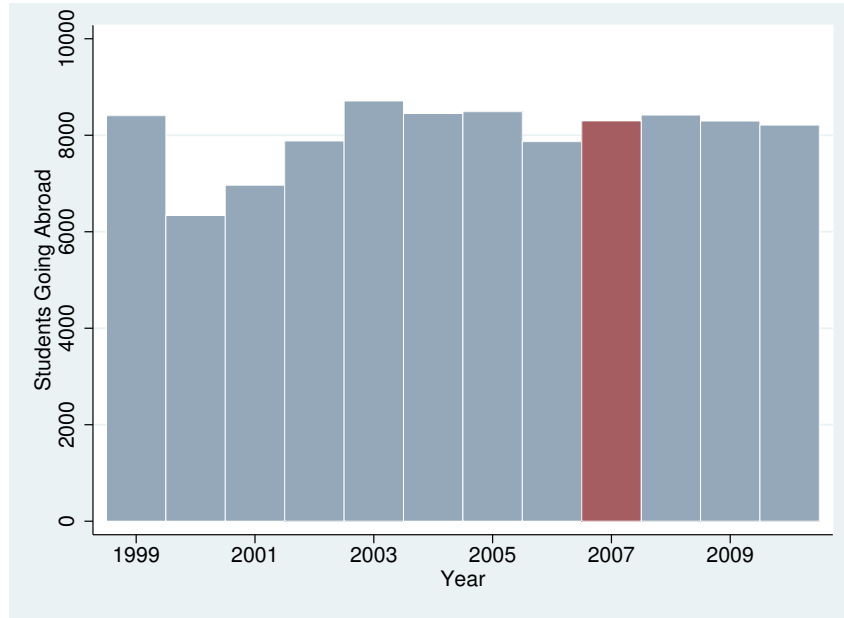
One concern is that even though local universities do not increase intake, Hong Kong students may substitute to universities overseas. If the supply of university spots abroad is elastic to the increase in local cohort size, then Assumption D may not hold. Using data from the World Bank’s Education Statistics series, Figure 1.8 shows the number of Hong Kong students going abroad annually by year.³⁴ Although there is a significant drop in 2000, the number of students seems to be fairly stable after 2003, the time period of interest. Again, the changes are small relative to the dragon cohort size increase.

Taken together, these pieces of descriptive evidence suggest that the overall number of Hong Kong students entering tertiary education, be it at local universities or institutions abroad, is fairly stable despite cohort size changes, implying that Assumption D does seem

³³For intake figures at local universities broken down by program of study, see Appendix 1.10.

³⁴I use the measure “Outbound mobile students: Tertiary” in the data, which is a count of “students from a given country studying abroad” for tertiary education. Assuming that all such students study in four year programs, I can divide this statistic by 4 to obtain the number of students going abroad annually.

Figure 1.8: Hong Kong Students Going Abroad Annually by Year



Note: These numbers are for students going abroad for tertiary education only.

Source: World Bank Education Statistics series.

reasonable.

1.4 Empirical Setup

1.4.1 Data

Though this paper uses data from several sources, the majority of analysis is carried out using data from the Programme for International Student Assessment (PISA). Administered by the Organisation for Economic Cooperation and Development (OECD), the PISA comprises a battery of tests (in math, reading, and science) and surveys conducted every three years in both OECD and select non-OECD countries, one of the latter being Hong Kong.

Each round of PISA consists of a cross section of randomly sampled 15-year-old students in different schools within a country, who complete the tests and surveys. School administrators and (in certain years) parents of students are also asked to complete additional surveys, from which school and household characteristics are derived. Data are available for

the rounds in 2000/2002³⁵, 2003, 2006, and 2009, but because of variation and inconsistencies in survey questions and variable definitions over the years, as well as other coinciding policy changes, the main analysis will only use data from 2003 and 2006.

The unit of observation is the student, whose grade level, birth year, and birth month are observed. For each observation, the outcome variables of interest are effort (hours per week studying math) and normalized test scores in math, reading, and science. Additional controls are also available at the student, parent, and school level. Student-level controls include sex, age, age in grade, and a foreign birth indicator. Parent-level controls include each parent's occupation status and highest education obtained between both parents. School-level controls include school type (public/private), school size, and student-teacher ratios (STR). The student's self-reported class size and expected educational attainment are also used in some of the secondary analyses. Full definitions of all variables listed above, as well as details of their construction, can be found in Appendix 1.8.

Summary statistics (by PISA round and grades) for the Hong Kong sample for a subset of these variables are shown in Table 1.3. (Only the statistics for the sample used for the main analysis, 2003 and 2006, are shown.) In terms of student- and school-level characteristics, the dragon cohort does not seem that different from other cohorts in the data. I investigate differences in cohort characteristics further in Appendix 1.9.

Additional data are obtained from various other sources. Hong Kong cohort sizes (aggregate enrollment by year and grade) and monthly birth figures are obtained from the the Hong Kong Census and Statistics Department³⁶, and annual first-year undergraduate university intake figures (by university and course of study) are obtained from the Hong Kong

³⁵This round of the PISA was peculiar in that it was the first round in which non-OECD countries participated. For OECD countries, the PISA was conducted in 2000 per usual. For non-OECD countries, it was conducted under the separate PISA-plus program between 2000 and 2002. For this particular round, PISA was conducted in Hong Kong in 2002; moreover, unlike other rounds which consist of Hong Kong students in grades 9 and 10, this round of PISA consists of Hong Kong students in grades 10 and 11. As such, the main analysis to follow will not use data from this round, though some secondary analysis, such as the teacher reallocation analysis, may rely on this data for certain inferences.

³⁶These figures were extracted from the Hong Kong Annual Digest of Statistics from 1988 to 2011, and the January and February editions of the Hong Kong Monthly Digest of Statistics between 1988 and 1994.

Table 1.3: Summary Statistics

PISA Round:	2003		2006		All years
Variable	Grade 9	Grade 10	Grade 9	Grade 10	All grades
Effort (hours)	2.85 (2.19)	3.15 (2.22)	1.96 (1.72)	2.42 (1.92)	2.65 (2.08)
Math Score	-0.25 (1.05)	0.17 (1.00)	-0.46 (0.94)	0.12 (0.94)	0.00 (1.00)
Reading Score	-0.47 (1.03)	-0.09 (0.93)	-0.19 (1.00)	0.26 (0.93)	-0.03 (0.99)
Science Score	-0.29 (1.04)	0.10 (0.97)	-0.39 (1.04)	0.14 (0.98)	-0.01 (1.02)
Sex (=Male)	0.51 (0.50)	0.48 (0.50)	0.51 (0.50)	0.48 (0.50)	0.49 (0.50)
Age (months)	187.4 (3.36)	190.7 (2.95)	187.7 (3.39)	190.2 (3.15)	189.6 (3.42)
HK Born	0.74 (0.44)	0.92 (0.27)	0.64 (0.48)	0.90 (0.3)	0.85 (0.36)
High School (either parent)	0.28 (0.45)	0.26 (0.44)	0.31 (0.46)	0.33 (0.47)	0.30 (0.46)
School Type (=private)	0.09 (0.29)	0.11 (0.31)	0.07 (0.26)	0.08 (0.28)	0.09 (0.29)
STR	18.13 (2.48)	18.28 (2.34)	18.03 (2.06)	18.23 (2.03)	18.2 (2.21)
<i>N</i>	1132	2692	1134	2978	7936

Note: Standard deviations in parentheses.

University Grants Committee³⁷. Data on student mobility are obtained from the World Bank Education Statistics series.³⁸

1.4.2 Identification Strategy

The analysis to follow uses exogenous variation in cohort size induced by the dragon cohort to identify changes in effort and test scores in response to an increase in cohort size. The “treatment effect” estimated is a cumulative peer effect on a student, in that it is the effect of experiencing the larger cohort of peers, and of following this larger group year after year as the student progresses through the grades. This effect is interpreted as the competitive behavioral response to a larger cohort size. Experiencing this larger cohort size is exogenous because birth timing cannot be controlled precisely, especially around the policy boundary date determining which cohort a student ends up in.³⁹

For the main analysis, I use a difference-in-difference estimation strategy. I use only students from the 2003 and 2006 PISA rounds, and in each round, only the students in grades 9 or 10. The larger dragon cohort comprises the students in grade 9 in the 2003 PISA round; this cohort is larger because the majority of its students are dragon babies born in the dragon year starting February 17th, 1988 (and ending February 5th, 1989). The other three comparison “cells” (grade 10 in 2003, grades 9 and 10 in 2006) are non-dragon cohorts and do not experience abnormal size increases. In using the difference-in-difference approach, I assume that one additional year of schooling for the students in grade 10 has the same effect across all PISA rounds, and that any within-year effects (from time trends or the way a PISA round is administered) affect the two grades in the same PISA round similarly.

A major issue encountered in papers where identification is based on superstitions is selection bias. Selection into the dragon year (not the dragon cohort) arises when parents

³⁷These are available for download in the statistics section of the Committee’s website at <http://www.ugc.edu.hk/eng/ugc/index.htm>.

³⁸This is available on the World Bank’s DataBank website at <http://databank.worldbank.org/>.

³⁹Leung et al. (2001) estimate that only 16.6% of all births in 1987 involved C-sections.

target childbirth past the Chinese lunar new year. If these parents are inherently different from parents who do not target, then selection bias will arise.⁴⁰ In order to avoid this bias, I exploit the misalignment of the school cohort assignment policy boundary date and the Chinese lunar new year.

In Hong Kong, cohort assignment is done according to whether a student is born before or after the boundary date December 31st (i.e. by Gregorian calendar year of birth).⁴¹ However, since the dragon year only begins on February 17th, 1988, students born between December 31st and February 17th are in the dragon cohort but not born in the dragon year. As such, my sample of analysis will only include dragon cohort students born in January and February, whose parents are presumably not inherently different.⁴² This fix furthers the interpretation of the results as a within-cohort peer effect, since identification hinges on non-dragon babies who happen to end up in the dragon cohort and experience its larger size due to their birth timing. Figure 1.9 depicts a time-line to illustrate sample coverage.

1.5 Empirical Results

This section presents the main empirical findings. I first present empirical evidence that being in the larger dragon cohort has an effect on effort and test scores, and that this effect varies depending on predicted score percentile⁴³, a proxy for relative position on the ability distribution (Propositions 1.2 and 1.3). I then perform a falsification exercise to show that this pattern is not observed outside Hong Kong by repeating a similar analysis with data

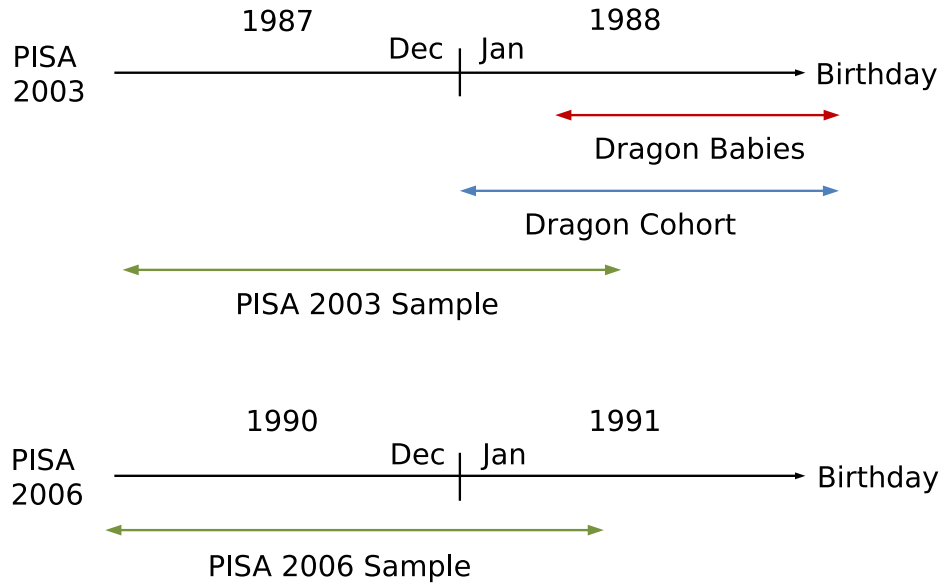
⁴⁰In Appendix 1.11, I provide evidence suggesting that dragon baby births are new births on the extensive margin rather than births shifted from other periods.

⁴¹Though compliance is imperfect because some parents may purposely enroll their children earlier or later than prescribed, these special cases consists of only 8.3% of observations across all PISA rounds, and their predetermined characteristics are very similar to those of their correctly assigned counterparts. However, this will not matter since I will control for age and age in grade.

⁴²I include all students born in February because only birth month is observed, not the exact date of birth, and excluding them reduces the sample size by so much that it becomes infeasible to carry out the analysis. I argue that this is reasonable because a majority of the days in this February were non-dragon baby birth-dates, and the inherently different parents who are intentionally targeting their childbirths seem to be “playing it safe” and not aiming for a date so close to the lunar new year cutoff, as Appendix 1.11 suggests. Other selection issues are also investigated further in this appendix.

⁴³Discussion to follow. See Appendix 1.8 for precise construction.

Figure 1.9: Sample Time-line



from other countries. The main findings are partially consistent with the predictions of my tournament model of academic competition, and suggest that they result from greater academic competition among the students in the larger dragon cohort. Lastly, I carry out additional analysis to show that reallocation of teachers across grades within schools occurred in response to the larger cohort size, addressing the issue of the role (or rather lack thereof) played by school resources.

1.5.1 Effort and Test Score Responses

According to Propositions 1.2 and 1.3, the exogenous increase in cohort size can have either positive or negative effects on effort and test scores, depending on a student's relative position on the ability distribution. Given Proposition 1.1, I could use test score percentiles as a proxy for ability. However, using test score percentiles as a right hand side variable may lead to confounding in that dependent variables and right hand side variables may be jointly determined. As such, I construct a proxy measure of relative ability which I call "predicted

score percentile” using observed student characteristics.⁴⁴

To document these effort and test score responses, I carry out three sets of analysis. The first set of regressions uses a quartiles specification where the effort and test score responses are estimated for each quartile of the predicted score distribution. The second set of regressions uses a polynomial specification where an interaction with a fifth-order polynomial in predicted score percentile enables me to plot the effort and test score responses at each point in the distribution. Finally, as a falsification exercise, I run a pooled quartile specification using data from a host of countries other than Hong Kong to estimate the nonexistent dragon cohort effect on students from these countries, to show that these results are specific to Hong Kong.

The quartiles specification estimates regressions of the form

$$\begin{aligned}
outcome_{igsy} = & \beta_0 + \sum_q \beta_q \left[\underbrace{\left(\mathbf{1}_{\{y=2003\}} \times \mathbf{1}_{\{g=9\}} \right)}_{\text{Dragon Cohort}} \times Q_{igsy}^q \right] \\
& + \sum_q \rho_{qY} \mathbf{1}_{\{y=2003\}} \times Q_{igsy}^q + \sum_q \rho_{qG} \mathbf{1}_{\{g=9\}} \times Q_{igsy}^q \\
& + \sum_q \rho_q Q_{igsy}^q + X_{igsy} \gamma_X + W_{sy} \gamma_W + \mu_g + \mu_y (+\mu_s) + \varepsilon_{igsy} \quad (1.7)
\end{aligned}$$

where

- $outcome_{igsy}$ is the outcome of interest (effort, test score) of student i in grade g in school s in the year y PISA round
- $\left(\mathbf{1}_{\{y=2003\}} \times \mathbf{1}_{\{g=9\}} \right)$ is an indicator variable for being in the dragon cohort (1 for students in grade $g = 9$ in the year $y = 2003$)
- Q_{igsy}^q is an indicator variable for student i being in the q th quartile of the predicted

⁴⁴See Appendix 1.8.3 for precise construction. In short, I use out-of-sample PISA 2009 data to regress average normalized test scores on control variables. Using these coefficient estimates, I predict the scores of students in the in-sample PISA 2003 and 2006 data, and convert these predicted test scores into percentiles within cohort.

score distribution among students in the cohort in grade g in year y

- X_{igsy} is a vector of student- and parent-level controls (listed in Subsection 1.4.1)
- W_{sy} is a vector of school-level controls (listed in Subsection 1.4.1)
- μ_g , μ_y and μ_s are grade, year (PISA round) and (in some specifications) school fixed effects⁴⁵

This regression separately estimates the “treatment effect” β_q of being in the larger dragon cohort for each quartile q .

In contrast, the polynomial specification which follows allows us to analyze the effort and test score response at every point on the distribution. Replacing the quartile indicators with fifth-order polynomials in predicted score percentile, the regression specification becomes

$$\begin{aligned}
outcome_{igsy} = & \beta_0 + \underbrace{\left(\mathbf{1}_{\{y=2003\}} \times \mathbf{1}_{\{g=9\}} \right)}_{\text{Dragon Cohort}} \times \pi(\hat{a}_{igsy}) \\
& + \mathbf{1}_{\{y=2003\}} \times \pi_Y(\hat{a}_{igsy}) + \mathbf{1}_{\{g=9\}} \times \pi_G(\hat{a}_{igsy}) \\
& + \pi_0(\hat{a}_{igsy}) + X_{igsy}\gamma_X + W_{sy}\gamma_W + \mu_g + \mu_y (+\mu_s) + \varepsilon_{igsy} \quad (1.8)
\end{aligned}$$

where

- $\pi(\cdot)$, $\pi_Y(\cdot)$, $\pi_G(\cdot)$ and $\pi_0(\cdot)$ are separate fifth-order polynomial functions in predicted score percentile $\hat{a}_{igsy}$

With this specification, the “treatment effect” of being in the dragon cohort (the response to larger cohort size) at a particular percentile $\hat{a}_{igsy}$ is given by the polynomial function $\hat{\pi}(\hat{a}_{igsy})$ evaluated at that percentile.

Table 1.4 shows the coefficient estimates for regression (1.7) with effort in hours as the outcome variable. The baseline specification shown in column (1) is for the regression without

⁴⁵Because the PISA data is cross sectional, school identifiers are not consistent across years. Thus, inclusion of school fixed effects subsumes year fixed effects as well as school-level controls W_{sy} .

Table 1.4: Changes in Effort of Dragon Cohort by Quartile

Dep. Var.:	(1)	(2)	(3)	(4)
Effort	Base	Controls	Dec/Jan	Sch. FE
D.C. $\times$ Q1	-0.050 (0.193)	-0.028 (0.192)	-0.354 (0.491)	0.045 (0.202)
D.C. $\times$ Q2	0.129 (0.211)	0.150 (0.212)	0.424 (0.552)	0.270 (0.234)
D.C. $\times$ Q3	0.188 (0.207)	0.190 (0.211)	-0.207 (0.449)	0.309 (0.218)
D.C. $\times$ Q4	0.499** (0.223)	0.517** (0.226)	0.647 (0.444)	0.653*** (0.220)
Constant	2.232*** (0.226)	-39.73*** (11.54)	-20.28 (19.93)	-40.87 (104.8)
N	7403	7403	1201	7403
R^2	0.068	0.105	0.195	0.205
Quartiles	Yes	Yes	Yes	Yes
Grade FEs	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	N/A
Controls	No	Yes	Yes	Most
Sch. FEs	No	No	No	Yes

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Effort is measured in hours per week studying math. Standard errors clustered at the year-school level are in parentheses. Including school fixed effects subsumes year fixed effects and school-level controls (see footnote 45). See Subsection 1.4.1 for list of controls, and Appendix 1.10 for their coefficient estimates.

any controls or school fixed effects. The specification in column (2), the preferred specification, adds student-, parent- and school-level controls. These two columns suggest that it is students in the top quartile (students with the highest relative test scores) who are reacting most significantly to being in the larger dragon cohort, increasing their effort (on average) by around 0.5 hours per week (with standard errors of about 0.2). For students in the lowest quartile however, the point estimates suggest that they decrease their effort exertion only slightly; effort, on average, decreases by 0.03 to 0.05 hours in response to being in the larger dragon cohort, though the estimates are not statistically significant.

In general, estimates of effort response are increasing in each successive quartile. These

results are largely consistent with the predictions of the model: that those higher in the test score distribution (above a certain threshold) respond to an increase in cohort size by exerting more effort. However, there does not seem to be a significant negative response in the lower quartiles. Moving down the quartiles, the point estimates become less statistically significant, suggesting perhaps that the variance of effort responses among successively less-able students becomes higher.

There are several possible explanations as to why a negative response is not observed for those lower in the distribution. One possibility is that lower ability students are not as salient about the change in cohort size (and hence competition) as their higher ability counterparts; thus, they do not adjust effort optimally to maximize utility.⁴⁶

Another possibility is that there is an offsetting positive peer effect in the dragon cohort which increases the effort levels and test scores of all students. This would be the result of lower ability students being encouraged by the more-hardworking higher ability students to increase their effort. However, this positive peer effect would have to work asymmetrically in only one direction, in that higher ability students are not correspondingly encouraged by slacking lower ability students to decrease their effort.

One last possibility as to why a negative response is not observed for low ability students is heterogeneous returns to tertiary education. Suppose the additional utility from attending university is a function of ability; that is, $\alpha = \alpha(\theta_i)$. There is reason to believe that the economic returns to a university degree are higher for high ability students. High ability may be correlated with higher productivity in the workplace, thus leading to higher wage offers from employers. On the other hand, the productivity level of low ability students may not command such wage premiums even with a university degree. So if $\alpha(\theta_i)$ is positive for the high ability, but close to zero for the low ability, then any effort response by low ability students will be dampened. (This can be seen from the expression for $\frac{de_i}{dn}$ in equation (1.4),

⁴⁶Alternatively, their parents are not as salient to the dragon cohort and do not push their children to respond.

the numerator of which contains an α term.)

Columns (3) and (4) present additional specifications to check the robustness of these results. Column (3) includes in the sample only students born in the months of December and January. This is to address two concerns. The first concern is that parents of students born in months just before the Chinese new year may be comparably more or less superstitious. They may be more superstitious than average if they are parents who attempted to select their child's birth into the dragon year, but failed to do so, thus ending up with a birth just before the Chinese New Year in the treatment group. Alternatively, they may be less superstitious than average if they are parents "left behind" in the treatment group by superstitious parents selecting out of these months and into the dragon year. By looking only at students born in the two months around the cohort assignment boundary date of December 31st, I am using the premise that birth timing cannot be controlled precisely across this boundary. Thus, even with the movement of superstitious parents into and out of these months, if parents of students born in the month just before the December 31st boundary are very similar to those of students born in the month just after, then there is cleaner identification of the effect of just falling into the larger dragon cohort. On the other hand, further analysis of selection issues in Appendix 1.11 suggest that these types of selection are not significant.

The second concern which column (3) addresses is that of more astute parents selecting out of the dragon cohort, due to concerns of fewer resources per child or greater competition. Thus, parents of students born before the cohort assignment boundary date of December 31st may be on average different from those whose children are born after the boundary date. Again, by looking only at students born two months around this boundary date, I am using the premise that birth timing cannot be controlled precisely across this boundary to identify the effect. However, suppose more astute parents practice this type of selection, and that having such astute parents is correlated with exerting more effort or better test scores. Therefore, this type of selection should favor the non-dragon cohort students just before the dragon cohort in the data. That there are still positive and significant effects on effort and

test scores for the on-average now less-able dragon cohort students (due to astute parents out-selecting) suggests that either the true effects should be larger, or this type of selection is not a major problem.

The results in column (3) suggest that the main estimates are robust to these concerns. The third quartile estimate does turn negative; however, this estimate is not significantly different from the one in column (2). These estimates are statistically imprecise due to the substantial reduction in sample size.

The specification in column (4) includes school fixed effects. This is to address sorting (or tracking) of students by ability between schools. In Hong Kong, schools are classified into tiers with better students assigned to higher tier schools, based on performance at the end of primary school. If more resources are allocated to, or different policies are implemented at, schools with lower-performing students, and such differences interact with the way dragon cohort students are assigned to schools (perhaps because of the larger cohort size), then estimates may be biased if these differences are not picked up entirely by the school-level controls.

Including school fixed effects does not change the estimates by much, as column (4) shows. The pattern whereby the point estimates are increasing in quartile remains. However, since most students in a particular school will fall within the same (Hong Kong-wide) quartile as their peers, school fixed effects will be highly correlated with the quartile categories which are included in regression (1.7). There is thus a trade-off in including school fixed effects between inducing collinearity and reducing bias from sorting.

To show that these effort responses actually translate into better educational outcomes, I repeat this exercise with test scores as outcome variables in regression (1.7). Table 1.5 shows results for math (columns (1) and (2)), reading (columns (3) and (4)), and science (columns (5) and (6)) scores. For each subject, the baseline specification and the specification with controls are presented. Again, the general pattern is similar to the results with effort as the outcome: estimates of test score responses are increasing in each successive quartile. In

the top quartile, test score improvements range from 0.21 to 0.27 standard deviations (with standard errors around 0.08), depending on the test subject and regression specification. To put these results in perspective, Krueger (1999) finds the effect on test score of being in a small class (approximately 15 instead of 25 students) in the Tennessee STAR experiment to be about 0.2 standard deviations.⁴⁷

While these results may seem large, it is important to remember that the test scores measure the stock of human capital acquired over a student's entire education, and if a dragon cohort student in the top quartile consistently studies 0.5 hours more per week (a flow variable) as estimated earlier, then this accumulates into possibly hundreds of additional study hours.⁴⁸ Furthermore, the results suggest that even though the effort measure concerns only math study time, it is nonetheless a good proxy for studying in general, since more time spent studying math (by those in the top quartile) correlates with better reading and science scores.⁴⁹

Figure 1.10 plots the effort response to a larger cohort across the entire distribution, using estimates from regression (1.8) with effort as the dependent variable. The effort response becomes more positive moving up the distribution, reaching a peak of almost 0.7 additional hours per week, before attenuating slightly towards the top of the distribution.

Figures 1.11, 1.12, and 1.13 show similar plots for changes in test scores in response to a larger cohort for math, reading, and science scores respectively. These are constructed using estimates from regression (1.8) with the respective test scores as dependent variables. While there is a slight dip in the middle of the distribution, the general upward trend observed in the effort response is still present in the changes in test scores: on average, students higher in the distribution experience more positive changes in test scores in response to the larger dragon cohort. In the top quartile of the distribution, students in the larger dragon cohort

⁴⁷Moreover, simple mean comparisons usually find the Black-White test score gap to be about 1 standard deviation.

⁴⁸And this is just hours studying math.

⁴⁹Moreover, these results also suggest that the additional time spent studying math is not crowding out the studying of other subjects.

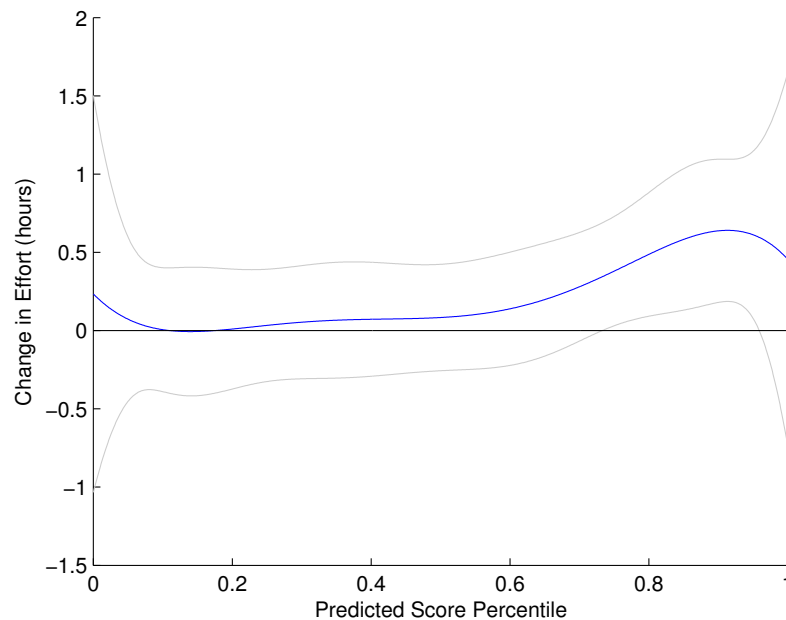
Table 1.5: Changes in Test Scores of Dragon Cohort by Quartile

Dep. Var.: Test Score	(1)		(2)		(3)		(4)		(5)		(6)	
	Math		Reading		Base		Controls		Base		Science	
D.C. $\times$ Q1	0.042 (0.126)		0.069 (0.114)		-0.058 (0.12)		-0.058 (0.117)		0.043 (0.126)		0.070 (0.115)	
D.C. $\times$ Q2	0.168* (0.089)		0.160* (0.088)		-0.003 (0.086)		0.000 (0.082)		0.055 (0.092)		0.036 (0.089)	
D.C. $\times$ Q3	0.192** (0.079)		0.177** (0.075)		0.117 (0.089)		0.137* (0.081)		0.198** (0.086)		0.189** (0.082)	
D.C. $\times$ Q4	0.250*** (0.088)		0.224*** (0.077)		0.269*** (0.085)		0.211*** (0.077)		0.262*** (0.088)		0.228*** (0.081)	
Constant	-0.992*** (0.156)		-35.92*** (8.394)		-1.105*** (0.149)		-29.29*** (7.398)		-1.058*** (0.157)		-34.03*** (7.955)	
N	7697		7697		7697		7697		7697		7697	
R^2	0.255		0.339		0.247		0.343		0.249		0.329	
Quartiles	Yes		Yes		Yes		Yes		Yes		Yes	
Grade FEs	Yes		Yes		Yes		Yes		Yes		Yes	
Year FEs	Yes		Yes		Yes		Yes		Yes		Yes	
Controls	No		Yes		No		Yes		No		Yes	
Sch. FEs	No		No		No		No		No		No	

Legend: Significance level: *** = 1% ; ** = 5% ; * = 10%.

Note: Test scores are normalized with mean 0 and standard deviation 1. Standard errors clustered at the year-school level are in parentheses. Coefficient estimates for other variables, as well as additional robustness specifications, are available in Appendix 1.10.

Figure 1.10: Effort Response Across Relative Test Score Distribution

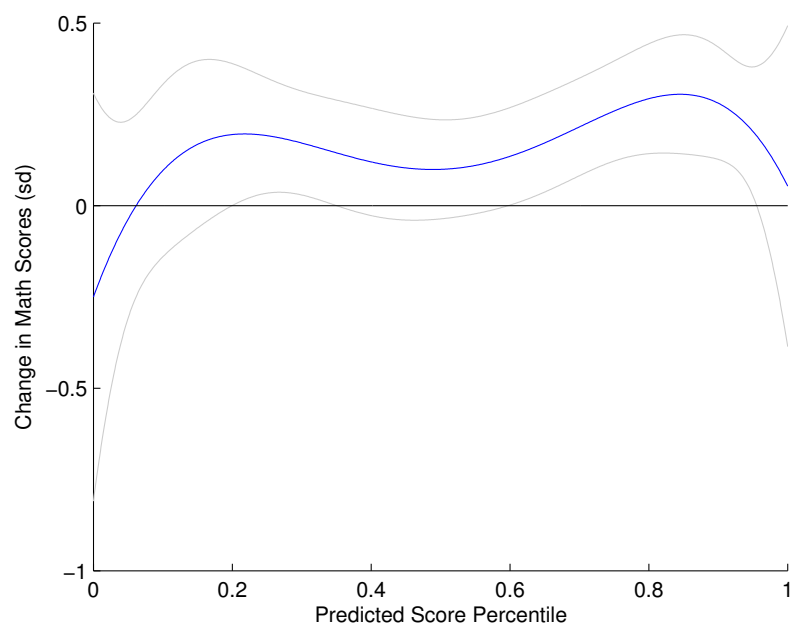


Note: Graph is fitted using a fifth-order polynomial. Dashed lines are 90% confidence intervals.

received test scores between 0.2 to 0.3 standard deviations higher than their counterparts in other cohorts.

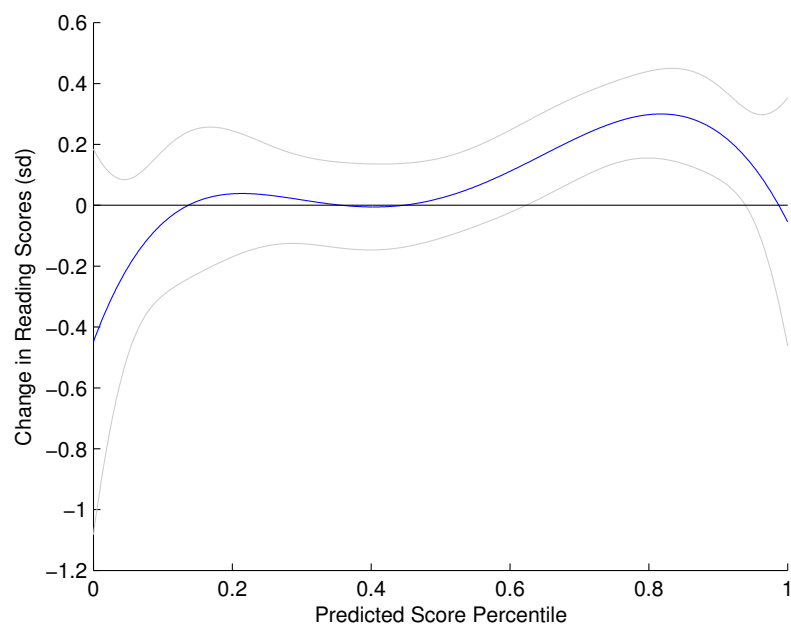
Though we do not observe much of a negative response towards the bottom half of the distribution, the quartile estimates and plots suggest that the threshold in Propositions 1.2 and 1.3 may be towards the middle of the distribution. While this may seem low when only 30% to 35% of students attend university (suggesting a p of around 0.65 to 0.70), one explanation for this is that student expectations of future university attendance may be playing a role. In the 2003 PISA survey, 64.2% of students surveyed expected to attend university, which is almost twice the percentage who actually end up matriculating. If such inflated expectations (of the value of p) are driving effort decisions, then a threshold towards the middle of the distribution may not seem that far-fetched. A second possibility is that the mean of the error distribution is not at zero. This will shift the thresholds in the

Figure 1.11: Change in Math Score Across Relative Test Score Distribution



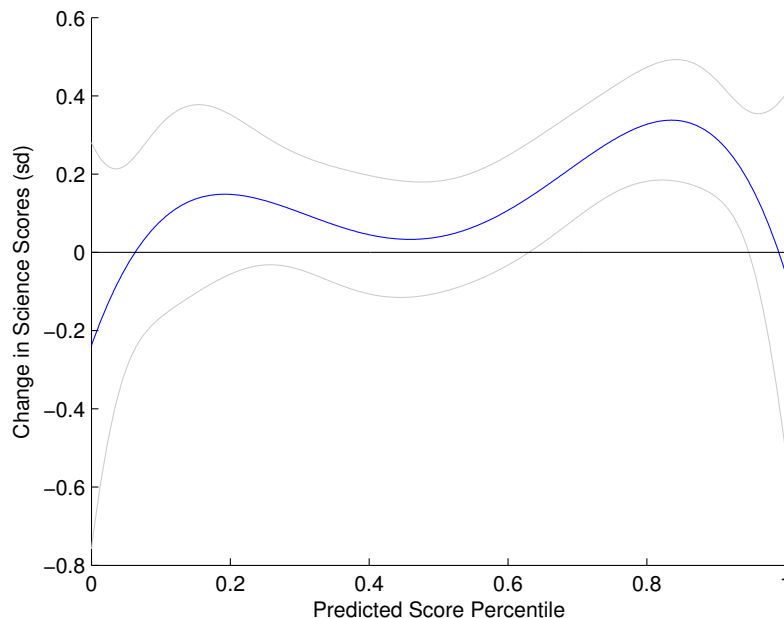
Note: Change is in standard deviations. See notes of Figure 1.10.

Figure 1.12: Change in Reading Score Across Relative Test Score Distribution



Note: Change is in standard deviations. See notes of Figure 1.10.

Figure 1.13: Change in Science Score Across Relative Test Score Distribution



Note: Change is in standard deviations. See notes of Figure 1.10.

propositions accordingly.⁵⁰ A third possibility (as considered in the extensions discussion) is that there may be multiple cutoffs for multiple rewards resulting from other post-secondary opportunities with different schooling requirements.

1.5.2 Falsification Exercise

To allay concerns that this may be a global phenomenon and not particular to Hong Kong's dragon cohort, I perform a falsification exercise by extending the analysis to pooled PISA data from additional countries other than Hong Kong who also participated in PISA in the same years as Hong Kong, in which students in both grades 9 and 10 are observed.⁵¹ Using

⁵⁰More generally, if I do not assume that the peak is at zero, this shift will depend on the distance between the peak and the mean.

⁵¹The countries used are Austria, Belgium, Brazil, Canada, Czech Republic, Germany, Greece, Hungary, Indonesia, Ireland, Italy, Liechtenstein, Luxembourg, Netherlands, Portugal, Russia, Slovak Republic, Spain, Switzerland, Thailand, and Uruguay.

this data, I run an altered version of the quartile specification of the form

$$\begin{aligned}
outcome_{igsysc} = & \beta_0 + \sum_q \beta_q \left[\underbrace{\left(\mathbf{1}_{\{y=2003\}} \times \mathbf{1}_{\{g=9\}} \right)}_{\text{Dragon Cohort}} \times Q_{igsysc}^q \right] \\
& + \sum_{c_0 \in \mathbb{C}} \sum_q \rho_{cqY} \mathbf{1}_{\{c=c_0\}} \times \mathbf{1}_{\{y=2003\}} \times Q_{igsysc}^q + \sum_{c_0 \in \mathbb{C}} \sum_q \rho_{cqG} \mathbf{1}_{\{c=c_0\}} \times \mathbf{1}_{\{g=9\}} \times Q_{igsysc}^q \\
& + \sum_{c_0 \in \mathbb{C}} \sum_q \rho_{cq} \left[\mathbf{1}_{\{c=c_0\}} \times Q_{igsysc}^q \right] + \sum_{c_0 \in \mathbb{C}} \left[\mathbf{1}_{\{c=c_0\}} \times (X_{igsysc} \gamma_{cX} + W_{sysc} \gamma_{cW}) \right] \\
& + \mu_{gc} + \mu_{yc} + \varepsilon_{igsysc} \quad (1.9)
\end{aligned}$$

where c indexes country in the set of countries $\mathbb{C}$, and $\mathbf{1}_{\{c=c_0\}}$ is an indicator variable for country being c_0 . The controls and fixed effects enter separately for each country, allowing for these variables to have different-sized effects for each country. β_q is the “treatment effect” of being in the (non-existent) dragon cohort in quartile q of the score distribution. Predicted score percentiles are estimated within each country individually.

Table 1.6 shows coefficient estimates for regression (1.9) with effort, math, reading and test scores as outcome variables (columns (1) through (4) respectively). These results are very different from the corresponding results in the Hong Kong regressions. Firstly, all point estimates except for those for reading scores are negative, and most are not statistically significant. Secondly, the pattern whereby coefficient estimates increase moving up the quartiles is now absent. Thirdly, almost all coefficient estimates are now much closer to zero, though there are still some cells (such as the first quartile estimate of the effort outcome) which are significantly more negative. Overall, these findings suggest that the pattern of effort responses and changes in test score to the larger dragon cohort found in Hong Kong are not present in other countries.

Table 1.6: Falsification Exercise Using Other Countries

	(1)	(2)	(3)	(4)
Dep. Var.:	Effort	Math	Reading	Science
D.C.×Q1	-0.12*** (0.043)	-0.004 (0.025)	0.049 (0.03)	-0.046* (0.026)
D.C.×Q2	-0.024 (0.041)	-0.015 (0.022)	0.042 (0.027)	-0.028 (0.024)
D.C.×Q3	-0.092** (0.042)	-0.036 (0.022)	0.031 (0.026)	-0.055** (0.024)
D.C.×Q4	-0.057 (0.045)	-0.023 (0.025)	0.027 (0.029)	-0.049* (0.027)
Constant	-10.41 (10.09)	9.439 (8.808)	13.54 (10.57)	7.31 (9.36)
<i>N</i>	245196	256852	256852	256852
<i>R</i> ²	0.162	0.421	0.387	0.37
Quartiles	Yes	Yes	Yes	Yes
Grade FEs	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Sch. FEs	No	No	No	No

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Effort is measured in hours per week. Test scores are normalized with mean 0 and standard deviation 1. Standard errors clustered at the year-school level are in parentheses. Coefficient estimates for other variables are available in Appendix 1.10.

1.5.3 Teacher Reallocation

An alternative consequence of an increase in cohort size is class size increases. Previous studies have found that larger class sizes⁵², and more generally, fewer resources per student⁵³, lead to worse educational outcomes. While the positive test score findings for the better students discount this explanation, it would still be interesting to consider whether class sizes changed in response to the cohort size increase. In this subsection, I summarize an analysis (presented in detail in Appendix 1.12) which suggests that teachers were reallocated

⁵²See Krueger (1999); Krueger and Whitmore (2001); Chetty et al. (2011); Angrist and Lavy (1999). Disruption in the classroom as a potential mechanism is discussed in Lazear (2001).

⁵³E.g., Card and Krueger (1992)

between grades to accommodate the larger dragon cohort, thus equalizing class sizes across grades.

I evaluate the degree of teacher reallocation between grades by estimating the identity

$$C_{gs} = \frac{B_{gs}}{T_{gs}(\rho) / TPC}$$

where

- C_{gs} is the class size in grade g at school s
- B_{gs} is the total number of students in grade g at school s (batch size)
- $T_{gs}(\rho)$ is the number of teachers allocated to grade g at school s
- ρ is the teacher reallocation factor (see below)
- TPC is the number of teachers per class

This identity states that the number of students in a class in grade g at school s must be equal to the total number of grade g students at that school divided by the number of grade g classes at that school, the latter of which is given by the number of teachers assigned to that grade divided by the number of teachers per class.

The teacher reallocation factor ρ is defined to be between 0 and 1, where $\rho = 1$ implies a policy of total reallocation of teachers across grades (i.e., teachers are flexible), and $\rho = 0$ implies no reallocation (i.e., teachers can only teach a certain grade, perhaps due to lack of training or institutional rules). Teacher allocation $T_{gs}(\rho)$ is a function of policy parameter ρ , the exact form of which is given by equation (1.11) in Appendix 1.12. The free parameters to be estimated are ρ and TPC , which are assumed to be similar across all grades and schools.⁵⁴

⁵⁴For ρ , this seems reasonable if school administrators are required by legislation or union rules to reallocate teachers in a similar way at every school. For TPC , this also seems reasonable if resources for hiring teachers at the school are allotted centrally in proportion to total number of students, and if all students take a similar number of subjects every grade year.

The teacher reallocation factor is estimated to be approximately $\rho = 0.902$ (s.e. 0.047), which implies a high degree of teachers being reallocated between grades. Another way to interpret this estimate is that if schools can only choose extreme policies of $\rho \in \{0, 1\}$, then on average, for every 10 students in Hong Kong, 9 of them will be in a school where total reallocation is carried out ($\rho = 1$), while the remaining one will be in a school where no reallocation occurs ($\rho = 0$).⁵⁵

1.6 Conclusion

Hong Kong’s 1988 dragon cohort experiences on average a 5% larger cohort size year-on-year, relative to other schooling cohorts, but university intake does not seem to respond to this phenomenon. The tournament model of academic competition presented predicts that this will factor into the effort decisions of students, and that the response to the larger cohort will depend on a student’s relative position on the ability distribution. The empirical results reflect this, with findings which suggest that students towards the top of the distribution “step up” and exert more effort in response to being in the larger dragon cohort. Students towards the bottom of the distribution also appear to “give up” and exert less effort, though only slightly. Similar changes are observed in test scores. Furthermore, this pattern of effects is not found in other countries.

While these findings suggest that cohort size can induce changes in competitive behavior, they also seem to indicate that the changes are mainly taking place at the top of the distribution, with these students working harder in response to increased competition. Students at the bottom of the distribution do not seem to work any less hard (or only slightly so) in response to increased competition. This may be good news to policy makers, because while they should still take competition into account when making related policy decisions, increased competition does not seem to discourage the students for whom the theoretical model predicts a decrease in effort, and hence detrimental effects such as worse educational

⁵⁵See Appendix 1.12 for precise meaning of ρ .

outcomes are not observed either.

This paper stands in contrast with the previous literature both theoretically and empirically. The tournament model of academic competition presented is different from previous models in two ways. The first difference is the inclusion of ability heterogeneity. Almost all previous tournament models have output (the latent variable determining the winner of the tournament) which depend only on effort⁵⁶, with heterogeneity between contestants (if considered) entering through the cost of effort function. I model output as a function of both effort and heterogeneous ability, which seems more realistic in the test score context (as well as other situations). The second difference is the emphasis on a cohort size change. Whereas previous models take tournament size as exogenously given or as a number to be solved for to achieve an optimal outcome, my model explicitly takes into account the comparative statics of a change in tournament size.^{57,58} Cohort size is also modeled as a measure rather than a natural number in my model, given the large number of students in any given cohort.

From an empirical standpoint, in many previous empirical studies of tournaments, effort is unobserved; often, only the effect on outcomes (i.e., outputs) is considered. The availability of not only test scores but also hours studying math from the PISA survey data adds to a small but growing empirical tournament literature where effort is observed (or at least inferred).⁵⁹ Also, studying competition of a schooling cohort is much larger-scale than other settings in both the empirical tournaments and education financial incentives literatures, though as the experimental results in Garcia and Tor (2009) suggest, mere perceptions of the size of the competitor pool may be enough to induce changes in competitive behavior. And while the number of competitors in most settings may be endogenously determined

⁵⁶For example, in the classic model of Lazear and Rosen (1981), workers produce output $q = \mu + \epsilon$, where μ is effort and ϵ is an error term. In this case, output is linear in effort exertion.

⁵⁷While others [e.g., Green and Stokey (1983)] have modeled tournaments with more than two contestants, to the best of my knowledge, this is the first model to carry out such an analysis to examine the comparative statics of a change in cohort size.

⁵⁸To this end, I also abstract away from the firm side of this problem, and take the number and size of rewards as externally determined.

⁵⁹See Ehrenberg and Bognanno (1990); Casas-Arce and Martinez-Jerez (2009); and Boudreau et al. (2012), among others.

or in some cases treated as a choice variable, the cohort size in this setting is exogenously determined.

The results in this chapter point to further avenues for research. What other effects does this enlarged dragon cohort have on itself and other cohorts? Are there other policy variables which remain unresponsive to such a cohort size increase? And what of other countries with significant Chinese populations which experience similar cohort size variation?

Other dragon cohort effects are potentially harder to assess. Because this study is interested in what is essentially a peer effect (of experiencing more students around you), the context of the situation lends itself particularly well to my identification strategy and research question. In other contexts, selection bias and general equilibrium effects may confound estimates. Furthermore, external validity is an issue which needs to be addressed. Whereas the increased size of dragon cohorts and double cohorts are very apparent to the competitors, smaller and more gradual demographic changes may not induce the same kind of reaction in competitive behavior. And given that such changes happen over longer time-spans, policies may be more responsive, and additional aspects of the economy must be considered and accounted for. Nevertheless, this study offers a first step in examining the potential effects of increased competition, arising from an enlarged schooling cohort, on effort decisions and educational outcomes.

Appendices

1.7 Proofs

To facilitate the proofs, the first and second order conditions of the model are reprinted below.

The FOC is

$$\left[\phi + \alpha h \left(S(\theta_i, e_i) - \bar{S}_p \right) \right] \frac{\partial S(\theta_i, e_i)}{\partial e_i} - c'(e_i) = 0$$

The SOC is

$$\begin{aligned} \alpha h' \left(S(\theta_i, e_i) - \bar{S}_p \right) \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 \\ + \left[\phi + \alpha h \left(S(\theta_i, e_i) - \bar{S}_p \right) \right] \frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} - c''(e_i) < 0 \end{aligned}$$

1.7.1 Proof of Proposition 1.1

The total derivative of $S(\theta_i, e_i)$ with respect to θ_i is

$$\frac{dS(\theta_i, e_i)}{d\theta_i} = \frac{\partial S(\theta_i, e_i)}{\partial \theta_i} + \frac{\partial S(\theta_i, e_i)}{\partial e_i} \frac{de_i}{d\theta_i} \quad (1.10)$$

Applying the implicit function theorem to the FOC,

$$\begin{aligned} \frac{de_i}{d\theta_i} &= - \frac{\frac{\partial}{\partial \theta_i} \left\{ \left[\phi + \alpha h \left(S(\theta_i, e_i) - \bar{S}_p \right) \right] \frac{\partial S(\theta_i, e_i)}{\partial e_i} \right\}}{\frac{\partial}{\partial e_i} \left\{ \left[\phi + \alpha h \left(S(\theta_i, e_i) - \bar{S}_p \right) \right] \frac{\partial S(\theta_i, e_i)}{\partial e_i} \right\} - c''(e_i)} \\ &= - \frac{\alpha h' \left(S(\theta_i, e_i) - \bar{S}_p \right) \frac{\partial S(\theta_i, e_i)}{\partial \theta_i} \frac{\partial S(\theta_i, e_i)}{\partial e_i} + \left[\phi + \alpha h \left(S(\theta_i, e_i) - \bar{S}_p \right) \right] \frac{\partial^2 S(\theta_i, e_i)}{\partial \theta_i \partial e_i}}{\alpha h' \left(S(\theta_i, e_i) - \bar{S}_p \right) \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 + \left[\phi + \alpha h \left(S(\theta_i, e_i) - \bar{S}_p \right) \right] \frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} - c''(e_i)} \end{aligned}$$

Substituting this into equation (1.10) and omitting some arguments in $h(\cdot)$,

$$\begin{aligned}
\frac{dS(\theta_i, e_i)}{d\theta_i} &= \frac{\partial S(\theta_i, e_i)}{\partial \theta_i} - \frac{\alpha h'(\cdot) \frac{\partial S(\theta_i, e_i)}{\partial \theta_i} \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 + [\phi + \alpha h(\cdot)] \frac{\partial^2 S(\theta_i, e_i)}{\partial \theta_i \partial e_i} \frac{\partial S(\theta_i, e_i)}{\partial e_i}}{\alpha h'(\cdot) \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 + [\phi + \alpha h(\cdot)] \frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} - c''(e_i)} \\
&= \frac{[\phi + \alpha h(\cdot)] \left[\frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} \frac{\partial S(\theta_i, e_i)}{\partial \theta_i} - \frac{\partial^2 S(\theta_i, e_i)}{\partial \theta_i \partial e_i} \frac{\partial S(\theta_i, e_i)}{\partial e_i} \right] - c''(e_i) \frac{\partial S(\theta_i, e_i)}{\partial \theta_i}}{\alpha h'(\cdot) \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 + [\phi + \alpha h(\cdot)] \frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} - c''(e_i)}
\end{aligned}$$

The SOC in the denominator is negative. Hence, $\frac{dS(\theta_i, e_i)}{d\theta_i} > 0$ if and only if the numerator is negative. That is,

$$[\phi + \alpha h(\cdot)] \left[\frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} \frac{\partial S(\theta_i, e_i)}{\partial \theta_i} - \frac{\partial^2 S(\theta_i, e_i)}{\partial \theta_i \partial e_i} \frac{\partial S(\theta_i, e_i)}{\partial e_i} \right] - c''(e_i) \frac{\partial S(\theta_i, e_i)}{\partial \theta_i} < 0$$

This is true under Assumptions A and B, and the fact that $h(\cdot) > 0$.

$$\begin{aligned}
&\underbrace{[\phi + \alpha h(\cdot)]}_{+ \text{ by def.}} \left[\underbrace{\frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2}}_{- \text{ by A}} \underbrace{\frac{\partial S(\theta_i, e_i)}{\partial \theta_i}}_{+ \text{ by A}} - \underbrace{\frac{\partial^2 S(\theta_i, e_i)}{\partial \theta_i \partial e_i} \frac{\partial S(\theta_i, e_i)}{\partial e_i}}_{+ \text{ by A}} \right] - \underbrace{c''(e_i)}_{+ \text{ by B}} \underbrace{\frac{\partial S(\theta_i, e_i)}{\partial \theta_i}}_{+ \text{ (or 0) by A}} < 0
\end{aligned}$$

The second inequality follows from the fact that

$$\frac{dE[S(\theta_i, e_i) - \eta_i]}{d\theta_i} = \frac{dS(\theta_i, e_i)}{d\theta_i} - \frac{dE[\eta_i]}{d\theta_i} = \frac{dS(\theta_i, e_i)}{d\theta_i}$$

since η_i is orthogonal to ability.

Q.E.D.

1.7.2 Proof of Proposition 1.2

From equation (1.1), using the implicit function theorem,

$$\frac{de_i}{dp} = -\frac{-\alpha h' \left(S(\theta_i, e_i) - \bar{S}_p \right) \frac{\partial \bar{S}_p}{\partial p} \frac{\partial S(\theta_i, e_i)}{\partial e_i}}{\alpha h' (.) \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 + [\phi + \alpha h (.)] \frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} - c''(e_i)}$$

Substituting into equation (1.3),

$$\frac{de_i}{dn} = -\frac{-\alpha h' \left(S(\theta_i, e_i) - \bar{S}_p \right) \frac{\partial \bar{S}_p}{\partial p} \frac{\partial S(\theta_i, e_i)}{\partial e_i} \frac{dp}{dn}}{\alpha h' (.) \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 + [\phi + \alpha h (.)] \frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} - c''(e_i)}$$

The denominator is the negative SOC, so $\frac{de_i}{dn} > 0$ if and only if the numerator

$$-\alpha h' \left(S(\theta_i, e_i) - \bar{S}_p \right) \frac{\partial \bar{S}_p}{\partial p} \frac{\partial S(\theta_i, e_i)}{\partial e_i} \frac{dp}{dn} > 0$$

Note that under Assumption A, $\frac{\partial S(\theta_i, e_i)}{\partial e_i} > 0$, and under Assumption D, $\frac{dp}{dn} > 0$. Also note that by chain rule

$$\frac{\partial \bar{S}_p}{\partial p} = \frac{\partial}{\partial p} S \left(F^{-1}(p) \right) = \frac{dS}{d\theta_i} \frac{\partial F^{-1}(p)}{\partial p} > 0$$

This is positive because of the following. Firstly, Proposition 1 states that under Assumptions A and B, $\frac{dS}{d\theta_i} > 0$. Secondly,

$$\frac{\partial F^{-1}(p)}{\partial p} = \left(\frac{\partial F^{-1}}{\partial n} \Big|_{p \text{ constant}} \right) / \frac{\partial p}{\partial n} + \frac{1}{f(F^{-1}(p))}$$

The first term represents a possible change in the ability distribution with the cohort size change; this is zero under Assumption C. The second term is positive. Hence, $\frac{\partial \bar{S}_p}{\partial p} > 0$.

Thus, $\frac{de_i}{dn} > 0$ if and only if

$$h' \left(S(\theta_i, e_i) - \bar{S}_p \right) < 0$$

Note that for a student with relative ability a

$$S_i = F_S^{-1}(a) = S \circ F^{-1}(a)$$

Using this formula to replace $S(\theta_i, e_i)$ and $\bar{S}_p$,

$$h'(S(\theta_i, e_i) - \bar{S}_p) = h'(S \circ F^{-1}(a_i) - S \circ F^{-1}(p) + E(\eta_i))$$

where $a_i = F(\theta_i)$ is the relative ability of student i . Under Assumption η , the error distribution is single peaked at $\eta_{peak} = 0$, so

$$h'(S \circ F^{-1}(a_i) - S \circ F^{-1}(p) + E(\eta_i)) < 0$$

if and only if

$$\begin{aligned} S \circ F^{-1}(a_i) - S \circ F^{-1}(p) + E(\eta_i) &> \eta_{peak} = 0 \\ a_i &> F \circ S^{-1}(S \circ F^{-1}(p) - E(\eta_i)) \end{aligned}$$

Note that if the error distribution is single peaked at somewhere other than zero, then the threshold relative ability level will be shifted accordingly, since

$$a_i > F \circ S^{-1}(S \circ F^{-1}(p) - E(\eta_i) + \eta_{peak})$$

Q.E.D.

1.8 Data Appendix

Tables 1.7, 1.8, and 1.9 define and document the construction of variables in the data.

Table 1.7: Details for Student-level Variables

Variable	Definition [Availability] (Construction)
Effort	Self-reported hours per week studying math [2003, 2006 only] (See Subsection 1.8.1 of this appendix)
Test Score (outcome)	Normalized test scores for math, reading, science (See Subsection 1.8.2 of this appendix)
Predicted Score Percentile	Percentile of predicted test score using PISA 2009 data (See Subsection 1.8.3 of this appendix)
Sex	Male / Female Indicator
Age	Age in months (Derived from birth year, birth month and survey round)
Age in Grade	Relative age in student's grade in months (Derived from birthday (month and year) and grade, based on distance in months of student's birthday to youngest students born in December assigned by cohort assignment policy)
Foreign Born	Indicator for whether student is born in country of test or not
Class Size	Student-reported number of students in class for a particular subject [2002 all subjects, 2003 math only, 2009 reading only]
Expected Education	Self-reported expectation of future highest educational attainment in ISCED Levels [2003, 2009 only]

Table 1.8: Details for Parent-level Variables

Variable	Definition (Construction)
Occupation Status (each parent)	International Socioeconomic Index (ISEI) (Based on ISCO occupation codes of each parent ⁶⁰ ; N/A treated as separate category to account for stay-at-home parents)
Highest education (between both parents)	International Standard Classification of Education (ISCED) Level (Maximum of both parents assigned. 0 = no or pre-primary education; 1 = primary education; 2 = lower secondary education; 3 = upper secondary education; 4 = post-secondary (usually vocational) education; 5/6 = tertiary education (Associates or Bachelors degree respectively))

Table 1.9: Details for School-level Variables

Variable	Definition (Construction)
School Type	Indicator for school being public, private government funded, or private independent
School Size	Number of student across all grades at school
Student-Teacher Ratio (STR)	Number of students per teacher at school (Calculated by dividing total number of students at school by total number of teachers. Part time teachers count for 0.5 teachers.)

1.8.1 Effort Comparability

Differences in the way students were asked to report effort (hours per week studying math) were taken into account. In 2003, the survey question asks students for integer number of hours. In 2006, however, the survey question asks students to choose from a set of multiple responses. These responses consist of the following categories: No time; Less than 2 hours a week; 2 or more but less than 4 hours a week; 4 or more but less than 6 hours a week; and 6 or more hours a week.

To ensure comparability, I construct the effort variable as follows. First, integer hours from the 2003 responses were mapped into the categories from 2006. Next, an integer number of hours was assigned to each category. For “No time,” 0 was assigned. For ranged categories with two endpoints, the mean of the two endpoints was assigned. (For example, “2 or more but less than 4 hours...” was assigned 3 hours.). For the “6 or more hours...” category, 7 hours was assigned. This was done so that the estimates of changes in effort would be on the conservative side (compared to assigning, say, 8 hours instead). The analysis was repeated with different integer assignments for this last category, producing qualitatively similar results.

1.8.2 Test Score Normalizations

Raw test scores are obtained by taking the average of the plausible values generated in PISA. Normalized test scores for each subject are normalized such that the mean and standard deviation over Hong Kong students of the final score is 0 and 1 respectively. This is done by subtracting from the raw score the mean of raw scores, and then dividing by the standard deviation of the raw scores. This applies to normalizations for test scores of other countries as well, which are normalized by subtracting the Hong Kong mean and then dividing by the Hong Kong standard deviation.

1.8.3 Predicted Score Percentile

To obtain predicted score percentile, I do the following.

1. Using out-of-sample PISA 2009 data (restricted to students in grade 9 and 10), I calculate the mean of math, reading, and science normalized test scores for each student in this sample. Then, I regress this average test score on a set of covariates which are also balanced across PISA 2003, 2006, and 2009. These are sex, age, age in grade, a foreign birth indicator, parent's highest education level, school type, school size and student teacher ratio. (Parent's occupation status is excluded because there are too many categories in this variable which do not appear in any one of the years.) This regression is of the form

$$avescore_{igs09} = \varphi_0 + X_{igs09}\varphi_X + W_{s09}\varphi_W + \mu_g + \varepsilon_{igs09}$$

and gives coefficient estimates $\hat{\varphi}_0$, $\hat{\varphi}_X$, $\hat{\varphi}_W$ and $\hat{\mu}_g$.

2. With the coefficient estimates, I predict the scores in the in-sample PISA 2003 and 2006 data by taking each student's covariates and substituting the values into the equation

$$score_{igsy} = \hat{\varphi}_0 + X_{igsy}\hat{\varphi}_X + W_{sy}\hat{\varphi}_W + \mu_g$$

to obtain a predicted score value $score_{igsy}$.

3. I then convert this $score_{igsy}$ into a within-cohort percentile by calculating it's percentile among a particular student's cohort of peers in the same grade and PISA year. Dummies for quartiles are created using this percentile.

The correlation between predicted score percentile and actual test score percentile (within cohort) is 0.48.

1.9 Differences in Cohort Characteristics

To test whether predetermined characteristics of the dragon cohort are significantly different from other cohorts, I can run the regression

$$char_{isgy} = \beta_0 + \beta dragoncohort_{isgy} + \mu_g + \mu_y + \varepsilon_{isgy}$$

where

- $char_{isgy}$ is the characteristic being tested
- $dragoncohort_{isgy}$ is an indicator variable for being in the dragon cohort (1 for students in grade $g = 9$ in the year $y = 2003$)
- μ_g and μ_y are grade and year fixed effects

The estimates of β indicate how different the dragon cohort is in terms of that characteristic. Table 1.10 shows coefficient estimates for this regression, where each row is a specification with a different dependent variable.

Students in the dragon cohort are on average 0.74 months younger; this is because the 2003 PISA was administered approximately one month later in the year compared to the 2006 PISA. A dragon cohort student is 7.8 percentage points more likely to have been born in Hong Kong; this may be due to crowding out by the larger cohort if preference is given to local-born students and foreign-born students substitute towards international schools. Fathers of the dragon cohort group are also slightly more likely to have attended high school, by almost 4 percentage points, though this β coefficient is significant only at a 10% level. In the regressions specifications, I will control for these as well as all other variables. The β coefficients for the remaining variables are not significantly different from zero.

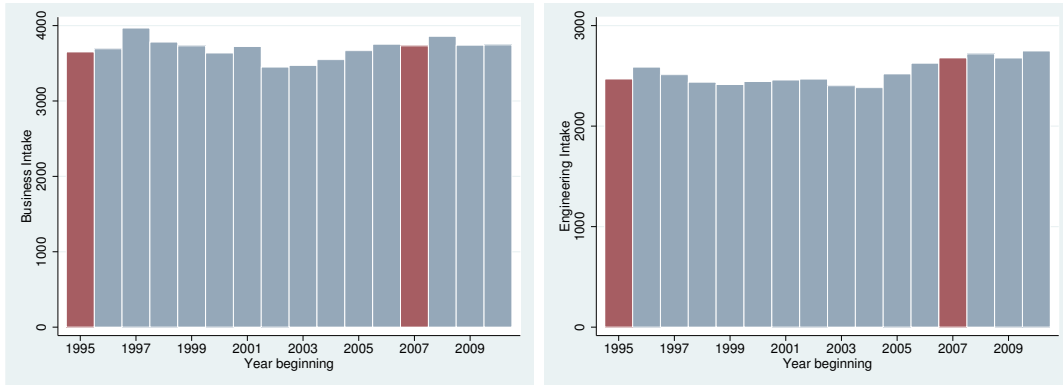
Table 1.10: Differences in Predetermined Characteristics

		β	Constant	N
	Dep. Var.	Coef. (S.E.)	Coef. (S.E.)	$[R^2]$
(1)	Male	-0.001 (0.027)	0.512*** (0.03)	7936 [0.001]
(2)	Age (months)	-0.741*** (0.202)	188*** (0.165)	7936 [0.152]
(3)	Mother H.S.	0.010 (0.02)	0.149*** (0.018)	7936 [0.006]
(4)	Father H.S.	0.039* (0.022)	0.192*** (0.02)	7936 [0.002]
(5)	HK Born	0.078*** (0.024)	0.662*** (0.019)	7936 [0.079]
(6)	School Private	-0.004 (0.013)	0.096*** (0.028)	7936 [0.002]
(7)	School Size	-12.59 (12.3)	1050*** (15.36)	7906 [0.004]
(8)	Student/Teacher Ratio	0.052 (0.132)	18.08*** (0.202)	7861 [0.001]

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Standard errors clustered at the year-school level are in parentheses. See Appendix 1.10 for variable definitions.

Figure 1.14: Intake for Business and Engineering Programs by Year



Note: Intake figures are for undergraduate business and engineering programs are among local universities which offer such programs only.

1.10 Supplemental Results

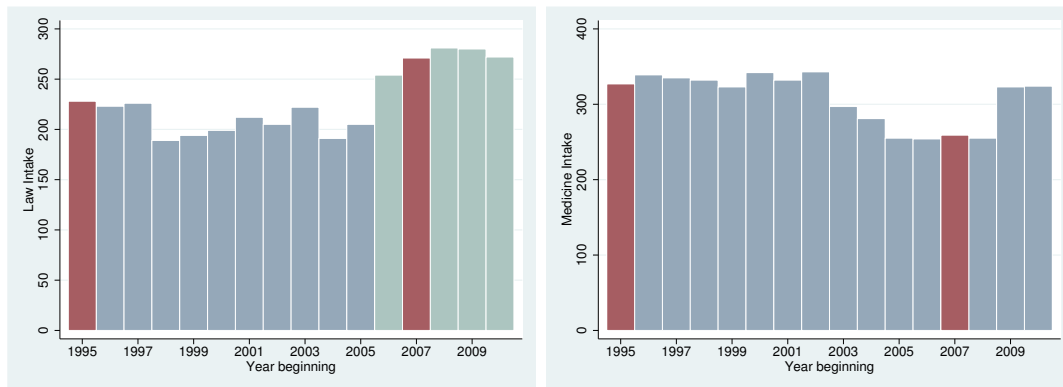
1.10.1 University Intake by Program

Figure 1.14 shows the total intake for undergraduate business and engineering programs at local universities by year.⁶¹ These represent the two most popular programs by absolute student numbers at local universities. Figure 1.15 shows the total intake for undergraduate law and medicine programs at local universities by year. Anecdotally, these represent the two most admissions-competitive programs at local universities. Note that in 2006, a new law school was founded at the Chinese University of Hong Kong, accounting for the jump in intake for undergraduate law programs that year.⁶² Again, there does not appear to be any discernible difference between intake for dragon cohort years and that of surrounding years for any of the individual programs of study.

⁶¹These numbers are among those universities offering undergraduate business or engineering programs. The same applies to subsequent figures which show intake for certain programs only.

⁶²There may be the concern of the endogeneity of the opening of this new law school. If the school was timed to open in order to accommodate the larger dragon cohort, then the argument is weakened. However, I do not believe this to be so for several reasons. Firstly, the planning for this law school's founding was initiated well in advance. The Chinese University's application for government funding for the new law school was submitted in 2004, and there was no guarantee that it would be accepted, and if so, when funding would start. Moreover, the law school opened in 2006, one year before the dragon cohort intake. Lastly, the yearly intake of this new law school of around 60 students would have done little to alleviate the large increase in the dragon cohort size.

Figure 1.15: Intake for Law and Medicine Programs by Year



Note: Intake figures are for undergraduate law and medicine programs among local universities which offer such programs only. In 2006, a new law school was found at the Chinese University of Hong Kong, accounting for the jump in intake for undergraduate law programs that year.

1.10.2 Regression of Outcomes on Controls Only

Regressions of the different outcomes on only the set of controls (i.e. excluding the interaction terms with the dragon cohort indicators) can be informative about the baseline relationship between predetermined controls and the outcomes in question. Table 1.11 shows coefficient estimates of such regressions with the controls listed in Subsection 1.4.1.

Column (1) shows the results for effort as the dependent variable. On average, girls work harder than boys, by almost 0.3 hours per week. After taking into account both the coefficient estimates on age and age in grade, the net effect of being one month older on effort exertion is close to zero, controlling for grade fixed effects. In general, students with more educated parents exerted more effort. Students in larger schools and in schools with higher student-teacher ratios tend to exert more effort on average. This may be suggestive of increased competition, but may also be attributed to a policy of weaker students are assigned to schools with fewer students and smaller classes.

Columns (2), (3) and (4) show the results for math, reading and science test scores (respectively) as the dependent variable. On average, boys perform better than girls by 0.2 to 0.3 standard deviations in science and math, but worse by 0.3 standard deviations in reading; this is consistent with anecdotal evidence. The net effect of being one month older is slightly negative for all three subjects, with each additional month of age decreasing test scores by 0.01 to 0.02 standard deviations. Foreign born students also perform better than Hong Kong born students by slightly more than 0.1 standard deviations across all three subjects. In general, having more educated parents has a positive effect on test scores. Test scores at larger schools and schools with higher student-teacher ratios are on average higher as well, potentially for the same reasons as effort stated above. Other than for reading, school type has no appreciable effect on test scores.

Table 1.11: Regressions of Effort and Test Score on Controls Only

Dep. Var.:	(1) Effort	(2) Math	(3) Reading	(4) Science
Male	-0.286*** (0.060)	0.280*** (0.033)	-0.262*** (0.031)	0.195*** (0.032)
Age (months)	0.232*** (0.064)	0.184*** (0.046)	0.144*** (0.040)	0.171*** (0.043)
Age in Grade (months)	-0.234*** (0.064)	-0.203*** (0.046)	-0.156*** (0.040)	-0.185*** (0.044)
Foreign Born	0.066 (0.068)	0.108*** (0.032)	0.133*** (0.032)	0.142*** (0.035)
Parent Edu.				
0	-0.168 (0.145)	-0.027 (0.063)	-0.122* (0.066)	-0.144** (0.068)
1	-0.275*** (0.104)	-0.195*** (0.043)	-0.184*** (0.044)	-0.291*** (0.044)
2	-0.287*** (0.096)	-0.144*** (0.042)	-0.138*** (0.041)	-0.193*** (0.041)
3	-0.176** (0.085)	-0.129*** (0.038)	-0.118*** (0.037)	-0.206*** (0.039)
4	-0.071 (0.089)	-0.048 (0.039)	-0.086** (0.038)	-0.071* (0.038)
School Size (1000)	0.942*** (0.337)	0.878*** (0.254)	0.783*** (0.238)	0.807*** (0.253)
STR	0.085*** (0.027)	0.146*** (0.020)	0.140*** (0.019)	0.154*** (0.021)
School Type				
1: Gov. Aided	-0.054 (0.259)	0.323 (0.359)	0.424* (0.237)	0.295 (0.274)
2: Private	-0.235 (0.271)	0.398 (0.37)	0.490** (0.247)	0.356 (0.283)
Constant	-39.37*** (11.03)	-36.12*** (7.941)	-29.12*** (6.951)	-33.99*** (7.485)
<i>N</i>	7403	7697	7697	7697
<i>R</i> ²	0.102	0.334	0.341	0.326
Occ. Status	Yes	Yes	Yes	Yes
Grade FEs	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Standard errors clustered at the year-school level are in parentheses. Parent's education expressed in ISCED categories; the omitted category is college (some/completed, ISCED 5 or 6). The omitted category for school type is public.

1.10.3 Additional Coefficient Estimates of Main Results

The following tables show coefficient estimates of all controls which were suppressed in the main results.

Table 1.12 (and its continuation) shows the coefficient estimates corresponding to the regression specifications in Table 4 of the paper. Occupation status category dummies are not shown in this table because of space constraints.

The coefficient estimates show a similar pattern to those found in the “controls only” regressions. Boys in general exert less effort than girls. The net effect of one additional month in age is close to zero. Students with more educated parents exert more effort. Students in larger schools and schools with higher student teacher ratios work harder, potentially for the same reasons as stated above.

Table 1.13 (and its continuation) shows the coefficient estimates corresponding to the regression specifications in Table 5 of the paper. As before, occupation status category dummies are not shown in this table because of space constraints.

Again, the coefficient estimates show a similar pattern to those found in the “controls only” regressions. Boys score higher in math and science, whereas girls perform better in reading. The net effect of an additional month in age is slightly negative. Foreign born students and students with more educated parents in general score higher in all subjects. Lastly, students in larger schools and schools with higher student teacher ratios are observed to receive higher scores.

Table 1.12: Changes in Effort of Dragon Cohort by Quartile

Dep. Var.: Effort	(1) Base	(2) Controls	(3) Dec/Jan	(4) Sch. FE
D.C.×Q1	-0.050 (0.193)	-0.028 (0.192)	-0.354 (0.491)	0.045 (0.202)
D.C.×Q2	0.129 (0.211)	0.150 (0.212)	0.424 (0.552)	0.270 (0.234)
D.C.×Q3	0.188 (0.207)	0.190 (0.211)	-0.207 (0.449)	0.309 (0.218)
D.C.×Q4	0.499** (0.223)	0.517** (0.226)	0.647 (0.444)	0.653*** (0.220)
Q2	0.414 (0.288)	-0.040 (0.292)	-0.378 (0.690)	-0.525* (0.309)
Q3	0.746** (0.296)	-0.022 (0.333)	0.190 (0.724)	-0.967*** (0.361)
Q4	0.643** (0.293)	-0.397 (0.358)	-0.591 (0.756)	-1.110*** (0.395)
Gr. 10×Q2	0.249 (0.181)	0.247 (0.186)	0.539 (0.487)	0.279 (0.194)
Gr. 10×Q3	0.225 (0.175)	0.256 (0.178)	0.238 (0.417)	0.286 (0.179)
Gr. 10×Q4	0.185 (0.18)	0.214 (0.183)	0.856** (0.425)	0.322* (0.192)
2006×Q2	-0.283 (0.256)	-0.246 (0.228)	0.077 (0.599)	0.168 (0.302)
2006×Q3	-0.434 (0.271)	-0.344 (0.255)	-0.386 (0.621)	0.611* (0.364)
2006×Q4	0.002 (0.263)	0.143 (0.245)	0.400 (0.529)	1.041*** (0.397)
Grade 10	0.288** (0.114)	-2.601*** (0.816)	-1.735 (2.334)	-2.860 (7.240)
Year 2006	-0.541** (0.218)	-0.357** (0.179)	-0.580 (0.390)	
<i>N</i>	7403	7403	1201	7403
<i>R</i> ²	0.068	0.105	0.195	0.205
Controls	No	Yes	Yes	Most
Occ. Status	No	Yes	Yes	Yes
Sch. FEs	No	No	No	Yes

Table continues...

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Effort is measured in hours per week. Standard errors clustered at the year-school level are in parentheses.

Including school fixed effects subsumes year fixed effects and school-level controls.

Table 1.12: Changes in Effort of Dragon Cohort by Quartile (Continued)

Dep. Var.: Effort	(1) Base	(2) Controls	(3) Dec/Jan	(4) Sch. FE
Male		-0.281*** (0.059)	-0.509*** (0.125)	-0.137*** (0.051)
Age (months)		0.234*** (0.066)	0.120 (0.114)	0.255 (0.603)
Age in Grade (months)		-0.236*** (0.067)	-0.133 (0.201)	-0.255 (0.603)
Foreign Born		0.082 (0.072)	0.271 (0.185)	0.121* (0.066)
Parent Edu.				
0		-0.211 (0.158)	-0.544 (0.361)	-0.082 (0.167)
1		-0.317*** (0.119)	-0.142 (0.312)	-0.254** (0.125)
2		-0.328*** (0.110)	-0.124 (0.267)	-0.247** (0.116)
3		-0.212** (0.100)	-0.016 (0.242)	-0.168 (0.109)
4		-0.087 (0.093)	-0.164 (0.248)	-0.048 (0.099)
School Size (1000)		1.059*** (0.348)	1.218** (0.564)	
STR		0.085** (0.033)	0.000 (0.061)	
School Type				
1: Gov. Aided		-0.063 (0.252)	0.992** (0.493)	
2: Private		-0.211 (0.261)	0.997* (0.511)	
Constant	2.232*** (0.226)	-39.73*** (11.54)	-20.28 (19.93)	-40.87 (104.80)
<i>N</i>	7403	7403	1201	7403
<i>R</i> ²	0.068	0.105	0.195	0.205
Occ. Status	No	Yes	Yes	Yes
Sch. FEs	No	No	No	Yes

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Effort is measured in hours per week. Standard errors clustered at the year-school level are in parentheses.

Including school fixed effects subsumes year fixed effects and school-level controls. Parent's education expressed in ISCED categories; the omitted category is college (some/completed, ISCED 5 or 6). The omitted category for school type is public.

Table 1.13: Changes in Effort of Dragon Cohort by Quartile

Dep. Var.:	(1)	(2)	(3)	(4)	(5)	(6)
Test Score	Math	Math	Reading	Reading	Science	Science
D.C.×Q1	0.042 (0.126)	0.069 (0.114)	-0.058 (0.12)	-0.058 (0.117)	0.043 (0.126)	0.07 (0.115)
D.C.×Q2	0.168* (0.089)	0.160* (0.088)	-0.003 (0.086)	0.000 (0.082)	0.055 (0.092)	0.036 (0.089)
D.C.×Q3	0.192** (0.079)	0.177** (0.075)	0.117 (0.089)	0.137* (0.081)	0.198** (0.086)	0.189** (0.082)
D.C.×Q4	0.250*** (0.088)	0.224*** (0.077)	0.269*** (0.085)	0.211*** (0.077)	0.262*** (0.088)	0.228*** (0.081)
Q2	0.466*** (0.178)	-0.048 (0.174)	0.549*** (0.173)	-0.027 (0.161)	0.602*** (0.179)	0.077 (0.174)
Q3	0.868*** (0.172)	0.101 (0.19)	0.853*** (0.17)	-0.034 (0.178)	0.906*** (0.176)	0.097 (0.194)
Q4	1.076*** (0.18)	-0.048 (0.209)	0.947*** (0.171)	-0.133 (0.195)	1.128*** (0.181)	-0.013 (0.212)
Gr. 10×Q2	0.208** (0.102)	0.197* (0.101)	0.066 (0.102)	0.098 (0.099)	0.068 (0.106)	0.051 (0.106)
Gr. 10×Q3	0.082 (0.094)	0.073 (0.09)	0.011 (0.099)	0.103 (0.095)	0.031 (0.103)	0.033 (0.1)
Gr. 10×Q4	0.175* (0.102)	0.193** (0.093)	0.113 (0.099)	0.154* (0.092)	0.097 (0.105)	0.11 (0.099)
2006×Q2	-0.008 (0.174)	0.066 (0.153)	-0.027 (0.172)	0.047 (0.142)	-0.03 (0.174)	0.046 (0.15)
2006×Q3	-0.094 (0.178)	0.034 (0.162)	-0.075 (0.178)	0.032 (0.156)	-0.034 (0.179)	0.094 (0.163)
2006×Q4	-0.03 (0.184)	0.148 (0.164)	0.108 (0.177)	0.197 (0.155)	0.031 (0.181)	0.184 (0.16)
Grade 10	0.458*** (0.082)	-1.966*** (0.597)	0.394*** (0.078)	-1.558*** (0.53)	0.469*** (0.085)	-1.762*** (0.569)
Year 2006	-0.031 (0.158)	0.092 (0.127)	0.335** (0.158)	0.439*** (0.121)	0.028 (0.16)	0.132 (0.126)
<i>N</i>	7697	7697	7697	7697	7697	7697
<i>R</i> ²	0.255	0.339	0.247	0.343	0.249	0.329
Controls	No	Yes	No	Yes	No	Yes
Occ. Status	No	Yes	No	Yes	No	Yes
Sch. FEs	No	No	No	No	No	No

Table continues...

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Test scores are standardized with mean 0 and standard deviation 1. Standard errors clustered at the year-school level are in parentheses. Including school fixed effects subsumes year fixed effects and school-level controls.

Table 1.13: Changes in Effort of Dragon Cohort by Quartile (Continued)

Dep. Var.:	(1)	(2)	(3)	(4)	(5)	(6)
Effort	Math	Math	Reading	Reading	Science	Science
Male		0.280*** (0.033)		-0.262*** (0.031)		0.197*** (0.032)
Age (months)		0.184*** (0.049)		0.147*** (0.043)		0.173*** (0.046)
Age in Grade (months)		-0.203*** (0.049)		-0.158*** (0.043)		-0.186*** (0.047)
Foreign Born		0.101*** (0.034)		0.127*** (0.034)		0.134*** (0.037)
Parent Edu.						
0		0.026 (0.08)		-0.091 (0.081)		-0.092 (0.083)
1		-0.155** (0.061)		-0.159** (0.062)		-0.253*** (0.06)
2		-0.106* (0.058)		-0.112* (0.059)		-0.157*** (0.058)
3		-0.097* (0.052)		-0.096* (0.052)		-0.176*** (0.052)
4		-0.033 (0.043)		-0.074* (0.042)		-0.058 (0.041)
School Size (1000)		0.801*** (0.263)		0.749*** (0.246)		0.730*** (0.266)
STR		0.122*** (0.026)		0.126*** (0.023)		0.131*** (0.026)
School Type						
1: Gov. Aided		0.382 (0.357)		0.467* (0.243)		0.349 (0.277)
2: Private		0.443 (0.368)		0.518** (0.252)		0.396 (0.286)
Constant	-0.992*** (0.156)	-35.92*** (8.394)	-1.105*** (0.149)	-29.29*** (7.398)	-1.058*** (0.157)	-34.03*** (7.955)
<i>N</i>	7697	7697	7697	7697	7697	7697
<i>R</i> ²	0.255	0.339	0.247	0.343	0.249	0.329
Occ. Status	No	Yes	No	Yes	No	Yes
Sch. FEs	No	No	No	No	No	No

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Test scores are standardized with mean 0 and standard deviation 1. Standard errors clustered at the year-school level are in parentheses. Including school fixed effects subsumes year fixed effects and school-level controls. Parent's education expressed in ISCED categories (see variable definitions in Appendix); the omitted category is college (some/completed, ISCED 5 or 6). The omitted category for school type is public.

1.11 Selection Issues

In order to further investigate the phenomenon of parents selecting childbirth into the dragon year, I obtain monthly Hong Kong birth data from various issues of the Hong Kong Monthly Digest of Statistics published by the Hong Kong Census and Statistics Department.

Figure 1.16 shows births by month for the dragon year beginning in February 1988 (bars), and the average of other months for surrounding years (line). The x-axis runs from February to January so as to align it with the Chinese lunar calendar. The number of births in months towards the beginning of the dragon year are very similar to those of surrounding years; these are the months used in the sample of analysis. The number of births in months towards the end of the dragon year are higher than those of surrounding years. This suggests that selection into the dragon year occurs mainly for births towards the end of the dragon year, probably because (superstitious) discerning parents want to “play it safe” and ensure their child is a dragon baby. This also suggests that selection of births into the dragon year are “new” births on the extensive margin rather than shifting of births from surrounding years into the dragon year on the intensive margin.

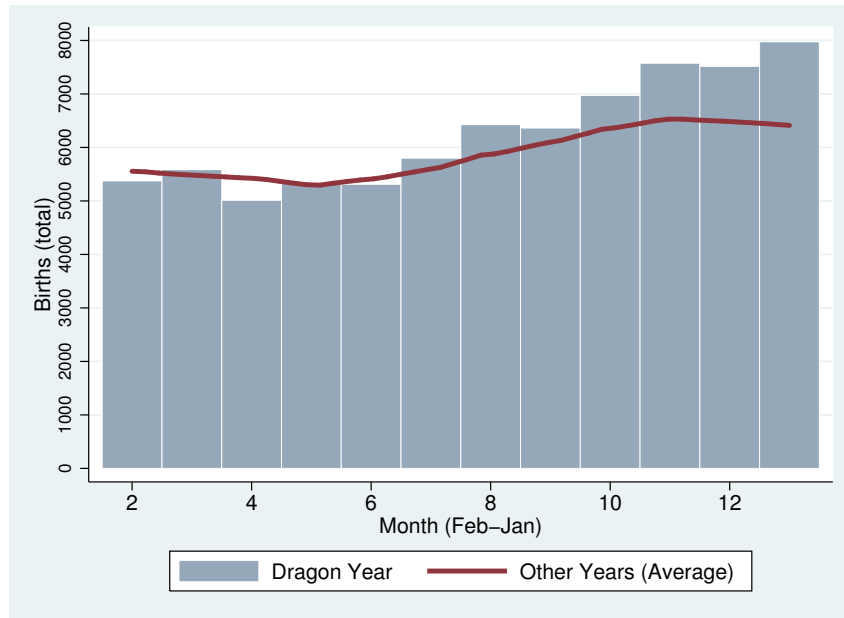
Figure 1.17 shows similar trends for female and male births separately.

1.12 Teacher Reallocation Details

This model estimates the following two parameters: the teacher reallocation factor and the number of teachers per class. Recall that ρ denotes the teacher reallocation factor, where $\rho = 1$ implies a policy of total reallocation of teachers across grades (i.e., teachers are flexible) and $\rho = 0$ implies no reallocation (i.e., teachers can only teach a certain grade, perhaps due to lack of training or institutional rules). Let TPC denote the number of teachers per class.⁶³ As stated, I assume that ρ and TPC are parameters which are similar across all grades and

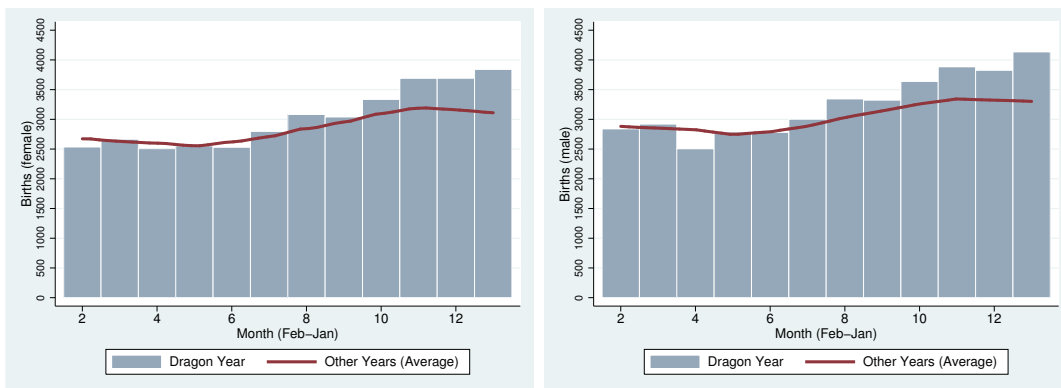
⁶³For example, if a class is assigned a single teacher to teach all subjects, then $TPC = 1$. If two classes share four teachers who each teach separate subjects, then $TPC = 2$.

Figure 1.16: Dragon Year Births by Month (Total)



Note: Dragon year bars are for dragon year starting in February 1988. Average of other years is over 1985-1987, 1989-1991, and January 1988.

Figure 1.17: Dragon Year Births by Month (Female and Male)



Note: Dragon year bars are for dragon year starting in February 1988. Average of other years is over 1985-1987, 1989-1991, and January 1988.

schools.⁶⁴

Let the number of teachers allocated to grade $g \in \mathbb{G}$ at school s be given by $T_{gs}(\rho)$, which is a function of the reallocation factor ρ . If there is total reallocation of teachers, then

$$T_{gs}(\rho = 1) = \frac{B_{gs}}{\sum_{g \in \mathbb{G}} B_{gs}} T_s^{tot}$$

where

- B_{gs} is the total number of students in grade g at school s (batch size)
- T_s^{tot} is the total number of teachers at school s

That is, the teachers are divided according to the proportion of students in each grade.⁶⁵

On the other hand, if there is no reallocation of teachers ($\rho = 0$), then

$$T_{gs}(\rho = 0) = \frac{1}{G_s} T_s^{tot}$$

where $G_s = |\mathbb{G}|$ is the number of grades at school s . That is, teachers are divided evenly across the grades. If the actual teacher reallocation policy is somewhere in between, then

$$\begin{aligned} T_{gs}(\rho) &= \rho T_{gs}(\rho = 1) + (1 - \rho) T_{gs}(\rho = 0) \\ &= T_s^{tot} \left[\rho \left(\frac{B_{gs}}{\sum_g B_{gs}} \right) + (1 - \rho) \frac{1}{G_s} \right] \end{aligned} \quad (1.11)$$

Let C_{gs} be the class size in grade g in school s . Assume that within a school, students in a particular grade are evenly distributed among the classes. The following identity holds for any grade and school.

$$C_{gs} = \frac{B_{gs}}{T_{gs}(\rho) / TPC}$$

That is, the number of students in a class in grade g at school s must be equal to the total

⁶⁴See footnote 54.

⁶⁵I will abstract away from the fact that this and other numbers will not be an exact integer.

number of students in grade g at that school divided by the number of grade g classes at that school (which is given by the number of teachers assigned to that grade in that school divided by the number of teachers per class). Substituting in equation (1.11) gives

$$C_{gs} = \frac{B_{gs} \times TPC}{T_s^{tot} \left[\rho \left(\frac{B_{gs}}{\sum_g B_{gs}} \right) + (1 - \rho) \frac{1}{G_s} \right]}$$

The data contain C_{gs} , T_s^{tot} , $\sum_g B_{gs}$, and G_s at the student level for each PISA round. However, the exact number of students in each grade at a particular school (B_{gs}) is not observed in the data. Using the overall Hong Kong cohort size for each grade (n_g) in each PISA round year, a “best guess” of this number is

$$B_{gs} \approx \frac{n_g}{\sum_g n_g} \left(\sum_g B_{gs} \right)$$

Thus

$$C_{gs} = \frac{\frac{n_g}{\sum_g n_g} \left(\sum_g B_{gs} \right) \times TPC}{T_s^{tot} \left[\rho \left(\frac{n_g}{\sum_g n_g} \right) + (1 - \rho) \frac{1}{G_s} \right]}$$

I estimate the parameters ρ and TPC in this last equation using non-linear least squares. Table 1.14 shows the results for two samples. Column (1) uses only survey responses from students in grades 9 and 10 in PISA round 2003, the most similar sample to the previous analysis. Column (2) uses survey responses from students across all available grades in PISA rounds 2002, 2003 and 2009. (No responses for class size were available for 2006.) The results indicate that the teacher reallocation factor is approximately $\rho = 0.9$, suggesting a high degree of teacher reallocation between grades for every year in the data. The estimate of $TPC = 2$ suggests that there are on average two teachers per class, which seems reasonable.

Table 1.14: Teacher Reallocation Parameter Estimates

	(1)	(2)
Sample:	Grades 9-10, 2003	2002 + 2003 + 2009
ρ	0.902*** (0.047)	0.941*** (0.027)
TPC	1.997*** (0.015)	2.034*** (0.009)
N	3589	13060
R^2	0.973	0.973

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Robust standard errors are in parentheses.

2 Tournament Structure and Effort in an Experimental Setting

Abstract

Tournament structure, in terms of the prize amount and the location of the relative winning cutoff, may affect effort decisions. This chapter presents a tournament model which predicts that receiving larger prizes and being nearer to the relative cutoff increases effort exertion. I test these predictions experimentally by recruiting participants to perform real effort memorization tasks over multiple rounds. Between rounds, I manipulate the tournament structure in order to induce effort changes. I find that increasing the prize amount is effective in incenting additional effort. I also find that low-performing participants increase (decrease) effort when the relative winning cutoff is shifted nearer to (further from) them, but high-performing participants do not respond in this manner when the cutoff is shifted nearer (further).

I would like to thank Alexandre Mas for his guidance and feedback. The Princeton Laboratory for Experimental Social Science (PLESS) played a crucial part in this project, providing the space and resources necessary to conduct the experiments. I am especially grateful to J. Baxter Oliphant, the lab manager at PLESS, whose dedication and advice ensured the smooth running of the experiments. I would also like to thank the attendees of the Public Finance Working Group and the Summer Labor Lunch Series for their comments and suggestions. Funding for this project was generously provided by the Industrial Relations Section at Princeton University. Lastly, I would like to thank all the participants of the experiment. All experimental procedures were approved by Princeton's Institutional Review Board for Human Subjects (No. 6230).

2.1 Introduction

Tournaments are reward mechanisms in which relative—rather than absolute—position determines the winner. Tournaments can be found in many everyday settings; for instance, in sports (and other competitions), in firms (personnel promotions, retail sales contests), and in education (university admissions). In all of these settings, tournaments can be defined by three basic structural parameters: the prize amount, the relative cutoff past which winners receive the prize, and the absolute number of players competing in the tournament.

The structure of a tournament can affect effort decisions and outcomes. Considerations as to how to parameterize the structure of a tournament are especially relevant in settings in which tournament organizers want to induce effort exertion or specific outcome goals. For example, educators may structure a tournament in order to incent (certain) students to study harder or obtain higher test scores. Knowing how tournament structure affects these effort decisions and outcomes will be useful in formulating tournaments.

This chapter uses experimental tournaments in a controlled lab setting to investigate how adjusting tournament structure affects the effort decisions of participants. In particular, I focus on manipulating the prize amount and the relative winning cutoff. In each round of the experiment, groups of participants were asked to memorize and recall lists of word-number pairs on a computer screen. There are a total of five rounds in each experimental session, and a different list of word-number pairs is presented in each round. At each participant's own discretion, an amount of time is spent memorizing this list. When satisfied, the participant clicks to the next screen, where the same list of words (but reordered randomly) is presented with blank text boxes next to each word. The goal of the task is to recall as many numbers as possible (each corresponding to separate words). A subset of participants with the most number of correct answers (relative to all participants in the session) wins a monetary prize. The prize dollar amount and proportion of participants receiving the prize is changed from round to round, thus varying the tournament structure. Effort is measured as the amount

of time spent memorizing the list chosen by each participant.

In the economics literature, tournament theory was brought to prominence by the seminal work of Lazear and Rosen (1981), and later expanded upon by Green and Stokey (1983), among others. Since then, many empirical papers have analyzed tournaments in various natural experiment settings, including software programming contests⁶⁶, retail sales⁶⁷, personnel economics within the firm⁶⁸, and university admissions⁶⁹.

A separate strand of literature has focused on using experimental methods to analyze tournaments in controlled settings. Most of this research has focused on how using a tournament reward mechanism compares to using a piecerate one, in which prizes are awarded proportional to outcomes. Bull et al. (1987) randomize participants into two such reward mechanisms to study how effort choices differ under the two schemes. They find that mean effort is similar in both cases, but variance in effort is higher in tournaments. Their study elicits effort by having pairs of participants choose “decision numbers” from cost and payoff tables; the “effort decisions” of competing pairs (plus a random component) are then compared against one another with the winner receiving a prize, less an effort cost. Higher decision numbers are associated with increasing costs, but offer higher expected payoffs. Use of such “cost and payoff tables” are common in the experimental tournament literature. In a subsequent paper, Schotter and Weigelt (1992) add unfair aspects to similar experimental tournaments and analyze the impact of fairness-restoring policies.

Research by Orrison et al. (1997) and Harbring and Irlenbusch (2003) focus on tournament (as opposed to piecerate) reward structures and investigate how varying the number of competitors and the prize structure affect effort decisions. Both these papers use cost and payoff tables in their experimental setup. Orrison et al. (1997) find that altering the number of competitors from 2 to 4 to 6 does not change mean effort. They also obtain results for

⁶⁶Boudreau et al. (2012)

⁶⁷Casas-Arce and Martinez-Jerez (2009)

⁶⁸Main et al. (1993); Eriksson (1999)

⁶⁹See Chapter 1.

changing the proportion of relatively larger prizes, and using discriminatory cost and payoff tables. Harbring and Irlenbusch (2003) find that mean effort increases as the proportion of prize winners increases, but that variance in effort decreases.

Using cost and payoff tables, however, has the disadvantage of external validity. In many real-life tournaments, the decision to exert effort often incurs more tangible costs of the physically and cognitively taxing sort. More recent experimental research involving human subjects devise what experimentalists call “real effort tasks” in order to elicit effort. Dijk et al. (2001) use a grid computer game task in which participants search for prizes within a grid of cells as a real effort task to compare how different compensation schemes—including individual (i.e. piece-rate), team, and tournaments—incent effort exertion. They find similar effort levels in individual and team compensation schemes, but higher and more variable effort levels in tournaments.

The contribution of this study is to use a real effort task to investigate the effects of changing tournament structure parameters—namely the prize amount and the relative winning cutoff—on effort decisions. To the best of my knowledge, previous papers looking at tournament structure have only used cost and payoff tables or similar non-real effort measures to elicit effort. In the experiment, I use time spent memorizing the list of word-number pairs as my measure of effort. This innovation allows me to measure effort precisely without the risk of significant Hawthorne effects, as it is not immediately apparent during the experiment that this is the outcome of interest.

The remainder of the chapter is presented as follows. In Section 2.2, I present a tournament model and discuss the comparative statics of changing tournament structure. Next, I describe the experimental procedure in Section 2.3. I then present the results in Section 2.4 and conclude in Section 2.5.

2.2 Model

In this section, I explore how the structure of a tournament—in particular the (relative) location of the relative winning cutoff and the prize amount—influences effort decisions. I will set up a tournament model to describe the effort decisions of experiment participants. In this framework, the score a participant obtains will be a function of ability and effort. Participants choose effort to maximize their expected prize less an effort cost.⁷⁰ Furthermore, a constraint will be added to account for the fact that there is a maximum obtainable score which cannot be surpassed—corresponding to the experiment conducted, in which the maximum score attainable was 18.

Let θ_i denote the ability of participant i , drawn from the ability distribution $F(\theta_i)$. Let e_i denote the effort choice of participant i . Let $(1 - p)$ be the proportion of participant who win the prize α . A participant wins only if his or her score is at least as high as a certain absolute score cutoff $\bar{S}_p$.⁷¹ This $\bar{S}_p$ depends on the p which is set in each round of the experiment. The participant's realized score is described by the function

$$S(\theta_i, e_i) - \eta_i$$

which contains the production function component $S(\theta_i, e_i)$, and a random error component η_i which is i.i.d. and orthogonal to effort and ability. The score production function depends on the participant's ability and effort choice as inputs.

A participant's utility depends on whether the prize is won, less an effort cost which must

⁷⁰This model will be similar to the tournament model of academic competition presented in Chapter 1.

⁷¹Note that this absolute score cutoff, which is the numerical score a participant needs to surpass in order to win the prize, while related, is not the same as the relative winning cutoff, which is the rank a participant needs to surpass in order to win the prize.

be paid prior to observing the realized score. This is given by

$$u(e_i) = \begin{cases} \alpha - c(e_i) & \text{if the participant wins} \\ -c(e_i) & \text{otherwise} \end{cases}$$

I make the following assumptions.

Assumption A. Score production is increasing in effort and non-decreasing in ability. That is,

$$\frac{\partial S(\theta_i, e_i)}{\partial e_i} > 0 \text{ and } \frac{\partial S(\theta_i, e_i)}{\partial \theta_i} \geq 0$$

However, score returns on effort are non-increasing in effort and increasing in ability⁷².

That is,

$$\frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} \leq 0 \text{ and } \frac{\partial^2 S(\theta_i, e_i)}{\partial \theta_i \partial e_i} > 0$$

Assumption B. Effort costs are increasing and strictly convex. That is,

$$c'(e_i) > 0 \text{ and } c''(e_i) > 0$$

These are reasonable assumptions, considering that common functional forms of $S(\theta_i, e_i)$ (e.g., Cobb-Douglas) and $c(e_i)$ (e.g., quadratic costs) satisfy them.

The participant's expected utility conditional on ability and effort choice is

$$\begin{aligned} E[u(e_i) | e_i, \theta_i] &= \alpha \Pr(S(\theta_i, e_i) - \eta_i \geq \bar{S}_p) - c(e_i) \\ &= \alpha H(S(\theta_i, e_i) - \bar{S}_p) - c(e_i) \end{aligned}$$

where $H(\eta_i)$ is the c.d.f. of η_i (and $h(\cdot)$ is the corresponding p.d.f.).

Assumption η . Let the distribution of η be single peaked. Let this peak be (without loss

⁷²This would be so if higher ability participants make better use of their effort exertion.

of generality) at zero.

Furthermore, because of the existence of maximum obtainable score S_{max} , any optimization problem will be subject to the constraint

$$S(\theta_i, e_i) - \eta_i \leq S_{max} \quad (2.1)$$

In choosing effort to maximize expected utility subject to the above constraint, the participant's first order condition (FOC) is

$$\alpha \left[h(S(\theta_i, e_i) - \bar{S}_p) - \lambda \right] \frac{\partial S(\theta_i, e_i)}{\partial e_i} - c'(e_i) = 0 \quad (2.2)$$

where λ is the Lagrange multiplier for the constraint. This equation defines the participant's optimal effort decision e_i^* as a function of ability θ_i . In particular, the closer a participant is to the absolute score cutoff (i.e., the higher $h(S(\theta_i, e_i) - \bar{S}_p)$ is), the more effort he or she will exert in order to overcome the potential of an adverse error, get above the score cutoff, and receive the prize α . This generates a “bump” in effort near the absolute score cutoff.

Moreover, as ability increases and the realized score approaches the maximum score S_{max} , these increasingly higher performing participants exert less and less effort. This is because the λ term increases until $[h(S(\theta_i, e_i) - \bar{S}_p) - \lambda]$ becomes lower (possibly negative), and optimal effort decreases due to the convexity of $c(e_i)$ supposed in Assumption B.⁷³

The second order condition (SOC) is given by

$$\begin{aligned} \alpha h'p(S(\theta_i, e_i) - \bar{S}_p) \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 \\ + \alpha \left[h(S(\theta_i, e_i) - \bar{S}_p) - \lambda \right] \frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} - c''(e_i) < 0 \end{aligned} \quad (2.3)$$

⁷³I abstract away from the fact that the distribution of the error term $h(\cdot)$ will become positively skewed for participants with high score production $S(\theta_i, e_i)$ near the maximum S_{max} . Only the magnitude, but not the sign, of the comparative statics changes in this case.

This is negative as I will only consider interior solutions with $e_i^* > 0$.⁷⁴

Using the implicit function theorem,

$$\frac{de_i}{dp} = - \frac{-\alpha h' (S(\theta_i, e_i) - \bar{S}_p) \frac{\partial S(\theta_i, e_i)}{\partial e_i} \frac{d\bar{S}_p}{dp}}{\alpha h' (S(\theta_i, e_i) - \bar{S}_p) \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 + \alpha [h (S(\theta_i, e_i) - \bar{S}_p) - \lambda] \frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} - c''(e_i)} \quad (2.4)$$

This equation describes the comparative statics of an increase in the proportion of non-winners p . For illustration purposes, I will assume for now that $\frac{d\bar{S}_p}{dp}$ is positive; that is, that the absolute score cutoff increases with an increase in the proportion of non-winners.⁷⁵ (Later, I will discuss the possibility that $\frac{d\bar{S}_p}{dp}$ is negative.) Note also that the denominator is the SOC, which is negative.

Since $\frac{\partial S(\theta_i, e_i)}{\partial e_i} > 0$ under Assumption A, $\frac{de_i}{dp}$ in equation (2.4) is positive if and only if $h' (S(\theta_i, e_i) - \bar{S}_p)$ is negative. Recall that under Assumption η , $h(\cdot)$ is single-peaked at zero. This implies that initially (i.e. prior to the change in p), high-performing participants (with values of $h(S(\theta_i, e_i) - \bar{S}_p)$ to the right of the peak) will increase their effort in response to an increase in the proportion of non-winners; correspondingly, initially low-performing participants (with values of $h(S(\theta_i, e_i) - \bar{S}_p)$ to the left of the peak) will decrease their effort in response to an increase in the proportion of non-winners.

The intuition behind this result is that when the proportion of non-winners increases, initially high-performing participants see the absolute score cutoff move closer to them, thus increasing their risk of falling below the score cutoff due to the error term; in response, they “step up” effort to compete for the relatively fewer number of prizes. On the other hand, initially low-performing participants see the absolute cutoff score move further away from them, decreasing the probability of them ever surpassing it; in response, they “give up” competing for the relatively fewer number of prizes as effort exertion is costly.

⁷⁴In the case where the only solution to the FOC has a positive SOC (i.e., the FOC solution is utility-minimizing), then the corner solution $e_i^* = 0$ is utility-maximizing.

⁷⁵In a related model in Chapter 1, I show that under certain additional assumptions and propositions (relating to the distribution of realized scores) $\frac{d\bar{S}_p}{dp} > 0$.

Similarly,

$$\frac{de_i}{d\alpha} = - \frac{[h(S(\theta_i, e_i) - \bar{S}_p) - \lambda] \frac{\partial S(\theta_i, e_i)}{\partial e_i}}{\alpha h'(S(\theta_i, e_i) - \bar{S}_p) \left(\frac{\partial S(\theta_i, e_i)}{\partial e_i} \right)^2 + \alpha [h(S(\theta_i, e_i) - \bar{S}_p) - \lambda] \frac{\partial^2 S(\theta_i, e_i)}{\partial e_i^2} - c''(e_i)} \quad (2.5)$$

This equation describes the comparative statics of an increase in the prize amount α . For participants not experiencing censoring of scores (i.e. λ is zero), the derivative is positive.⁷⁶ This implies that all participants increase effort in response to an increase in the prize amount. The intuition is that the marginal benefit of winning (for every unit increase in the probability of winning) is now higher, so participants are more willing to exert additional effort and incur a higher marginal cost of effort. In particular, participants around the absolute score cutoff $\bar{S}_p$ will have more to gain (or lose) if they just surpass (or fall below) the cutoff. Hence, their effort increase will be highest (as seen by higher values of $h(S(\theta_i, e_i) - \bar{S}_p)$) compared to participants further away from the absolute score cutoff.

Should top performing participants be capped by S_{max} , then λ becomes positive and $\frac{de_i}{d\alpha}$ decreases. The effort response to the increase in prize amount is smaller because there is no benefit to the capped participant's score in exerting additional effort.

These two sets of predictions from the separate changes in tournament structure relating to $\frac{de_i}{dp}$ and $\frac{de_i}{d\alpha}$ are tested in the experiment to follow.

2.3 Experimental Procedure

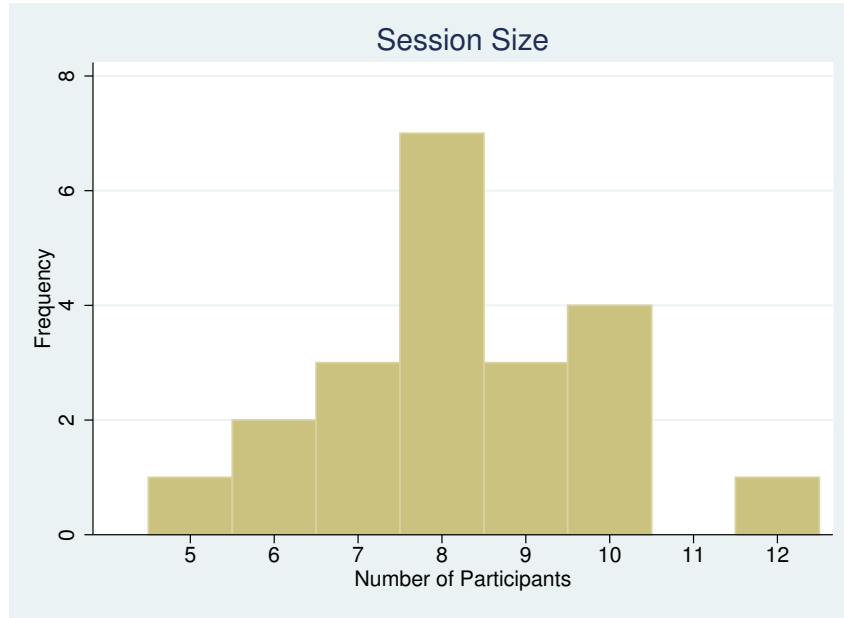
Experiment sessions were conducted at the Princeton Laboratory for Experimental Social Sciences (PLESS) over the course of two weeks in May 2013. A total of 173 participants were recruited through the PLESS mailing list⁷⁷ to participate in 21 sessions.⁷⁸ Participants

⁷⁶The denominator is the SOC, which is negative. Furthermore, $\frac{\partial S(\theta_i, e_i)}{\partial e_i} > 0$ is assumed, and $h(S(\theta_i, e_i) - \bar{S}_p)$ is a p.d.f., so it is positive. With the negative sign in front of the fraction, this implies the derivative is positive.

⁷⁷The exact text of the recruitment email can be found in Appendix 2.6.

⁷⁸These numbers exclude two sessions which were dropped from the data. In chronological order, the very first session was dropped because of a change in experimental procedure implemented in all subsequent sessions. (The prompts after 7 minutes were added.) The third session in chronological order were also dropped because two of the participants cheated, and ejecting them in the middle of the experiment would

Figure 2.1: Histogram of Session Size



included undergraduates, graduate students, and staff members at Princeton University, as well as local community members not necessarily affiliated with the institution. Each session consisted of multiple participants competing against one another for prizes. Additionally, participants were paid a base amount of \$15 for showing up, regardless of performance. The number of participants in each session varied between five and twelve.⁷⁹ Figure 2.1 is a histogram of the size of the 21 sessions.

As they arrived, participants were seated at computer terminals with a screen, keyboard, and mouse. Cubicle-like blinders in between each terminal ensured privacy. Once most participants who had signed up for that timeslot had arrived, the session began. Those who had signed up but were not there were assumed to be no-shows; latecomers were informed that the experiment had already started and encouraged to sign up for another timeslot. Participants who arrived on time and completed any remaining part of the session were disqualified from signing up for another subsequent timeslot by the computer system.

have meant only 4 participants would remain in that session.

⁷⁹This variation in session size arose because for each timeslot, ten or twelve seats were made available for sign up, but either not all were filled, or participants who had signed up did not show up. For all intents and purposes, the session size can be considered random.

A session consisted of five rounds of memorization tasks. Prior to these rounds, instructions were given to participants regarding their task, which is as follows.⁸⁰ In each round, participants are asked to memorize a list of 18 word-number pairs on the first screen. (These numbers are three digits long.) Participants are allowed to choose the amount of time spent on memorizing this list, after which they would click to the next screen. In this second screen, the list of 18 words (reordered randomly) are shown next to text boxes. Participants are asked to fill in as many of the corresponding 3-digit numbers into the text boxes. After submitting their answers, participants are then taken to a third “waiting” screen where they wait for the results of that round to be graded.⁸¹

While participants were free to choose the amount of time spent memorizing and responding to the questions, prompts were given to all participants at the 7, 9, 10, 11, 12, 13 and 14 minute marks, or up until all participants had submitted their responses, whichever being sooner. (Less than 1% of observed time spent were more than 12 minutes.) These prompts were necessary to ensure that the sessions ran on time, because grading of responses could not proceed until answers from all participants had been submitted.

The participants were ranked by the number of correctly recalled word-number pairs; ties were broken based on time spent on the entire task, in favor of faster participants. There was no penalty for guessing. Each participant was then informed privately of their rank (but not the absolute number of correct responses) using small slips of papers. After passing out these “result slips” the participants were then asked to click on a link on the waiting screen to proceed to the next round.

Out of the total of five rounds in each session, only in the last four were participants given the chance to win monetary prizes. The first round (Round 0) was an example round in which no prize was awarded; this round was used to allow participants to familiarize themselves with the interface. In the later four actual rounds (Rounds 1 through 4), prizes

⁸⁰The exact wording of the instructions given to participants at the beginning of each session can be found in Appendix 2.7.

⁸¹Screen-shots of the computer interface can be found in Appendix 2.8.

of \$X were given to the top Y participants, where X and Y varied with each round. These two parameters were also projected onto a screen at the front of the room during the rounds so participants could remind themselves of the current tournament structure in the middle of any round. The result slips also contained this information for the round just completed. Prior to the beginning of the next round, new values of X and Y were announced verbally and then projected on the screen. Changing X and Y from round to round constitutes changing the parameters of the tournament structure. Data from example Round 0 will not be used in the analysis except in some cases as lags.

The integer Y is the relative winning cutoff, and varies both between sessions and within sessions between rounds. Table 2.1 lists the relative winning cutoffs used in each session depending on its size.⁸² For each session, there were two cutoffs: the low cutoff and the high cutoff. Within a session, two out of four rounds were assigned the low cutoff, while the other two rounds were assigned the high cutoff. For example, suppose at the beginning of the session, 9 participants show up. Then during the rounds assigned the low cutoff, the top 6 out of 9 participants receive the prize; and during the rounds assigned the high cutoff, the top 3 out of 9 participants receive the prize. The row for session size of 11 is struck through because none of the sessions had 11 participants.⁸³

As for the size of the prize \$X, two amounts were used: \$5 and \$10. Again, two out of four rounds were assigned the \$5 prize, while the other two rounds were assigned the \$10 prize. As such, each participant (in addition to the \$15 “show-up” payment) had the chance to win an additional \$30, making the maximum possible total payment \$45. Table 2.2 summarizes the four possible combinations of prize amounts and cutoff levels. Since each session comprised four rounds (excluding the example round), there was one round for each combination, as represented by one cell in the table.

⁸²These are different from the absolute score cutoffs, which will vary from session to session and round to round depending on the absolute performance of participants.

⁸³There is a “break” in the assignment of the cutoffs as session size increases from 9 to 10. This is done because of funding limitations.

Table 2.1: Relative Winning Cutoffs by Session Size

Session Size	Low Cutoffs	High Cutoffs
5	4 / 5	2 / 5
6	4 / 6	2 / 6
7	5 / 7	2 / 7
8	5 / 8	3 / 8
9	6 / 9	3 / 9
10	6 / 10	3 / 10
~~11~~	~~7 / 11~~	~~3 / 11~~
12	8 / 12	3 / 12

Note: No sessions had 11 participants. “3/10” means that the top 3 out of 10 participants received prizes.

Table 2.2: Possible Prize-Cutoff Combinations

	Low Cutoff	High Cutoff
\$5 Prize	Low \$5	High \$5
\$10 Prize	Low \$10	High \$10

As the experiment transitioned from one round to the next, I manipulated (only) one of the two tournament structure parameters—either \$X or Y—holding the other parameter constant. A “treatment” is the change in the tournament structure experienced by participants from one round to the next. There are a total of eight possible treatments. Table 2.3 lists these eight possibilities, which I label as +1, -1, +2, -2, etc. I refer to these as treatment codes.

Within a session, the experiment was structured such that participants experienced exactly three treatments as follows:

- From Round 1 to Round 2, I moved the cutoff Y (either up or down)
- From Round 2 to Round 3, I moved the prize \$X (either up or down)
- From Round 3 to Round 4, I moved the cutoff Y again (in the opposite direction

Table 2.3: Treatments

Treatment Code	Description
+1	Cutoff from low to high @ \$5 prize
-1	Cutoff from high to low @ \$5 prize
+2	Prize from \$5 to \$10 @ low cutoff
-2	Prize from \$10 to \$5 @ low cutoff
+3	Cutoff from low to high @ \$10 prize
-3	Cutoff from high to low @ \$10 prize
+4	Prize from \$5 to \$10 @ high cutoff
-4	Prize from \$10 to \$5 @ high cutoff

compared to the first treatment)

The decision to move the cutoff twice per session and the prize amount only once is intended to give such cutoff-movement treatments a larger sample size. This is done because the theoretical model predicts that high-performing participants will respond in the opposite direction compared to low-performing participants. This pattern will require additional precision to identify, compared to the comparative statics of an increase in the prize amount, which is in the same direction for both low- and high-performing participants.

In this manner, there are only four possible sequences of treatments. Table 2.4 lists these four permutations of treatment sequences. A session was assigned to one of four “treatment groups”—A through D—and each group experienced a corresponding treatment sequence, as defined in the table.⁸⁴

Table 2.5 shows summary statistics (means and standard deviations) of participants by treatment group. While sample sizes for treatment groups are small, other than an unusually low number of whites for group A, the statistics are very similar across the groups, suggesting that randomization was successfully carried out.

⁸⁴For example, after the example round, participants in group A first experienced a low cutoff with a \$5 prize in Round 1, then a high cutoff with a \$5 prize in Round 2, next a high cutoff with a \$10 prize in Round 3, and finally a low cutoff with a \$5 prize in Round 4.

Table 2.4: Treatment Sequences and Groups

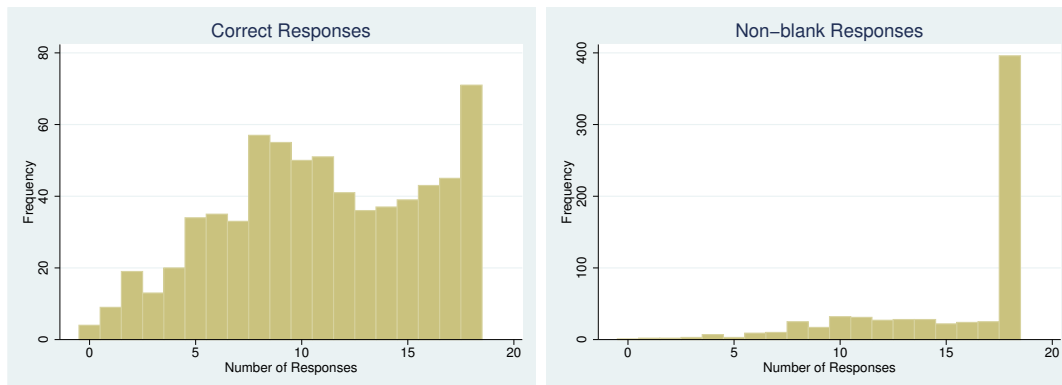
Group	A	B	C	D
Round 0	Example round, no prize			
Round 1	Low \$5	Low \$10	High \$5	High \$10
Round 2	High \$5	High \$10	Low \$5	Low \$10
Round 3	High \$10	High \$5	Low \$10	Low \$5
Round 4	Low \$10	Low \$5	High \$10	High \$5
Treatment Sequence	+1, +4, -3	+3, -4, -1	-1, +2, +3	-3, -2, +1

Table 2.5: Summary Statistics by Treatment Group

Group:	Overall	A	B	C	D
Male	0.384 (0.488)	0.422 (0.499)	0.35 (0.483)	0.333 (0.477)	0.422 (0.499)
White	0.6 (0.491)	0.395 (0.495)	0.757 (0.435)	0.703 (0.463)	0.579 (0.5)
Black	0.11 (0.314)	0.14 (0.351)	0.054 (0.229)	0.135 (0.347)	0.105 (0.311)
Age	25.06 (10.5)	23.04 (6.54)	27.59 (13.63)	25.36 (10.98)	24.57 (9.97)
% College	0.831 (0.376)	0.864 (0.347)	0.886 (0.323)	0.805 (0.401)	0.775 (0.423)
Father	0.747 (0.436)	0.778 (0.42)	0.743 (0.443)	0.756 (0.435)	0.707 (0.461)
% College					
Mother					
N	173	45	40	42	46

Note: Means and proportions within groups shown. Standard deviations in parentheses.

Figure 2.2: Histograms of Responses



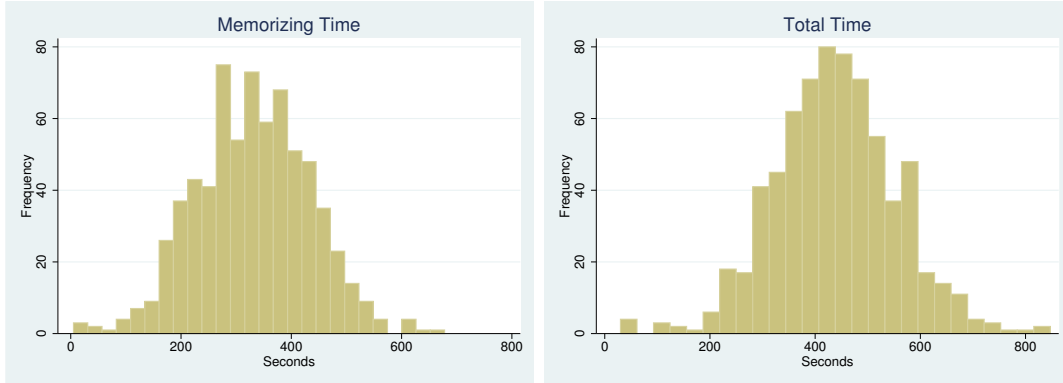
2.4 Results

The ability of participants to successfully complete the memorization and recall task varied greatly. The first panel in Figure 2.2 shows the distribution of correct responses in all rounds. There does seem to be top-coding towards the top of the distribution, which suggests that it may be important to consider a constraint such as the inequality in (2.1) in any theoretical framework. However, a majority of participants obtained results well below this, and adding further word-number pairs beyond the 18 would have increased the time necessary to conduct each experimental session, possibly to an amount that would have been infeasible. The second panel in Figure 2.2, a histogram of non-blank responses, supports this; because there was no penalty for guessing, almost all participants filled in all 18 text boxes in the second screen.⁸⁵

The main outcome of interest is effort as measured by the time taken by participants to memorize the word-number list. Timestamps of exactly when participants clicked from screen to screen are used to obtain the time spent memorizing the list, as well as the time spent filling in the text boxes in the next screen. I call the sum of these two times the total time taken by the participant to complete a round. The two panels in Figure 2.3 show histograms of memorizing time and total time spent in seconds by all participants. Despite

⁸⁵This lack of variation in non-blank responses also invalidates the variable as a measure of effort.

Figure 2.3: Histograms of Time Spent

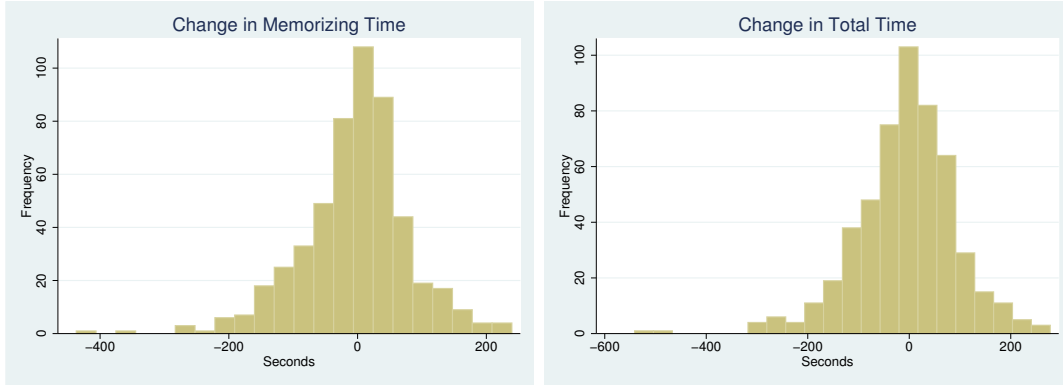


the first prompt given at 7 minutes (420 seconds), and subsequent prompts every minute starting at 9 minutes (540 seconds), there does not seem to be distinct mass points in either distribution at these two particular points in time.⁸⁶

From round to round, as participants are treated with changes in tournament structure, there is considerably variation in the change in time spent. The two panels in Figure 2.4 show histograms of these changes in both memorizing time and total time spent. This variation, which allows for the identification of the estimates to follow, can arise from two sources. Firstly, the variation can come from the treatments (changes in tournament structure) induced in the experimental procedure, which is what we are trying to identify. Secondly, as participants proceed from round to round, there may be learning in the sense that they get better at recalling the word-number pairs or more comfortable with the interface and environment. While the former source of variation will cause both positive and negative changes in time spent, the latter source causes only negative changes as participants require less time to complete the same task. Thus, in all the regressions to follow, I include round fixed effects to control for such learning.

⁸⁶Since prompts were consistently applied to all groups, it can merely be thought of as an additional cost of effort for high levels of effort (longer time spent), insofar as the prompts create discomfort for those still working on their responses.

Figure 2.4: Histograms of Change in Time Spent



2.4.1 Does effort affect outcomes?

Additional effort exertion can affect outcomes in two ways: spending more time on the task increases the number of correct responses; it can also potentially improve a participant's relative ranking. In order to investigate whether these effects exist, and whether memorizing time or total time spent (in seconds) is a better measure of effort, I regress both score and percentile rank on these two time measures, using specifications of the form

$$y_{it} = \alpha + \beta Time_{it} + \mu_t + \mu_i + \varepsilon_{it}$$

where

- y_{it} is the score (correct responses out of 18) or percentile rank (out of 100) of participant i in round t
- $Time_{it}$ is one or more measures of time
- μ_t are round fixed effects
- μ_i are participant fixed effects
- ε_{it} is an error term

The inclusion of round fixed effects controls for learning effects in the coefficient estimates of the time measures⁸⁷, and the relative size of the fixed effects can also shed light on how learning evolves as the rounds progress. Furthermore, including participant fixed effects accounts for individual ability as well as any effects which are fixed within session.

Table 2.6 shows the results for regressions of score (i.e. number of responses correct) on the time measures. The specification of column (1) regresses score on memorizing time. As expected, increasing memorizing time by 1 second increases a participant's score by 0.012 on average (significant at 1% level). The specification of column (2) regresses score on total time. Again, increasing total time spent by 1 second increases a participant's score by 0.009 on average (significant at 1% level).

In order to determine which time measure is a better measure of effort, the specification in column (3) regresses score on both memorizing and total time. Here, we see that only the coefficient of 0.016 on memorizing time is statistically significant (at a 1% level), implying that it is really time spent memorizing—and not the time spent entering responses into the text boxes—that has an effect on score.⁸⁸ As such, using memorizing time as the measure of effort seems most appropriate.

To understand learning effects from round to round, we look at the coefficient estimates on the round fixed effects. Note that Round 1 is the omitted category.⁸⁹ Estimates in columns (1) through (3) are qualitatively similar, all indicating that there is an initially steep learning curve which tapers off in later rounds. Moving from Round 1 to 2, participants' scores improved by about 0.7 to 0.8 responses on average, *ceteris paribus*. However, moving from Round 2 to 3 only improves scores by about 0.01 to 0.06 responses on average. There seems to be a further improvement from Round 3 to 4 of about 0.3 to 0.4 responses. All

⁸⁷This assumes that learning improves outcomes in level terms (e.g. each additional round increases the number of correct responses by Z responses), rather than improving outcomes through increases in the marginal benefit of each additional second of time spent on the task.

⁸⁸This also suggests that the significant coefficient in column (2) arises from the correlation between memorizing and total time.

⁸⁹Also recall that there is an example round, so Round 2 is actually the third round encountered by the participant.

Table 2.6: Score Regressions

Dep. Var.:	(1)	(2)	(3)	(4)
Score	Mem. Time	Total Time	Both	Top / Bottom
Mem. Time	0.012*** (0.001)		0.016*** (0.004)	0.012*** (0.001)
Total Time		0.009*** (0.001)	-0.003 (0.004)	
Mem. Time × Top Round				-0.001 (0.001)
2	0.758*** (0.204)	0.834*** (0.210)	0.732*** (0.202)	0.737** (0.334)
3	0.778*** (0.202)	0.892*** (0.205)	0.740*** (0.199)	0.505 (0.313)
4	1.167*** (0.201)	1.213*** (0.206)	1.144*** (0.200)	0.955*** (0.309)
Round×Top				
2×Top				0.038 (0.447)
3×Top				0.531 (0.431)
4×Top				0.435 (0.400)
Constant	6.217*** (0.475)	6.247*** (0.541)	6.465*** (0.536)	6.399*** (0.474)
N	692	692	692	692
R-square	0.877	0.871	0.877	0.878

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Time variables are in seconds. All specifications include participant fixed effects. Robust standard errors are in parentheses.

these significantly positive coefficients indicate that there is a good degree of learning as the rounds progress.⁹⁰

We can also check whether there is heterogeneity in learning or returns to effort. To do so, I create a “top” indicator which takes on the value of 1 when a participant’s percentile rank from the previous round is above 50%. Henceforth, I will refer to such “top” participants as high-performing participants (as opposed to low-performing participants). I then interact this top indicator with all the right hand side variables of interest. Column (4) in Table 2.6 shows the coefficient estimates for the specification where score is regressed on memorizing time, its interaction with the top indicator, round fixed effects, and their interactions with the top indicator.

The baseline coefficient on memorizing time is identical to the one in column (1), the specification without the added interaction terms. The coefficient on the interaction of memorizing time and the top indicator is small and statistically insignificant. This implies that the returns to effort of low-performing participants (as reflected in results from the previous round) is no worse than that of high-performing participants.

Furthermore, while the coefficients on the baseline round fixed effects are quantitatively similar to those in column (1) (except for the one for Round 3, which becomes statistically insignificant), none of the coefficients on the interacted fixed effects are statistically different from zero. This implies that both high- and low-performing participants learned at similar rates as the rounds progressed.

Instead of focusing on absolute score, we can also examine relative percentile rank as an outcome of effort exertion and learning. Table 2.7 shows the results for regressions of percentile rank (out of 100) on the time measures. The specification of column (1) regresses percentile rank on memorizing time. The coefficient estimate suggests that spending 1 more

⁹⁰Notice that the estimates are average learning effects averaged over all treatments and treatment groups. Also note that these estimates are conditional on memorizing time, which means that if memorizing time is affected by the various treatments (as will be shown later), then these coefficients are measuring learning independent of any round to round treatment effects.

Table 2.7: Percentile Rank Regressions

Dep. Var.:	(1)	(2)	(3)	(4)
Percentile Rank	Mem. Time	Total Time	Both	Top / Bottom
Mem. Time	0.030*** (0.010)		0.118*** (0.030)	0.081** (0.034)
Total Time		0.012 (0.008)	-0.080*** (0.025)	-0.048* (0.027)
Mem. Time × Top				0.101** (0.040)
Total Time × Top				-0.085*** (0.030)
<u>Round</u>				
2	-0.013 (1.589)	0.094 (1.599)	-0.678 (1.569)	0.436 (2.347)
3	-0.020 (1.460)	0.139 (1.478)	-1.005 (1.472)	0.353 (2.161)
4	0.330 (1.628)	0.236 (1.657)	-0.290 (1.617)	0.560 (2.281)
<u>Round × Top</u>				
2 × Top				-2.441 (3.184)
3 × Top				-2.585 (2.897)
4 × Top				-1.899 (3.091)
Constant	45.901*** (3.464)	50.705*** (3.824)	52.352*** (3.748)	52.499*** (3.943)
N	692	692	692	692
R-square	0.811	0.808	0.817	0.823

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Time variables are in seconds. Percentile rank is out of 100. All specifications include participant fixed effects. Robust standard errors are in parentheses.

second on memorizing the word-number list increases a participant's percentile rank by 0.03% on average.

However, regressing percentile rank on total time instead in column (2) yields a statistically insignificant coefficient estimate. The reason for this is explained by the results in

column (3), for the specification in which percentile rank is regressed on both memorizing time and total time. The coefficient on memorizing time is positive and statistically significant at 0.12. Combining this with the coefficient on total time, this implies that a 1 second increase in memorizing time increases percentile rank by 0.04%, similar to the estimate in column (1). While spending more time memorizing does seem to move participants up the rankings, spending more time in total (memorizing and responding) has an averse effect on percentile rank. The negative and statistically significant coefficient suggests that 1 more second of total time (or, equivalently, time spent filling in text boxes) reduces percentile rank by 0.08%. This negative effect is most likely the result of ties being broken in favor of participants who complete the entire task (memorizing and responding) faster.

Column (4) in Table 2.7 shows results for the specification where all the right hand side variables in column (3) are interacted with the top indicator. The coefficient on the interaction term between memorizing time and the top indicator is positive and statistically significant (at a 5% level). Moreover, the interaction term between total time and the top indicator is negative and statistically significant (at a 1% level). These results suggest two things.

First, high-performing participants (as reflected in results from the previous round) have better relative ranking returns to effort compared to low-performing ones—by about 0.015% per second of memorizing time in percentile rank terms. The fact that this difference is not observed in returns to effort in score may imply that high-performing participants (especially the top-performing ones) are being censored at the top by the maximum score of 18. Thus, while they are not able to increase their score anymore through additional effort exertion, they are still able to improve their relative ranking through being faster and hence breaking ties in their favor.

This latter point is the second point suggested by the results. The -0.085 statistically significant coefficient on total time interacted with the top indicator implies that each additional second of total time (or, equivalently, response time) hurts high-performing partic-

ipants much more in relative ranking terms than low-performing counterparts. This means that high-performing participants are more likely to encounter ties with other participants (in part due to top censoring), so those additional seconds in total time spent matter more for their relative ranking.

Throughout all the specifications, coefficients on round fixed effects, as well as their interaction with the top indicator, are statistically insignificant. This implies that learning does not affect relative ranking. This would be expected since we concluded previously that there seems to be no heterogeneity in learning. If every participant learns at about the same rate, regardless of relative position, then even as participants improve as the rounds progress, there is no relative gain for any one participant compared to other participants who improve at the same rate.

2.4.2 Effort Response to Changes in Tournament Structure

Having settled on using time spent memorizing the word-number list as the measure of effort, this section now examines the effect of tournament structure on effort. To recap, the tournament model in Section 2.2 predicts the following:

1. Having the relative winning cutoff shift towards (away from) a participant's own high or low location in the score distribution increases (decreases) effort.
2. Increasing (decreasing) the prize amount increases (decreases) effort.

Equivalently,

1. Participants near (far from) the relative winning cutoff (i.e. high-performing when high cutoff, low-performing when low cutoff) exert more (less) effort
2. Participants competing for higher (lower) prizes exert more (less) effort

A simplified way to test these predictions is to assume symmetry in “opposite” treatments. By opposite treatments, I am referring to treatment +1 being the opposite of treatment -1,

in that the former treatment is a movement of the cutoff from low to high at a \$5 prize, while the latter treatment is the movement of the cutoff in the opposite direction from high to low at a \$5 prize, and similarly for other treatment pairs. To assume symmetry in opposite treatments means that the treatment effect of a movement in a tournament structure parameter is exactly the negative of the treatment effect of a movement in the opposite direction of the same tournament structure parameter.⁹¹ This will be true if, for instance, there are no behavioral patterns which predict dissimilar treatment effects based on directionality (e.g. loss aversion, endowment effects, reference points, etc.).

To test the two predictions above under symmetry in opposite treatments, I run pooled regressions of the form

$$Effort_{it} = \beta_0 + \beta_1 LowCutoff_{it} \times Bottom_{i(t-1)} + \beta_2 HighCutoff_{it} \times Top_{i(t-1)} + \beta_3 HighPrize_{it} + X_{it}\gamma + \mu_t + \mu_i + \varepsilon_{it} \quad (2.6)$$

where

- $Effort_{it}$ is the memorizing time in seconds of participant i in round t
- $LowCutoff_{it}$ and $HighCutoff_{it} \equiv 1 - LowCutoff_{it}$ are indicators for whether the relative winning cutoff is low or high respectively for participant i in round t
- $Top_{i(t-1)}$ and $Bottom_{i(t-1)} \equiv 1 - Top_{i(t-1)}$ are indicators for whether participant i was ranked in the top or bottom half (respectively) among all participants in the previous round $(t - 1)$
- $HighPrize_{it}$ is an indicator for the prize being \$10 (as opposed to \$5) for participant i in round t

⁹¹For instance, if increasing the prize from \$5 to \$10 when the cutoff is low induces effort to increase by 10 seconds, then decreasing the prize from \$10 to \$5 (at the same low cutoff) should make effort decrease by 10 seconds.

- X_{it} is a set of other possible controls which may be included in some specifications

All regressions also contain round and participant fixed effects.

Column (1) of Table 2.8 shows coefficient estimates of the regression using specification (2.6) where no additional controls (X_{it}) are included. Low-performing (i.e. bottom) participants spend on average of 24.4 seconds more memorizing the list during rounds where the cutoff is low. On the other hand, high-performing (i.e. top) participants do not seem to be induced by the nearness of the cutoff to exert more effort. Furthermore, increasing the prize amount from \$5 to \$10 induces all participants to exert on average 33.0 more seconds of effort.

Two possible concerns arise from this first specification. First, high-performing participants (as determined by the previous round) may be encouraged by their performance independent of any change in the tournament structure. Second, high-performing participants may respond differently to a higher (\$10) prize compared to low-performing participants. In order to address these two concerns, I augment the specification in column (1) with two controls: the top indicator, and its interaction with $HighPrize_{it}$.

Column (2) shows the results of this regression. The coefficient estimates of interest (β_1 , β_2 , and β_3) are qualitatively similar to those of column (1). Being low-performing near a low cutoff increases one's effort by 29.8 seconds. Being high-performing near a high cutoff has no significant effect on effort. Lastly, the larger \$10 prize induces 40.6 seconds more of effort. Interestingly, high-scorers spend an average of 27.4 seconds more memorizing. (This estimate is significant at a 5% level.) Since participant fixed effects are included, this estimate represents the "encouragement" effect of being a high-performer from the last round, independent of any time-invariant "ability" factors.

Column (3) shows the same regression except now, the dependent variable is the logarithm of memorizing time. The pattern of coefficient estimates, now interpretable as percentage changes, remains similar. In particular, being low-performing near a low cutoff increases one's effort by 13%, whereas being high-performing near a high cutoff has no significant

Table 2.8: Pooled Regressions (Assumes Symmetry)

Dep. Var.:	(1)	(2)	(3)	(4)	(5)
Mem. Time	Pooled	Pooled Top	Log	No Round 1	No Middle
Low Cutoff	24.448***	29.759***	0.127***	25.706**	30.381**
×Bottom	(8.15)	(8.953)	(0.047)	(11.603)	(12.878)
High Cutoff	0.983	-3.307	0.003	-0.271	-1.397
×Top	(5.567)	(5.379)	(0.018)	(6.182)	(6.848)
\$10 Prize	32.97***	40.621***	0.101**	32.349***	33.555**
	(4.996)	(8.888)	(0.048)	(10.694)	(13.782)
\$10 Prize		-14.705	-0.030	-2.373	-7.139
×Top		(10.589)	(0.053)	(12.672)	(15.517)
Top		27.356**	0.095**	14.713	19.023
		(11.641)	(0.047)	(14.035)	(22.121)
Round					
2	0.296	0.216	-0.003		-2.435
	(6.955)	(6.919)	(0.029)		(10.388)
3	0.324	0.029	0.008	0.004	-2.096
	(6.514)	(6.428)	(0.026)	(6.231)	(9.523)
4	-11.005	-11.219	-0.090**	-11.303	-21.996*
	(8.17)	(8.128)	(0.041)	(7.456)	(12.236)
Constant	311.109***	296.866***	5.637***	304.704***	304.551***
	(6.322)	(9.476)	(0.043)	(10.085)	15.838
N	692	692	692	519	472
R-Square	0.697	0.700	0.577	0.756	0.719

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Memorizing time in seconds. All specifications include participant fixed effects. Robust standard errors are in parentheses.

effect. A larger prize of \$10 increases effort by 10%.

Specifications in columns (1) through (3) contain top and bottom indicator variables coming from Round 0. However, since this was an example round without monetary prizes, participants may not have taken this round seriously⁹², and the percentile rank outcomes reported to them may not reflect their true belief about their relative position. To address this concern, column (4) uses the same specification as column (2), except observations from Round 1 (which use lagged dependent variables from Round 0) have been excluded. The coefficient estimates are qualitatively similar to those in column (2), suggesting that this should not be too big of an issue.

One additional concern is that the group in between the low and high cutoffs (before and after treatment) are affected by two opposing forces as the cutoff moves “over” them. On the one hand, the cutoff is moving towards them, inducing them to increase effort. On the other hand, the cutoff then passes them and moves away from them, inducing them to decrease effort. Depending on the error structure (the shape of $h(.)$), their effort response may be positive or negative. Column (5) uses the same specification as column (2), except the middle third of participants (based on the percentile rank from the previous round) has been excluded. Again, the coefficient estimates are qualitatively similar, suggesting that any irregularity in effort response by this middle group is not a significant concern.

Learning through the rounds in terms of reductions in effort exerted does not seem to be taking place. This can be seen from the insignificant coefficient estimates on the round fixed effects in all three columns.⁹³ This is nonetheless consistent with the findings from the previous section. On average, from the results in this subsection, participants are not learning how to memorize the lists faster as they progress through the rounds; however, from the results previously, they do seem to be learning how to memorize the lists with better

⁹²An attempt was made to encourage participants through the instructions to take this round seriously as practice for the real thing. However, some participants immediately clicked through the example round screens without completing a single response.

⁹³These specifications treat the rate of learning as the same for all participants.

accuracy as they progress through the rounds, thereby increasing their scores—though again, in relative ranking terms, there is no improvement. And since only relative rank was reported to the participants in the result slips, they would believe they were not improving through learning, thus maintaining similar effort levels, *ceteris paribus*.

To summarize, these results suggest that all participants (regardless of performance) are induced by larger prizes to increase effort. However, only low-performing participants are incented by being near a low cutoff to work harder to surpass the nearby cutoff and win a prize. High-performing participants, it seems, are not threatened by the risk of falling below a nearby high cutoff, and do not alter their effort choice. I will discuss possible explanations for these findings in the discussion subsection.

But before that, it would be interesting to know whether these findings remain as they are without making the symmetry in opposite treatments assumption. To do this, I estimate the treatment effects for each of the 8 possible treatments (+1, -1,...,-4), and I do so separately for high-performing (i.e. top) and low-performing (i.e. bottom) participants. The regression specification I use is of the form

$$\begin{aligned} \Delta Effort_{it} = & \sum_{\tau} \beta_{Top,\tau} Top_{i(t-1)} \times Treatment(\tau)_{it} \\ & + \sum_{\tau} \beta_{Bottom,\tau} Bottom_{i(t-1)} \times Treatment(\tau)_{it} + \mu_t + \varepsilon_{it} \quad (2.7) \end{aligned}$$

where

- $\Delta Effort_{it} = Effort_{it} - Effort_{i(t-1)}$ is the change in effort from the previous round
- $Treatment(\tau)_{it}$ is a set of dummies indicating that treatment τ was administered to participant i in period t

and other notation are similarly defined as before. Note that because the perfectly colinear top and bottom indicators are both included with a full set of treatment dummies, no constant is necessary. For this specification, because neither lagged effort nor the top and

Table 2.9: Regressions with Separate Treatment Effects

Dep. Var.:	(+1)	(-1)	(+3)	(-3)
Δ Effort	Treatment	Treatment	Treatment	Treatment
Top	-4.106 (11.201)	5.400 (8.809)	6.110 (9.035)	12.773 (11.812)
Bottom	-23.690* (13.64)	23.468* (13.706)	-44.928** (21.472)	23.288 (15.895)

Dep. Var.:	(+2)	(-2)	(+4)	(-4)
Δ Effort	Treatment	Treatment	Treatment	Treatment
Top	22.405 (14.165)	-12.312 (10.732)	35.871*** (10.871)	-35.281*** (13.12)
Bottom	24.881 (19.81)	-28.174 (20.93)	33.633* (17.757)	-46.305** (18.383)

Legend: Significance level: *** = 1% ; ** = 5%; * = 10%.

Note: Change in effort measured in seconds. All specifications include round fixed effects (see footnote 94). Robust standard errors are in parentheses. N=519; R-square is 0.096.

bottom indicators in Round 0 can be taken seriously (for the same reasons as stated before for the top indicator), only observations from Round 2 onwards (with lag dependent variables from Round 1) are used. The inclusion of round fixed effects again controls for learning through the rounds.^{94,95}

The two panels in Table 2.9 present the coefficient estimates from the regression using the above specification (2.7). Columns are labeled with treatment codes. The first panel shows treatments in which the cutoff is shifted; the second panel shows treatments in which the prize amount is changed. The sample size (N) for this regression is 519, and the R-square is 0.096.

⁹⁴However, because Round 2 observations are used as the omitted category, and Round 3 treatments are always movements in the prize amount, and therefore omitted when including a full set of treatment dummies, only a coefficient estimate for Round 4 can be identified. This estimate is -11.170 with a standard error of 9.056, which is not statistically significant.

⁹⁵I do not include participant fixed effects because the dependent variable is the first difference of effort, so any participant specific component of effort is differenced out already.

The coefficient estimates for each separate treatment effect, while not as statistically significant as the corresponding estimates in the pooled regression⁹⁶, reveal similar findings. For treatments in which the cutoff shifts from low to high (+1, +3), effort for low-performers decreases. Correspondingly, for treatments in which the cutoff shifts from high to low (-1, -3), effort for low-performers increases (though only treatment -1 has a statistically significant effect at the 10% level). And similar to earlier findings, high-performers do not respond significantly to any shift in the cutoff, be it towards or away from their high relative position (see the top row of the first panel).

For treatments in which the prize amount increases (+2, +4), effort for all participants increases. (However, only for treatment +4 are the estimates statistically significant.) Correspondingly, for treatments in which the prize amount decreases (-2, -4), effort for all participants decreases. (Again, however, only for treatment -4 are the estimates statistically significant.) In general, the coefficient estimates for high-performers are more or less similar in magnitude to those of low-performers for each column of prize-amount moving treatments in the second panel.

From this set of estimates, it is hard to tell if the symmetry assumption holds. The joint significance test of the hypothesis that every pair of opposite treatment effects is the negative of one another cannot be rejected. However, given the small sample size of each cell, this is not all that surprising. In Appendix 2.9, I present non-parametric local polynomial regressions which show the treatment effect across the entire relative percentile rank distribution.

2.4.3 Discussion

The results in the previous section suggest that all participants (regardless of performance) are induced by larger prizes to increase effort. However, only low-performing participants are induced by being near a low cutoff to work harder to surpass this cutoff and win a prize.

⁹⁶This is because by not assuming symmetry, each “cell” within each treatment by top/bottom combinations is identified over a smaller number of observations.

High-performing participants do not seem to be affected by being near a high cutoff, and are not changing their effort choice to avoid falling below this cutoff. Relating these results back to the model, there are several possible explanations as to why this is so.

First, are the high-performers not responding because they believe they are being capped? While it is clear from the distribution of scores in Figure 2.2 that there are top performers who hit the maximum score of 18, since the absolute scores are never reported to them, their effort decisions may not take this into account. Equation (2.5) of the model suggests that if top performers believed they were capped, the constraint on score would bind, λ would be greater than zero, and the positive $\frac{de_i}{d\alpha}$ would be lower (and potentially be negative even). However, high performers and low performers responded similarly to an increase in the prize amount; the coefficient estimate on the interaction term between the high prize and top indicators, while negative, is not statistically different from zero. This suggests that at least a large fraction of top performers are not aware of hitting the cap (and thus consider $\lambda = 0$).

Second, should we be concerned about how the tie-breaking rule may affect behavior, in particular those high performers who are non-responsive to cutoff shifts? Ties were pretty common: of the 692 observations from Rounds 1 to 4 (173 participants multiplied by 4 rounds per participant), a total of 204 of these constituted two-way ties (102 pairs), and a total of 63 constituted three-way ties (21 triples). Recall that from the discussions regarding Table (2.7), the application of the tie-breaking rule did seem to affect the relative ranking of participants.

The tie-breaking rule basically generates an incentive for participants to finish faster (i.e. exert less effort), *ceteris paribus* (conditional on score), because doing so increases the probability of winning any tie. While it is possible to add in an additional term into the utility function to account for this, it can equivalently be thought of as an additional cost of exerting too much effort. From equations (2.4) and (2.5), such a cost would not affect $\frac{de_i}{dp}$ and $\frac{de_i}{d\alpha}$ differentially, since the cost function appears identically in the denominator. Thus, even if the tie-breaking rule were to affect high-performing participants differentially somehow, this

would still not explain their significant effort response to a change in α , versus no response to a change in p .

Looking more closely at equations (2.4) and (2.5), we note that one difference between their numerators is the additional $\frac{d\bar{S}_p}{dp}$ term in $\frac{de_i}{dp}$. Hence, one possible explanation for the high-performing participants' non-response is heterogeneous expectations of the sign of $\frac{d\bar{S}_p}{dp}$. Since the exact location of the absolute score cutoff (or even one's own score) is not revealed, each participant will have individual beliefs as to what the expected value of $\frac{d\bar{S}_p}{dp}$ is. That it does not appear in the equation (2.5) explains why there is no difference between high- and low-performers when the prize amount is changed.

Suppose low performers always expect $\frac{d\bar{S}_p}{dp}$ to be positive, whereas some high performers, being possibly more sophisticated, expect $\frac{d\bar{S}_p}{dp}$ to be negative as well with some non-zero probability. Such negative expectations may arise because high-performers (correctly) believe that when p increases, some low performers give up and reduce their effort, thus exerting downward pressure on the cutoff $\bar{S}_p$. As such, $\frac{de_i}{dp}$ will be negative for these high performers, and any estimate of the average effort response will be less positive, and less statistically significant due to the greater variation in effort responses. Low performers, on the other hand, will always reduce effort when p increases, if they believe $\frac{d\bar{S}_p}{dp} > 0$ with probability 1. And because $\frac{d\bar{S}_p}{dp}$ does not appear in $\frac{de_i}{d\alpha}$, there would not be any heterogeneity in effort response to an increase in the prize amount.

2.5 Conclusion

The findings in this chapter are opposite of what is found in Chapter 1, in which top performers in the dragon cohort, who see the relative cutoff move towards them, actually increase effort in response, as opposed to bottom performers in the dragon cohort, who do not seem to reduce effort. Of course, this could be because of the differences in the size of the competition group in question. On a more optimistic note, the results in this study are similar to those of Harbring and Irlenbusch (2003), who find that as the relative winning cutoff is

lowered, mean effort increases.

The policy implication of these findings is that if one wants to incent everyone to work harder, the best way to do this would be to increase prize amounts (i.e. dangle a bigger carrot). On the other hand, if one's goal is to encourage only low-performers to work harder, then it seems like the way to go about doing this is to shift the cutoff closer to the lower end of the distribution (i.e. dangle the carrot further down). The experimental results suggest that this latter method would not discourage high-performers from slacking off, though by lowering the relative winning cutoff, more prizes have to be offered, the number of which may be constrained by some fixed budget.

One solution to overcome such a budgetary constraint is to create an eligibility criteria which excludes high-performers from winning prizes. However, this criteria cannot be a function of the score (or high-performers will be encouraged to “fake” a low score), so it must be based on some other factor correlated with performance. In this way, low-performing students are still encouraged by the low relative cutoff near them to step up effort, but there is no need to allocate scarce resources to prizes for high-performing students. Examples of this include merit scholarships given only to low-income students, or ones given only to students in below-average schools (with a higher incidence of low-performing students).

One concern regarding identification in this experiment is that there is no control group of participants who remain untreated during any of the round transitions. Moving from one round to the next, one of the tournament structure parameters is always changed. (That is, there is no round transition in which X and Y number of winners is held constant.) This was done to ensure fairness across sessions, so that in every session, the maximum amount of prize money (on top of the \$15 base amount) any participant could earn stays the same at \$30.⁹⁷ Thus, an underlying assumption which is being made in identification is that conditional on round fixed effects (possibly from learning), effort (or other outcomes)

⁹⁷This feature of not having a control group actually occurs more often than it should in the experimental literature. For example, in Chapter 3, I examine an experiment involving bicycle messengers in Fehr and Goette (2007) which also has this feature.

remain constant from round to round when the tournament structure remains unchanged.

While the sample size of this experiment is relatively small and unrepresentative, the findings are partially supported by tournament theory. As with any experiment, there is the question of external validity. But using a real effort task and a novel yet unobtrusive measure of effort hopefully addresses at least some of these concerns, especially those relating to the ability of this experiment to reflect the real world, or the possibility of Hawthorn effects.

Appendices

2.6 Recruitment Email

Subject: Sign up for our Economics Experiment!

Hi!

We are recruiting participants for an economics experiment to be carried out at the Princeton Laboratory for Experimental Social Science (PLESS) located in Green Hall. The experiment will involve several recall tasks. You will also be asked to complete a short survey. You will receive a base amount \$15 for your participation, as well as the potential to win up to an additional \$30 depending on your performance on the tasks. The entire experiment should take approximately 60 minutes.

If you are interested in helping to further the frontiers of economics research, please visit [web sign up form] to sign up.

Thanks,

Yan Lau

Project PI

2.7 Instructions Provided to Participants

The following instructions (in point form) were projected on a screen in front of the room at the beginning of each session. They were also read out to participants in full, who were given the chance to interrupt and ask clarifying questions.

- In this study, you will be asked to complete memorization tests involving words and numbers.
- There will be 5 rounds of these tests. In each round, the following will occur:

1. You will be asked to memorize a list of 18 pairs of words and 3-digit numbers.
 2. You may choose the amount of time you spend memorizing this list. (You will be prompted at 7 min.) After you are satisfied, you will then click to the next screen.
 3. In the next screen, you will be presented with a list of the 18 words (reordered randomly) with text boxes next to them. Your task is to fill in the 3-digit numbers associated with each word.
 4. Your score (out of 18) will be the number of 3-digit numbers you are able to recall correctly.
 5. You don't have to answer all of them. (Most participants don't.) There is no penalty for guessing.
 6. After all participants have finished, everyone will be ranked based on their score. Ties (which are common) will be broken based on who took less time to complete the entire task.
 7. Based on this ranking, the top X participants will be awarded a monetary prize. (PLESS is non-deception) The number X and the size of the monetary prize will be announced at the beginning of each round.
- The initial round (round 0) is an example round to introduce the interface. As such, there will be no prize.
 - However, we highly suggest you complete this round as if it were the real thing, as many have found it helpful for practicing their memory techniques.
 - The remaining four rounds (rounds 1 to 4) will have prize rules announced at the beginning of the round.
 - During each round when you are interacting with the computer, please

- refrain from using electronic devices (including cell phones), and any writing or reading materials
 - do not communicate with other participants
 - do not use the “back” feature on the web browser (an error page will appear)
 - excludes “rest time” in between rounds
- Cheating, or any attempt to gain an unfair advantage, is grounds for disqualification.
 - Please raise your hand at any point if you encounter issues or if you have any questions.
 - You will now receive Participant ID numbers on sticky-notes. Please be sure to enter these correctly.

2.8 Screen-shots of Computer Interface

In each round, the first screen encountered by participants was the list of word-number pairs shown in figure 2.5. After memorizing the list, participants click the “next” button to proceed to the second screen shown in figure 2.6, which contains text boxes along with the reordered list of words. Having filled in their responses, they click “submit” to proceed to the third screen shown in figure 2.7. This last screen of the round informs participants that they should wait for their results to be graded and relax until instructed to proceed to the next round.

2.9 Local Polynomial Regressions

While the regressions with separate non-symmetric treatment effects in Section 2.4 consider differences between the top- and bottom-performing participants using “top” and “bottom” indicator variables, it would be interesting to see how the treatment effect varied across the entire performance distribution. To graph this out, I run a non-parametric local polynomial regression with residual change in effort on the y-axis and relative percentile rank on the

Figure 2.5: First Screen: Word-Number Pairs

Round 0 (Example) - Mozilla Firefox

https://yanlau.mycpanel.princeton.edu/lime/index.php/survey/index

Round 0 (Example)

If you experience any technical difficulties, please raise your hand for assistance.

• Participant ID

Please memorize the following list of word-number pairs. In the next screen, you will be given the list of the words only (in a different order) and asked to fill in the associated numbers.

Word	Number
Squash	824
Microwave	239
Green	698
January	873
Chair	307
Chicken	810
Blue	973
Cow	710
Grape	421
Red	819
Lemon	317
Corn	278
September	912
Horse	101
Onion	797
Desk	311
Coconut	980
Dog	801

When you feel sufficiently ready, click next to begin the response part of the round.

Next > Exit and clear survey

Figure 2.6: Second Screen: Words with Text Boxes

The screenshot shows a web browser window titled "Round 0 (Example) - Mozilla Firefox". The address bar displays the URL "https://yanlau.mycpanel.princeton.edu/lime/index.php/survey/index". The main content area has a light blue background and contains the following text:

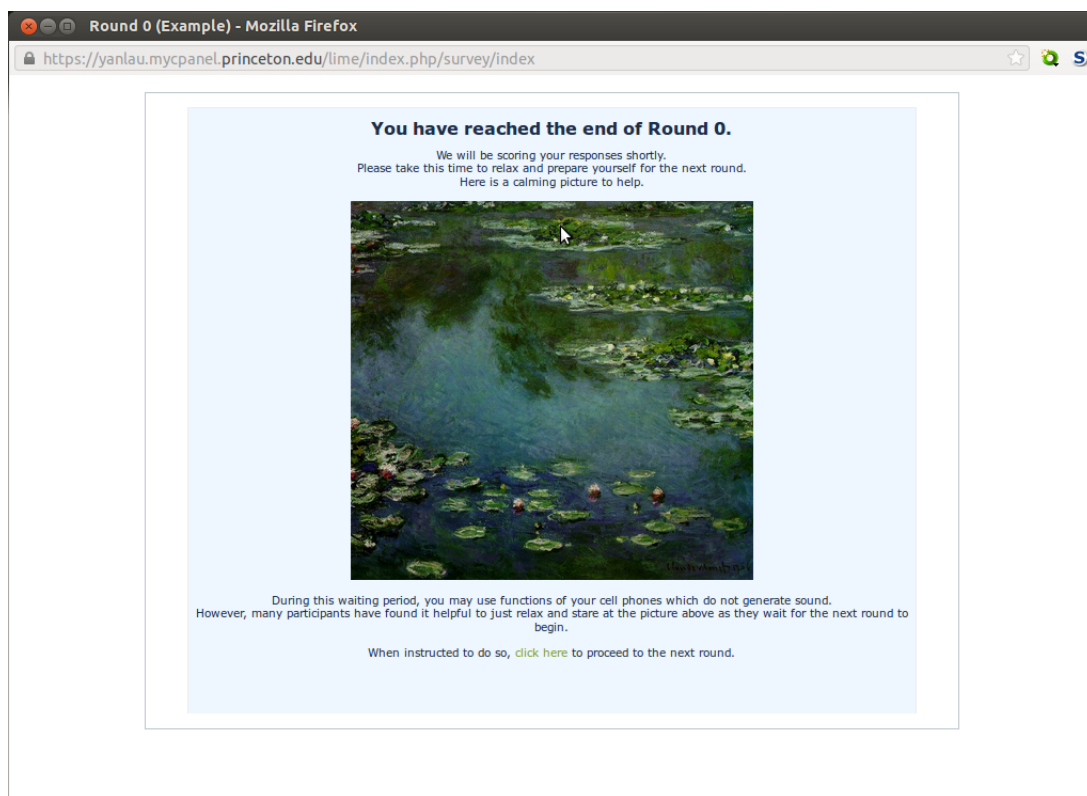
Round 0 (Example)
If you experience any technical difficulties, please raise your hand for assistance.

Please fill in the number associated with each word. There is no penalty for guessing.

Onion	
Squash	
Chair	
Coconut	
Lemon	
January	
Cow	
Corn	
Microwave	
Dog	
Horse	
September	
Grape	
Green	
Chicken	
Blue	
Desk	
Red	

At the bottom of the form, there are two buttons: "Submit" and "Exit and clear survey".

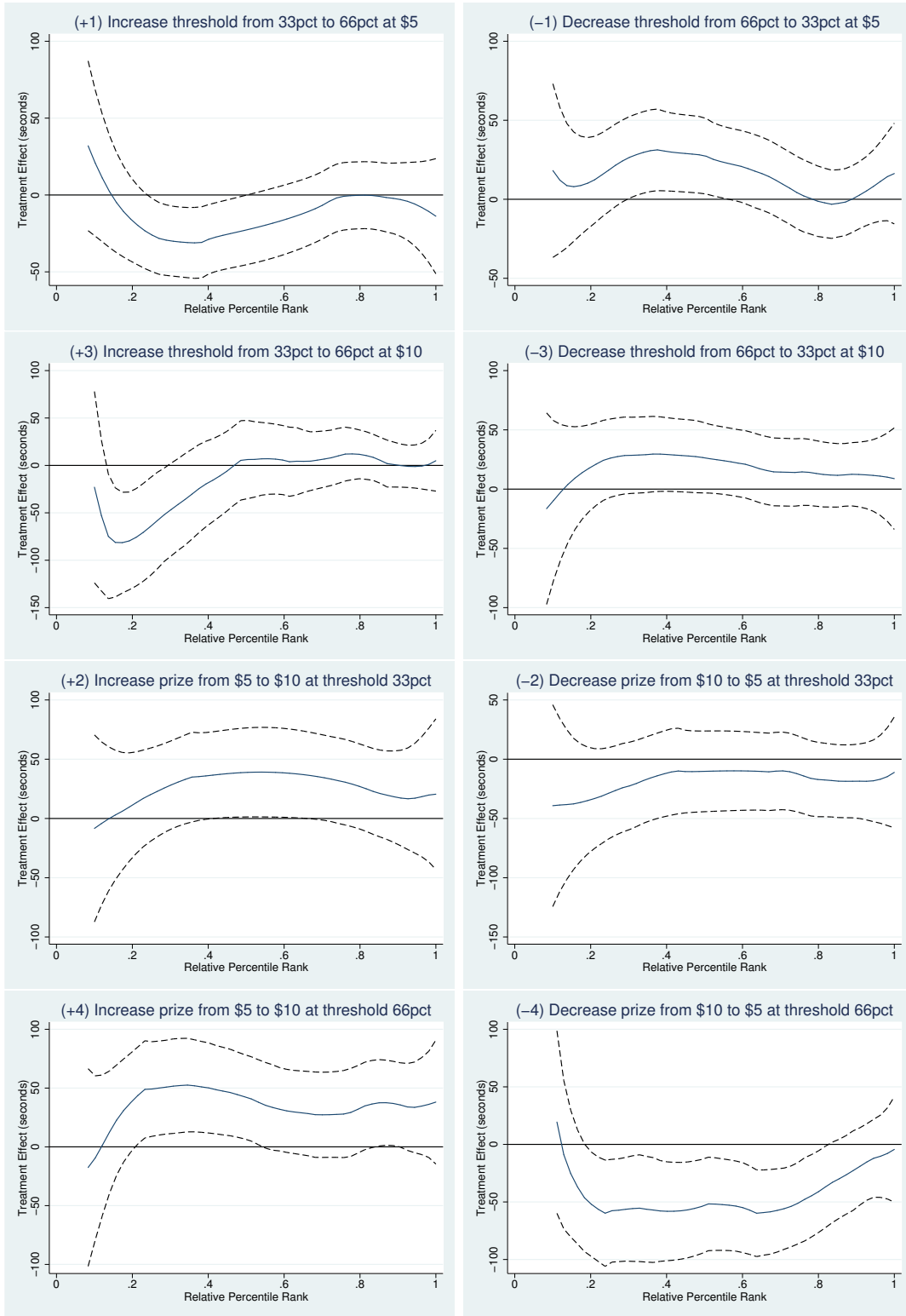
Figure 2.7: Third Screen: End-of-Round Waiting



x-axis. (Residual change in effort is obtained by regressing change in effort on round fixed effects and a constant and then backing out the residuals using the coefficient estimates.)

I run these non-parametric regressions separately for each treatment. Figure 2.8 shows 8 panels corresponding to each of these regressions. The dashed lines are 95% confidence intervals. The graphs reinforce the results discussed in Section 2.4, and are qualitatively similar to those described in the previous regressions.

Figure 2.8: Non-parametric regressions



3 A Dynamic Model of Labor Supply for Bicycle Messengers: A Re-examination of Fehr and Goette's Experiment

Abstract

This chapter develops a dynamic neoclassical model of labor supply in which 1) effort decisions of previous periods affect the cost of effort exertion in the current period (i.e. non-time-separability); 2) workers select beforehand which periods to commit to working; and 3) there is uncertainty in work availability, expressed as a random cap on possible effort exertion each period. I argue that this framework is sufficient in explaining the results found in Fehr and Goette (2007) without relying on reference dependent preferences. I perform further empirical analyses on their bicycle messenger experiment data to test additional predictions of both the neoclassical and reference dependent preferences models.

I am indebted to Henry Farber and Alexandre Mas for their suggestions throughout the process of bringing this study to fruition. Diogo Guillen, Nicholas Lawson and Jay Lu gave numerous constructive comments on drafts, for which I am grateful. I would also like to acknowledge the useful feedback garnered from participants of the Public Finance Working Group at Princeton University. A portion of the empirical analysis was made possible by data from the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss).

3.1 Introduction

Traditional labor supply models describe workers who choose consumption and labor (indirectly through leisure) to maximize utility subject to a budget constraint. In reality, a majority of occupations do not let workers choose their labor supply exactly, opting instead for a standard 40-hour work week. To account for this, some economists focus on occupations which do let workers choose labor supply, be it in terms of hours or some measure of effort. By taking advantage of certain variations in the wage of such workers, these papers have attempted to empirically estimate the labor supply elasticity of taxi drivers⁹⁸, stadium vendors⁹⁹, and bicycle messengers¹⁰⁰.

In recent years, models of labor supply incorporating reference dependent preferences have gained considerable attention. Inspired by behavioral economics, these models assume that workers have an income reference point at which marginal utility of income drops discretely. This discrete jump arises from the notion that incomes below the reference point are treated as losses while those above it are treated as gains. A worker who is loss averse and at the reference point will experience a larger utility loss from a decrease in wealth than the utility gain associated with an equivalent increase in wealth.¹⁰¹

One prominent paper which considers the reference dependent preferences model of labor supply is Fehr and Goette (2007). The authors generate exogenous variation in the wage of bicycle messengers in Zurich through an experiment and use this to measure labor supply elasticity, which they find to be negative at the (messenger delivery) shift level. They also conduct a survey to measure each messenger's degree of loss aversion. Finding that the negative daily labor supply response to an increase in wage is primarily driven by the loss averse messengers, Fehr and Goette (2007) conclude that a model with reference dependent utility better explains these empirical findings than a neoclassical labor supply model.

⁹⁸Camerer et al. (1997); Farber (2005, 2008); Ashenfelter et al. (2010)

⁹⁹Oettinger (1999)

¹⁰⁰Fehr and Goette (2007)

¹⁰¹For the theoretical aspects of loss aversion, see Tversky and Kahneman (1991). For a survey of work applying reference dependent preferences to labor supply, see Goette et al. (2004).

In the remainder of this chapter, I propose an alternative neoclassical model which can account for their observed empirical results, and provide further analyses to argue in favor of my model. I first summarize their findings in Section 3.2. I then set up my alternative neoclassical framework in Section 3.3, and briefly discuss why I believe Fehr and Goette’s measure of loss aversion can also be interpreted as a measure of risk aversion. Using Fehr and Goette’s data, I conduct additional analyses in Section 3.4 to consider other aspects of the data which seem to confirm the predictions of my neoclassical framework. Section 3.5 concludes.

3.2 Fehr and Goette (2007)

In their paper entitled “*Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment*,” Fehr and Goette estimate the intertemporal labor supply elasticity of a group of bicycle messengers in Zurich. Most authors exploit natural variation in wage to identify the elasticity estimate; the novel approach taken by Fehr and Goette (2007) is to generate the wage variation themselves in the form of a randomized temporary wage increase.

Before continuing, it is useful to understand how the bicycle messenger service operates. Within the company where they work, messengers work shifts lasting five hours. There are two types of shifts: “fixed shifts” and “sign-up shifts”. Fixed shifts are designated by the messenger in advance (at the beginning of a block of shifts) and (in effect) impossible to get out of once the messenger has committed to them. Sign-up shifts, on the other hand, are “signed up” for on the day of the shift itself, granted there are still slots available. In their empirical analysis, Fehr and Goette (2007) use only fixed shifts data; my empirical analysis will also only use fixed shifts as it was not possible to obtain sign-up shifts data.¹⁰²

While on a shift, messengers carry a radio to communicate with the dispatcher; any

¹⁰²According to Fehr and Goette (2007), this is done in order to avoid contaminating estimates with “selection effects”, because the increase in shifts is likely to have been made up entirely of additional sign-up shifts.

communication via radio is heard by all messengers. When a delivery job arrives at the company, the dispatcher will radio to the messenger closest to the job’s point of origin to see if that messenger is willing to take the job. At this point, several things can happen. The messenger being addressed can accept the job. Alternatively, if the job is rejected by the messenger being addressed, other messengers (who have been listening in on the communications) can claim the job for themselves. Moreover, if other messengers feel that they are nearer to the point of origin than the first messenger being addressed, they can radio back and “intercept” the job for themselves. In this way, even though messengers commit to five hour shifts per se, they are free to choose exactly how much effort to exert—that is, how much of the five hours is actually spent delivering packages—by accepting or rejecting jobs. And because messengers are paid solely on a commission basis (they receive a percentage of the revenues they generate that shift and no fixed amount for showing up to a shift), a reasonable measure of effort exerted during a particular shift is the revenue generated that shift.¹⁰³

In order to estimate the labor supply elasticity, Fehr and Goette (2007) induce wage variation by conducting a “randomized field experiment” in which they temporarily and exogenously raise the commission rate (henceforth “wage”) of a subset of messengers by about 25%.¹⁰⁴ The experiment is conducted during several four-week “blocks” of five-hour shifts, of which there are two per day, and these only occur from Monday to Friday. All messengers are randomly assigned to one of two groups: group A and group B. During the first four-week block in September 2000, messengers in group A experience the 25% wage increase (“treatment”) while those in group B receive normal wages (“control”). During the

¹⁰³In their paper, Fehr and Goette sometimes refer to shifts more generally as “hours,” so when they mention an increase in “hours,” they mean an increase in the number of five hour shifts. They also use revenue and effort interchangeably to refer to the amount of revenue generated by messengers within these five hour shifts. To reduce confusion, I will henceforth use “shift” to refer to the number of five hour segments a messenger shows up for, and “effort” to refer to the amount of revenue generated for the company by a messenger (before commission calculations).

¹⁰⁴For men, the commission rate was increased from 0.39 to 0.49. For women, it was increased from 0.44 to 0.54.

next block in October 2000, neither group is treated. Then, during the block in November 2000, messengers in group B are now treated while those in group A now act as controls. Like the in-between October block, both groups receive normal wages during the surrounding blocks before and after the experiment. The messengers are told ahead of time about the timing and magnitude of these wage increases, though not the purpose of the experiment. The extra wages earned during treatment are paid out only in December 2000, after both groups have been treated.

Throughout this period, data on how much revenue is generated by each messenger during each shift is collected. Analyzing this data, Fehr and Goette (2007) establish the following empirical results regarding the messengers' labor supply response to a wage increase:

1. Total effort (as measured by revenues) per block increases in response to a wage increase
2. Number of shifts per block increases in response to a wage increase
3. Effort per shift decreases in response to a wage increase

The authors note later on that the third negative result is driven mainly by those messengers deemed to be loss averse.

The authors argue that because the wage increase is exogenous, temporary, and fully anticipated by the messengers, marginal utility of lifetime wealth remains constant and the elasticities calculated can be interpreted as intertemporal substitution elasticities, devoid of any income effects. However, the third result above implies a negative elasticity of effort per shift with respect to wage, contradicting the always-positive nature of intertemporal substitution elasticities, which by assumption contain no negative income effect.¹⁰⁵ To reconcile this negative elasticity, the authors suggest two potential competing models of labor supply.

The first potential model is a simple example of a neoclassical model with non-time-separable utility; in particular, the non-time-separability stems from the fact that the pre-

¹⁰⁵The authors present a standard neoclassical model with time-separable utility to this effect, showing that, under constant marginal utility of lifetime wealth, effort per period must increase in periods experiencing wage increases.

vious period's effort affects the (dis)utility function this period. This captures the intuition that working yesterday will cause the messenger to have higher marginal cost of effort today. A wage increase induces messengers to exert more effort in total, either through increasing shifts or effort per shift (or both). The negative elasticity of effort per shift is a result of messengers who work a greater number of shifts yet reduce their effort per shift over all shifts, while still increasing total effort. (This will be expanded upon subsequently in Section 3.3.)

The second potential model is based on time-separable reference dependent utility. Reference dependent preferences imply that there is a discrete jump in marginal utility at a certain reference point income. Below this reference point, marginal utility is high, strongly inducing the messenger to exert additional effort to earn more income. Upon reaching the reference point, marginal utility drops discretely, thus reducing the incentive to exert additional effort. In behavioral economics vernacular, incomes below the reference point are treated as losses while those above it are treated as gains. When faced with a wage increase, messengers are able to reach the reference point with less effort than before and therefore reduce effort per shift. Furthermore, Fehr and Goette (2007) argue that

“...workers increase the number of shifts when they are temporarily paid a higher wage: a rise in wages increases the utility of working on a given day. Thus, at higher wages it is more likely that the utility of working... exceeds the fixed costs of working.” [p. 306]

Without further evidence, it is unclear which of these two models is in play.

To discriminate between them, a survey is conducted whereby the messengers are asked whether they would accept or reject either of the following lotteries:

Lottery A: Win \$8 with probability $\frac{1}{2}$ or lose \$5 with probability $\frac{1}{2}$.

Lottery B: Six independent repetitions of Lottery A.

Rejection of either lottery yields a default payoff of zero. Messengers can choose to participate in neither of, one of, or both of the lotteries. Having made their choice, the lottery

Table 3.1: Reproduction of Table 6 from Fehr and Goette (2007)

TABLE 6—DOES LOSS AVERSION EXPLAIN THE TREATMENT EFFECT? (Dependent variable: <i>log (revenues per shift)</i> during fixed shifts, OLS regressions)		
	(1)	(2)
Treatment effect $\times$ not loss averse	−0.0273 (0.033)	−0.027 (0.032)
Treatment effect $\times$ rejects lottery A	−0.105** (0.046)	
Treatment effect $\times$ rejects one lottery		−0.0853* (0.062)
Treatment effect $\times$ rejects both lotteries		−0.12** (0.053)
Log(tenure)	0.00152 (0.061)	0.0074 (0.060)
Day fixed effects	Yes	Yes
Individual fixed effects	Yes	Yes
R-Squared	0.243	0.26
N	1137	1137

Note: Robust standard errors, adjusted for clustering on messengers, are in parentheses.

*** Indicates significance at the 1-percent level.

** Indicates significance at the 5-percent level.

* Indicates significance at the 10-percent level.

Source: Own calculations.

(or lotteries, or lack thereof) is executed. Participation in a greater number of lotteries is interpreted as a lower degree of loss aversion for that messenger.

With this measure of loss aversion for each messenger, (log) effort per shift is regressed on an interaction between treatment and degree of loss aversion, and it is found that the coefficient on this interaction term is negative and significant. These results in Table 6 of Fehr and Goette (2007) are reproduced in Table 3.1. Both specifications suggest that the negative effort per shift elasticity with respect to wage is driven primarily by those messengers with a high degree of loss aversion. Fehr and Goette (2007) conclude that this is evidence in favor of the reference dependent preferences model.

3.3 Neoclassical Model of Labor Supply

In this section, I build a neoclassical intertemporal model of labor supply with non-time-separable utility which will attempt to explain the findings in Fehr and Goette (2007) and

Section 3.4 later in this chapter, without appealing to reference dependent preferences. The uncertainty in the full model stems from the fact that there is limited demand for bicycle messenger delivery services in each period; thus, effort is capped by a random variable. The full model will consist of a two-step decision process, in which messengers first choose the particular periods to commit to showing up for at the beginning of each block of shifts, and then choose effort decisions in each of the committed periods subject to the uncertain effort cap. Commitments to show up for specific periods are made in expectation. The effort decisions depend on the history of realized effort caps so far. This two-step setup closely reflects the realities of employment as one of the sampled bicycle messengers.

The intuition behind the mechanics of the model is that because of the uncertainty in the effort cap, each shift is a gamble. During “good” shifts, the effort cap is high (lots of business) and a messenger’s optimal effort choice is unlikely to hit the cap. During “bad” shifts, the effort cap is low (very little business) and a messenger’s optimal effort choice is likely to hit the cap.

When the wage is increased during a particular block of shifts, all messengers want to increase total effort to earn more income.¹⁰⁶ Messengers can increase total effort either on the intensive margin (exert more effort each shift) or on the extensive margin (commit to a greater number of shifts). However, when a messenger signs up for additional shifts at the extensive margin, while total effort increases, it is spread out over a greater number of periods, thus causing effort per shift to potentially decrease.¹⁰⁷ Messengers prefer to spread out effort in this way because of the non-time-separable nature of their cost of effort functions.

There is heterogeneity in the response to the wage increase depending on a messenger’s risk aversion. More risk averse messengers will want to utilize to a greater extent the extensive margin in increasing total effort. This is because by adding to number of shifts, thus

¹⁰⁶As in Fehr and Goette (2007), I will assume constant marginal utility of lifetime wealth; hence, there is no negative income effect, and total effort must increase.

¹⁰⁷Although in a separate case examined later, an increase is possible as well.

spreading out and reducing effort per shift, they face a lower risk of encountering shifts for which the effort caps become binding. In doing so, risk averse messengers pay additional showing-up-for-a-shift fixed costs (e.g. time spent getting to work) to avoid this risk of hitting the effort caps. On the other hand, not so risk averse messengers would be less willing to pay these fixed costs and more willing to endure the risk of being limited by uncertain effort caps.

3.3.1 Measuring Risk Aversion Versus Loss Aversion

These considerations involving risk aversion bring us to an important question, which is: What exactly is the lottery survey measuring? I would argue that the lottery survey is a proxy for the degree of concavity in a messenger's overall utility function.

To understand why this argument is plausible, it is useful to look at how Fehr and Goette (2007) link loss aversion in the lottery survey to loss aversion in effort choices. Note that the lottery survey is administered eight months after the conclusion of the experiment, so it is likely that the reference point at the time of the lottery survey is not the same reference point during while the messenger is working a shift, making effort decisions. It would seem that if the lottery survey were measuring loss aversion, it is doing so only insofar as it pertains to the lotteries presented in the survey. Fehr and Goette (2007) link loss aversion in this lottery survey setting to loss aversion in the effort choice setting by claiming that, in fact,

“...individuals who are loss averse in one type of decision are also loss averse in other domains of life (see Gaechter, Herrmann, and Johnson 2005).” [p. 306]

Yet both of the tasks presented in Gaechter, Johnson and Herrmann (2007), over which a correlation in loss aversion is found, are one-off low-effort-stakes tasks administered in a controlled, experimental setting. Whether such tasks (similar to the lottery survey) are correlated with real-world non-experimental decisions (such as labor supply decisions) is debatable.

Even if we were to grant that loss aversion measured in the lottery survey is correlated with loss aversion across some potential reference point, it still does not explain why loss averse messengers in particular would be more likely to subscribe to an income reference point policy for determining effort. A messenger could be very loss averse in all aspects of his or her life, but that does not necessitate that he or she sets a reference point and uses it when making effort decisions while on the clock.¹⁰⁸

To advance the subsequent neoclassical framework, I propose that the lottery survey is a proxy for risk aversion. That is, while the lottery survey measures the local concavity of the utility function, I argue that this is correlated to the overall (global) concavity of the utility function. Thus, messengers who are loss averse in this type of (lottery) decision are also risk averse in other domains of life, such as when making labor supply decisions.

One concern that brought up by Fehr and Goette is that interpreting the lottery survey as risk aversion may not be realistic because the lottery prize structure would imply ridiculously high levels of risk aversion for even slightly higher stakes, when applying the calibration theorem in Rabin (2000).¹⁰⁹ However, I would argue that using loss aversion in the lottery survey to infer loss aversion in the unrelated effort decisions during delivery shifts, as argued by Fehr and Goette (2007), is no less unreasonable than using loss aversion in the lottery survey to infer risk aversion in the utility function more generally, as I propose. In essence, Fehr and Goette (2007) argue that concavity in one local region of the utility function implies concavity in another local region of the utility function, whereas I argue that the former implies concavity in the utility function more globally. In the end, both arguments are using the lottery survey as a proxy for the degree of concavity somewhere along a messenger's utility function.

¹⁰⁸To further this line of inquiry, if a loss averse person were sophisticated enough to be aware of his or her high degree of loss aversion, then he or she could take advantage of this knowledge by setting a reference point—perhaps as a commitment device. However, this raises more questions as to how the reference point is endogenously set in the first place, and whether the wage rate could affect the chosen reference point. Koszegi and Rabin (2006) present one such model in which reference points are endogenously determined based on the potential of expected earnings. I discuss this possibility further in Subsection 3.4.3.

¹⁰⁹See p. 313 of Fehr and Goette (2007) and their Appendix B.

3.3.2 Deterministic Framework

Before presenting the two-step decision framework with uncertainty in the effort caps, I will present a deterministic version of a neoclassical model with non-time-separability as a baseline. This baseline model is similar to the one presented in Fehr and Goette (2007) and will help ground understanding in what happens when the wage for a particular block of shifts increase. As in their model, I also assume constant marginal utility of lifetime wealth, and hence no income effects.

Under this assumption¹¹⁰, a messenger's per-period utility is given by a linear function in income, minus the cost of exerting effort, or

$$\lambda w_t e_t - g_t(e_t, e_{t-1}) \quad (3.1)$$

where

- λ is the constant marginal utility of lifetime wealth
- $w_t > 0$ is the wage in period t
- $e_t \geq 0$ is effort in period t
- $g_t(e_t, e_{t-1})$ is a period- t -specific non-time-separable cost of effort function

Note that the cost function depends on both current (e_t) and lagged (e_{t-1}) effort. I also model the functions as period- t -specific because period-specific factors—such as weather, personal schedules, birthdays, etc.—affect a messenger's effort cost. Assume that $\frac{d}{de_t} g_t(\cdot) > 0$ and $\frac{d^2}{de_t^2} g_t(\cdot) > 0$ for all t . Also, since effort in the previous period makes working this period more tiring, assume that $\frac{d}{de_{t-1}} g_t(\cdot) > 0$ and $\frac{d^2}{de_{t-1}^2} g_t(\cdot) > 0$ for all t . Furthermore, assume that $g(0, e_{t-1}) = 0$, but that $\lim_{e_t \rightarrow 0} g_t(e_t, e_{t-1}) > 0$, as there are fixed costs to showing up for a shift, such as time and resources spent on getting to work.

¹¹⁰This assumption allows us to use Fehr and Goette's static one period utility function (see their equation (3) [p. 304]), but augment it with lagged effort (similar to their equation (4) [p. 305]) more generally.

Abstracting away from discounting, for a particular block of T periods (from period 1 to period T), the per-block utility is given by

$$\sum_{t=1}^T \lambda w_t e_t - g_t(e_t, e_{t-1}) \quad (3.2)$$

Each block represents the four-week blocks of shifts of the bicycle messengers in the sample, and are inclusive of weekends. To abstract away from effort decisions in one block being affected by those of another, let $t = 1$ be a Sunday and $t = T$ be a Saturday, so that $e_0 = e_1 = e_T = e_{T+1} = 0$; this is true of the weekends in between as well, since the messengers only work weekdays. The subsequent four-week block of shifts will be represented by a similar summation, except it will be from period $T + 1$ to period $2T$, and so forth. Because wages within a block are constant, $w_t = w$ for a single block, and since I will only examine effort decisions within a particular block, I drop time subscripts for the wage variable henceforth.

A messenger solves the following utility maximization problem for a particular block of periods.

$$\max_{\{e_t\}} \sum_{t=1}^T \lambda w e_t - g_t(e_t, e_{t-1}) \quad (3.3)$$

subject to the constraint that $e_t \geq 0$ for every period t . The first order conditions (FOCs) are given by the equations

$$\lambda w \leq \frac{d}{de_t} g_t(e_t, e_{t-1}) + \frac{d}{de_t} g_{t+1}(e_{t+1}, e_t) \quad (3.4)$$

for every period t .¹¹¹ Some of these FOCs will be binding, and in these periods, the messenger shows up for work and has effort $e_t > 0$. For other periods, the FOCs will not be binding because of high fixed costs, and in these periods, the messenger has effort $e_t = 0$. The messenger will show up for the set of periods with the lowest fixed costs amongst all periods in the block.

¹¹¹It is easy to check that given the assumptions made, the second order conditions are negative.

Consider a period during which the FOC is binding. For steeper cost functions (e.g. larger $\frac{d}{de_t}g_t(e_t, e_{t-1})$ at every e_t), the chosen effort level will be lower, *ceteris paribus*. The shape of this cost function in a given period is affected by many factors, one of which is weather. Inclement weather represents an additional and intimate cost to every unit of effort exerted; in the empirical section, I will test for such weather effects.

When the wage w is increased within a block of periods, given that there are no income effects, messengers necessarily increase the total effort exerted during this block. This is because a positive substitution effect will cause messengers to choose additional consumption (a result of the income generated from effort exertion) over leisure (negative effort). Upon closer evaluation, two distinct forces are at work.

On the extensive margin, messengers will work more (or if not, at least as many) periods. This arises because the higher wage causes a potentially greater number of the inequality FOCs to be binding. With the higher wage, messengers are more likely to be able to overcome the fixed costs of showing up for a shift, and will thus choose more effort decisions where $e_t > 0$ in equilibrium.

On the intensive margin, the response in effort per period depends on the extensive margin response. If the wage increase did not induce the messenger to show up for additional periods, then on the whole, effort increases during the (same) periods in which the messenger does work.^{112,113}

However, if the wage increase does induce the messenger to show up for additional periods, then effort per period potentially decreases. For this case to occur, effort per period decreases during periods in which the messenger originally worked (prior to the wage hike), but increases from zero to some positive amount for the new infra-marginal periods. This

¹¹²Recall that there are no income effects, so the negative component of the Slutsky equation in response to a wage increase is zero.

¹¹³Of course, there may be some periods in which the lagged effort (now higher due to the wage increase) has such a big effect on the cost function that the worker reduces effort. But this is a special case which we can abstract away from without much consequence. In this case with no increase in the number of periods, the point that effort per period will increase on average still holds.

arises because the jump from $e_t = 0$ to some $e_t > 0$ for the new infra-marginal periods will set off a chain reaction of effort reductions in other periods due to the non-time-separable nature of the cost of effort functions. Messengers in essence spread out their effort exertion over a greater number of periods in order to take advantage of the lower marginal costs offered by the new infra-marginal days.¹¹⁴ For example, if the messenger previously worked on Wednesday but not on Tuesday, and the wage increase induced this messenger to now work on Tuesday, then he or she will reduce effort on Wednesday because having worked Tuesday, her cost of effort on Wednesday is now higher—from being tired as a result of working on Tuesday.¹¹⁵

Of course, there is also the possibility that average effort per shift increases, with an increase in the number of shifts. This can occur in the following situation. Suppose the wage increase induces the addition of one new infra-marginal day. If the wage increase is just enough to induce effort to increase during the periods in which the messenger originally worked, and also increase from zero to an amount higher than the original average for the new infra-marginal period, then the overall average effort per shift in that block will increase. However, this must all occur without pushing total effort too far such that a second infra-marginal day is needed, thus resulting again in a spreading out of effort which would then decrease the average overall, as in the first case.¹¹⁶

To summarize the comparative statics of a non-time-separable neoclassical model, it is not unusual at all to observe an increase in the number of shifts worked and decreases in

¹¹⁴A good analogy for this is a firm allocating production across two factories, and equalizing marginal costs across them. When opening a third factory, the firm reallocates some of the production from the two initial factories to the third new factory. Building a new factory is analogous to showing up for an additional period.

¹¹⁵This chain reaction may not be the case (or potentially be less efficacious) if the working periods are spaced out far enough. However, given that there are 20 work days in a four week block of shifts, and untreated messengers work on average 11.8 shifts per block, an average (statistically significant) increase of 3.68 shifts per block in response to the high-wage treatment leaves little spacing between days. Furthermore, while my model only accounts for lagged effort from the period immediately prior, it could be the case that additional lags in effort (from two or more periods prior) play a role in determining cost of effort.

¹¹⁶Think of this in terms of a sawtooth graph with total effort on the x-axis and average effort per shift on the y-axis, analogous to Maimonides' Rule.

effort per shift following an increase in the wage within a block of periods. And this is indeed what is observed empirically in Section 3.4.

However, one of the main findings in Fehr and Goette (2007) is that lottery-rejecting messengers respond to the wage increase to a greater extent (greater increases in total effort and number of shifts per block, and greater decreases in effort per shift) compared to non-lottery-rejecting messengers. I have argued that what Fehr and Goette measure as loss aversion, I will think of as risk aversion. In the next subsection, I extend the model further to account for these response differences between messengers of varying degrees of risk aversion.

3.3.3 Two-Step Decision Framework with Uncertain Effort Caps

Two important considerations reflecting the bicycle messenger scenario motivate this framework. In reality, messengers must commit to shifts in advance.¹¹⁷ Messengers also face the uncertainty of limited demand for bicycle messenger delivery services each period, which can be represented as a random effort cap.¹¹⁸ To model these two considerations, I set up a two-step decision process in which messengers first choose the particular set of periods, denoted by the set of indicator variables $\{D_t\}$ for all t , to commit to showing up for at the beginning of each block. Then, having shown up in a particular committed period, the messenger must choose effort e_t subject to the history of realized effort caps.

Before proceeding, I define the following additional notation.

- $\hat{e}_t$ is the effort cap in period t drawn from a period-independent distribution $f(\hat{e}_t)$ with support on the interval $[0, \infty)$
- $D_t \in \{0, 1\}$ is an indicator with a value of 1 when the messenger commits to working

¹¹⁷In the experiment, Fehr and Goette note that messengers must commit to fixed shifts four-weeks in advance, which is exactly the timespan of one block. In this model, we must necessarily abstract away from the sign up shifts, which do not show up in the data anyway. This is because if sign up shifts were available, then no messenger would commit to fixed shifts because the increased flexibility would always be better. This is not unrealistic because according to Fehr and Goette (2007), all messengers at the company must sign up for at least a minimum number of fixed shifts as a condition for employment.

¹¹⁸The better business is, the higher the cap will be.

in period t

- $s^t = \{\hat{e}_1, \dots, \hat{e}_t\}$ denotes a particular history of effort cap realizations up to and including period t , where S^t is the set of all s^t for each t , and where $s^t \mid s^{t'}$ (with $t' > t$) is the “sub-history” up to period t within history $s^{t'}$
- $\pi(s^t) = f(\hat{e}_1) \cdots f(\hat{e}_t)$ is the probability of history s^t occurring, where $\pi(s^t \mid s^{t'}) = f(\hat{e}_{t'+1}) \cdots f(\hat{e}_t)$ (with $t' < t$) is the conditional probability of history s^t occurring given that history $s^{t'}$ has already been realized

To solve this model, I will work backwards and solve for the second step first. (There will be similarities in the solutions corresponding to the deterministic model presented in the previous subsection.) Suppose the set of committed periods $\{D_t\}$ has already been chosen in the first step. Conditional on these choices, the messenger will solve the following expected utility maximization problem for all potential $s^t \in S^t$ for every t in the second step.

$$\max_{\{e_t(s^t)\}} \sum_{t=1}^T \int_{S^t} \left[\lambda w D_t e_t(s^t) - g_t(D_t e_t(s^t), D_{t-1} e_{t-1}(s^{t-1} \mid s^t)) \right] \pi(s^t) ds^t \quad (3.5)$$

subject to the constraints $e_t(s^t) \geq 0$ and $e_t(s^t) \leq \hat{e}_t(s^t)$ for all potential $s^t \in S^t$ for every t .¹¹⁹ The FOCs of this problem are

$$\begin{aligned} \lambda w D_t - D_t \frac{d}{de_t(s^t)} g_t(D_t e_t(s^t), D_{t-1} e_{t-1}(s^{t-1} \mid s^t)) \\ - D_t \int_{S^{t+1} \mid s^t} \frac{d}{de_t} g_{t+1}(D_{t+1} e_{t+1}(s^{t+1}), D_t e_t(s^t \mid s^{t+1})) \pi(s^{t+1} \mid s^t) ds^{t+1} \\ + \phi_t(s^t) D_t - \rho_t(s^t) D_t = 0 \end{aligned} \quad (3.6)$$

which holds for any potential $s^t \in S^t$ for every t , and where $\phi_t(s^t)$ and $\rho_t(s^t)$ are the

¹¹⁹Though I write effort caps as a function of the whole history, they in fact only depend on the realization in the history's final period. However, to avoid defining unnecessary notation, I will keep it as such.

Lagrange multipliers for the two constraints respectively. Note that if $D_t = 0$ had been chosen for a particular t , then the FOCs for all states s^t that period collapses and are not taken into consideration in equilibrium.

On the other hand, if $D_t = 1$ had been chosen, then there are three possibilities, expressed in the inequality

$$\begin{aligned} \lambda w \leq & \frac{d}{de_t(s^t)} g_t(e_t(s^t), e_{t-1}(s^{t-1} | s^t)) \\ & + \int_{S^{t+1}|s^t} \frac{d}{de_t} g_{t+1}(e_{t+1}(s^{t+1}), e_t(s^t | s^{t+1})) \pi(s^{t+1} | s^t) ds^{t+1} \end{aligned} \quad (3.7)$$

for all potential $s^t \in S^t$ for every t . If $\lambda w < \dots$, then $e_t(s^t) = 0$ because the wage is not high enough to overcome marginal costs and make the FOC bind.¹²⁰ If $\lambda w = \dots$, then $0 < e_t(s^t) < \hat{e}_t(s^t)$, which represents the binding FOC case where the messenger works but is not limited by the effort cap. If $\lambda w > \dots$, then $e_t(s^t) = \hat{e}_t(s^t)$ because even though the wage is high and the worker wants to work more, the effort cap is now binding. Conditional on the chosen set $\{D_t\}$, the solution to this problem comprise a mix of binding and non-binding FOCs, and is represented by the set of optimal effort choices $e_t^*(s^t)$ for all potential $s^t \in S^t$ for every t .

Having solved the second step for every possible realization of effort cap histories, we now move back to the first step, where the messenger chooses the set of periods $\{D_t\}$ to commit to in order to maximize the expected utility subject to the FOCs in the second step. That is, the messenger will solve

$$\max_{\{D_t\}} \sum_{t=1}^T \int_{S^t} [\lambda w D_t e_t(s^t) - g_t(D_t e_t(s^t), D_{t-1} e_{t-1}(s^{t-1} | s^t))] \pi(s^t) ds^t \quad (3.8)$$

subject to the set of equations (3.6) for all potential $s^t \in S^t$ for every t . Let $D \equiv \sum_t D_t$ be

¹²⁰For periods where $D_t = 1$, since the messenger is committed to showing up for them, $g(0, \cdot) > 0$ in this second step and fixed costs must be incurred. The argument for $e_t(s^t) = 0$ occurring is that the marginal costs are negative even at zero effort.

the number of shifts committed to. In this particular setup, since the effort caps are drawn from the same distribution each period, the solution set $\{D_t^*\}$ consists of the D periods with the lowest combination of costs amongst all periods. Having chosen this solution set $\{D_t^*\}$, history plays itself out and the messenger chooses the appropriate $e_t^*(s^t)$ as planned, according to the FOCs and the path of realized effort caps.

As with the comparative statics of before, when the wage w increases, total effort over the entire block increases, the number of shifts D chosen in the first step increases, and average effort per shift within the block can decrease (or increase, under the second case). However, in this extension, an additional prediction arises. In response to a wage increase, more risk-averse messengers, who dislike the uncertainty of the effort caps, choose a higher increase in D so as to lower the risk of these effort caps binding.

To see why this is the case, I rewrite the objective function in the preceding utility maximization problem in expectations (omitting history dependence terms) as follows. Let $x_t \equiv we_t$ be the amount of income earned in period t (i.e. take-home pay). Then

$$D_t u_t(x_t, x_{t-1}) \equiv \lambda D_t x_t - g_t\left(\frac{D_t}{w} x_t, \frac{D_{t-1}}{w} x_{t-1}\right) \quad (3.9)$$

Note that it is possible to factor the D_t term out of $g_t(\cdot)$ if $D_t = 1$, because $g_t(0, \cdot) = 0$. Thus,

$$u_t(x_t, x_{t-1}) \equiv \lambda x_t - g_t\left(\frac{1}{w} x_t, \frac{D_{t-1}}{w} x_{t-1}\right)$$

With this expression, the expected utility objective function in the first step becomes

$$EU = \sum_{t=1}^T D_t E[u_t(x_t, x_{t-1})] \quad (3.10)$$

where expectations are over possible realizations of effort caps. The x_t terms are treated as random variables because of uncertainty over effort caps.

The concavity of the per-period utility functions $u_t(\cdot)$ will vary among messengers.¹²¹ A second order Taylor approximation of this utility function around expected values gives

$$\begin{aligned} u_t(x_t, x_{t-1}) \simeq & u(\cdot) + (x_t - E[x_t]) \frac{du}{dx_t} + (x_{t-1} - E[x_{t-1}]) \frac{du}{dx_{t-1}} \\ & + \frac{1}{2} (x_t - E[x_t])^2 \frac{d^2u}{dx_t^2} + \frac{1}{2} (x_{t-1} - E[x_{t-1}])^2 \frac{d^2u}{dx_{t-1}^2} \\ & + \frac{1}{2} (x_t - E[x_t]) (x_{t-1} - E[x_{t-1}]) \frac{d^2u}{dx_t dx_{t-1}} \end{aligned} \quad (3.11)$$

where all utility functions and its derivatives on the right hand side are evaluated at the expected values $(E[x_t], E[x_{t-1}])$. Taking expectations, the equation becomes¹²²

$$E[u_t(x_t, x_{t-1})] \simeq u(E[x_t], E[x_{t-1}]) + \frac{1}{2} Var[x_t] \frac{d^2u}{dx_t^2} + \frac{1}{2} Var[x_{t-1}] \frac{d^2u}{dx_{t-1}^2} \quad (3.12)$$

The second derivatives of the utility function evaluated at the expected values are negative¹²³—and they will be more negative for messengers deemed to be “risk averse” by the lottery survey.

With this Taylor approximation, the messenger’s objective function becomes

$$EU = \sum_{t=1}^T D_t \left\{ u(E[x_t], E[x_{t-1}]) + \frac{1}{2} Var[x_t] \frac{d^2u}{dx_t^2} + \frac{1}{2} Var[x_{t-1}] \frac{d^2u}{dx_{t-1}^2} \right\} \quad (3.13)$$

From this equation, we can see that the more risk averse a messenger is (i.e. the more

¹²¹This concavity manifests itself through differences between messengers in the cost of effort function. Note that for committed periods (i.e. $D_t = 1$)

$$\begin{aligned} \frac{d}{dx_t} u_t(x_t, x_{t-1}) &= \lambda - \frac{dg_t}{de_t} \left(\frac{1}{w} \right) > 0 \\ \frac{d^2}{dx_t^2} u_t(x_t, x_{t-1}) &= -\frac{d^2g_t}{de_t^2} \left(\frac{1}{w^2} \right) < 0 \end{aligned}$$

where the first term is positive for $e_t > 0$, because if it were negative, then $e_t = 0$ as it would not make sense for the messenger to exert any positive effort to earn income x_t that period. Similar equations can be derived for concavity with respect to x_{t-1} .

¹²²The covariance term between x_t and x_{t-1} is zero because the effort caps (the random component) are independent draws across periods.

¹²³See footnote 121.

negative the second order derivatives $\frac{d^2 u}{dx_t^2}$ and $\frac{d^2 u}{dx_{t-1}^2}$ are), the more he or she dislikes variations in per-period incomes (i.e. the $Var[x_t]$ terms).

When the wage increases, the messenger will want to increase¹²⁴ total effort and income, and can either respond on the intensive or extensive margin (or some combination of both). But more risk averse messengers will respond in such a way so as to minimize the variance of daily incomes, as it enters more negatively into their expected utility functions. And the way to do this is to focus more on increasing the extensive margin (i.e. adding more periods by increasing D). Of course, by committing to additional periods, the messenger will have to pay additional fixed costs, but this additional cost might be worth it, being less costly than an increase in negative utility due to relatively higher variance terms had additional periods not been added.

To illustrate this point, suppose the choice for a messenger is between the options of

1. Either not add one more period and increase effort per period by a large magnitude in existing periods of work [Intensive margin approach]
2. Or add exactly one more period and decrease effort per period, spreading out effort into the newly added period (or increase effort per period only slightly, a possibility explained previously) [Extensive margin approach]

In Appendix 3.6, I show that the more positive the effort response $\frac{de_t}{dw}$ is, the more positive the change in $E[x_t]$ and $Var[x_t]$ will be, for any period t . I also note that for certain cases where $\frac{de_t}{dw}$ is negative, $E[x_t]$ and $Var[x_t]$ decrease.

Using this relationship, if the messenger chooses option #1, then $\frac{de_t}{dw}$ will be large and positive for most periods, and there will be relatively large increases in $E[x_t]$ and $Var[x_t]$ (as well as $E[x_{t-1}]$ and $Var[x_{t-1}]$) for all the periods t in which $D_t = 1$. This messenger gains utility as each summand's first term $u(E[x_t], E[x_{t-1}])$ increases, but loses utility from the latter two negative terms in the summand becoming more negative. More risk averse

¹²⁴Recall there is no income effect.

messengers will feel the losses from the latter two negative terms to a greater extent.

On the other hand, if the messenger chooses option #2, then $\frac{de_t}{dw}$ will either be negative, or small and positive, for most periods, and there will be relatively small increases (or possibly even decreases) in $E[x_t]$ and $Var[x_t]$ (as well as $E[x_{t-1}]$ and $Var[x_{t-1}]$) for all the periods t in which $D_t = 1$ originally. There will also be one additional period of work, in which both $E[x_t]$ and $Var[x_t]$ for that period jumps from zero to a positive amount. This messenger may gain or lose utility from the changes in $E[x_t]$ and $Var[x_t]$ from the periods in which $D_t = 1$ originally. But a more sizable utility gain can come from the addition of the summand from the single new period of work. Only if this gain from the additional summand is large enough will option #2 be viable. (If the summand is small, or negative even, it means that the fixed costs inside the $u(E[x_t], E[x_{t-1}])$ term are too high.) If viable, more risk averse messengers will prefer this option because either the increases in $Var[x_t]$ over all periods are smaller (relative to option #1), or the $Var[x_t]$ terms actually decrease.

Which option is adopted in the end in response to the wage increase will depend on the size of the utility gain from working the additional period (net fixed costs) in option #2, relative to the size of the net utility loss associated with the larger increases in risk from variation in income over multiple periods experienced in option #1. There is, in essence, a trade-off between fixed costs of additional periods and greater risks from uncertain effort caps. The extent to which using the extensive margin (option #2) is preferred over the intensive margin (option #1) is increasing in the messenger's degree of risk aversion.

To reiterate the intuition behind this result, risk averse messengers place a more negative weight on variation in income per period. When the higher wage increases the desire for more effort exertion, these messengers are more willing to pay the fixed costs of committing to additional periods, and spread out effort exertion. This is because by spreading out effort exertion, there is less risk that the chosen effort amounts in each period will hit the random effort caps, thus reducing variation in income per period—a reduction which is preferred by risk averse messengers.

Moreover, because $x_t \equiv we_t$ is the take-home pay earned by the messenger (i.e. the income after commission calculations, as opposed to the revenue earned by the company), its level and variation ($E[x_t]$ and $Var[x_t]$) can be examined empirically, for both risk averse and non-risk averse messengers, before and after a wage increase. In particular, take-home pay per shift increases during periods of higher wage, but this increase will be larger for non-risk averse messengers who prefer option #1 over option #2. Similarly, risk averse messengers prefer to avoid binding effort caps, so the variation in their take-home pay per shift ($Var[x_t]$) will be lower compared to their non-risk averse counterparts, who are limited more often by binding effort caps. Holding the wage rate constant, weather conditions have similar effects on take-home pay as they do on effort choice—again, through changes in the shape of the effort cost function. Uncomfortable weather conditions which increase the marginal cost of effort will reduce the level of effort exerted, *ceteris paribus*.

Columns (1) and (2) of Table 3.2 summarize the predictions of the two-step neoclassical model for messengers of differing degrees of risk aversion, with respect to various factors as discussed throughout this section. For the outcome variables revenue per shift (prediction [1]), number of shifts per block (prediction [2]), total revenue per block (prediction [3]), and the level of take-home pay per shift (prediction [4]), the arrows indicate the predicted direction of that outcome's movement in response to a wage increase. Take-home pay variance within messenger (prediction [5]) refers to the amount of variation in take-home pay for the two types of messengers as described in the previous paragraph. Whether or not effort decisions are correlated across periods (prediction [6]) speaks to the potential non-time-separable aspect of the model. For weather conditions affecting revenue per shift (prediction [7]) and take-home pay per shift (prediction [8]), the arrows indicate the predicted direction of that outcome's movement in response to worsening weather (i.e. an increase in the marginal cost of effort), holding wage and all other factors constant.

Table 3.2: Summary of Model Predictions

Model:	Two-step Neoclassical This study		Reference Dependent Preferences Fehr and Goette (2007)	
	(1)	(2)	(3)	(5)
Messenger:	-Risk Averse	Risk Averse	-Loss Averse	Loss Averse
			Stays on RP	Moves off RP
	<u>Wage Effects (in Response to a Wage Increase)</u>			
[1] Rev. / Shift	↓* or ↑	Larger ↓* or Smaller ↑	↑	↓* Smaller ↓*
[2] Shifts / Block	↑*	Larger ↑*	↑*	↑*
[3] Rev. / Block	↑*	↑*	↑*	↑*
[4] Take-home Pay / Shift	↑*	Smaller ↑*	↑*	0 Smaller ↑*
[5] Variance of Take-home Pay	Higher*	Lower*	Higher*	0 Lower*
[6] Effort Decisions Across Periods	Correlated*	Correlated*	Uncorrelated	Uncorrelated
	<u>Weather Effects (in Response to Worse Weather Conditions)</u>			
[7] Rev. / Shift	↓*	↓*	↓*	0 Smaller ↓
[8] Take-home Pay / Shift	↓*	↓*	↓*	0 Smaller ↓

Notes: Asterisks (*) denote predictions which are accounted for in the empirical analyses. Column (4) is for the case where the messenger stays on the reference point (RP) in response to a variable change. Column (5) is for the case where the messenger move off the RP in response to a variable change. Comparisons made in column (2) (e.g. “smaller”) are with respect to predictions in column (1). Comparisons made in columns (4) and (5) are with respect to predictions in column (3).

3.3.4 Predictions of Reference Dependent Preferences Model in Fehr and Goette (2007)

The reference dependent preferences model in Fehr and Goette (2007) will have somewhat different predictions regarding these 8 aspects of the data. Their model captures reference dependent preferences by using the kinked one-period time-separable utility function

$$v(e_t) = \begin{cases} \lambda(we_t - \tilde{y}) - g_t(e_t) & \text{if } we_t \geq \tilde{y} \\ \gamma\lambda(we_t - \tilde{y}) - g_t(e_t) & \text{if } we_t < \tilde{y} \end{cases}$$

where $\tilde{y}$ is the reference point and $\gamma > 1$ represents the degree of loss aversion.¹²⁵

For non-loss averse messengers, $\gamma = 1$ and the utility function degenerates into the standard non-time-separable neoclassical model, where $\tilde{y}$ is merely a monotonic shift in utility level. In this case, their FOCs for each period are

$$\lambda w \leq g'(e_t)$$

where the binding case is if optimal effort choice is positive. When the wage w increases, the FOC is more likely to be binding for all periods; thus, the number of shifts worked per block increases. For the case where the FOC is already binding, under assumptions made on the cost of effort function, an increase in wage will result in an increase in effort per shift. Total revenue per block will necessarily increase since messengers must necessarily benefit from the higher wage regardless of how they choose to reoptimize effort levels across periods.

Since effort per shift and wage both increase, take-home pay (we_t , the product of the two) must also increase. With differences in the period-specific cost of effort functions, there will be variation in effort levels across periods, and thus variation in take-home pay. Given that

¹²⁵In their formulation of the model, Fehr and Goette use the cost function $g(e_t, x_t)$ where x_t captures other period-specific factors affecting disutility in effort. In my formulations, I have combined this into a (period-specific) cost of effort function, denoted as $g_t(e_t)$.

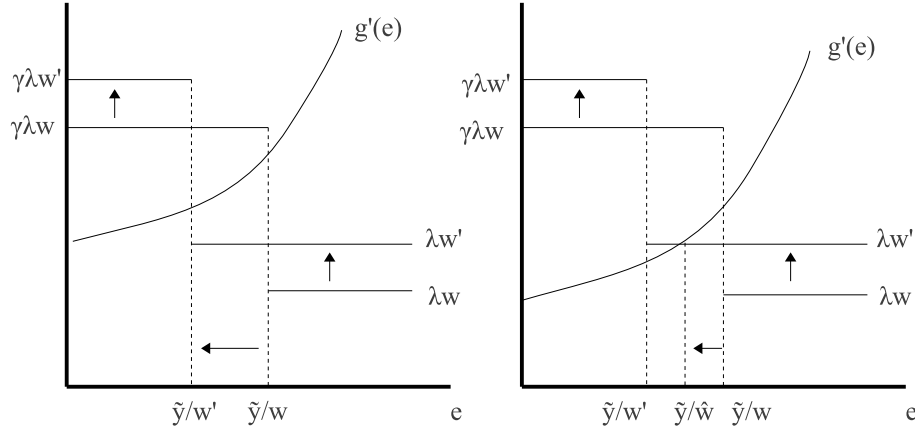
this is a time-separable model, effort decisions are uncorrelated across periods. Furthermore, when weather conditions worsen in a period, the marginal cost of effort increases. With no change in the wage, both effort exertion and take-home pay decrease as a result. These predictions for the non-loss averse messenger are summarized in column (3) of Table 3.2.

For loss averse messengers, with $\gamma > 1$, the comparative statics of the reference dependent preferences model differ. To analyze them, I will assume that prior to any change in variables, the original effort level chosen by the loss averse messenger is such that take-home pay is at the reference point $\tilde{y}$ (i.e. $e_t = \frac{\tilde{y}}{w}$).¹²⁶

Firstly, consider the comparative statics of a wage increase for loss averse messengers. When the wage increases from w to w' , there are two possible cases of effort responses to consider. In the first case, the loss averse messenger's final take-home pay "stays on" the reference point $\tilde{y}$, so the final effort choice is $e_t = \frac{\tilde{y}}{w'}$. This is depicted in the left panel of Figure 3.1, which graphs the marginal cost curve $g'(e)$, as well as the piecewise marginal benefit levels on either side of the reference point level of effort, before (λw and $\gamma \lambda w$) and after ($\lambda w'$ and $\gamma \lambda w'$) the wage increase. Before the wage increase, the marginal cost curve "crosses" the marginal benefit curve through the gap at effort $e_t = \frac{\tilde{y}}{w}$; this implies that the messenger's take-home pay is originally at the reference point level $\tilde{y}$. After the wage increase, the marginal cost curve again "crosses" the marginal benefit curve through the gap at $e_t = \frac{\tilde{y}}{w'}$; this implies that the messenger's take-home pay stays on the reference point. Note that take-home pay per shift remains the same at $\tilde{y}$ before and after the wage increase. This also implies that there should be no variation in take-home pay. However, effort per shift decreases because the wage is now higher (i.e. $\frac{\tilde{y}}{w'} < \frac{\tilde{y}}{w}$). As with non-loss averse messengers, total revenue per block and the number of shifts per block increase, and effort decisions are uncorrelated across periods. These responses to a wage increase are summarized in the top

¹²⁶If the original effort level were not at $\frac{\tilde{y}}{w}$, then such "loss averse" messengers will behave as if they were not loss averse, since the marginal benefit of effort (λw) is constant near this original effort level, and the utility function is no longer kinked. Furthermore, this somewhat defeats the purpose of using reference dependent preferences to describe labor market decisions in the first place, if reference points do not play a binding role in such decisions.

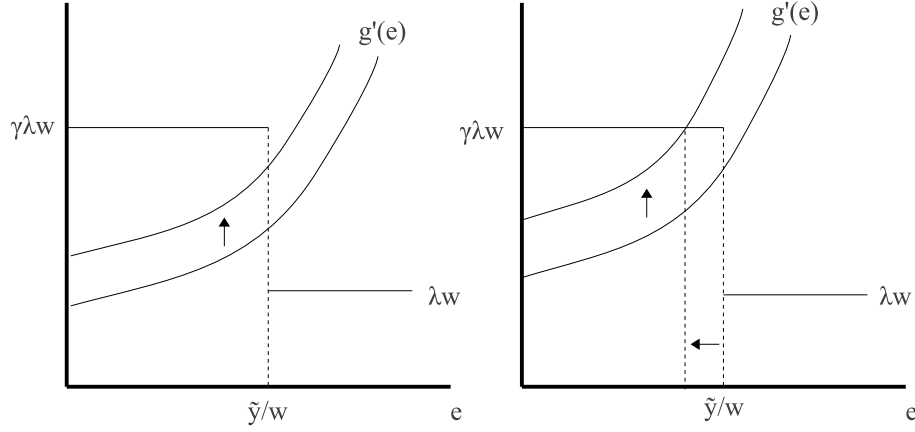
Figure 3.1: Wage Effects in Reference Dependent Preferences Model



half of column (4) in Table 3.2.

The second case is where the loss averse messenger's final take-home pay “moves off” the reference point $\tilde{y}$, so the final effort choice is $e_t = \frac{\tilde{y}}{\hat{w}}$, where $\frac{\tilde{y}}{w'} < \frac{\tilde{y}}{\hat{w}} < \frac{\tilde{y}}{w}$. This is depicted in the right panel of Figure 3.1. As in the previous case, before the wage increase, the marginal cost curve $g'(e)$ “crosses” the marginal benefit curve through the gap at effort $e_t = \frac{\tilde{y}}{w}$; this implies that the messenger's take-home pay is originally at the reference point level $\tilde{y}$. However, after the wage increase, the marginal cost curve now intersects the lower $\lambda w'$ segment of the marginal benefit curve, resulting in a decreased effort level of $e_t = \frac{\tilde{y}}{\hat{w}}$. The magnitude of this effort decrease is smaller than the magnitude of the decrease in the “stays on reference point” case. To compare take-home pay before and after the wage increase, note that the initial take home pay amount is $\tilde{y} = w e_t$. After the wage increase, final take-home pay becomes $w' \left(\frac{\tilde{y}}{\hat{w}} \right) = \left(\frac{w'}{\hat{w}} \right) \tilde{y}$. Since $w' > \hat{w}$, the final take-home pay will be higher than the initial amount (i.e. $\left(\frac{w'}{\hat{w}} \right) \tilde{y} > \tilde{y}$), despite the effort decrease. However, the magnitude of this increase is smaller than the magnitude of the increase in take-home pay experienced by the non-loss averse messenger, since the latter's increase is a result of an increase in both wage and effort. This also implies that while variation in take-home pay within messenger (for

Figure 3.2: Weather Effects in Reference Dependent Preferences Model



this group of “moves off reference point” loss averse messengers) is non-zero, the variation is lower than that of non-loss averse messengers. These responses to a wage increase are summarized in the top half of column (5) in Table 3.2; other predictions are similar to those of previous cases.

Secondly, consider the comparative statics of changes in weather conditions for loss averse messengers. Again, there are two cases to consider: loss averse messengers who “stay on” the reference point, and loss averse messengers who “move off” the reference point. Suppose weather conditions worsen and the marginal cost of effort increases. The marginal cost of effort and marginal benefit curves of loss averse messengers who “stay on” the reference point are depicted in the left panel of Figure 3.2. When the marginal cost of effort curve shifts up, it still “crosses” the marginal benefit curve at the gap at effort level $\frac{\tilde{y}}{w}$. In this case, effort level, and thus take-home pay, remains the same.

For the loss averse messengers who “move off” the reference point, their marginal cost of effort and marginal benefit curves are depicted in the right panel of Figure 3.2. Here, the marginal cost of effort curve shifts up, but now intersects the top $\gamma\lambda w$ segment of the marginal benefit curve. Optimal effort choice decreases, and so does take-home pay. However, these

decreases will be smaller than the corresponding decreases in effort and take-home pay of the non-loss averse messenger with the same increase in marginal cost of effort. This is because for loss averse messengers starting at the reference point effort level $\frac{\tilde{y}}{w}$, marginal cost at that point must increase until the level $\gamma\lambda w$ before having an impact on optimal effort choice.

The predictions for these two cases of loss averse messengers responding to worsening weather conditions are summarized in the last two rows of columns (4) and (5) in Table 3.2.

A general comparison of the predictions of the two-step neoclassical model presented in the previous subsection, and those of the reference dependent preferences model in Fehr and Goette (2007), reveals several differences.

- For non-risk averse messengers in the former framework, effort (as measured by revenue) per shift can decrease or increase, whereas the reference dependent preferences model would predict that non-loss averse messengers would necessarily increase their effort per shift.
- The two-step neoclassical model predicts that the number of shifts per block response by risk averse messengers is larger than that of non-risk averse messengers. However, the reference dependent preferences model does not make any specific prediction about the relative size of these responses.
- With respect to the response of take-home pay per shift to a wage increase, differences between messenger types will depend on whether the loss-averse messengers in the Fehr and Goette (2007) model “stay on” or “move off” their reference point. If they “move off”, then there is no difference in the two models’ predictions. However, if the loss averse messenger “stay on” the reference point, then one would expect no change in take-home pay for these messengers in the reference dependent preferences model, compared to risk averse messengers in the neoclassical model, who see an increase in take-home pay.
- A similar point is true for comparisons between the two models with respect to variance

of take-home pay. If loss averse messengers “move off” the reference point, then the predictions are identical. Otherwise, there is a difference if loss averse messengers “stay on” the reference point.

- Given that the neoclassical model I present is non-time separable, effort decisions are correlated between periods. On the other hand, the model in Fehr and Goette (2007) is time separable, which implies that effort decisions are uncorrelated between periods. (Of course, this does not preclude extending the Fehr and Goette (2007) model to make it non-time separable.)
- The two-step neoclassical model does not say anything regarding the relative sizes of weather effects between risk averse and non-risk averse messengers—just that effort responds negatively to worse weather conditions. However, in the reference dependent preferences framework, non-loss averse messengers respond to worse weather conditions with much larger decreases in effort compared to loss averse messengers. And if loss averse messengers “stay on” the reference point, they should not respond at all to changes in weather conditions.

These differences allow us to examine which set of predictions from the two models is more aligned with patterns in the data.

3.4 Additional Empirical Analyses

This section describes the results from further analyses utilizing the bicycle messenger data from Fehr and Goette (2007).¹²⁷ I test the 8 predictions from the previous section in order to see which model is better able to describe pattern observed in the data.

The units of analysis are at the messenger-block and messenger-shift levels, depending on the specification. The data contain variables for each messenger which include

¹²⁷These are available online at the AER website accompanying the paper.

- Total revenues per block
- Number of shifts per block
- Revenues per shift
- Messenger characteristics (sex, tenure, treatment group, degree of loss aversion)
- Indicator for receipt of high wage during a particular period
- Messenger, block and date identifiers
- Weather variables (temperature, relative humidity) merged by date¹²⁸

All revenue amounts are in Swiss Francs (SFr) and all weather data are in metric units (degrees Celsius and % relative humidity). Table 3.3 reports summary statistics for the data broken down by treatment group. The numbers suggest that randomization between groups A and B was carried out adequately.¹²⁹

While the analyses are divided into multiple subsections, most of the regressions have the form

$$y_{it} = \alpha + \beta_0 (high_{it} \times R0_i) + \beta_1 (high_{it} \times R1_i) + \beta_2 (high_{it} \times R2_i) + X_{it}\gamma + \mu_i + \mu_t + \varepsilon_{it} \quad (3.14)$$

where

- y_{it} is the outcome of interest for messenger i at period t (depending on the context, a period is either a (four-week) block or a shift)
- α is a constant term

¹²⁸These weather data for the Zurich-Affoltern meteorological station from August 2000 to November 2000 were obtained separately, courtesy of the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss).

¹²⁹Additional details regarding the dataset can be found in Fehr and Goette (2007).

Table 3.3: Summary Statistics

	(1)	(2)	(3)
Variable	Overall	Group A	Group B
Total Revenue / Block	3517.85	3613.82	3411.22
(in SFr)	(2455.01)	(2677.79)	(2201.34)
Number of Shifts / Block	11.76	12.12	11.37
	(7.62)	(7.77)	(7.51)
Revenue / Shift	316.56	316.7	316.37
(in SFr)	(126.72)	(127.66)	(125.62)
Tenure	382.71	374.67	393.18
(in shifts)	(240.38)	(256.87)	(216.86)
High wage shifts	0.406	0.395	0.421
(proportion)	(0.491)	(0.489)	(0.494)
Female	5	1	4
	[13.2%]	[5%]	[22.2%]
Rejects Neither Lottery	23	13	10
	[60.5%]	[65%]	[55.6%]
Rejects One Lottery	7	4	3
	[18.4%]	[20%]	[16.7%]
Rejects Both Lotteries	8	3	5
	[21.1%]	[15%]	[27.8%]
N (messengers)	38	20	18
N (messenger-shifts)	1137	643	494

Note: Standard deviations in parentheses. Percentage shares of all messengers in square brackets.

- $high_{it}$ is an indicator of whether messenger i receives the high wage treatment in period t
- $R0_i$, $R1_i$ and $R2_i$ are mutually exclusive indicators denoting whether messenger i rejects zero, one or two of the lotteries respectively
- X_{it} is a vector of other covariates
- μ_i and μ_t are messenger i and period t fixed effects, and
- ε_{it} is an error term

This is the same specification used in the regressions in Table 6 of Fehr and Goette (2007), where outcome y_{it} the revenue per shift. β_0 , β_1 and β_2 measure the “treatment effects” of a high wage for messengers who reject zero, one or two of the lotteries respectively. Depending on which are appropriate, μ_t are fixed effects at the date or block level. Logged tenure is included in X_{it} for most specifications. Note that base level indicators for $R0_i$, $R1_i$ and $R2_i$ are omitted because of the messenger fixed effects.

Furthermore, two restrictions to this regression are sometimes considered. They are

$$y_{it} = \alpha + \beta_0 (high_{it} \times R0_i) + \beta_A (high_{it} \times RA_i) + X_{it}\gamma + \mu_i + \pi_t + \varepsilon_{it} \quad (3.15)$$

where a single measure $RA_i \equiv R1_i + R2_i$ indicates rejection of any lottery, and

$$y_{it} = \alpha + \beta_H high_{it} + X_{it}\gamma + \mu_i + \pi_t + \varepsilon_{it} \quad (3.16)$$

where the $high_{it}$ indicator is no longer interacted with aversion measures.¹³⁰

3.4.1 Revenue Per Shift (Prediction [1])

As noted in Fehr and Goette (2007), revenue (effort) per shift responds to wage increases negatively—and more so for messengers rejecting lotteries. Columns (1) through (4) of Table 3.4 shows regression results for the outcomes y_{it} of logged revenues per shift (as in Fehr and Goette (2007)) and (level) revenues per shift. The first two columns replicate the regressions in Fehr and Goette’s Table 6.¹³¹ Column (1) shows logged revenues per shift using the equation (3.14) specification; column (2) shows them using the equation (3.15) specification. Both sets of estimates show that it is the messengers rejecting lotteries who are driving the negative effort response to the wage increase. Columns (3) and (4) use the

¹³⁰This specification in equation (3.16) is used in several of the specifications in Fehr and Goette (2007). I present these along with related results for completeness.

¹³¹The results differ slightly due to data cleaning and dissimilar exclusion rules.

Table 3.4: Revenue (Effort) per Shift Regressions

Dep. Var. (/ shift):	(1) Log Rev.	(2) Log Rev.	(3) Revenue	(4) Revenue	(5) Take-home
High×R0	-0.028 (0.032)	-0.028 (0.033)	-12.887 (9.493)	-12.951 (9.478)	26.699*** (4.14)
High×R1	-0.085 (0.064)		-19.604 (12.286)		19.175*** (5.552)
High×R2	-0.12** (0.055)		-27.324* (14.583)		20.967*** (6.717)
High×RA		-0.104** (0.046)		-23.962** (10.051)	
Constant	5.521*** (0.332)	5.555*** (0.327)	256.727*** (83.433)	264.202*** (80.466)	95.006** (36.835)
N	1137	1137	1137	1137	1137
R-square	0.26	0.26	0.312	0.312	0.328

Note: * = $p < 0.10$, ** = $p < 0.05$, *** = $p < 0.01$. Standard errors clustered at the date level in parentheses. All outcomes are per shift. All specifications include logged tenure, as well as date and messenger fixed effects.

same respective specifications, except the outcome is now the (level) revenue per shift.¹³² On average, messengers who reject any lottery reduce effort as measured by revenue by SFr 23.96 per shift.

One surprising result is that the coefficient estimates for treatment interacted with the indicator $R0_i$ is negative and insignificant in columns (1) through (4), suggesting that revenue per shift of these messengers did not respond much (or responded with a slight decrease) to the wage increase.¹³³ This differs from what the reference dependent preferences model would predict (see prediction [1] in Table 3.2).¹³⁴ This negative though insignificant estimate seems more amiable to the two-step neoclassical model, which predicts that the effort response can be either positive or negative, as well as the relatively more negative response of risk averse messengers.

¹³²I present these estimates for the levels for consistency with my later analyses on revenues per block, which are also in levels.

¹³³Although with these standard errors, a slight positive response cannot be ruled out.

¹³⁴Alternatively it could be the case that there is a considerable negative income effect.

Table 3.5: Number of Shifts per Block Regressions

Dep. Var.: Shifts / Block	(1) FG	(2) Reject 1 or 2	(3) Reject Any
high	3.683*** (1.047)		
high×R0		3.151** (1.371)	3.163** (1.362)
high×R1		3.036 (2.096)	
high×R2		5.738*** (2.038)	
high×RA			4.473*** (1.489)
Constant	11.684*** (1.029)	11.684*** (1.025)	11.684*** (1.029)
N	114	114	114
R-square	0.695	0.699	0.697

Note: * = $p < 0.10$, ** = $p < 0.05$, *** = $p < 0.01$. Robust standard errors in parentheses.
All specifications include block and messenger fixed effects.

3.4.2 Number of Shifts and Total Revenue (Predictions [2] and [3])

The increase in the number of shifts per block in response to higher wages is different for messengers who responded to the lottery survey differently. Table 3.5 shows regressions with number of shifts per block as the outcome variable. Column (1) reports the coefficient estimate from regressions using the equation (3.16) specification, equivalent to the specification in column (4) of Table 3 in Fehr and Goette (2007).¹³⁵ On average, messengers work 3.68 shifts more during high-wage treatment blocks. (This estimate is statistically significant at a 1% level.)

Breaking this estimate down by lottery survey response group paints a more complex picture. Columns (2) and (3) report coefficient estimates from regressions using equation (3.14) and (3.15) specifications respectively. These estimates suggest that the increase in

¹³⁵The estimates are not identical because of differences in sample restrictions and data cleaning rules.

the number of shifts in response to the wage increase varies. For instance, from column (3), messengers who rejected zero lotteries worked on average 3.16 shifts more during high-wage treatment blocks. On the other hand, messengers who rejected any lottery worked on average 4.47 shifts more during high-wage treatment blocks—a relatively larger response. These estimates are statistically significant at 5% and 1% levels respectively. While the joint significance test with the null that the two coefficient estimates are equal cannot be rejected at a 5% level, the point estimates are nonetheless in line with prediction [2] for the two-step neoclassical model, which states that risk averse messengers increase the number of shifts by more, compared to non-risk averse messengers.

A similar difference is observed with regard to the increase in total revenue per block in response to higher wages for messengers who responded to the lottery survey differently—a difference not predicted by either model (see prediction [3]). Table 3.6 shows similar regressions with total revenue per block as the outcome variable. Column (1) reports the coefficient estimate from regressions using the equation (3.16) specification, equivalent to the specification in column (1) of Table 3 in Fehr and Goette (2007).¹³⁶ On average, messengers generate SFr 924.19 more in revenue (effort) during high-wage treatment blocks.¹³⁷

Breaking this estimate down by lottery survey response group, columns (2) and (3) report coefficient estimates from regressions using equation (3.14) and (3.15) specifications respectively. These estimates suggest that the increase in total revenue (effort) in response to the wage increase varies. For instance, from column (3), messengers who rejected zero lotteries have total revenues which are on average SFr 809.15 higher during high-wage treatment blocks. On the other hand, messengers who rejected any lottery have total revenues which are on average SFr 1098.77 higher during high-wage treatment blocks—again a relatively larger response. These estimates are statistically significant at 10% and 5% levels respectively; however, the joint significance test with the null that the two coefficient estimates are

¹³⁶As before, the estimates are not identical because of differences in sample restrictions and data cleaning rules.

¹³⁷This estimate is statistically significant at a 1% level.

Table 3.6: Total Revenue per Block Regressions

Dep. Var.: Rev. / Block	(1) FG	(2) Reject 1 or 2	(3) Reject Any
high	924.191*** (320.207)		
high×R0		804.66* (420.946)	809.147* (418.32)
high×R1		547.107 (571.072)	
high×R2		1584.265** (633.664)	
high×RA			1098.769** (444.782)
Constant	3444.19*** (298.507)	3444.19*** (296.855)	3444.19*** (299.121)
N	114	114	114
R-square	0.739	0.744	0.74

Note: * = $p < 0.10$, ** = $p < 0.05$, *** = $p < 0.01$. Robust standard errors in parentheses.
All specifications include block and messenger fixed effects.

equal cannot be rejected at a 5% level.

3.4.3 Take-home Pay (Predictions [4] and [5])

Whereas revenue per shift generally decreases with the high wage treatment, evidence suggests that take-home pay is positively correlated with wage. In order to analyze this, I first create a take-home pay variable by multiplying the revenue per shift with the commission rate (wage) received by the messenger, depending on whether that shift was a “treated” shift with a high wage.¹³⁸

Column (5) of Table 3.4 shows the coefficient estimates for the regression using the equation (3.14) specification and with take-home pay as the outcome variable. The coefficient estimates suggest that across all three survey response categories, take-home pay increased by between SFr 19 and SFr 27 during treatment periods of high wages. And while the coefficient estimates for both those rejecting one or two lotteries are slightly lower than those rejecting none, these differences are not statistically significant.¹³⁹

This increase in take-home pay for those who rejected one or two lotteries is inconsistent with the prediction of loss averse messengers who “stay on” the reference point, for whom take-home pay does not stray from the reference point income level (see prediction [4]). This suggests that if messengers did have reference dependent preferences, the loss averse ones would be the “move off” the reference point type of messengers who experience some increase in take-home pay. If this were the case, then these estimates for take-home pay are consistent with both models, which both predict that take-home pay increases, but to a lesser degree for risk/loss averse messengers.

Next, I examine variation in take-home pay within messenger (see prediction [5]). Both models predict that variance in take-home pay is lower for messengers rejecting more lotteries (i.e. more risk/loss averse); moreover, for messengers who “stay on” the reference point, there

¹³⁸The commission rate was increased from 0.39 to 0.49 for men, and from 0.44 to 0.54 for women.

¹³⁹In the joint significance test with the null hypotheses that $\text{high} \times \text{R0}$ is different from both the coefficients on $\text{high} \times \text{R1}$ and $\text{high} \times \text{R2}$, the p-value is 0.514.

Table 3.7: Standard Deviation of Take-home Pay

Number Rejected:	0 lotteries	1 lottery	2 lotteries
Mean (of s.d.)	48.43	41.84	42.88
s.d. (of s.d.)	20.00	11.26	14.80
Minimum	21.46	22.15	22.53
Maximum	101.36	57.93	67.68
N	23	7	8

should be near-zero variation in take-home pay, if they successfully target the reference point income level. I measure the within-messenger variation in take-home pay by calculating the standard deviation of take-home pay across periods for each of the 38 messengers in the sample.

Table 3.7 reports means (over messengers) of the standard deviations of take-home pay by lottery-rejection groups, as well as other statistics. While slightly smaller, the differences between the means of messengers rejecting one or two lotteries, and to those rejecting neither lotteries, are not statistically significant. But what is striking is that there is still much within-messenger variation in take-home pay for messengers who reject one or two lotteries. In the reference dependent preferences model, these messengers would have very low variation in take-home pay if they managed to “stay on” the reference point. This piece of evidence again suggests that if we were to use a reference dependent preferences model, the loss averse messengers would likely be the type that “move off” the reference point.¹⁴⁰

3.4.4 Weather Effects (Predictions [7] and [8])

Both the two-step neoclassical model and the reference dependent preferences model have certain predictions regarding the relationship between weather conditions and effort decisions, as well as between weather conditions and take-home pay (see predictions [7] and [8])

¹⁴⁰In Appendix x, I consider one additional possibility as to why there is so much variation in take-home pay within messenger by using Koszegi and Rabin’s (2006) model, which treats reference points as endogenously determined.

Table 3.8: Weather Summary Statistics

4-week Block Starting:	Aug 14	Sept 11	Oct 30	Overall
Temperature	17.22	14.28	6.61	12.70
(in Celsius)	(3.23)	(2.88)	(2.07)	(5.27)
Relative Humidity	78.00	84.36	82.61	81.66
(in %)	(6.46)	(4.31)	(8.84)	(7.21)
N	20	20	20	60
Treatment Group	Neither	A	B	-

Note: Weekdays only. Standard deviations in parentheses.

in Table 3.2). To test them, I obtain weather data from the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) for the Zurich-Affoltern meteorological station from August 2000 to November 2000. The main variables of interest are daily mean temperature (in degrees Celsius) and relative humidity (in %), which are merged into the messenger data by date. Table 3.8 contains summary statistics for the weather variables by block.¹⁴¹

Weather conditions—which I take to be exogenously determined and essentially random— affect effort decisions because inclement weather increases the marginal cost to effort exertion. In particular, literature on heat indices and human perception of temperature suggests that discomfort experienced will depend on the interplay between temperature and relative humidity.¹⁴²

Table 3.9 reports the coefficient estimates from regression specifications which now include in X_{it} the five weather variables of (daily mean) temperature, its square, relative humidity, humidity interacted with temperature, and humidity interacted with temperature squared.¹⁴³ Equation (3.15) specifications, which include the high-wage indicator interacted

¹⁴¹These statistics pertain only to the weekdays on which messengers were observed to have worked in the data.

¹⁴²Humidity accentuates discomfort at extreme temperatures: high temperatures feel hotter in humid weather because sweat evaporates at a slower rate; conversely, low temperatures feel colder in humid weather because humid air's specific heat capacity is higher, thus conducts heat away from the body at a faster rate.

¹⁴³A quadratic in temperature is used to reflect the fact that temperature extremes generate more disutility. Temperature interactions with humidity are included because of reasons given in footnote 142.

with two lottery survey result groupings (R0 and RA), are used in order to obtain weather effects conditional on wage. These regressions exclude date fixed effects, which are co-linear with the weather variables; they instead include block fixed effects in their place. Column (1) shows the regression with revenue per shift as the outcome variable, but without weather variables or date fixed effects; this is for comparison purposes.¹⁴⁴

The specification in column (2) adds in weather variables. From the statistically significant coefficients on the weather variables, it is clear that weather conditions do affect the effort decisions of messengers in general.¹⁴⁵ The joint significance test on these five coefficients reports a p-value of 0.0001.

Prediction [7] says that under a reference dependent preferences framework, weather affects the effort decisions of loss averse messengers to a lesser extent. To test whether this is true, I interact all the weather variables with (mutually exclusive) indicators for rejection of any lottery in the survey (R0 and RA). Column (3) reports these results, which indicate that weather affects effort no differently between loss averse messengers (who reject any lottery) and non-loss averse ones. The joint significance test for the null that each of the five corresponding coefficient pairs (for R0 and RA) are equal cannot be rejected, with a p-value of 0.664. On the other hand, the joint significance test for the null that all ten coefficients are jointly equal to zero is rejected with a p-value of 0.0003. These results suggest that loss averse messengers are just as easily persuaded by the elements to change their effort decisions as non-loss averse messengers. This finding is more in line with a two-step neoclassical model, which does not make this distinction in weather effects between risk averse and non-risk averse messengers.

¹⁴⁴While smaller and no longer statistically significant, these coefficient estimates are qualitatively similar to those found in column (4) of Table 3.4.

¹⁴⁵Note that effort as a function of temperature is an inverted U shape. The coefficient values suggest that at the mean relative humidity level of 81.66%, the temperature at which messengers exert the most effort is 14.8 degrees Celsius (or 58.6 degrees Fahrenheit). Also, at the mean temperature level of 12.70 degrees Celsius, each percentage point increase in relative humidity reduces effort (as measured by revenue) by SFr 2.62. (The positive coefficient on humidity is rather misleading; one must remember to include in the calculations the interaction terms which contain humidity as well.)

Table 3.9: Weather Regressions

Dep. Var. (/ shift):	(1) Revenue	(2) Revenue	(3) Revenue	(4) Take-home
High×R0	-12.335 (10.664)	-10.331 (10.434)	-9.759 (10.676)	28.165*** (4.67)
High×RA	-18.156 (11.257)	-17.822 (11.056)	-17.689 (11.086)	22.802*** (5.033)
<i>Temp</i> [×R0]		<i>203.893***</i> (46.654)	186.991*** (65.795)	75.513*** (26.92)
[Temp×RA]			226.811*** (58.205)	92.593*** (25.663)
<i>Temp</i> ² [×R0]		-7.157*** (1.675)	-6.292*** (2.405)	-2.492** (0.979)
[Temp ² ×RA]			-8.44*** (2.107)	-3.422*** (0.925)
<i>Humid</i> [×R0]		<i>12.241***</i> (3.149)	11.022** (4.422)	4.497** (1.827)
[Humid×RA]			13.773*** (3.924)	5.584*** (1.751)
<i>Humid</i> × <i>Temp</i> [×R0]		-2.123*** (0.526)	-1.885** (0.754)	-0.75** (0.307)
[Humid×Temp ×RA]			-2.454*** (0.661)	-0.995*** (0.293)
<i>Humid</i> × <i>Temp</i> ² [×R0]		<i>0.075***</i> (0.02)	0.063** (0.028)	0.024** (0.011)
[Humid×Temp ² ×RA]			0.093*** (0.025)	0.038*** (0.011)
Constant	263.045*** (74.015)	-942.876*** (300.967)	-956.219*** (302.848)	-410.249*** (131.061)
N	1137	1137	1137	1137
R-square	0.214	0.239	0.241	0.258

Note: * = $p < 0.10$, ** = $p < 0.05$, *** = $p < 0.01$. Robust standard errors in parentheses.

All outcomes are per shift. For column (2), the coefficients are for variable names excluding terms in square brackets (e.g. 203.893 is the coefficient for just temperature, or “Temp”). For columns (3) and (4), the coefficients are for variable names including terms in square brackets (e.g. 186.991 is the coefficient for temperature interacted with an indicator for rejecting zero lotteries, or “Temp×R0”). To demarcate this, in column (3), I have italicized the estimates corresponding to the italicized variables. All specifications include logged tenure, as well as block and messenger fixed effects. Date fixed effects cannot be included because they would be co-linear with the weather variables.

Prediction [8] makes similar claims regarding differences in weather effects (or lack thereof) on take-home pay between the competing models. Column (4) of Table 3.9 report coefficient estimates for regressions similar to the specification in column (3), but with take-home pay (i.e. the supposed reference point for loss averse messengers) as the outcome. The statistically significant estimates for the messengers who reject any lottery show that, in fact, weather conditions do shift the reference point around substantially. The joint significance test for the null that all ten coefficients are jointly equal to zero is rejected with a p-value of 0.0006.

Similar to before, the coefficients for messengers rejecting any lottery and for those rejecting zero lotteries are very similar. The joint significance test for the null that each of the five corresponding coefficient pairs (for R0 and RA) in column (4) are equal cannot be rejected, with a p-value of 0.716. The similarity of weather effects between these two groups goes against the case for a reference dependent preferences model, given that it would predict smaller weather effects for the loss averse messengers.

3.4.5 Effort of Surrounding Periods (Prediction [6])

In a non-time-separable model, the cost of effort function for this period could depend on the effort exertion from previous periods. This is because having worked harder previously, a messenger may experience increased marginal costs to exerting effort this period. This seems especially relevant in the context of the hard work of bicycle messengers. I check for non-time-separability by including in X_{it} covariates relating to effort decisions of surrounding periods. Table 3.10 shows results from regressions using the equation (3.14) specification with the change in revenue per shift (i.e. effort), from the previous period to this period, as the outcome variable. With the inclusion of lagged dependent variables as covariates, messenger fixed effects are no longer included in these regressions, but date fixed effects still are.¹⁴⁶

¹⁴⁶The coefficient estimates of the interaction variables (between the high indicator and each of R0, R1 and R2) have been suppressed for space reasons. However, these coefficient estimates are qualitatively similar to

Table 3.10: Prior Effort Regressions

Dep. Var.:	(1)	(2)	(3)	(4)
Rev. / Shift	Indicators	OLS	OLS (close)	IV
Worked Yesterday	5.731 (18.885)			
Two-shift day	-48.708*** (17.238)			
Lagged Rev. / Shift		-0.961*** (0.042)	-0.969*** (0.039)	-0.779*** (0.191)
Lagged Rev. / Shift × Close			0.012 (0.026)	
Constant	138.634 (183.252)	117.554*** (30.363)	116.521*** (30.996)	-501.698 (323.686)
N	900	900	900	900
R-square	0.077	0.532	0.532	0.471

Note: * = $p < 0.10$, ** = $p < 0.05$, *** = $p < 0.01$. Standard errors clustered at the date level in parentheses. All specifications include the high indicator interacted with R0, R1, and R2; logged tenure; as well as date fixed effects. See footnote 146 regarding suppressed coefficients.

One way to measure surrounding-period effort is to see whether a messenger worked on the immediately previous day. A second way to measure surrounding-period effort is to see whether a messenger worked two five-hour shifts on the same day. Two indicator variables for these measures are generated and included as covariates in the specification in column (1). The coefficient estimate on the first indicator is not significantly different from zero, suggesting that having worked yesterday does not affect today's effort decision. The coefficient estimate on the second indicator suggests that messengers who work two shifts in a single day exert on average SFr 48.71 less effort (in revenue terms) during those busy-day shifts. (This estimate is statistically significant at a 1% level.)

Another way to test for non-time-separability is to include lagged effort measures in the regression. In column (2), I include the lagged revenue per shift of the last worked shift as a covariate. The 1%-level statistically significant coefficient estimate suggests that for those of previous specifications which exclude the lagged effort variables.

every additional Swiss Franc of effort exerted the previous shift, a messenger will on average subsequently reduce effort exertion this period by SFr 0.96.

One concern may be that the last worked shift need not have been in the immediate past. For example, the last worked shift could have been days ago, especially if the current shift in question is on a Monday. One would expect the negative correlation between effort to be stronger for shifts which are closer to the current shift. To investigate this, I generate a “close” indicator variable which takes the value of one when the last worked shift is either on the same day or the day immediately before. Indeed, only about 52% of last worked shifts were “close” to the current shift in this sense. I then interact this “close” indicator with the lagged effort measure and present the results in column (3). The base coefficient estimate for (uninteracted) lagged revenue per shift is similar at -0.97 (and statistically significant at a 1% level). The coefficient estimate on the interacted term insignificant and close to zero, suggesting that the lagged effort effect is just as strong for “close” shifts.

I further take an instrumental variables (two stage least squares) approach to estimate the effects of lagged effort on changes in effort. I instrument lagged effort with lagged weather variables¹⁴⁷, which I argue are exogenous and random conditional the date fixed effects which are included. The second stage coefficient estimates are presented in column (4). The coefficient on lagged revenue per shift is -0.78 (statistically significant at 1% level), which is qualitatively similar (though slightly smaller in magnitude) compared to the OLS estimate. Unfortunately, the F-stat for the first stage (coefficient estimates not shown) is 3.53, suggesting that while jointly significant at a 1% level, the instruments are weak. Nonetheless, I include these results for completeness.

The findings presented in this subsection suggest that any model of labor supply should take into account the potentially non-time-separable nature of labor supply decisions.

¹⁴⁷The five instruments used are lagged temperature, its square, lagged humidity, lagged humidity interacted with lagged temperature, and lagged humidity interacted with lagged temperature squared.

3.5 Conclusion

In this chapter, I argue that a two-step neoclassical model with uncertainty arising from effort caps is able to account for the empirical findings in Fehr and Goette (2007), as well as findings from the additional analyses carried out in this study. And while the framework with reference dependent preferences does perform well in explaining certain aspects of the data (as long as the loss averse messengers are the type who “move off” the reference point), some of its predictions—such as an increase in effort per shift for non-loss averse messengers or smaller weather effects—are not observed empirically in the bicycle messenger sample.

To summarize, the two-step neoclassical model presented in this chapter is able to explain the observed decrease in effort per shift and increase in number of shifts as arising from messengers spreading an increased amount of total effort over a greater number of shifts. This arises from the non-time-separable nature of the model, which is observed when higher lagged effort makes messengers want to reduce effort this period (relative to the last).

Furthermore, this model is able to distinguish between the different observed responses of risk averse versus non-risk averse messengers. As the model predicts, risk averse messengers respond in a way which focuses more on the extensive margin, seeing larger shift per block increases and greater revenue per shift decreases, when compared to non-risk averse counterparts. The resulting avoidance by risk averse messengers of hitting effort caps also results in slightly lower variance in take-home pay, as observed in the data. Moreover, the extended two-step model has all the hallmarks of a regular neoclassical model, and is therefore able to explain the correlation between specific outcomes and changing weather conditions (via changes in cost of effort).

Of course, both models contain simplifications to the realities of the bicycle messenger services market. Neither model accounts for competition between messengers for delivery jobs. Fehr and Goette (2007) suggest that the demand by messengers for additional work is slack, since there are often unfilled shifts. A more sophisticated model, though, would

account for the demand side from firms needing delivery services in this market market, and the interplay between agents on both sides of the market. This would be especially relevant for my extension involving effort caps, since the size of these caps will invariably depend on the number of messengers who show up during a shift.

From an empirical standpoint, the experimental approach taken by Fehr and Goette (2007) in randomly assigning wage variation is highly innovative and empirically compelling, and may be an indication of the direction in which future research will thrive, especially research focusing on experimentally estimating parameters of the labor market. Their use of the real world setting of bicycle messengers is especially helpful in addressing external validity issues, which are often of concern in experimental economics research involving controlled laboratory settings.

Appendices

3.6 $E[x_t]$ and $\text{Var}[x_t]$

We can think of x_t as the either the income the messenger would receive if limited by the effort cap, or the income received when not limited and exerting the chosen optimal effort. Thus, $x_t = w \min[e_t, \hat{e}_t]$ where $\hat{e}_t$ is drawn from a distribution $f(\hat{e}_t)$ (with cumulative distribution $F(\hat{e}_t)$). Therefore,

$$E[x_t] = w \times E[\min[e_t, \hat{e}_t]]$$

$$\text{Var}[x_t] = w^2 \times \text{Var}[\min[e_t, \hat{e}_t]]$$

When the wage increases,

$$\frac{d}{dw} E[x_t] = E[\min[e_t, \hat{e}_t]] + w \left(\frac{d}{de_t} E[\min[e_t, \hat{e}_t]] \right) \frac{de_t}{dw} \quad (3.17)$$

$$\frac{d}{dw} \text{Var}[x_t] = 2w \text{Var}[\min[e_t, \hat{e}_t]] + w^2 \left(\frac{d}{de_t} \text{Var}[\min[e_t, \hat{e}_t]] \right) \frac{de_t}{dw} \quad (3.18)$$

Since

$$E[\min[e_t, \hat{e}_t]] = \int_0^{e_t} \hat{e}_t dF(\hat{e}_t) + [1 - F(e_t)] e_t \quad (3.19)$$

$$\text{Var}[\min[e_t, \hat{e}_t]] = \int_0^{e_t} (\hat{e}_t - E[x_t])^2 dF(\hat{e}_t) + [1 - F(e_t)] (e_t - E[x_t])^2 \quad (3.20)$$

differentiating with respect to e_t , we obtain

$$\frac{d}{de_t} E[\min[e_t, \hat{e}_t]] = [1 - F(e_t)] > 0 \quad (3.21)$$

$$\frac{d}{de_t} \text{Var}[\min[e_t, \hat{e}_t]] = 2[1 - F(e_t)] \int_0^{e_t} F(\hat{e}_t) d\hat{e}_t > 0 \quad (3.22)$$

thus showing that both moments are increasing in e_t .^{148,149}

This means that in general, the more positive $\frac{de_t}{dw}$ is, the more positive $\frac{d}{dw}E[x_t]$ and $\frac{d}{dw}Var[x_t]$ will be, for any period t . This can be seen from equations (3.17) and (3.18), which are linear in $\frac{de_t}{dw}$. For $\frac{de_t}{dw} \geq 0$, both $E[x_t]$ and $Var[x_t]$ necessarily increase, and the magnitude will be large. However, for $\frac{de_t}{dw} < 0$, there are cases where $E[x_t]$ and $Var[x_t]$ will decrease (if each of the latter terms in equations (3.17) and (3.18) are sufficiently large relative to the first terms), or in other cases, they will increase, but by a smaller magnitude compared to when $\frac{de_t}{dw} \geq 0$. Note also that it could be the case that one increases while the other decreases; the signs of these two derivatives need not be the same.

In turn, the direction and magnitude of $\frac{de_t}{dw}$, which constitutes the intensive margin decision, depends on whether the messenger decides to increase the number of periods to commit

¹⁴⁸The algebra for the second equation is

$$\begin{aligned}
& \frac{d}{de_t} Var[\min[e_t, \hat{e}_t]] \\
&= (e_t - E[x_t])^2 f(e_t) + \int_0^{e_t} \frac{d}{de_t} \left\{ (\hat{e}_t - E[x_t])^2 \right\} dF(\hat{e}_t) \\
&\quad + \frac{d}{de_t} \left\{ [1 - F(e_t)] (e_t - E[x_t])^2 \right\} \\
&= (e_t - E[x_t])^2 f(e_t) - \int_0^{e_t} 2(\hat{e}_t - E[x_t]) \frac{dE[x_t]}{de_t} dF(\hat{e}_t) \\
&\quad - f(e_t) (e_t - E[x_t])^2 + [1 - F(e_t)] 2w(e_t - E[x_t]) (1 - 1 + F(e_t)) \\
&= - \int_0^{e_t} 2(\hat{e}_t - E[x_t]) [1 - F(e_t)] dF(\hat{e}_t) \\
&\quad + 2[1 - F(e_t)] F(e_t) (e_t - E[x_t]) \\
&= 2[1 - F(e_t)] \left\{ F(e_t) (e_t - E[x_t]) - \int_0^{e_t} (\hat{e}_t - E[x_t]) dF(\hat{e}_t) \right\} \\
&= 2[1 - F(e_t)] \left\{ e_t F(e_t) - \int_0^{e_t} \hat{e}_t dF(\hat{e}_t) \right\} \\
&= 2[1 - F(e_t)] \int_0^{e_t} F(\hat{e}_t) d\hat{e}_t
\end{aligned}$$

where the last equality comes from reversing integration by parts.

¹⁴⁹Intuitively, when a messenger chooses a higher optimal effort decision, expected effort for that period will increase. Less intuitive is the fact that variance in effort will also increase. To quote an extreme example, if the messenger chooses a negligible optimal effort amount very close to zero, then for almost all realizations of the effort cap, final effort that period will be very close to zero, with little variation in different states of the world. On the other hand, if the messenger chooses a very high optimal effort amount, then there will be a range of possible effort amounts in different states of the world (some capped, some not), but with much more variation.

to on the extensive margin.

3.7 Endogenous Reference Points

One possible alternate explanation for the rather high variation in take-home pay in a reference dependent preferences framework, even for loss averse messengers supposedly targeting the reference point, is that the reference point in fact varies from period to period, thus moving take-home pay along with it. Recent literature suggests that this may be the case. For instance, Farber (2008) finds that New York City taxi drivers almost never reach the reference point and that there is a large amount of unpredictable day to day variation in the reference point within a given driver. But what determines this variation in the reference point?

Koszegi and Rabin (2006) consider a model in which reference points are determined endogenously, depending on a worker's expectations about earnings potential for that day. Their framework would ascribe an upward shift of reference points in response to a higher wage to loss averse messengers' expectations of predictably higher earning opportunities during these high-wage treatment periods. This would mean that variation in take-home pay exists when variance is calculated over both high- and normal-wage periods taken together because of changing reference points. However, calculating variance within intervals of high- and normal-wage periods, this variation in take-home pay should be much lower, since reference points would be somewhat constant across these periods with unchanging earnings expectations.

Yet the evidence suggests otherwise. Table 3.11 reports means of the standard deviations of take-home pay calculated by treatment status and lottery survey response.¹⁵⁰ With

¹⁵⁰To clarify the numbers in the table using an example, for messengers who rejected zero lotteries in the survey, the mean of their standard deviations over periods when they were part of the control group (experiencing normal wages) is 43.19.

More specifically, these are calculated by first finding for each messenger the standard deviation of take-home pay within high-wage treatment periods, and then the same within normal-wage control periods. Next, the means of these standard deviations over all messengers are calculated for each treatment status by lottery survey response. (In certain cells, there were messengers with zero or only a single observation;

Table 3.11: Standard Deviation of Take-home Pay by Treatment

Number Rejected:	0 lotteries	1 lottery	2 lotteries
Mean (of s.d.), high=0	43.19 (21.33)	39.60 (10.43)	32.50 (19.42)
Mean (of s.d.); high=1	45.45 (23.75)	40.15 (21.87)	48.44 (15.93)
N	22	7	8

Note: Standard deviations of take-home pay for each messenger are calculated within treatment status first, before means of these standard deviations are taken by lottery survey response (see footnote 150). Standard deviation (of s.d.) in parentheses.

the model of Koszegi and Rabin (2006), messengers rejecting lotteries would have stable reference points within each treatment status; hence, we would expect to see significantly lower variation in take-home pay for these messengers in the last two columns. However, the means of the standard deviation in take-home pay for these two groups are very similar to those of messengers rejecting zero lotteries, whether during treatment or control periods; moreover, these means are very different from zero, which is what would be expected if loss averse messengers hit their reference points at least most of the time.

Of course, we have implicitly assumed in this appendix that the messengers “stay on” their reference points when the wage increases. However, this may not be the case if they “move off” the reference point, in which case we revert back to similar predictions in the main section of the chapter. While the extension to the reference dependent preferences framework in Koszegi and Rabin (2006) is interesting and offers much theoretical insight into other real-world situations, it does not seem to be enough for explaining the large variation in take-home pay in the data.

these messenger-treatment pairs were excluded in calculating the mean of the standard deviation as standard deviation is undefined and zero respectively.) This method of calculation explains why the weighted averages of the two means in Table 3.11 are not close to the mean in the corresponding column in Table 3.7.

References

- Angrist, Joshua D. and Victor Lavy (1999) "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *The Quarterly Journal of Economics*, Vol. 114, No. 2, pp. 533–575.
- Angrist, Joshua and Victor Lavy (2009) "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial," *The American Economic Review*, Vol. 99, No. 4, pp. 1384–1414.
- Angrist, Joshua, Eric Bettinger, and Michael Kremer (2006) "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia," *The American Economic Review*, Vol. 96, No. 3, pp. 847–862.
- Angrist, Joshua, Daniel Lang, and Philip Oreopoulos (2009) "Incentives and Services for College Achievement: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*, Vol. 1, No. 1, pp. 136–163.
- Ashenfelter, Orley, Kirk Doran, and Bruce Schaller (2010) "A Shred of Credible Evidence on the Long-run Elasticity of Labour Supply," *Economica*, Vol. 77, No. 308, pp. 637–650.
- Behrman, Jere R., Piyali Sengupta, and Petra Todd (2005) "Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico," *Economic Development and Cultural Change*, Vol. 54, No. 1, pp. 237–275.
- Berelson, B. (1979) "Romania's 1966 Anti-Abortion Decree: The Demographic Experience of the First Decade," *Population Studies*, Vol. 33, No. 2, pp. 209–222.
- Bettinger, Eric P. (2010) "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores," Working Paper 16333, National Bureau of Economic Research.
- Bishop, John (2006) "Chapter 15 Drinking from the Fountain of Knowledge: Student Incentive to Study and Learn - Externalities, Information Problems and Peer Pressure,"

- in E. Hanushek and F. Welch eds. *Handbook of the Economics of Education*, Vol. 2 of *Handbook of the Economics of Education*: Elsevier, pp. 909 – 944.
- Boudreau, Kevin J., Constance E. Helfat, Karim R. Lakhani, and Michael Menietti (2012) “Field Evidence on Individual Behavior & Performance in Rank-Order Tournaments,” *Harvard Business School Working Paper*, No. 13-016, August.
- Bound, John and Sarah Turner (2007) “Cohort crowding: How resources affect collegiate attainment,” *Journal of Public Economics*, Vol. 91, No. 5-6, pp. 877 – 899.
- Bull, Clive, Andrew Schotter, and Keith Weigelt (1987) “Tournaments and Piece Rates: An Experimental Study,” *Journal of Political Economy*, Vol. 95, No. 1, pp. pp. 1–33.
- Camerer, Colin, Linda Babcock, George Loewenstein, and Richard Thaler (1997) “Labor Supply of New York City Cabdrivers: One Day at a Time,” *The Quarterly Journal of Economics*, Vol. 112, No. 2, pp. 407–441.
- Card, David and Alan B. Krueger (1992) “Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States,” *Journal of Political Economy*, Vol. 100, No. 1, pp. 1–40.
- Casas-Arce, Pablo and F. Asis Martinez-Jerez (2009) “Relative Performance Compensation, Contests, and Dynamic Incentives,” *Management Science*, Vol. 55, No. 8, pp. 1306–1320.
- Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan (2011) “How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star,” *The Quarterly Journal of Economics*, Vol. 126, No. 4, pp. 1593–1660.
- Dept, Hong Kong Census & Statistics (1988-1994) *Hong Kong Monthly Digest of Statistics*: Govt. Print.

- (1988-2011) *Hong Kong Annual Digest of Statistics*: Govt. Print.
- Ehrenberg, Ronald G. and Michael L. Bognanno (1990) “Do Tournaments Have Incentive Effects?” *Journal of Political Economy*, Vol. 98, No. 6, pp. 1307–1324.
- Eriksson, Tor (1999) “Executive Compensation and Tournament Theory: Empirical Tests on Danish Data,” *Journal of Labor Economics*, Vol. 17, No. 2, pp. 262–280.
- Farber, Henry S. (2005) “Is Tomorrow Another Day?: The Labor Supply of New York City Cabdrivers,” *The Journal of Political Economy*, Vol. 113, No. 1, pp. 46–82.
- (2008) “Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers,” *The American Economic Review*, Vol. 98, No. 3, pp. 1069–1082.
- Fehr, Ernst and Lorenz Goette (2007) “Do Workers Work More if Wages Are High?: Evidence from a Randomized Field Experiment,” *The American Economic Review*, Vol. 97, No. 1, pp. 298–317.
- Fryer, Roland G. (2011) “Financial Incentives and Student Achievement: Evidence from Randomized Trials,” *The Quarterly Journal of Economics*, Vol. 126, No. 4, pp. 1755–1798.
- Gaechter, Simon, Eric J. Johnson, and Andreas Herrmann (2007) “Individual-Level Loss Aversion in Riskless and Risky Choices,” *IZA Discussion Paper*, Vol. No. 2961, July 2007.
- Garcia, Stephen M. and Avishalom Tor (2009) “The N-Effect,” *Psychological Science*, Vol. 20, No. 7, pp. 871–877.
- Goette, Lorenz, David Huffman, and Ernst Fehr (2004) “Loss Aversion and Labour Supply,” *Journal of the European Economic Association*, Vol. 2, No. 2/3, pp. 216–228.
- Green, Jerry R. and Nancy L. Stokey (1983) “A Comparison of Tournaments and Contracts,” *Journal of Political Economy*, Vol. 91, No. 3, pp. 349–364.

- Harbring, Christine and Bernd Irlenbusch (2003) “An experimental study on tournament design,” *Labour Economics*, Vol. 10, No. 4, pp. 443 – 464, August. Special Issue on Labour Market Experiments.
- Konrad, Kai A. (2009) *Strategy and Dynamics in Contests*: Oxford University Press.
- Koszegi, Botond and Matthew Rabin (2006) “A Model of Reference-Dependent Preferences,” *The Quarterly Journal of Economics*, Vol. 121, No. 4, pp. 1133–1165.
- Kremer, Michael, Edward Miguel, and Rebecca Thornton (2009) “Incentives to Learn,” *Review of Economics and Statistics*, Vol. 91, No. 3, pp. 437–456, 7.
- Krueger, Alan B. (1999) “Experimental Estimates of Education Production Functions,” *The Quarterly Journal of Economics*, Vol. 114, No. 2, pp. 497–532.
- Krueger, Alan B. and Diane M. Whitmore (2001) “The Effect of Attending a Small Class in the Early Grades on College-test Taking and Middle School Test Results: Evidence from Project Star,” *The Economic Journal*, Vol. 111, No. 468, pp. 1–28.
- Lazear, Edward P. (2001) “Educational Production,” *The Quarterly Journal of Economics*, Vol. 116, No. 3, pp. 777–803.
- Lazear, Edward P. and Sherwin Rosen (1981) “Rank-Order Tournaments as Optimum Labor Contracts,” *Journal of Political Economy*, Vol. 89, No. 5, pp. 841–864.
- Leung, Gabriel M., Tai-Hing Lam, Thuan Q. Thach, Simmy Wan, and Lai-Ming Ho (2001) “Rates of Cesarean Births in Hong Kong: 1987-1999,” *Birth*, Vol. 28, No. 3, pp. 166–172.
- Leuven, Edwin, Hessel Oosterbeek, and Bas van der Klaauw (2010) “The effect of financial rewards on students achievement: Evidence from a randomized experiment,” *Journal of the European Economic Association*, Vol. 8, No. 6, pp. 1243–1265.

- Main, Brian G. M., Charles A. O'Reilly III, and James Wade (1993) "Top Executive Pay: Tournament or Teamwork?" *Journal of Labor Economics*, Vol. 11, No. 4, pp. 606–628.
- Morin, Louis-Philippe (forthcoming) "Estimating the Benefit of High School for University-Bound Students: Evidence of Subject-Specific Human Capital Accumulation," *Canadian Journal of Economics*.
- Nalebuff, Barry J. and Joseph E. Stiglitz (1983) "Prizes and Incentives: Towards a General Theory of Compensation and Competition," *The Bell Journal of Economics*, Vol. 14, No. 1, pp. 21–43.
- OECD (2003) "Literacy Skills for the World of Tomorrow: Further Results from PISA 2000," Technical report, Organization of Economic Co-operation and Development, Paris, France.
- Oettinger, Gerald S. (1999) "An Empirical Analysis of the Daily Labor Supply of Stadium Vendors," *The Journal of Political Economy*, Vol. 107, No. 2, pp. 360–392.
- Orrison, Alannah, Andrew Schotter, and Keith Weigelt (1997) "On the Design of Optimal Organizations Using Tournaments: An Experimental Examination," Working Papers 97-26, C.V. Starr Center for Applied Economics, New York University.
- Rabin, Matthew (2000) "Risk Aversion and Expected-Utility Theory: A Calibration Theorem," *Econometrica*, Vol. 68, No. 5, pp. 1281–1292.
- Rohlf, Chris, Alexander Reed, and Hiroyuki Yamada (2010) "Causal effects of sex preference on sex-blind and sex-selective child avoidance and substitution across birth years: Evidence from the Japanese year of the fire horse," *Journal of Development Economics*, Vol. 92, No. 1, pp. 82 – 95.
- Schotter, Andrew and Keith Weigelt (1992) "Asymmetric Tournaments, Equal Opportunity

- Laws, and Affirmative Action: Some Experimental Results,” *The Quarterly Journal of Economics*, Vol. 107, No. 2, pp. 511–539.
- Skinner, Ellen A. and Michael J. Belmont (1993) “Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year.,” *Journal of Educational Psychology*, Vol. 85, No. 4, pp. 571–581.
- Stapleton, David C. and Douglas J. Young (1984) “The Effects of Demographic Change on the Distribution of Wages, 1967-1990,” *The Journal of Human Resources*, Vol. 19, No. 2, pp. 175–201.
- (1988) “Educational Attainment and Cohort Size,” *Journal of Labor Economics*, Vol. 6, No. 3, pp. 330–361.
- Tversky, Amos and Daniel Kahneman (1991) “Loss Aversion in Riskless Choice: A Reference-Dependent Model,” *The Quarterly Journal of Economics*, Vol. 106, No. 4, pp. 1039–1061.
- van Dijk, Frans, Joep Sonnemans, and Frans van Winden (2001) “Incentive systems in a real effort experiment,” *European Economic Review*, Vol. 45, No. 2, pp. 187 – 214.
- Vere, James P. (2008) “Dragon Children: Identifying the Causal Effect of the First Child on Female Labour Supply with the Chinese Lunar Calendar*,” *Oxford Bulletin of Economics and Statistics*, Vol. 70, No. 3, pp. 303–325.
- Welch, Finis (1979) “Effects of Cohort Size on Earnings: The Baby Boom Babies’ Financial Bust,” *Journal of Political Economy*, Vol. 87, No. 5, pp. S65–S97.
- Wentzel, Kathryn R. (1997) “Student motivation in middle school: The role of perceived pedagogical caring.,” *Journal of Educational Psychology*, Vol. 89, No. 3, pp. 411–419.
- Yip, Paul S.F, Joseph Lee, and Y.B Cheung (2002) “The influence of the Chinese zodiac on fertility in Hong Kong SAR,” *Social Science & Medicine*, Vol. 55, No. 10, pp. 1803 – 1812.