Summer 8-1-2020

Essays in Behavioral and Experimental Economics

Puneet Arora
Georgia State University, parora2@student.gsu.edu

ABSTRACT

ESSAYS IN BEHAVIORAL AND EXPERIMENTAL ECONOMICS

BY

PUNEET ARORA

AUGUST 2020

Committee Chair: Dr. Alberto Chong

Major Department: Economics

This dissertation comprises of essays in the field of development economics. Leveraging insights from psychology, mingling them in economic theory, and testing them through experimental (and quasi-experimental) methods, I study three policy questions: Does grading system affect student performance? Does symbolic incentive offered in a competitive game increase student attendance? Does institutional quality affect the provision of public goods and the perception of tax as a burden?

In the first chapter, I study whether level of discretization of reported grades to students affect their academic performance? This is the first experimental study to test how performance reporting using a coarse grading scale: {A,B,C,D,F} (Letter Grading System or LGS - the treatment group) or a very fine grading scale: [0,100] (Numerical Grading System or NGS - the control group) affects student performance. While I find no difference in average student performance between LGS and NGS, this negligible effect, however, masks important gender-differences in student performance with female students responding more positively to NGS and male students responding more positively to LGS. The theoretical model presented in the paper throws light on a possible mechanism (risk-attitudes) causing these gender-differences. This evidence is informative of the role that the chosen grading

scale may have played in the widely seen gender-gap in student learning.

In the second chapter, I report the results of a field experiment conducted to study how student attendance changes when transformed into a strategic-decision using a competitive game setting. The game awarded points to students based on their daily attendance. Each student could track the weekly status of his/her game points, along with the game points of every other student in the class, thereby bringing an explicit strategic element to the intervention. The scoring rule had a novel weighting mechanism, designed to lower absenteeism by a greater degree in the weeks when students are more likely to be absent. This experiment was conducted with 217 classrooms, divided into two treatments (Group Game and Classroom Game) and one control group of a not-for-profit educational institution in India. Symbolic rewards were provided to the winners of the game in the treatment classrooms. The results show a significant positive effect on students whose attendance was greater than 75% while not much effect on the rest. This provides evidence of how the use of a simple competitive game with low-cost, symbolic rewards can efficiently and positively impact student attendance.

In the third chapter, coauthor Alberto Chong and I study the role that institutional quality plays in determining government's effectiveness in delivering public goods and in, therefore, mediating the effects of higher taxation in an economy. This study is inspired from the observation that poorer countries, despite having a much smaller public sector and correspondingly a smaller tax burden than richer countries, have mostly displayed a weaker economic performance than richer countries. Using a simple theoretical model, we show that the provision of public goods and optimal tax levels increase with improved institutional quality. Using firm level perceptions data on the quality of public services and the tax burden, consistent with the predictions of our model, we find that a higher level of institutional quality bolsters positive perception of the quality of public services while at the same time moderating the view of the taxes as an obstacle to growth.

ESSAYS IN BEHAVIORAL AND EXPERIMENTAL ECONOMICS

BY

PUNEET ARORA

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree
of
Doctor of Philosophy
in the
Andrew Young School of Policy Studies
of
Georgia State University

GEORGIA STATE UNIVERSITY

2020

ACCEPTANCE

This dissertation was prepared under the direction of the candidate's Dissertation Committee. It has been approved and accepted by all members of that committee, and it has been accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Economics in the Andrew Young School of Policy Studies of Georgia State University.

Committee Chair: Dr. Alberto Chong

Committee Members: Dr. Thomas Mroz

Dr. Jonathan Smith

Dr. Gustavo Bobonis

Electronic Version Approved:

Dr. Sally Wallace, Dean

Andrew Young School of Policy Studies Georgia State University

August 2020

Acknowledgement

It has personally been one lifetime of a journey working towards my dissertation. Starting from the early days of head scratching and coming up with umpteen potential research ideas to the latter stages of roadside discussions with friends to question and flatten many of those ideas, and then to the ultimate stages of taking the remaining ones to the faculty members in the department for them to take away any life that may be left in the surviving ones. Following this process over and over eventually culminated in finding those three "good and doable" ideas, the secret sauce to a completed dissertation, that went on to become the three chapters of my dissertation.

I cannot sufficiently express my gratitude towards my advisors at Georgia State University who guided me through all the ups and downs that I went through - during the conception to implementation to the analytical stages of the research projects. Alberto Chong took me under his tutelage and motivated me early on to learn the skills required for a good dissertation. Walking into his office was always the magic pill to any research concerns that I used to have, or alternatively, he offered my brain with food for thought when it lacked some. He inspired and infused energy in me to undertake projects that I may have, otherwise, dropped for the difficulties that they faced. One such project eventually became my job market paper. Tom Mroz gave me the mantra of a successful researcher: "answer one question with good precision and confidence than many with little power and policy implication". Jon Smith, veiled in his pleasing personality, always acted as a devil's advocate and raised points that significantly contributed to several improvements in my papers. I

would also like to thank my external advisor from University of Toronto, Gustavo Bobonis, whose critical comments on research kept me grounded and always working, and whose motivating words kept my spirits high.

I would like to express my gratitude towards individuals who were not my committee, yet contributed significantly in ensuring that I become a better researcher. To start with, it was Vjollca Sadiraj and James Cox who taught me the nitty-gritty of the methods of experiments in research, and hand-held me through my initial days of research career. A special thanks to Dan Kreisman for conducting research seminars that contributed to our improved presentation skills and to Garth Heutel for keeping close tabs on our job market preparation. My research and studying at GSU could not have been made this smooth without the help and prompt responsiveness of the administrative staff, especially, Bess Byler and Kristy Hill.

My research received excellent support from research assistants: Aditi Virmani (coincidently my wife too), Anjuri Das, Bhaumik Chabbra, Devanshi Malhotra, Navni Kothari and Sri Vaishnavi Varma. These projects would not have been possible without their extensive help at all stages. My research benefited significantly from the quick and helpful comments and feedback from Tareena Musaddiq, Siyu Pan, Ishtiaque Fazlul and Nicholas Wright. I certainly cannot thank them enough for their altruistic support throughout my graduate studies.

I appreciate the financial support provided for this dissertation through Andrew Young Dissertation Fellowship. In addition, I am also grateful for several other scholarships/grants that I received from GSU, CEAR and other organizations that helped me sail through this five year period. I thank Tom Mroz and Barry Hircsh for their generous support towards conference travel grants.

Finally, I would like to thank my family for it is their emotional and financial support that gave me the luxury to enjoy the student life for thirty years of my life. I thank my mother, Veena Arora, my father, Bharat Bhushan Arora, my brother, Ritesh Arora, sister-in-law,

Taruna Pahuja, sister, Tanya Arora, and brother-in-law, Saurabh Nanda. While they may not understand much about my research or what it may lead to, that did not hold them back from extending their relentless support, love and motivation in ensuring that I cross the line and be called Dr. Puneet Arora one day. It is their belief in me and my abilities that I chose to resume my Ph.D. once again, after choosing to drop out from Ph.D. after one year into the program. I thank my wife, Aditi Virmani, whom I married during my Ph.D. studies, used her help as a research assistant while on the field and as a friend and life partner while off the field. She stood by me, and motivated me even when our married life moved to a virtual mode with she working on her job in India and me working on my research in the United States. It is her love and unwavering support that helped me channel my energies correctly and complete the dissertation in time. I also thank my in-laws whose belief in me nudged me towards the completion of a journey that many feel scared to even start. I thank my roommate and dearest friend, Prakash Chourasia, without whom these five years would have been an immense struggle. He cooked delicious food for me when I couldn't, did groceries for me when I couldn't, took me out for movies when I needed a break, and was there whenever I needed to talk. I thank Navni Kothari who started as a student, later became a friend, a sister, a research assistant and then a co-author and has contributions in almost all my projects. Lastly, I thank my three year old nephew, Seyan Arora, for bringing in the much needed non-linearity in an otherwise fairly linear life. It is because of all you people that I am an economist today and I will be indebted to your support all my life.

Contents

List of Tables

List of Figures

Chapter 1

Grading, Incentives and Student Performance

1.1 Introduction

There is a constant debate in developing and developed countries on how to pursue educational reform to improve student learning. Policies often tested work with school finances, teacher incentives, parental incentives, instructional methods, health and nutritional assistance programs, among others[1]. However, policymakers tend to forget that incentives to students are also an important, if not, the most important factor that may help improve student learning. While there are some policies that deal with student incentives, they are usually indirect, for instance, in the form of conditional cash transfers that work their way towards learning through greater enrollment and attendance. The most direct incentives to students is very straightforward - grades and prizes[2]. In this context, it is important to consider whether how educators grade students may have any bearing on student performance and thus, learning. More specifically, does the coarseness or fineness of the grading scale

[1]See McEwan (2015) for a review of experimental literature on these policies. See Glewwe and Muralidharan (2016) and Muralidharan (2017) for supply-side policies.

[2]Studies that have tested effectiveness of direct prizes have mostly come from developed countries (Angrist and Lavy, 2009; Fryer, 2011; Bettinger, 2012)

affect student performance? Interestingly, this question has not been studied earlier causally in the education literature.

In this paper, I am concerned with the universe of students who attend grade schools (more specifically, K-11). The grading systems that are most relevant to them fall in the realm of criterion-referenced grading. Also known as absolute grading, criterion-referenced grading grades students based on their own absolute performance and not based on their relative performance in the class. In contrast, norm-referenced grading (also known as relative grading or grading-on-a-curve) evaluates a student's performance in comparison to his/her peers. Norm-referenced grading system is often used in a university setting or in highly competitive entrance examinations. Several studies compare criterion-referenced grading with norm-referenced grading theoretically (Becker and Rosen, 1992; Dubey and Geanokopolos, 2010) and empirically (Czibor et al, 2014; Paredes, 2017). My paper, however, deals specifically with the criterion-referenced versions of coarse and fine grading systems only, which are usually found in K-12 education system world over[3].

There is currently no existing research, to the best of my knowledge, that explicitly compares student performance under a very fine grading system like [0,100] (numerical grading system or NGS, henceforth) with a very coarse grading system like {A,B,C,D,F} (letter grading system or LGS, henceforth). There are studies that use survey data to discuss student and faculty perception (Baker and Bates, 1999; Fries et al., 2013) or present a qualitative argument of comparison (Bressette, 2002) about different grading scales. An ideal causal study would require (a) students in same cohorts to be randomly assigned to classrooms with different grading scales, (b) same teacher teaching each of these classrooms, (c) same teacher grading students in each of these classrooms without being aware of the treatment

[3]Becker and Rosen (1992) and Dubey and Geanokopolos (2010) make assumptions that are not necessarily true about all category of students, especially for the students studying in K-11 grades. Becker and Rosen (1992) assumes students only care about pass-fail; Dubey and Geanokopolos (2010) assumes students only care about their relative ranking. For students in grades K-11 who are almost always in a criterion-referenced grading system, considerations to grades may be beyond just pass-fail or status which are more relevant in norm-referenced grading. My paper, thus, presents a model with assumptions that are more relevant to a grade school student.

assignment of each student, and (d) no or low attrition and spill-overs. McClure and Spector (2005) come closest to causally estimating the difference in student performance (but is not causal) between coarse LGS {A,B,C,D,F} and a lesser coarse plus/minus system {A, A-, B+ ... D-, F} and finds no difference in student performance between the two systems. While the study controls for between-teacher differences by having same teacher teaching and same teacher grading, the study suffers from selection-bias since it allowed students to self-select into one of the two grading systems and the study also does not control for spill-over effect. My paper corrects for these concern and designs an experiment that takes care of all four elements of estimating the effect causally. Apart from being the first causal study to estimate the effect of different grading scales on student performance, it deviates from prior literature in three other ways: (1) while prior studies have overwhelmingly dealt with university students, this paper conducts the experiment with grade 8 students whose incentives to exert effort may be different from university students, (2) most prior studies compare coarse LGS {A,B,C,D,F} with a lesser coarse plus/minus system {A, A, B+ ... D, F}, this paper instead compares coarse LGS {A,B,C,D,F} with fine NGS [0,100], and most importantly, (3) most prior studies compare criterion-referenced grading system with norm-referenced grading system, this paper focuses only on different grading scales within the criterion-referenced grading system, which is what is most relevant to the grade-school students world over.

There are other studies that test the threshold effect in LGS (or other lesser coarse grading systems) but due to unavailability of a comparable group, they make no simultaneous comparisons with a finer grading system (Oettinger , 2002; Grant and Green, 2013; Main and Ost, 2014; Grant, 2016). This paper can be considered as an extension of the threshold literature with a comparative study of the optimal efforts as we transition from a coarse to a fine grading system.

The purpose of this paper is to present a theoretical model that compares the incentive structures underlying criterion-referenced grading systems as we move from very coarse

grading scale to very fine grading scale. For the sake of simplicity, I restrict the model to two specific grading scales (LGS and NGS), and determine the optimal student effort level under each of these two systems. However, the findings of the model can be generalized to all possible grading scales from the coarsest {Fail, Pass} to the finest [0,100]. In addition, this paper also tests the predictions of the model through a one-day long field experiment that was conducted with 438 grade 8 students of public and private schools in Delhi, India. This is the first paper in the strand of grading literature to explicitly model and test empirically these two widely observed criterion-referenced grading scales: LGS and NGS.

In the first part of the paper, I present a theoretical model which finds that, for any given student, there is no one grading scale that always elicits greater effort than all other grading scales. It is, instead, the risk-attitude of a student that plays the pivotal role in determining which grading scale dominates. In the simplified world with LGS and NGS, I find that increasing student risk-aversion increases the likelihood of NGS eliciting greater effort, thus, dominating LGS. This result implies that if we compare two identical and highly risk-averse students, one under NGS and one under LGS, then we should expect student under NGS to exert greater effort. If those students have low risk-averse attitude instead, then we should expect student under LGS to exert greater effort.

In the second part of the paper, I reports results from a randomized controlled trial conducted with 438 grade 8 students of 10 public and private schools in New Delhi, India. The experiment, conducted for one day, assigned students into two groups randomly - Treatment (LGS) and Control (NGS) - and then evaluated student performance through their test scores received on tests conducted during the day. The aggregate result from this experiment finds close to zero average treatment effect, possibly, indicative of no real differences between LGS and NGS. However, masked behind the zero average treatment effect, the study finds meaningful gender differences. On average, a female student in LGS group performs worse than a female student in NGS group by 0.14σ, and a male student in LGS group performs better than a male student in NGS group by 0.12σ. In other words, NGS dominates among

female students and LGS dominates among male students.

A direct test of the theoretical prediction about risk-attitudes and student performance under NGS and LGS is made impossible by the inability to conduct a risk-aversion task with participating students. This is due to the restrictions placed by the partnering schools to keep the engagement with students to strictly educational activities. I instead conduct an indirect test of the theory using gender as a proxy for risk-attitudes, a detailed discussion about which has been conducted in section 1.5. Using female students as proxy for high risk-aversion and male students as proxy for low risk-aversion, I find that NGS dominates among the former and LGS dominates among the latter.

The contribution of this paper is, thus, two fold. Firstly, this is the first empirical paper that experimentally compares LGS and NGS, two of the most observed grading scales in grade schools around the world [4]. Secondly, this paper presents a theoretical model about how these grading scales incentivize students differently and discusses how risk-attitudes play a pivotal role in determining dominance of one grading scale over other. This is an important contribution, more broadly, in the entire student-incentive literature which attributes the frequently observed gender-differences in student responses to diferences in motivation, study habits, self-discipline, competitiveness, etc[5]. This model brings to the fore an additional mechanism flowing through gender-differences in students' risk-attitudes.

The remainder of this paper is organized as follows. Section 2 presents the theoretical model. Section 3 introduces the research design and the estimation strategy. Section 4 presents the main results of the experiment. Section 5 interprets the findings and talks

[4]Countries like Canada, Kenya, Hong Kong, South Korea, United Kingdom, United States, Sweden, among others, majorly use 5-point LGS while many others like Costa Rica, Nicaragua, Chile, Venezuela, India, Pakistan, China, Israel, Indonesia, Poland, among others, have opted for the NGS. The list of countries presented overwhelmingly follow the mentioned grading scale in K-12 school system. There may, however, still be some schools within these countries which follow some other grading scale.

[5]Different responses to incentives by males and females is very common in the strand of education literature that studies incentives like monetary rewards, tuition waivers, vouchers, varying stakes in an exam, among others (Angrist et al., 2002; Dynarski, 2008; Angrist and Lavy, 2009; Fryer, 2011; Ors, Palomino and Peyrache, 2013; Jalava, Joensen and Pellas, 2015; Katreniak, 2018). The reasons often attributed to these gender differences include differences in motivation, study habits, self-discipline, competitiveness, etc., while not considering the differences in risk-attitudes between males and females.

about their policy implications. Section 6 concludes the paper. Appendix A in the paper gives the proofs to main theoretical results of the model, and Appendix B gives a history of how grading systems evolved over the past couple of centuries.

1.2 The model

This sections presents a theoretical model of the incentive structures underlying grading systems from very coarse to very fine. For simplicity, I write the model specifically for LGS and NGS versions of the coarse and fine grading systems, respectively. The model solves for and compares the optimal effort that a student exerts under LGS and NGS grading systems, given his cost and reward from exerting effort. The main finding of the model is that student's risk-attitude plays a pivotal role in determining the grading system that elicits greater effort. While the result is presented for LGS and NGS, the findings of the model can be generalized to other grading systems too. The set-up of the model is as follows:

1.2.1 Test score function

Assume that student $i's$ test score (q_{it}) in test t is a function of his unobserved ability (a_i), effort ($\mu_{it} \in [0, \bar{\mu}]$), test fixed effects ($\nu_t$) and idiosyncratic error (ϵ_{it}). Let $\epsilon_{it} \sim N(0, \sigma^2)$, and $G(.)$ and $g(.)$ be its distribution function and density function, respectively, which is known to all students. ϵ_{it} depends on factors like whether the student was lucky enough to have studied the material precisely relevant to the questions or how he felt on the day of test.

The test score $q_{it} \in [0, 100]$ under NGS with the corresponding function:

$$q_{it}(\mu_{it}, a_i) = \tau(a_i) + \mu_{it} + \nu_t + \epsilon_{it} \tag{1.1}$$

There is a perfect 1 to 1 transformation of effort μ_{it} into the score q_{it}, given ability a_i, and the error structure ν_t and ϵ_{it}. Absent any effort, test fixed effects and idiosyncratic error,

$\tau(a_i)$ is the minimum score that student i will get due to his sheer ability a_i.

Under LGS, I assume that students' letter grades come from the set $\{A, B, C, D, F\}$. For a given numerical score $q_{it} \in [0, 100]$, the function f returns a letter grade, i.e.,

$$f : [0, 100] \to \{A, B, C, D, F\}$$

where A is the best letter grade and F is the worst letter grade.

Under LGS, there is still a perfect 1 to 1 transformation of effort μ_{it} into the numeric score q_{it}, given the error structure ν_t and ϵ_{it} as was the case in NGS. However, q_{it} here is an intermediate score and $f(q_{it})$ is the final grade. The effective test score function for students under LGS, therefore, is:

$$f(q_{it}(\mu_i, a_i)) = f(\tau(a_i) + \mu_{it} + \nu_t + \epsilon_{it}) \tag{1.2}$$

where f is a step-function which does not change its value $f(q_{it})$ across several ranges of test scores. Let's consider the functional form of f to be[6]:

$$f(q_{it}) = \begin{cases} A, & \text{if } q_{it} \in (90, 100] \\ B, & \text{if } q_{it} \in (80, 90] \\ C, & \text{if } q_{it} \in (70, 80] \\ D, & \text{if } q_{it} \in (60, 70] \\ F, & \text{if } q_{it} \in [0, 60] \end{cases} \tag{1.3}$$

For rest of the theoretical analysis, I will assume test fixed effects ν_t to be zero.

[6]For the ease of exposition, I consider this specific functional form. However, the theory is applicable to various other functional forms adopted worldwide under the letter grading system umbrella.

1.2.2 Cost of effort

Let the cost of effort for a student of ability a be represented by function $C_a : \mathcal{R} \to \mathcal{R}$, with $C_a{}'(.) > 0$ and $C_a{}''(.) > 0$. This suggests that cost increases in effort and it increases at an increasing rate for a student with ability a. This is because students' trade-off is between studying and leisure activities. More time spent studying raises the marginal value of leisure which causes an increase in marginal cost of effort spent studying.

1.2.3 Reward structure

Students' interest in getting a higher test score comes from the rewards associated with it. Those rewards to students can be broadly categorized into (a) societal factors which include the incentives given by parents' to their children, and recognition and praise by teachers and friends, (b) warm glow which includes the intrinsic motivation of the student to learn, and (c) status which emanates from competition among peers which is the strongest in norm-referenced grading.

Rewards (pecuniary or non-pecuniary) to students due to societal factors are a function of student ability and effort (reflected in test-scores). Parents and teachers often have an idea of the ability of different students, and recognize and reward them if they do well given a student's own ability level. This concept of reward due to societal factors is based on student's own meritocracy.

Rewards due to warm glow are expected to be independent of the choice of grading system, and thus, I ignore them in the theoretical model. Rewards due to status are based on comparison between students and are usually given to top performers by school, teachers or parents. This kind of reward is significantly important only at specific junctures of education like during senior grades of high school or the senior years of college where competition and ranking matter most. This is because better performance among peers get them better letter of recommendations from their teachers and subsequent admission offers from higher ranked colleges and universities. Thus, status matters most as an incentive in the context

of relative meritocracy which falls under the purview of norm-referenced (relative) grading system. This paper, however, is about criterion-referenced (absolute) grading system where a student is graded based on his own meritocracy. Therefore, I consider rewards only due to societal factors in the model.

My model focuses on the population of young learners for whom the difference between NGS and LGS could emanate only from the societal factors (for instance, for students until grade 11 in a developing country context. Admissions to colleges are based on grade 12 scores only, thus making status reward extremely important in grade 12 but not as important before that). While status cannot be completely ruled out even for this set of student population, its role is not as significant as it can be assumed under norm-referenced grading system. Therefore, for the sake of simplicity in our optimization problem, I am assuming that this minimalist role of status among these early graders is accommodated within the rewards due to societal factors.

I assume the reward structure for NGS to be represented by function[7] $W_a^N(q_{it})$, i.e.,

$$W_a^N : [0, 100] \rightarrow \mathcal{R}$$

where N denotes NGS, a denotes ability level and $q_{it} \in [0, 100]$.

I assume the reward structure for LGS to be represented by function $W_a^L(f(q_{it}))$, i.e.,

$$W_a^L : \{A, B, C, D, F\} \rightarrow \mathcal{R}$$

where L denotes LGS, a denotes ability and $f(q_{it}) \in \{A, B, C, D, F\}$. $W_a^L(f(q_{it}))$ increases as $f(q_{it})$ moves from letter grade F to A. The reward function W_a^L takes the shape of a step-function while increasing in letter grades.

[7] $W_a(.)$ which I will call "Reward function" in this paper, represents a utility function in numerical scores received under NGS or letter grades received under LGS. $W_a^N(.)$ and $W_a^L(.)$ in the context of student risk-attitudes and their implications for optimal effort choice are shown in Figure 1.2 and Figure A.4, respectively.

1.2.4 Student Maximization Problem

Students under both NGS and LGS will choose the effort levels that will maximize their net expected rewards (expected rewards net of the cost of effort exerted)[8]. A student under NGS will exert effort μ that maximizes:

$$EB_a^N(\mu) = \int_\epsilon [W_a^N(\tau(a) + \mu + \epsilon)g(\epsilon)]d\epsilon - C_a(\mu) \tag{1.4}$$

A student under LGS will exert effort μ that maximizes:

$$EB_a^L(\mu) = \int_\epsilon [W_a^L(f(\tau(a) + \mu + \epsilon))g(\epsilon)]d\epsilon - C_a(\mu) \tag{1.5}$$

Optimization and a comparison of optimal efforts under NGS and LGS requires an assumption on how society perceives and rewards each letter grade in relation to numerical scores. For analytical purpose, I assume reward function under LGS to be constructed from the reward function under NGS in the following way[9]:

A1: $W^L(f(q_{it})) = W^N(\bar{q})$ if $q_{it} \in (\underline{q}, \bar{q}]$

1.2.5 Comparison of optimal effort levels under LGS and NGS

For the sake of simplicity, I assume letter grades to be partitioned by students' ability groups[10]. This implies that for ability $a \in \{1, 2, 3, 4, 5\}$ and $\forall\ q_{it} \in (\underline{q}_a, \bar{q}_a]$, there is a unique letter grade $f(q_{it}) \in \{A, B, C, D, F\}$. $\underline{q}_a$ and $\bar{q}_a$ represent the lowest and the highest numerical score that a student with ability a can get when there is no error ϵ_{it}. Lowest

[8] I assume test fixed effects ν_t to be 0.

[9] $W_a^L(A) = W_a^N(100)$, $W_a^L(B) = W_a^N(90)$,.., $W_a^L(F) = W_a^N(60)$. This implies that reward for letter grade A under LGS is identical to the reward for numerical score 100 under NGS for a student with ability a. Similarly, reward for B is equivalent to reward for 90, reward for 80 is equivalent to rewards for C, and so on. I can alternately assume $W_a^L(f(q_{it}))$ in LGS to be equal to any other weighted average of $W_a^N(q_{it})$ $\forall q_{it} \in (\underline{q}, \bar{q}]$ under NGS.

[10] The main theoretical result presented in Proposition 1 holds true even when this assumption is relaxed.

ability $5's$ test score range [0, 60] is assigned the lowest letter grade F and highest ability $1's$ test score range (90, 100] is assigned the highest letter grade A. This can be seen more clearly in Figure 1.1:

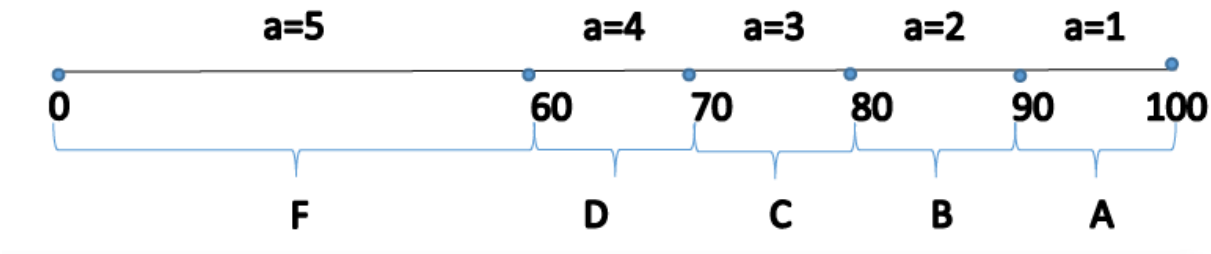


Figure 1.1: Grading Scale under LGS

In other words, letter grades (and the consequent reward) does not increase in effort for any ability group $a \in \{1, 2, 3, 4, 5\}$ (in the absence of any performance measurement error). This leads us to our next assumption:

A2: $f(\underline{q_5}) = f(\bar{q_5}) < f(\underline{q_4}) = f(\bar{q_4}) < f(\underline{q_3}) = f(\bar{q_3}) < f(\underline{q_2}) = f(\bar{q_2}) < f(\underline{q_1}) = f(\bar{q_1})$

Perfect Measurement

Consider a simple case where there is no error ϵ_{it} in measurement of student effort[11]. Also, assume test scores due to sheer ability, $\tau(a)$, to be equal to $\underline{q}_a$ which is the lowest score that a student with ability a will get. Under such a simplistic case, a student with ability a under NGS will exert effort μ such that his net reward is maximized:

$$NB_a^N = W_a^N(\underline{q_a} + \mu) - C_a(\mu) \tag{1.6}$$

which yields optimal effort $\mu_P^{N*} \geq 0$ such that $W_a^{N\prime}(\mu) = C_a'(\mu)$.[12]

A student with ability a under LGS will exert effort μ such that his net reward is maxi-

[11]This alludes to the extreme case where $\epsilon_{it} = 0$ and $\sigma^2 = 0$. In other words, student effort is recognized and rewarded perfectly without any error.

[12]Subscript P denotes perfect measurement.

mized:

$$NB_a^L = W_a^L(f(\underline{q_a} + \mu)) - C_a(\mu) \tag{1.7}$$

which yields zero optimal effort choice by ability type a, i.e., $\mu_P^{L^*} = 0$.[13]

Lemma 1: Under the assumption of perfect measurement of student effort, $\mu_P^{N^*} \geq \mu_P^{L^*}$. In other words, student effort under NGS will always be at least as much as that under LGS, ceteris paribus.[14]

Imperfect Measurement

Consider the case when student efforts cannot be measured precisely and thus, ensue an error, ϵ_{it}. Let $\epsilon_{it} \sim N(0, \sigma^2)$ with $G(\epsilon)$ and $g(\epsilon)$ being its distribution and density functions, respectively, which are known to all students. ϵ_{it} depends on factors like whether the student was lucky enough to have studied the material precisely relevant to the questions or how he felt on the day of test. Under such a scenario, a student with ability a under NGS will exert effort μ such that his net expected reward is maximized:

$$E(NB_a^N|\mu) = \int_\epsilon [W_a^N(\underline{q_a} + \mu + \epsilon)g(\epsilon)]d\epsilon - C_a(\mu) \tag{1.8}$$

Suppose the optimal effort level given by maximizing (1.8) for an NGS student is $\mu_I^{N^*}$.

Theorem 1: Under idiosyncratic error assumption and risk-neutrality, $\mu_P^{N^*} = \mu_I^{N^*}$. In other words, optimal student effort under NGS will be same whether measurement is perfect or imperfect[15].

[13]Effort level 0 does not mean no effort at all, it simply means the effort incentivized in a student exclusively by societal factors will be minimal. Student will still exert efforts that feed into his/her status and warm glow desires from learning.

[14]This result is robust to changes in societal perception of the relationship between NGS and LGS given by assumption A1; to different functional forms of reward function W - risk neutral, averse or loving; and to different ability groups.

[15]Proofs to all theorems are in appendix

This theorem implies that even under the noisy relation between effort and test scores due to measurement imprecision ϵ_{it}, the optimal effort of a risk-neutral students under NGS will not change as long as that noise ϵ_{it} is idiosyncratic.

Corollary 1: Effort level of a risk-neutral student under NGS is a constant function of measurement imprecision, ceteris paribus.

Corollary 1 implies that for a given risk-neutral student under NGS, expected rewards curve under imperfect measurement continues to be same as rewards curve under perfect measurement even when the measurement imprecision grows bigger, i.e., when error variance σ^2 grows while the error structure still stays idiosyncratic. Cost curve, on the other hand, stays unaffected by increasing imprecision. Thus, increasing imprecision does not change the optimal effort level of a risk-neutral student.

Theorem 2: Under idiosyncratic error assumption and risk-aversion, $\mu_P^{N*} > \mu_I^{N*}$. In other words, optimal student effort under NGS will decrease when measurement is imperfect. (For a risk-loving student, optimal effort under NGS will increase when measurement is imperfect.)

Corollary 2: Effort level of a risk-averse student under NGS is a decreasing function of measurement imprecision, ceteris paribus.

Corollary 2 implies that for a given risk-averse student under NGS, rewards from same efforts appear more uncertain and hazy, pivoting the expected rewards curve further down as imprecision increases, i.e., when error variance σ^2 grows while the error structure still stays idiosyncratic. Cost curve, on the other hand, stays unaffected by increasing imprecision. This reduces the optimal effort level exerted by the student. This argument can be pictured more clearly in Figure 1.2 where EW curve will pivot downward due to increased imprecision without any change in cost curve, thus, resulting in a lower optimal effort.

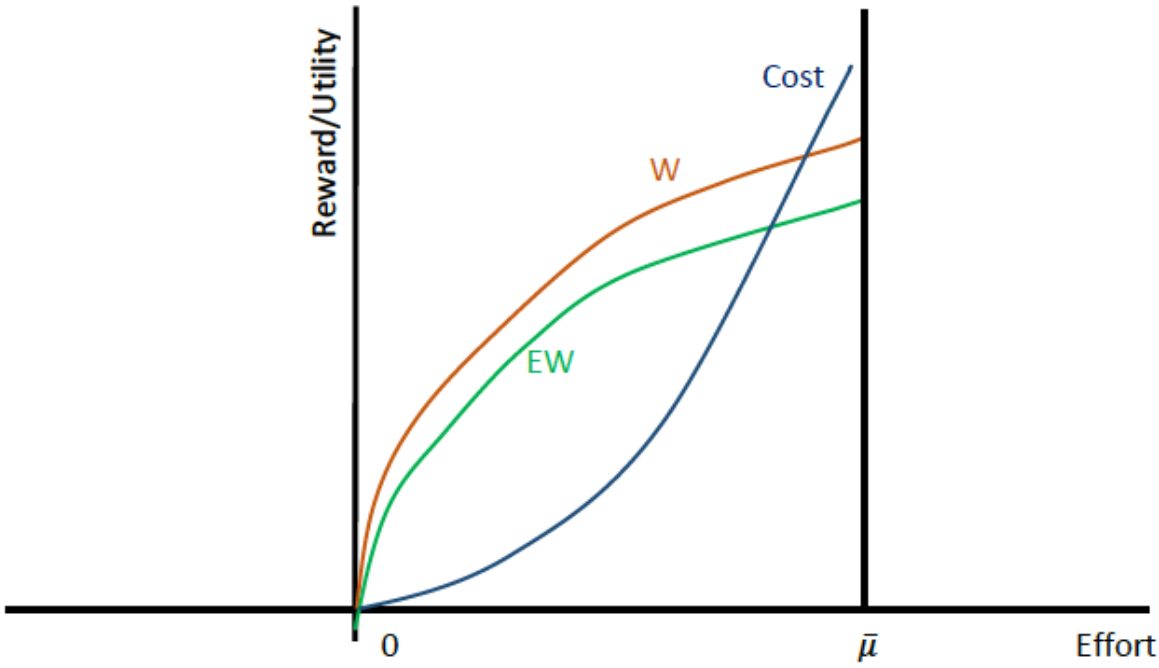


Figure 1.2: Reward function under NGS

Note: Rewards is equivalent to utility in our theoretical framework. This figure represents the student whose effort level cannot go below 0 and above $\bar{\mu}$, i.e., $\mu \in [0, \bar{\mu}]$. In a perfect measurement world of our theoretical model, this risk-averse student will get reward W(.) for exerted effort level μ. The optimal effort level can be found using marginal analysis. In an imperfect measurement world, W(.) shifts to EW(.). This will reduce the optimal effort. If risk-aversion were to increase further for this same student, then it is easy to see that the optimal effort level will go further down. The optimal effort level under NGS can lie anywhere between 0 and $\bar{\mu}$ which is not the case in LGS where the optimal effort level has one of the two local equilibrium solutions - precautionary effort which is closer to 0 or anticipatory effort which is closer to $\bar{\mu}$.

Corollary 3: Effort level under NGS is a decreasing function of risk-aversion, ceteris paribus.

Corollary 3 implies that, ceteris paribus, an increase in risk-aversion of a given student will pivot the expected rewards curve downwards while keeping the cost function unchanged, This will reduce the optimal effort level exerted by the student. This argument can be pictured more clearly in Figure 1.2.

A student under LGS will exert effort μ such that his net expected reward given below is maximized:

$$E(NB_a^L|\mu) = \int_\epsilon [W_a^L(f(\underline{q_a} + \mu + \epsilon))g(\epsilon)]d\epsilon - C_a(\mu) \tag{1.9}$$

Under LGS and perfect measurement case, a student with ability a faces constant $f(q_{it})$ $\forall q_{it} \in (\underline{q}_a, \bar{q_a}]$ and therefore, constant reward $W_a^L(f(q_{it}))$ $\forall q_{it} \in (\underline{q}_a, \bar{q_a}]$. The imprecision in measurement (ϵ_{it}), however, creates a possibility for a student with ability a to end up with a test score below $\underline{q}_a$ (the reward corresponding to which is lower than that of ability group a) or above $\bar{q_a}$ (the reward corresponding to which is higher than that of ability group a). The probability of attaining these different rewards inside or outside one's own ability group depends on $G(\epsilon)$ and $g(\epsilon)$ which determine the expected reward of effort μ for any student with ability a.

The probability of getting a higher grade (lower grade) increases (decreases) in effort level, given $g(\epsilon)$. This makes expected rewards $\int_\epsilon [W_a^L(f(\underline{q_a} + \mu + \epsilon))g(\epsilon)]d\epsilon$ a (weakly) increasing function in effort μ for given ability a. This is in contrast to perfect measurement case where rewards $W_a^L(f(\underline{q_a} + \mu))$ were constant in μ for given ability a. The cost function under imperfect measurement stays the same, $C(\mu)$, as under perfect measurement.

Theorem 3: Under ideosyncratic error assumption, $\mu_P^{L^*} \leq \mu_I^{L^*}$ irrespective of students' risk-attitudes. In other words, optimal student effort under LGS with imperfect measurement

of effort will be at least as much as that with perfect measurement of effort.

Optimization problem under LGS is similar to the problem under NGS whereby expected rewards and cost functions are rising in efforts under both the systems. While expected reward function under NGS is strictly increasing in effort, LGS has it weakly increasing in effort with steep slopes near the thresholds and flat stretches in between the two thresholds (see Figure A.2). Given identical cost function for a student under the two grading systems, any distinction in his optimal effort levels between these two systems will come from differences in these expected reward functions. While the optimal effort level under NGS will be uniquely determined anywhere in the range $\mu \in [0,]$, LGS' observed optimal effort level will be either Precautionary (i.e., very close to 0) or Anticipatory (i.e., very close to $\bar{\mu}$). This argument can be pictured more clearly in Figure 1.3. A step-by-step diagrammatic guide to the derivation of Figure 1.3 is given in the appendix through figures A.1, A.2, A.3 and A.4.

- **Precautionary Effort:** If a student with ability a increases effort starting from $\mu = 0$, it reduces the probability, $Pr(q_{it} < \underline{q}_a)$, sharply, however, it doesn't increase $Pr(q_{it} > \bar{q_a})$ by much due to $\bar{q_a}$ being very far away when μ is closer to 0. This can cause an optimal effort $\mu_I^{L^*}$ to be marginally greater than 0. This will be called Precautionary Effort acting as a precaution against scoring $q_{it} < \underline{q}_a$[16]. This effort assures that student with ability a does not get a letter grade and reward lower than what a's ability group deserves.

- **Anticipatory Effort:** If a student with ability a keeps increasing the effort far away from 0, $Pr(q_{it} < \underline{q}_a)$ is not decreasing sharply anymore and $Pr(q_{it} > \bar{q_a})$ is still not much affected. This implies expected rewards are increasing at a decreasing rate as effort increases. However, as effort level approaches $\bar{\mu}$, $Pr(q_{it} > \bar{q_a})$ increases sharply which causes expected rewards function to rise sharply. This can cause optimal effort

[16]The phrase Precautionary Effort has been borrowed from Grant (2106).

$\mu_I^{L^*}$ to be much closer to $\bar{\mu}$ or at $\bar{\mu}$. This will be called Anticipatory Effort raising the anticipation of scoring $q_{it} > \bar{q}_a$. This makes student with ability a highly likely to get a letter grade and reward higher than what a's ability group deserves.[17]

Theorem 4: Under assumption A1, optimal effort levels across students in LGS system will bunch right above the lower threshold, and at or right below the higher threshold, corresponding to their optimizing Precautionary or Anticipatory efforts, respectively.

Corollary 4: An increase in risk-aversion under LGS increases the likelihood of a student moving from anticipatory to precautionary effort, ceteris paribus.

Increasing risk-aversion, among otherwise identical students, shifts the expected rewards curve downwards under LGS and makes it flatter. This can be seen in the illustration given in Figure 1.3 for a given student who will score B in a perfect measurement world. In an imperfect measurement world, however, his expected reward from exerting effort shifts down from EW_0 to EW_1 as his risk-aversion increases.

The shift from EW_0 to EW_1 due to increased risk-aversion comes with an increased risk-premium. Risk-premium x represents additional precautionary effort required to minimize the chances of dropping to letter grade C and risk-premium y represents additional anticipatory effort required to create chances of getting letter grade A. In this figure, we can see that optimal choice of effort for the more risk-averse student with expected rewards curve EW_1 will be precautionary (i.e., closer to 0) while for the lesser risk-averse student with expected rewards curve EW_0, it may still be anticipatory effort. In other words, increase in risk aversion increased the probability of moving to precautionary effort from anticipatory effort. While Figure 1.3 shows this result for a specific ability group, the result holds true more generally for students from any ability group.

[17]These results parallel some of the findings of the threshold effects studied in Grant (2016) but Grant (2016) did not consider risk-attitudes of students or NGS in their setting which are the main contributions of this paper.

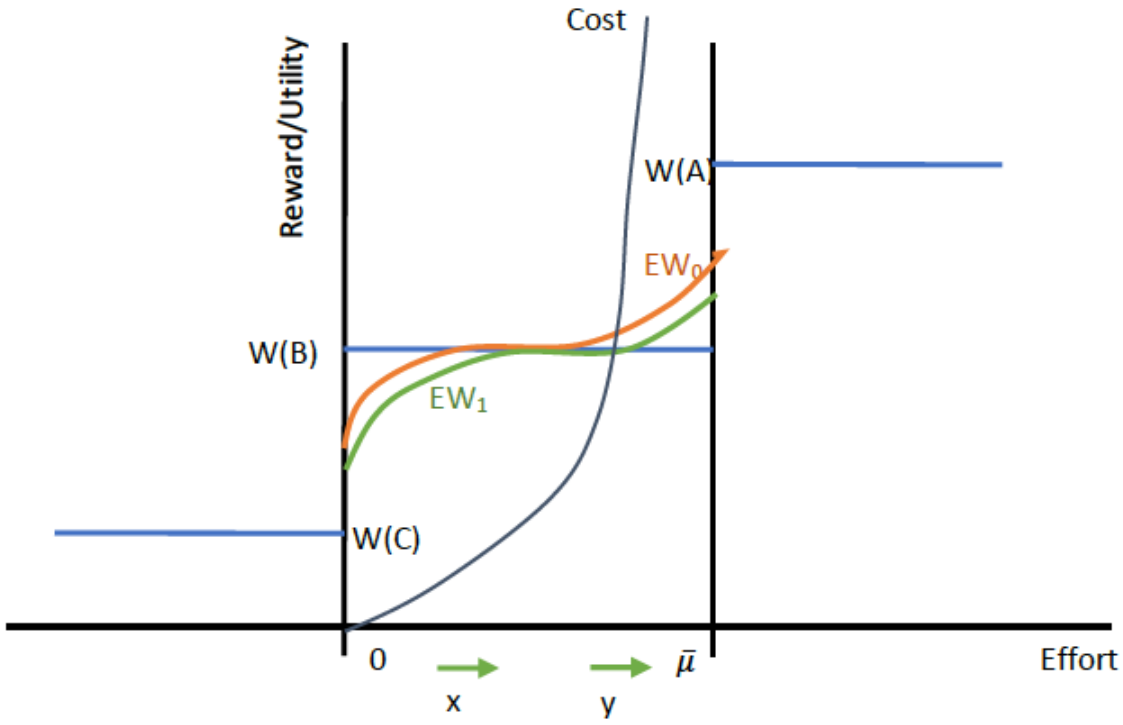


Figure 1.3: Expected Rewards Functions for LGS under different risk-attitudes (EW_1 more risk-averse than EW_0)

Notes:
1. Rewards is equivalent to utility and expected rewards is equivalent to expected utility in our theoretical framework. Consider a student who will receive W(B) in a perfect measurement world for all his efforts, $\mu \in [0, \bar{\mu}]$.
2. The expected rewards curve shifts from EW_0 to EW_1 as risk-aversion of this student increases, everything else staying constant.
3. x and y represent the risk-premium to be paid in terms of extra effort because of increased risk-aversion. x represents the extra effort required to minimize the possibility of letter grade C and corresponding reward $W(C)$. y represents the extra effort required to start anticipating letter grade A and corresponding reward $W(A)$. (EW_0 and EW_1 curves corresponding to effort below 0 and above $\bar{\mu}$ are never realized since $\mu \in [0, \bar{\mu}]$.)
4. For the shown cost function in this figure, we can see that with EW_0, student may (or may not) exert anticipatory effort close to $\bar{\mu}$ but with EW_1, student will definitely exert a precautionary (and not anticipatory) effort closer to 0.
5. This figure shows how increasing risk-aversion raises the possibility of exerting precautionary effort and diminishes the possibility of anticipatory effort, among otherwise identical students.
6. See figures A.1, A.2, A.3 and A.4 in the appendix for a step-by-step diagrammatic guide to the derivation of Figure 1.3.

Synthesising theoretical results

Risk-Neutral student: Lemma 1 shows that NGS elicits greater effort from students as compared to LGS in perfect measurement case. Theorem 1 and Theorem 3 show that, under imperfect measurement, efforts under NGS do not change and efforts under LGS increase, respectively, compared to their corresponding perfect measurement cases. Theorem 4 and Corollary 4 imply that a risk-neutral student is highly likely to exert anticipatory effort very close to $\bar{\mu}$. In purview of these results, LGS appears to be theoretically superior to NGS for a risk-neutral student under imperfect measurement scenario.

Risk-averse student: Lemma 1 shows that NGS elicits greater effort from students as compared to LGS in perfect measurement case. Theorem 2 and Theorem 3 show that, under imperfect measurement, efforts under NGS decrease and efforts under LGS increase, respectively, compared to their corresponding perfect measurement cases. Theorem 4 and Corollary 4 imply that a high risk-averse student is highly likely to exert precautionary effort in LGS while a low risk-averse student is highly likely to exert anticipatory effort. In purview of these results, LGS appears to be theoretically inferior to NGS for high risk-averse students and theoretically superior to NGS for low risk-averse students under imperfect measurement scenario[18].

Proposition 1 For a student with high (low) risk-averse attitude and imperfect measurement of effort, optimal effort exerted under LGS is highly likely to be lower (higher) than same student's effort exerted under NGS on account of precautionary (anticipatory) effort under LGS.

Proposition 1 implies that supremacy of LGS or NGS in an aggregated sense depends on the proportion of students exerting precautionary effort and the ones exerting anticipatory

[18]Corollary 2 suggests that students with high (low) risk-aversion will exert low (high) effort under NGS, however, that effort level will be on a continuous effort scale, unlike LGS.

efforts under LGS. A dominance of highly risk-averse students will imply more precautionary efforts under LGS, leading to supremacy of NGS over LGS in aggregate. Analogously, dominance of students with fairly low levels of risk-aversion will imply more anticipatory efforts under LGS, leading to supremacy of LGS over NGS in aggregate.

Although Proposition 1 has been derived from simplifying assumptions $A1$ and $A2$, the results presented are more general and are robust to the relaxation of both those assumptions. Relaxing assumption A1 will simply assign rewards $W(.)$ differently to letter grades, but as long as those rewards assigned are monotonically increasing with letter grades, the results do not change. Relaxing assumption A2 will shift the expected rewards curves $EW(.)$ leftward or rightward but an increase in risk-aversion will still lead to a flatter $EW(.)$ curve with higher risk-premium. This still implies a lower probability of exerting anticipatory effort with increasing risk-aversion, and thus, a lower likelihood of LGS dominating NGS for highly risk-averse students.

1.3 Research Design

1.3.1 Background information

The experiment is set in 10 CBSE (Central Board for Secondary Education) schools of New Delhi, India. CBSE is the most popular education board in India, run and governed by the union government, and operating 21,271 schools in India as of May 2019. These schools uniformly follow the educational strategies, curricula, pedagogical schemes and evaluation methodologies recommended by National Council for Educational Research and Training (NCERT).

In 2009, parliament of India passed the Right to Education (RTE) Act mandating free and compulsory education for all children between 6 and 14 years of age, requiring private schools to reserve 25% seats for students from economically weak and disadvantaged background, requiring schools to not hold back or expel failing students until they complete their

elementary schooling, providing special training for school drop-outs to bring them up to par with students of the same age, and making a move to continuous and comprehensive evaluation system (CCE), among others[19]. One other lesser discussed change brought about in RTE Act was a move from grading students using numbers (1 to 100) to grading students using letters (A to E). In this new system, students would only know their letter grades (LGS) and not their numerical scores. This was against the older evaluation system present before 2009 when students used to receive numerical scores on their performance (NGS).

In 2017, CBSE came out with a notice declaring a movement from CCE to a uniform system of assessment, examination and report card which was aimed at standardizing teaching and evaluation across schools. This change was also aimed at easy migration of students within the CBSE schools. This transformation was additionally accompanied by a movement back to the older system of grading students on the numerical scores (NGS). This switch from NGS to LGS after RTE Act 2009 and then back to NGS in 2017 makes CBSE schools in India a reasonable place to undertake this experiment. More specifically, I conducted this experiment with grade 8 students who had been a part of LGS system in grade 6th and moved to NGS in grade 7, thus, experiencing both NGS and LGS over past three years.

1.3.2 Intervention and Implementation

Schools in New Delhi often invite organizations and individuals to deliver workshops to students on various skills. My experiment too was embedded in the framework of a one-day workshop, conducted on separate days with only grade 8 students of 10 public and private schools. Students were not informed about this workshop until the experiment day, thus, preventing any systematic absenteeism on the day of the experiment[20].

Students in each school were divided into two groups randomly and each group was sent

[19]http://righttoeducation.in/know-your-rte/about

[20]They were introduced to the experiment as a workshop to be conducted with them and were informed of their option to quit the workshop at any stage. They were informed about the research nature of this study at the end of the project day and were informed of their right to deny us their assent to use their data if they so wished. All participating students gave their assent to use their data.

to a separate classroom. Each classroom had a projector using which an introductory video was played for students that provided instructions about the activities to follow during the day. These activities were separated into two sessions, Session 1 and Session 2[21]. Each session had three components - teaching, studying and assessment. In teaching, students were taught a mathematical topic through a recorded video lecture; in studying, students were given time and material to study on the topic taught; and in assessment, students were tested on the material taught and studied. The flow and content of activities were kept identical for both groups in both the sessions. The only difference between the two groups and thus, the intervention, was the choice of the grading system for their assessment tests.

Students in the treatment group received LGS, i.e., they were informed that their tests will be graded on letters $\{A, B, C, D, F\}$ while students in the control group continued with their business as usual with the default NGS, i.e., they received raw numerical scores in $[0, 100]$. Students in each group were introduced to their respective intervention through the introductory video played before the activities in session 1 and session 2 began. Treatment status of each group stayed same for both the sessions. Each group was unaware that the other group had a different grading system. Participation certificates mentioning the aggregate numerical scores or corresponding letter grades depending on the group's treatment status were promised to every participating student to be delivered within 7 days[22]. A detailed flow of the activities has been listed in Table 1.1.

In session 1, Lecture video 1 introduced them to the basic concepts of exponents and powers from their text book. This topic was scheduled to be covered in January, 2019 as part of their school curriculum and the experiment took place in December, 2018. This ensured minimum prior knowledge of students on this topic except what they would have gained in grade 7. Post Lecture video 1, each student was provided with a set of 40 printed questions (with solutions) to practice during their self-study session. Session 1 ended with

[21]Division of activities into two sessions was to ensure smaller duration for each activity to keep students' interest in the activities going.

[22]Prior interaction with teachers and students indicated that certificates are a significant symbolic incentive for students of this age-group to take the experimental task credibly.

S.No.	Activities	Duration
1	Information video	15 minutes
Session 1		
2	Lecture video 1 (Mathematics)	25 minutes
3	Self-Study Session 1	30 minutes
4	Assessment Test 1	30 minutes
5	Lunch	30 minutes
Session 2		
6	Lecture video 2 (Mathematics)	20 minutes
7	Self-Study Session 2	30 minutes
8	Assessment Test 2	30 minutes

Table 1.1: Project Day Routine

students writing Test 1. Post lunch break, students entered session 2 where they attended Lecture video 2 which was also conducted on exponents and powers introducing higher level concepts with negative bases and powers. Lecture video 2 was followed by self-study session on another set of 40 questions and wrote Test 2 afterwards. Introductory video and lecture videos were both recorded in a professional manner with a mathematics teacher who was unknown to the participating students[23].

1.3.3 Sample, Duration and Randomization

The experiment was administered with grade 8 students of 10 public and private schools in New Delhi, India that catered to low-income neighborhoods. Four schools were all-girls schools, four were all-boys schools and another two were co-educational schools. Each school participated in the experiment only for one day, with a maximum of two schools participating in the experiment on any specific day. The entire experiment was spread over a period of 15 days from 10th to 24th December, 2018. Access to these schools was provided through collaboration with an organization that has adopted one classroom of grade 8 in each of these 10 schools. All 459 students studying in those classes who were present on the day of the experiment became our sample for this study and were individually randomized into

[23]Both the classrooms were populated only by students and one research staff member who facilitated the flow of activities. I used only two facilitators for each of the 10 participating schools. These two facilitators would alternate between their assignment to treatment and control classrooms in each school.

treatment and control groups which were introduced to students as Group 1 and Group 2. After group assignment, students with their respective group members were sent to separate classrooms with a research staff member. The research staff member stayed inside the classroom for the entire school day and facilitated the flow of activities during the day. Introductory video introduced students about their respective assignment status, treatment (LGS) or control (NGS), and the set of activities to be conducted during the rest of the day[24].

Of the 459 students, 454 students (98.9%) participated in the entire experiment and the rest 5 (2 from treatment group and 3 from control group) did not write either Test 1 or Test 2 that were both conducted on the same day[25]. Further, 16 students (8 from treatment group and 8 from control group) of the 454 students had not written the mathematics midterm examination conducted in October, 2018, which I use as a measure of student's prior knowledge. This brings my effective sample size to 438 students (95%). 217 students (49.5%) were in control group and 221 students (50.5%) were in treatment group. Control group had 114 male (52.5%) and 103 female (47.5%) students, and treatment group had 107 male (48.4%) and 114 female (51.6%) students.

1.3.4 Data

The primary outcome of interest for this study is student's average of Test 1 and Test 2 percentage scores (Endline scores, henceforth). Same teacher designed both these tests[26]. The tests comprised of questions varying in difficulty from very easy to very difficult for a grade 8 student. They were paper-and-pen mode, and were administered and monitored by research staff members in December 2018 (no teacher from school was allowed to be present inside the classroom during the experiment). Answer scripts for all participating students

[24]Student-level randomization with assignment of students to different groups was conducted using slips mentioning one of the two group numbers, with each student picking up one slip from the well shuffled lot.

[25]Three students missed Test 1 and two students missed Test 2 because they came late or had to leave early for health reasons, or because they had to participate in their preparation for schools' annual festival.

[26]The teacher involved in designing the tests works at a school in Delhi and teaches mathematics to grade 8 students.

were corrected in the week following the experiment day and by only one teacher who did not belong to any of the participating 10 schools. She had no access to personalized information about these students and did not even know the names of these participating schools[27]. This makes the performance in the treatment and control groups comparable, devoid of any teacher-related effect.

From the schools' administrative records, I collected students' individual level mathematics test scores on the midterm examination conducted in October 2018 (Baseline scores, henceforth). This is used to test balance in the prior knowledge of students between the treatment and control groups. This also acts as a baseline measure of student knowledge to control for any biases in the treatment effect that might emerge from differences in students' prior knowledge. In addition, I use students' gender and school name information to control for any effects on students performance emanating from these factors. Table C.1 presents summary statistics of the described data.

1.3.5 Estimation

The basic framework is the linear regression model for student performance q for student i in school s, which can be specified as:

$$q_{is} = \beta_0 + X'_{is}\beta_1 + \beta_2 Treated_{is} + SchoolF.E. + \epsilon_{is} \tag{1.10}$$

X_{is} includes students' gender and baseline scores[28]. $Treated_{is}$ is an indicator variable that takes value 1 if student i in school s is a member of the treatment group, LGS, and 0, otherwise. School F.E. accounts for the time-fixed effects of the differences between schools

[27]The teacher involved in correcting these tests works at a college in Delhi and teaches mathematics to undergraduates.

[28]Due to confidentiality constraints, I was not allowed access to any other explanatory (socio-economic or demographic) variables by the partner institutions. This, however, should not affect the treatment effect by virtue of students' individual-level randomization between treatment and control groups. This is further vindicated in the results section where I see that inclusion of gender and/or school fixed effects do not change much the treatment effect or its significance, indicating that both groups would be well-balanced on socio-economic and demographic measures too.

	Mean(Treatment)	Mean(Control)	Difference	SE	N(Treatment)	N(Control)
Male	0.48	0.52	-0.04	0.04	221	217
Pre-intervention						
Baseline All	-0.05	0.05	-0.10	0.09	221	217
Baseline Females	-0.16	0.06	-0.22*	0.13	114	103
Baseline Males	0.06	0.05	0.01	0.13	107	114
Post-intervention						
Endline All	-0.04	0.04	-0.08	0.09	221	217
Endline Females	-0.09	0.20	-0.29**	0.13	114	103
Endline Males	0.02	-0.10	0.12	0.13	107	114

Standard errors in parentheses

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

Note: Treatment and control groups refer to students who were randomly assigned to LGS and NGS, respectively. The statistics presented show gender division (dummy male takes value 1 if male and 0 if female) and baseline test scores balance between treatment and control groups. Baseline test scores were the prior test scores on mathematics midterm examination that is a standardized test across the public schools in the city of New Delhi. Table also presents summary statistics of endline test scores for all students, for female students separately and for male students separately. Scores are standardized to have a mean of zero and standard deviation of one in the pre-intervention baseline test scores and also in the post-intervention endline test scores.

Table 1.2: Summary Statistics and Balance on Baseline Test Scores

that affect all students studying within a specific school uniformly and thus, could potentially bias the results.

I estimate this model for endline scores data (i.e., average of Test 1 and Test 2 percentage score for each student). The choice of presenting results for average scores as an outcome measure and not for each test separately is due to the fact that the certificates declared to students as an incentive for participating in the experiment were based on performance over both the tests. They were told that the certificates will mention average performance over Test 1 and Test 2 as a unique numerical score on the certificate if control group and the letter grade corresponding to that numerical score if treatment group. Additionally, research staff reported factors like gain or loss of interest between sessions, tiredness, lapses of concentration, etc., among students during the experiment. This, therefore, makes the choice of average performance over Test 1 and Test 2 for each student a more reasonable outcome measure of endline scores used for estimation[29]. Estimate of coefficient β_2 is the primary parameter of interest measuring the average treatment effect of being assigned to the treatment group, LGS.

1.4 Results

1.4.1 Balance test

Students' prior test scores data on mathematics midterm examination is used as a baseline measure to characterize the initial knowledge of students before they participated in the experiment. Tables C.1 reports the results on balance between treatment and control groups in students' distribution of these baseline scores. The table also presents the heterogeneity in baseline scores balance for separate gender. The results show that baseline scores are

[29]I also estimate this model with test scores pooled over Test 1 and Test 2 as a robustness check while controlling for baseline scores, school fixed effects, adding a test dummy and clustering standard errors at student level. The estimates do not change in magnitude or direction from those estimated with summed test scores data.

balanced, on average, between the two groups when both gender are considered together or when only male students are considered. The baseline scores balance is lost, however, when I consider only female students. These results are also highlighted in kernel density estimates of prior test scores shown in Figure 1.4, Figure 1.5 and Figure 1.6. The treatment effect calculated from a comparison of simple averages of the endline scores will be biased due to not so perfect randomization among females. A control for the baseline scores in our regression analysis, however, reduces or eliminates this bias.

1.4.2 Treatment effect

Table C.1 presents the difference in means for endline scores and finds that students in the treatment group performed 0.08sd worse than students in the control group. This difference, however, becomes close to zero (-0.01sd) when controlled for baseline scores and school fixed effect as shown in column 1 of Table 1.3 and stays close to zero (-0.02sd) when controlled for gender too as shown in column 2 of Table 1.3[30]. This negligible average treatment effect suggests that students who were assigned to the treatment group (LGS) performed no differently from students who were assigned to the control group (NGS). In other words, this analysis suggests that policy-makers need not worry about choosing between a coarse LGS or fine NGS on account of their effect on student performance.

In the presence of ample evidence on gender-differences in how students respond to incentives, I study if any such gender-related heterogeneity is present in grading context too. I include an interaction term between gender and treatment in my estimation model and find results that reaffirm prior evidence on gender-differences. Column 3 of Table 1.3 shows the estimates for the model specification with the interaction term. This specification finds the average treatment effect of -0.15sd which is much higher when compared with the -0.01sd or -0.02sd average treatment effect in columns 1 and 2, respectively, of Table 1.3. This sug-

[30]Note that baseline test scores explains most of the variation in dependent variable. Gender and school fixed effect do not change much the magnitude, direction or significance of the treatment effect. This is expected too because of individual-level randomization of students to treatment and control groups.

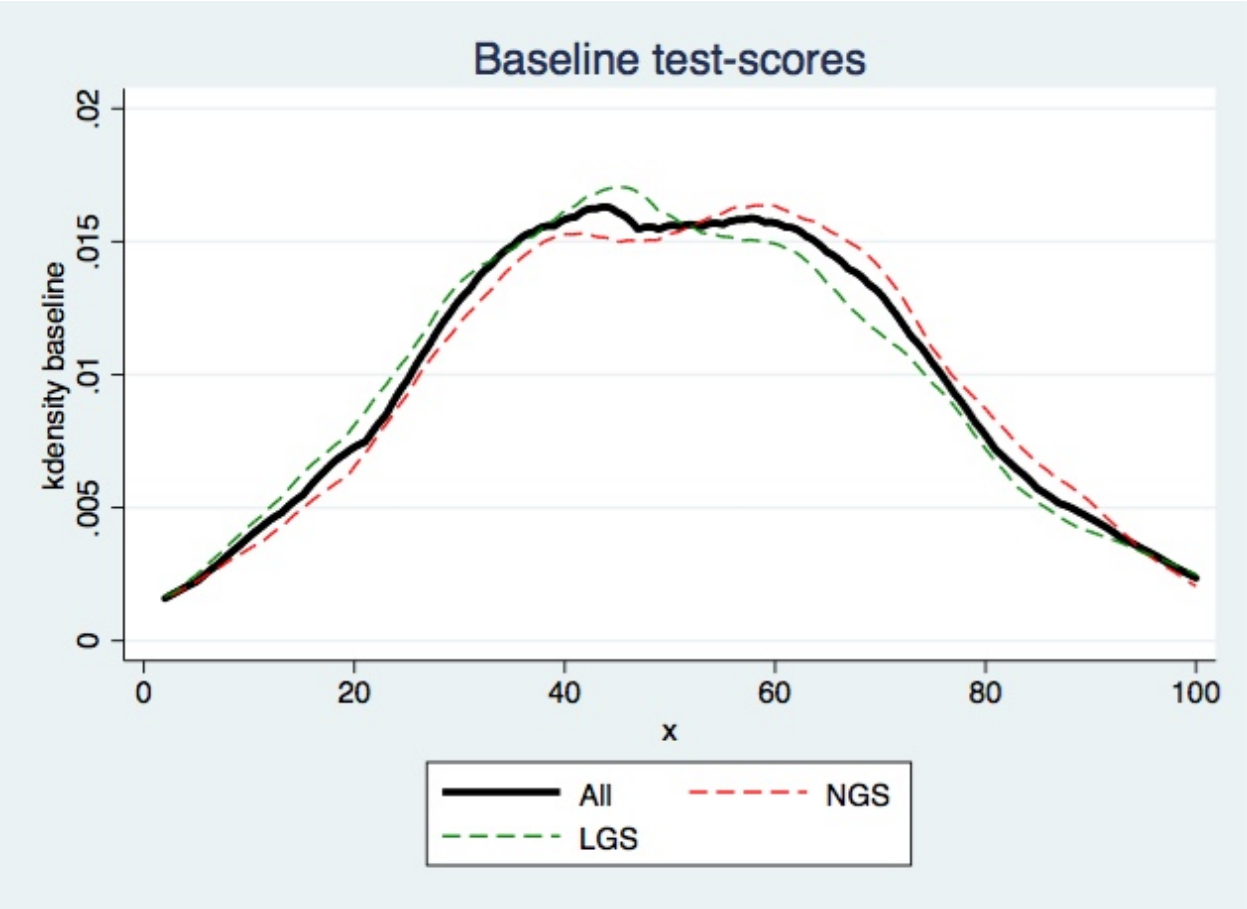


Figure 1.4: Kernel Density Plot - Basleine Test Scores All

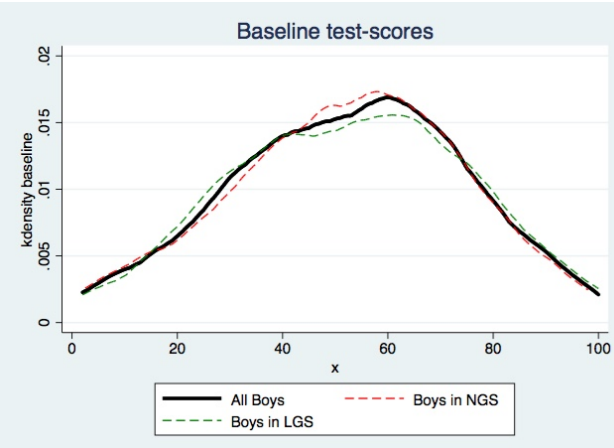


Figure 1.5: Kernel Density Plot Males

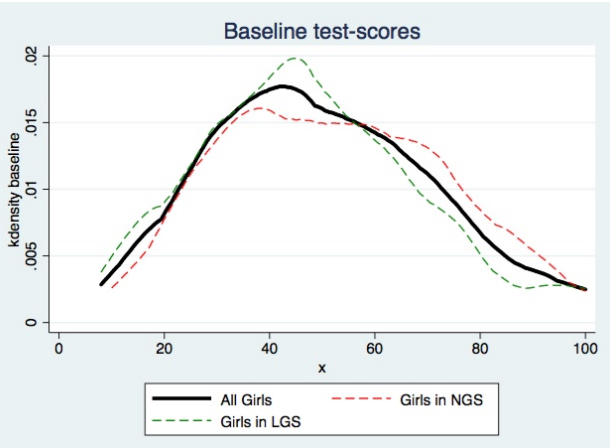


Figure 1.6: Kernel Density Plot Females

gests that treatment has, on average, a negative effect on student performance which implies that students under NGS perform better over LGS. However, coefficient estimate for the interaction term between gender and treatment is 0.27sd which indicates the difference in average treatment effect between male and female students. This implies that in comparison to female students, assignment to treatment affects male students by additional 0.27sd. This additional interaction effect makes the average treatment effect on male students positive while it stays negative for female students. This can be seen more clearly in Table 1.4 where I pursue this analysis by gender subgroups.

Columns 1 and 2 of Table 1.4 study how performance of female students in treatment group compare with female students in control group, and how performance of male students in treatment group compare with male students in the control group, respectively. Among female students, I find the treatment effect to be -0.14sd, i.e., students in treatment group performed 0.14sd worse than students in control group. Among male students, I find the treatment effect to be 0.12sd, i.e., students in treatment group performed 0.12sd better than students in control group. In other words, these estimates suggest that females deliver better performance when assigned to NGS and males deliver better performance when assigned to LGS. This opposite movement of average treatment effects (-0.14sd for female students and 0.12sd for male students) for the two gender groups explains why the average treatment effect was negligible in columns 1 and 2 of Table 1.3.

1.4.3 Heterogeneity

Regression estimates of the average treatment effect on student performance may have given an incomplete picture of the true effect of the grading systems. Using quantile regression plots, I investigate the possibility of the heterogeneity of treatment effect across different quantiles of the performance distribution on the endline scores. Since controlling for gender and including school fixed effects had no impact on the average treatment effect as shown in column 2 of Table 1.3, I pursue quantile regression analysis with baseline test scores as

	(1) Endline	(2) Endline	(3) Endline
Treatment	-0.0146	-0.0171	-0.153
	(0.0713)	(0.0711)	(0.107)
Baseline	0.618***	0.609***	0.604***
	(0.0397)	(0.0412)	(0.0415)
Male		-0.237	-0.376*
		(0.197)	(0.209)
Treatment*Male			0.267*
			(0.145)
School F.E.	Yes	Yes	Yes
Constant	0.356**	0.354**	0.425**
	(0.168)	(0.168)	(0.178)
Observations	438	438	438
R-Squared	0.462	0.465	0.469

Standard errors in parentheses

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

Note: Treatment and control groups refer to students randomly assigned to LGS and NGS, respectively. Treatment takes value 1 if student assigned to LGS, 0 if NGS. Baseline test scores were the prior test scores on mathematics midterm examination that is a standardized test conducted across the public schools in the city of New Delhi. Endline test scores are average of Test 1 and Test 2 percentage scores for each student. Both baseline and endline scores are standardized to have a mean of zero and standard deviation of one. Male dummy takes value 1 if male student and 0 if female student.

Table 1.3: Post-Intervention Regression Results

	(1) Endline Female	(2) Endline Male
Treatment	-0.145	0.118
	(0.108)	(0.0964)
Baseline	0.614***	0.608***
	(0.0743)	(0.0515)
School F.E.	Yes	Yes
Constant	0.423**	-0.189
	(0.179)	(0.156)
Observations	217	221
R-Squared	0.455	0.486

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Treatment and control groups refer to students randomly assigned to LGS and NGS, respectively. Treatment takes value 1 if student assigned to LGS, 0 if NGS. Baseline test scores were the prior test scores on mathematics midterm examination that is a standardized test conducted across the public schools in the city of New Delhi. Endline test scores are average of Test 1 and Test 2 percentage scores for each student. Both baseline and endline scores are standardized to have a mean of zero and standard deviation of one.

Table 1.4: Post-Intervention Regression Results- Male and Female

the only controls. Figure 1.7 presents three different panels with quantile regression estimate plots - first for intercept, second for treatment, and third for baseline scores. Each of these plots has quantile scale on horizontal axis and quantile treatment effect on the vertical axis. The dashed line depicts the ordinary least squares estimate of the conditional mean effect and the dotted lines around it show its 95% confidence interval. The solid line depicts conditional quantile regression estimates and grey area around shows its 95% confidence interval.

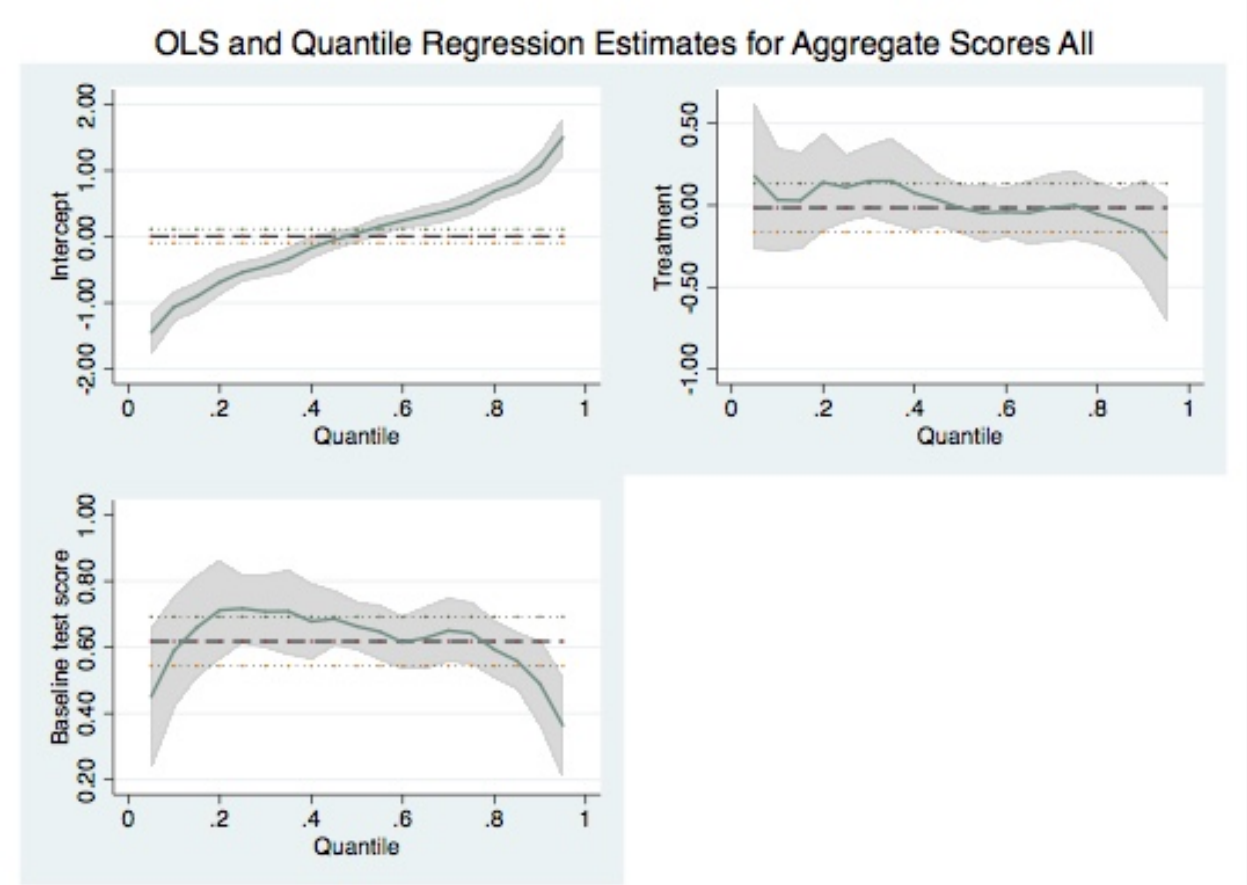


Figure 1.7: Quantile Regression - Endline Scores All

I will confine my discussion to the second panel in Figure 1.7 showing conditional quantile regression function for the treatment which gives the treatment effect at various quantiles of the endline scores' performance distribution, conditional on other covariates. With 95% confidence, this panel shows that assignment to treatment has a uniform effect, indifferent

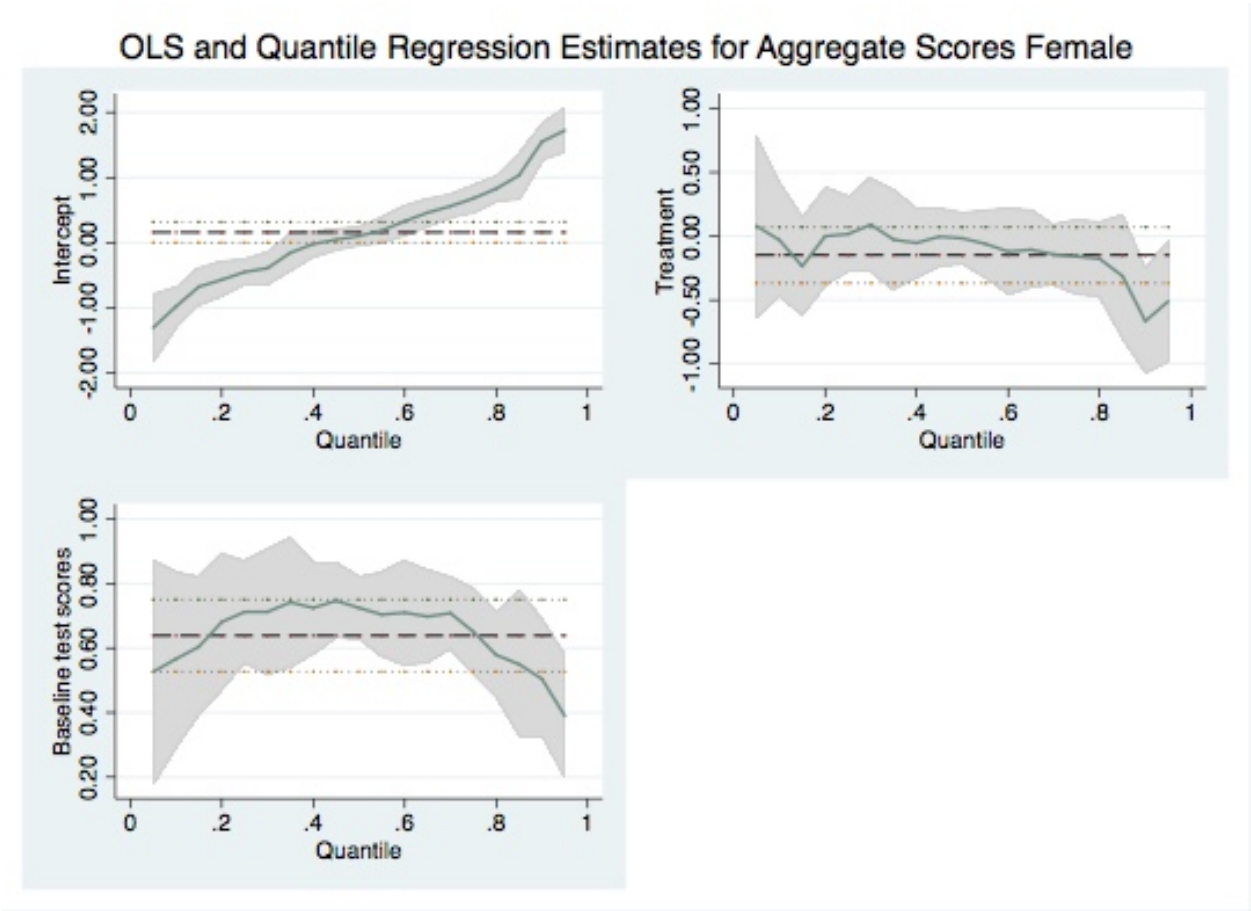


Figure 1.8: Quantile Regression - Endline Scores Females

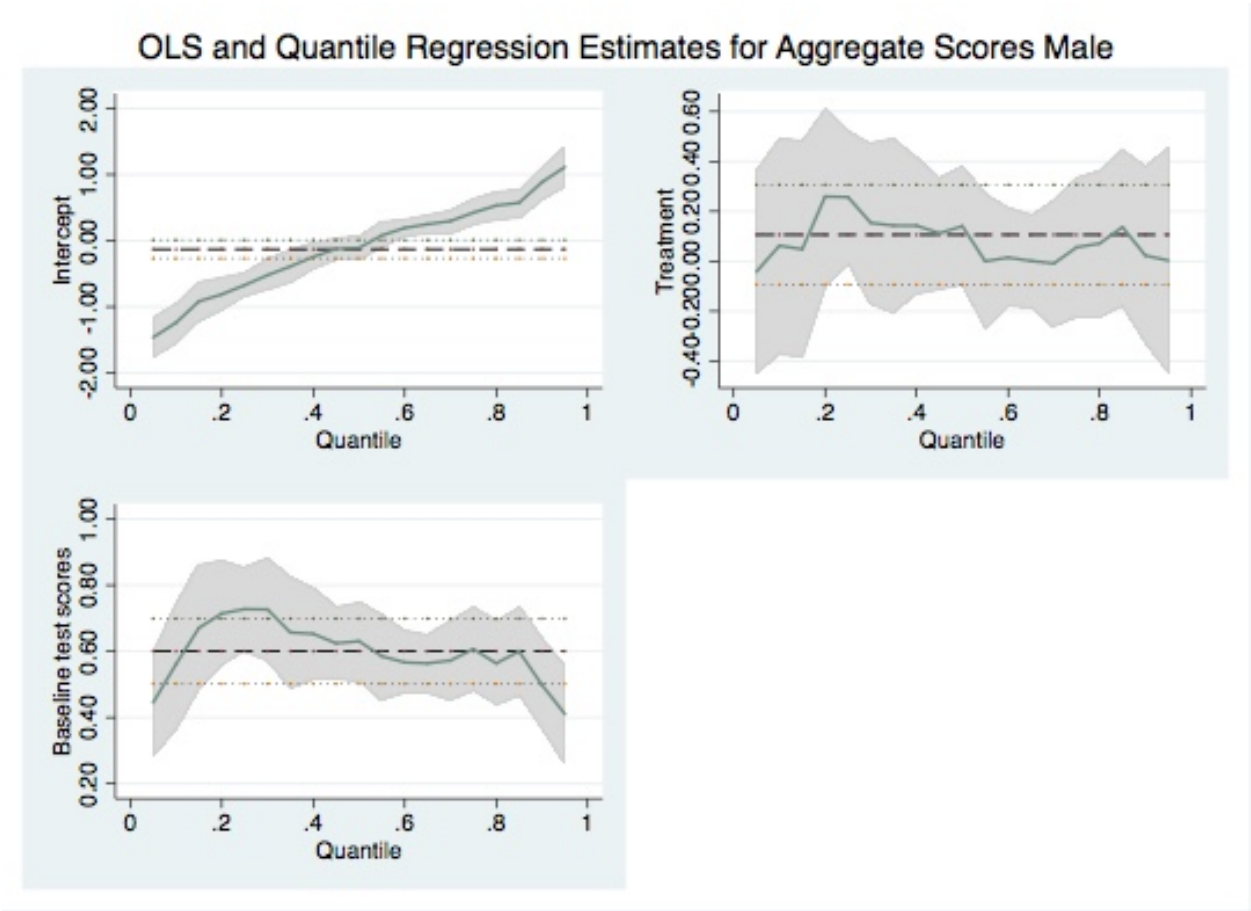


Figure 1.9: Quantile Regression - Endline Scores Males

from the average treatment effect of -0.01 sd, on the entire range of endline scores' distribution (except with extreme top quantiles). More importantly, the indifference of average treatment effect from conditional quantile treatment effect is preserved when I consider only female students in Figure 1.8 and only male students in Figure 1.9. Thus, in all three figures, I see that the quantile regression estimates lie within the 95% confidence interval of the ordinary least square estimates. This suggests that the effect of treatment is uniform across the performance quantiles.

1.4.4 Robustness check and other concerns

While the estimates reported in Table 1.3 give clear evidence of zero negligible effect masking the gender-differences in students' response to treatment, the sample size is not big enough (438 observations). As a robustness check, I double up on the sample size by pooling the test scores data over Test 1 and Test 2[31]. In doing so, each student counts as two separate observations, one for each test. To control for the correlation between these two observations for each student in the pooled data, I estimate the average treatment effect after clustering the standard errors at student level while controlling for baseline test scores, adding school fixed effects and a test dummy. Column 1 of Table 1.5 reports these results. The average treatment effect observed is -0.01 which is identical to the effect observed in column 1 of Table 1.3. Addition of a gender dummy does not change the treatment effect as seen in column 2 of Table 1.5. Column 3 of Table 1.5 presents the results after inclusion of interaction term between gender and treatment, and the results are a mirror reflection of the results observed in column 3 of Table 1.3.

Similarly, the average treatment effects from pooled data for female students only and for male students only observed in columns 1 and 2 of Table 1.6 are not very different from those observed in columns 1 and 2 of Table 1.4. Thus, estimates reported in Table 1.5 and Table 1.6 provide a robustness check for our estimates observed in Table 1.3 and Table 1.4.

[31]The treatment status of students did not change between Test 1 and Test 2.

	(1) Pooled Endline	(2) Pooled Endline	(3) Pooled Endline
Treatment	-0.0126	-0.0149	-0.132
	(0.0615)	(0.0613)	(0.0919)
Baseline	0.536***	0.528***	0.524***
	(0.0342)	(0.0355)	(0.0357)
Male		-0.205	-0.326*
		(0.170)	(0.180)
Treatment*Male			0.232*
			(0.125)
Test	Yes	Yes	Yes
School F.E.	Yes	Yes	Yes
Constant	1.192***	1.191***	1.252***
	(0.154)	(0.154)	(0.163)
Observations	876	876	876
R-Squared	0.434	0.436	0.439

Standard errors in parentheses

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

Note: Treatment and control groups refer to students randomly assigned to LGS and NGS, respectively. Treatment takes value 1 if student assigned to LGS, 0 if NGS. Baseline test scores were the prior test scores on mathematics midterm examination that is a standardized test conducted across the public schools in the city of New Delhi. Endline scores are the pooled scores over Test 1 and Test 2. This gives 876 observations since each test is written by 438 students. Both baseline and endline scores are standardized to have a mean of zero and standard deviation of one. Male dummy takes value 1 if male student and 0 if female student. Two tests were conducted during the experiment. Test dummy takes value 1 if Test 2 and 0 if Test 1.

Table 1.5: Post Intervention Pooled Test Scores Regression Results

	(1) Pooled Endline Female	(2) Pooled Endline Male
Treatment	-0.126	0.103
	(0.0928)	(0.0830)
Baseline	0.532***	0.527***
	(0.0640)	(0.0444)
School F.E.	Yes	Yes
Test	Yes	Yes
Constant	0.658***	0.135
	(0.156)	(0.135)
Observations	434	442
R-Squared	0.421	0.461

Standard errors in parentheses

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

Note: Treatment and control groups refer to students randomly assigned to LGS and NGS, respectively. Treatment takes value 1 if student assigned to LGS, 0 if NGS. Baseline test scores were the prior test scores on mathematics midterm examination that is a standardized test conducted across the public schools in the city of New Delhi. Endline scores are the pooled scores over Test 1 and Test 2. This gives 876 observations since each test is written by 438 students. Both baseline and endline scores are standardized to have a mean of zero and standard deviation of one. Two tests were conducted during the experiment. Test dummy takes value 1 if Test 2 and 0 if Test 1.

Table 1.6: Post Intervention Pooled Test Scores Regression Results

One concern that can bias the estimation of average treatment effect is different teachers teaching in treatment and control classrooms which may lead to different understanding of the same concepts for different students. To nullify this potential teacher effect from the experiment, I had opted for recorded video lectures to keep the teaching identical in both treatment and control classrooms. This ensured that any difference in performance found between the two groups cannot be biased due to different teachers teaching them. Another issue that can potential bias the estimates is if different teachers corrected the test scripts. To bring this between-teacher effect down to zero, only one teacher was recruited to correct all 876 test scripts (2 test scripts for each of the 438 participating students).

Another concern that needs to be addressed is sample attrition. 5 students (2 from treatment group and 3 from control group) either left the experiment after session 1 or joined the experiment directly in session 2. Attrition due to both those reasons were random since students who left early after session 1 were let go off of the experiment on the request of school teachers towards their preparation for school festival. Students who joined the experiment in session 2 had come late to school for health reasons and had no idea about this experiment before they arrived. Additionally, there were 16 other students (8 from treatment group and 8 from control group) who participated in the experiment in both the sessions but did not write the baseline mathematics mid-term examination conducted in October, 2018. Absence of these 16 students from their midterm examination may be suggestive of their differences from the other 438 students who attended the baseline. However, the fact that 8 of these students came from treatment group and another 8 from control group, once again, confirms that randomization exercise was done properly and therefore, does not bias the estimate of average treatment effect.

1.5 Discussion

One of the common findings in the student-incentive domain (including monetary rewards, tuition waivers, vouchers, varying stakes in an exam, among others) of education literature is about gender-differences in how students respond to those incentives (Angrist et al., 2002; Dynarski, 2008; Angrist and Lavy, 2009; Fryer, 2011; Ors, Palomino and Peyrache, 2013; Jalava, Joensen and Pellas, 2015; Katreniak, 2018). The experimental results presented in this paper find negligible difference in student performance between LGS and NGS as seen in column 2 of Table 1.3. However, Table 1.4 shows that this negligible difference is masking important gender-differences in students' response to the two grading systems well in alignment with the prior literature on student incentives. Female students performed better in NGS and male students performed better in LGS. These results suggest that in an education system that has opted for grading students with NGS (like in India), male students are under-incentivized and are performing below their potential. In an education system that has opted for grading students with LGS (like in the United States), female students are under-incentivized and are performing below their potential. This evidence also suggests that the choice of an appropriate grading system could be one of the possible ways to reduce gender-differences in student learning, which has been attracting a lot of attention in the education literature.

Policy makers can improve the performance of male students and female students through the choice of an optimal grading system at the least in single-sex schools. There are several developing countries that have historically had majority of their public schools as single-sex schools (Pakistan, Bangladesh, India, Iran, Syria, etc.). In addition, developed countries like New Zealand or Ireland have also had around 20% of students studying in single-sex schools while a lot of developed countries like United Kingdom, United States and Australia have started seeing increasing proliferation of single-sex schools (or single-sex classes within a co-educational school) in an attempt to close the increasing achievement gap between female and male students (Younger and Warrington, 2006; Salvi del Pero and Bytchkova, 2013;

Czibor et al, 2014; Else-Quest and Peterca, 2015). The existence (and increasingly so) of single-sex schools (or classes) in both developing and developed countries makes the results of this paper relevant more widely than to a specific Indian context.

To put things in perspective about how grading as a policy tool fares, I compare it with the treatment effects of other costly interventions undertaken to improve student learning (see Glewwe and Muralidharan (2016) for a literature review). The average treatment effect on students performance from the provision of scholarships is found to be up to 0.28σ (Kremer, Miguel and Thornton, 2009; Blimpo, 2014; Li et al., 2014), from conditional cash transfer to be up to 0.20σ (Baird, McIntosh and Ozler, 2011; Baez and Camacho, 2011; Barham, Macours and Maluccio, 2013; Benhassine et al., 2015), from class size reduction to be up to 0.09σ (Urquiola and Verhoogen, 2009; Duflo, Dupas and Kremer, 2015), and from educating parents to be up to 0.06σ (Banerji, Berry, and Shotland, 2013; Handa, 2002). In comparison to these interventions, choice of optimal grading system showed an average treatment effect of at least 0.12σ on student performance. However, an almost zero monetary cost nature of grading intervention makes this as one of the most cost-effective policy tools that must be given enough consideration by policy makers in the endeavor to improve student performance.

While the empirical finding about gender-differences in students' responses to different grading systems is an important take-away in itself, I discuss the possible mechanisms that may lead to such differences. Past literature has often attributed gender-differences in students' response to incentives to their differences in study habits, self-discipline, competitiveness, etc. One mechanism that this strand of literature has omitted is that of gender-differences observed in their risk-attitudes. Experimental studies have shown females to be more risk-averse than males, on average, in both developing and developed countries (see Charness and Gneezy (2012) for a literature review), in matrilineal and partiarchal societies (Gong and Yang, 2012, Mukherjee, 2017) and in lab and in field studies (see Eckel and Grossman (2008) for a literature review)[32].

[32]There are some studies that find no or insignificant evidence about females being more risk-averse than males, but those studies are an exception and not a norm.

The theoretical model presented in the paper shows how risk-attitudes play a pivotal role in determining the dominance of a grading system. It finds that the likelihood of NGS eliciting higher student effort (and performance) increases in students' risk-aversion, ceteris paribus. In light of the prior empirical evidence about gender-differences in risk-attitudes, we can consider female students as proxy for high-aversion and male students as proxy for low risk-aversion. The theoretical model, thus, predicts female students to perform better in NGS and male students to perform better in LGS. This theoretical prediction concurs with the experimental results we saw in Table 1.4 where female students perform better (-0.14sd) in NGS and male students perform better in LGS (0.12sd). While the risk-aversion argument produced by the theory and the empirical evidence produced by the experiment align well, I next discuss other possible mechanisms that may or may not be ruled out.

In our experimental design, students were randomly and individually assigned to either treatment (LGS) or control (NGS) group. Thus, by the definition of a randomized controlled trial, students in both groups can be assumed to be balanced in study habits, self-discipline and competitiveness before the experiment began. Factors like study-habits and self-discipline develop over longer period of time; are not likely to change in a short span of an experimental day; and are very likely to be independent of the chosen grading system. Thus, it can be argued that these factors did not change or change differently in the two groups to act as a mechanism behind the gender-differences observed in student performance. It is, however, possible that different grading systems may have primed different levels of competitiveness between the two groups of students. While the role of competitiveness cannot be completely ruled out, the choice of grade 8 students as the sample for the experiment ensured that this role is not as big as it would be for students in high schools or in college.

To summarize, I can say that while differences in risk-attitudes is an intuitive mechanism contributing to gender-differences in student response to the two grading systems, the role of differences in competitiveness cannot be ruled out completely and will require further investigations to disentangle the role of risk-attitudes and competitiveness. In our static ex-

perimental setting, gender-differences in study habits and self-discipline cannot be expected to influence the gender-differences in student performance in the two grading systems. However, they may gain significance in driving differences in student behavior in the two grading systems when it comes to a more dynamic multi-period setting. These may be some food for thought for future research in this area.

1.6 Conclusion

Different countries, school settings and examinations adopt grading systems that can be as coarse as a pass and a fail or as fine as any real number possible on a 0-100 scale. In this paper, I present the first extensive theoretical model that deals with the incentive structures of different criterion referenced grading systems, more specifically, numerical grading system (NGS) and letter grading system (LGS). The main result of the theoretical model finds NGS to be more likely to elicit greater effort than LGS as risk-aversion of students increase.

The paper presented the results from a field experiment conducted with 438 grade 8 students from schools of New Delhi, India. These students were individually and randomly assigned to treatment (LGS) and control (NGS) groups. The estimation found no aggregate differences between LGS and NGS, but uncovered significant gender-differences in how students responded to LGS and NGS. Female students in NGS performed better than female students in LGS, and male students in LGS performed better than male students in NGS. The effects were uniform across entire performance quantile range. I also discuss several mechanisms that could possibly explain these gender-differences, and risk-attitudes (and competitiveness) differences seem to play a major role.

This is, however, the first empirical investigation about the effectiveness of these two grading systems. More research needs to be conducted to replicate these findings. If these findings hold their ground in the future research, then apart from education, other areas will also stand to gain. It will nudge researchers to also investigate the optimal grading policies

for restaurants' food and hygiene, employees' work and commitment, individuals' credit-worthiness, uber riders' consumer etiquette and drivers' service, websites' privacy policies, among others.

Chapter 2

Competitive Games to Nudge Students: Experimental Evidence From India

2.1 Introduction

The causal impact that attending a higher number of school days (higher attendance, henceforth) has on improving a student's performance is well known and widely established in several contexts (Dobkin et al., 2010; Aucejo and Romano, 2014; Gershenson, Jacknowitz, and Brannegan, 2017). This is the core of the reason for policymakers to believe that higher attendance should be the focus area of all educational institutions. This is all the more important for developing countries where education status still paints a dismal picture. A 2019 ASER report in India suggests that only 16.2% grade 1 level students could read their own grade level text and only 50.8% of grade 3 students could read a grade 1 level text[1]. This shows the huge gap present between the real and intended knowledge of students. Policies aimed at increasing attendance could help bridge some of that knowledge gap. This paper

[1]http://img.asercentre.org/docs/ASER%202019/ASER2019%20report%20/aserreport2019earlyyearsfinal.pdf

experiments with one such low-cost policy intervention in a not-for-profit educational set-up of India.

Students (and parents) are often unable to appropriately evaluate the current costs of attending classes against the future benefits of better academic and labor outcomes. This could be due to their impatience, present-biasedness or access to incomplete information. To counter such sub-optimal decision-making with respect to enrolment, attendance and student learning, developing countries have tried both supply-side and demand-side interventions.

Glewwe and Muralidharan (2016) and Muralidharan (2017) give a detailed insight into the supply-side measures. This paper, however, is a contribution towards the demand-side interventions. Other such demand side interventions include, among others, the use of Conditional Cash Transfers to parents (Attanasio et al., 2012; Benhassine et al., 2013), provision of better information about returns to education (Jensen, 2012), notifying parents of their wards' learning outcomes (Bobba and Frisancho, 2016) , spreading information about school quality in competitive education markets (Andrabi et al., 2015), and student-level incentives for better academic performance (Kremer, Miguel, and Thornton, 2009; Blimpo, 2014).

Another strand of literature with which this paper connects most closely concerns the use of symbolic rewards to incentivize students towards higher attendance. Springler et al (2015) studied the effectiveness of such incentives by conducting an RCT with 300 middle grade students. They found that symbolic reward group attended 42.5% more allotted tutoring hours than those assigned to the control group. Robinson et al (2018), on the other hand, found that pre-announced symbolic awards had no impact on student attendance. Both these studies were pursued in developed countries and had offered rewards in non-competitive settings. While mixed results about the success of such symbolic rewards call for further research, this paper differs from the prior literature on two counts- (i) our experiment was set in a developing country, and we tested the effectiveness of such symbolic incentives for higher attendance in this specific setting, and (ii) our experiment was designed keeping in

mind the principles of game theory, and took the form of a competitive game.

Taking cue from the hypothesized link between attendance and learning outcomes, the paper developed a competitive game to enhance the extrinsic motivation among students towards maintaining higher attendance. Students could score points for every school day they attended. These scores were managed to reward consistency by awarding students bonus points for being present on all working days in a week. This gave them the incentive to attend all school days in any week. Also, the maximum points that a student could potentially score were doubled in the latter half of the experiment's duration (weeks 5 to 8). This was done because existing institutional trends indicated greater absenteeism in the second month of school. Thus, the experimental design accounted for existing attendance patterns, and moulded incentives accordingly. This weighing strategy ensured greater incentive in the latter half of school days. These novel aspects of the game make this paper different from, and a significant improvement upon existing literature.

In this paper, we report the results from a randomized control trial (RCT) conducted with 217 classrooms of a not-for-profit educational institution in India. We studied the effect of our intervention on student attendance in two contexts- (i) in a large group setting (Classroom Game or CG or Treatment 1 or T1) where all the students within a classroom competed for a total of four top positions; and (ii), in a small group setting (Group Game or GG or Treatment 2 or T2) where a classroom was randomly divided into four groups and the student within every one of these groups competed for one top position each. The randomization was done at classroom level with 78 classrooms allocated to T1, 63 classrooms allocated to T2, and 76 classrooms serving as the control group or the comparison group.

The study is unique based on the following aspects: (a) In the education space, it is one of the few studies to be conducted in a developing country which explores policies that target students, as opposed to parents and communities. (b) By focusing on short term, tangible incentives, the study significantly diminishes the time lag between student action (eg: attendance, effort, etc.) and the realization of incentives, making a case for policy

to reward attendance, as opposed to performance only. (c) The design uses a modified, weighted version of the standard attendance game and thus, proposes a tool that different institutions can explore to suit their specific contexts, making room for customisation of the experimental design. (d) The comparison of budget equivalent interventions (T1 and T2) will bring out the behavioral implications of varying group sizes in a competitive setting.

The remainder of this paper is organized as follows: Section 2 provides the experimental details including institutional information, data, econometric methodology and summary statistics. Section 3 presents the main results of the experiment. Section 4 interprets the findings from the study and comments on their policy implications. Section 5 concludes the paper.

2.2 Experiment Details

2.2.1 Background Information

Freedom English Academy (FEA) is an Indian not-for-profit organization operating more than 100 schools (learning centres) spread over 11 cities in the states of Delhi and Uttar Pradesh, India. It provides free English language and non-cognitive skill development classes to students in the age group of 15-22. Students enrol in Grade 1 and graduate after completing Grade 5 – FEA's final grade level.

There are multiple classrooms at each grade level, with classes being held at different times of the day, and at different schools of FEA. There are about 20 students in each classroom. Each grade level requires students to attend classes for 2 months (for 6 days in a week, and for 1 hour 45 minutes each day). In these 2 months, the FEA faculty delivers lectures from an English book (different for each grade level) designed by FEA. At the end of 2 months of classes on any level, students appear for an examination (components: reading, listening, speaking and writing). Students graduate to next grade level only if they clear these examinations. They are permitted a total of three attempts at the examination. If they

are still unable to attain a passing grade, they must repeat that grade level, and undertake classes for 2 more months. A student gets a graduation certificate from FEA on clearing all 5 grade levels.

2.2.2 Intervention Details

Classrooms were assigned to one of the following groups: Control Group, Treatment 1 (Classroom Game or CG) or Treatment 2 (Group game or GG). While all classrooms assigned to the Control Group continued as usual, classrooms assigned to the treatment groups were offered an explicit, symbolic reward for greater attendance. Classrooms assigned to CG were introduced to within-classroom competition based on students' attendance. Classrooms assigned to GG were first divided into four sub-groups formed most equitably (four groups of 5 students if 20 total students in a classroom; three groups of 5 students and one group of 4 students if 19 total students in a classroom; and so on[2]). The assignment of students to each of these four sub-groups was done randomly[3].

The competition continued for a duration of eight weeks where each student within the classroom scored points based on his class attendance during that period. The scoring rule for week 1 to 4 differed from week 5 to 8[4]. In week 1 to week 4, every attended class yielded 1 point and student was awarded additional 2 bonus points for each week when student attended all classes in that week. In week 5 to week 8, every attended class now yielded 2 points and student was awarded 4 bonus points for each week when student attended all classes in that week. Student was awarded 0 points for days when he was absent. At the end of eight week of classes, four highest scorers from the classroom were declared the winners

[2]We did not introduce GG game if the number of students in a classroom was lesser than 15 so as to have a decent number of students in every group. For consistency, we did not introduce CG as well if number of students was lesser than 15 in a classroom.

[3]Each student picked up one folded slip with group numbers written inside. Students from within the classroom picked up slips for students who were absent on the day of introduction of this game. The process was conducted with students by a territory manager for each region who were all trained by the researcher.

[4]Magnitude of decrease in student attendance and increase in dropout rate is usually found to be higher during the latter four weeks of the eight weeks of classes. The weight (points in the game) assigned to these days, therefore, was kept higher for these latter four weeks.

of the game from that classroom[5].

For classrooms allocated to CG, four winners were declared. These winners were to four students who scored the highest points. For the classrooms allocated to GG, one winner per small group (with one classroom consisting of four small groups) was announced. The difference between the two treatments stems from the experimental design – CG implies competition at the classroom level (approximately 20 students, with 4 winners), while GG implied within group competition, at the group level (a small group of about 5 people, with one winner). Across these, the total number of winners in the classroom stays the same (4 winners), but the number of people that a student is competing against is significantly different (approximately 19 in CG, approximately 4 in GG).

Thus, fundamentally, GG differed from CG based on the competition type, which was within-group (and not within-classroom as in CG). There were four winners in GG game-one from each small group. This budget constraint equality (4 certificates in a classroom) makes both CG and GG games comparable to each other, and to the control group which had no extrinsic motivator to attend more classes.

2.2.3 Sample Selection and Randomization

New classrooms at FEA start on a rolling basis, i.e., a new classroom starts whenever the previous classroom finishes with its grade 5 classes. For our project, we worked with students from 217 classrooms. The classrooms that were eligible to be part of our intervention were: (1) either new classrooms starting with Grade 1 or, (2) the existing classrooms that started afresh at grade 2, 3 or 4 level but were not part of this game during their previous grade level. At the beginning of every month, FEA provided us details of these eligible classrooms. On receipt of this information, we randomly assigned these classrooms to one of the following:

[5]An 8-layer tie breaker rule was used in case of a tie leading to more than 4 highest scorers and thus, winners in the game. The rule required considering only 8th week points to determine the winner only from these highest scorers. If 8th week's points fail to break the tie, then 7th week's points will determine the winner, and so on. If we failed to break the tie even until the 1st week of the game, then all the highest scores in the classroom were declared as winners.

(1) Control group, (2) Treatment 1 group (CG), or (3) Treatment 2 group (GG). While assignment, we stratified them by the city and school. At the end of their grade level, four winning students from each treatment classroom received a symbolic reward - a certificate and diary from FEA.

2.2.4 Data and Summary Statistics

We had a total of 217 classrooms that participated in our experiment with 3898 students. While we had attendance, student-related and household-related information on all these participating students, prior test scores (a standardized test on speaking, reading and writing conducted for all students before they join the academy) for 97 students could not be found in the administrative records[6]. Table 2.1 provides summary statistics of several student and household related variables separated by their treatment status. Most of the pre-intervention variables appear well-balanced across the groups, except the family-size variables where students allocated to T2 seem to come from larger families. This points towards a need to control for family size in the regression. Additionally, we find differences in student marital status, and other educational qualifications as well across the groups[7].

Our primary outcome variables for this study are student attendance and drop-out rate – two variables which are directly affected by our intervention. While we consider accumulated attendance over all weeks, i.e., week 1 to 8 as our main attendance outcome variable, we also pursue separate analysis for accumulated attendance over weeks 1 to 4 and over weeks 5 to 8. The interest in pursuing this partitioned analysis comes out of testing the effectiveness of the modified, weighted version of the standard attendance game. Against the expectation of increased likelihood of absenteeism in weeks 5 to 8 (from the institution's past experience), the game created stronger incentives and allowed students to score twice the points in weeks

[6]We do not include these 97 students when we present averages of past test scores but include them in our data analysis since the treatment status was randomly allotted to the classrooms they belonged

[7]On running both versions of the regression equation - with and without controls, it is observed that controlling for family size, other education, and marital status does not impact the size or significance of our estimates of interest. Thus, we present the results for a regression with school fixed effects and clustering standard error at classroom level. Results with control variables are presented in the appendix

5 to 8 as compared to weeks 1 to 4.

Table 2.2 below provides post intervention summary statistics of drop-out rates, accumulated attendance over all weeks 1 to 8, accumulated attendance over weeks 1 to 4, and accumulated attendance over weeks 5 to 8. In addition, we also provide summary statistics for week-wise attendance variables for each treatment group, a cursory glance at which informs us of the effectiveness of interventions in most school weeks. Another significant aspect to notice from Table 2.2 is the decrease of 6.66% in average attendance of control group from 74.80% in weeks 1-4 to 68.14% in weeks 5-8. This implies that, on average, a student attended 1.65 fewer school days in weeks 5-8, as compared weeks 1-4. This drop justifies our use of the weighted attendance game intended to provide stronger incentives to students during the latter half of the programme. However, both our intervention groups combined also observe a decrease of 6% in attendance from weeks 1-4 to weeks 5-8. Such similar figures indicate that despite the intentional design, our intervention may not have worked differentially between weeks 1-4 to weeks 5-8.

A preliminary evaluation of these summary attendance outcomes in Table 2.2 is suggestive of the effectiveness of both the treatment games. The exploratory analysis through the kernel density plot of post-intervention attendance shown in Figure 2.2 (also Figures 2.3 and 2.4) brings forth further evidence on the effectiveness of treatment games.

We observe that both the treatment groups led to higher attendance, and a movement to the highest quartile. In the treatment groups as against the control group, the number of students who attain an attendance that lies between 80% to 100% is visibly greater, while the number of students who attain attendance between 60% and 80% is lower. This indicates a potential movement of students from the 60% to 80% attendance bracket to the 80% to 100% attendance bracket, and thus, a positive impact of the intervention.

Additionally, the Kernel Density plots do not suggest any change in the attendance of the students who attain less than 60% attendance. These preliminary observations were verified by our regression results.

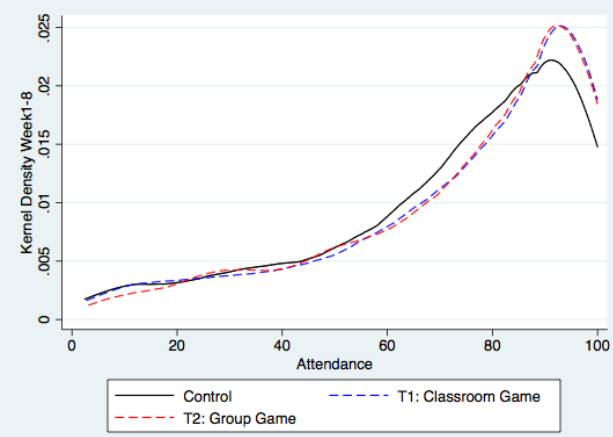


Figure 2.1: Kernel Density Plot: Week 1 to 8

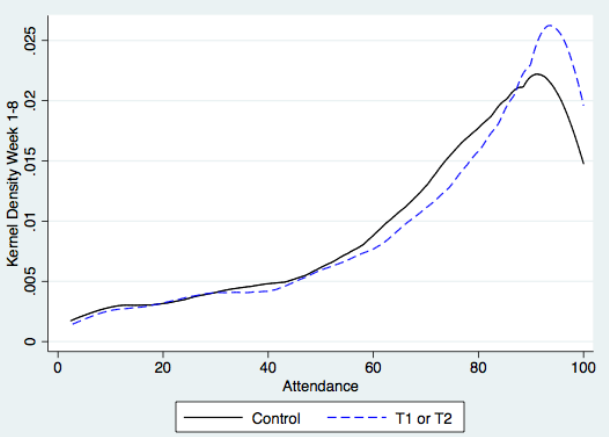


Figure 2.2: Kernel Density Plot: Week 1 to 8

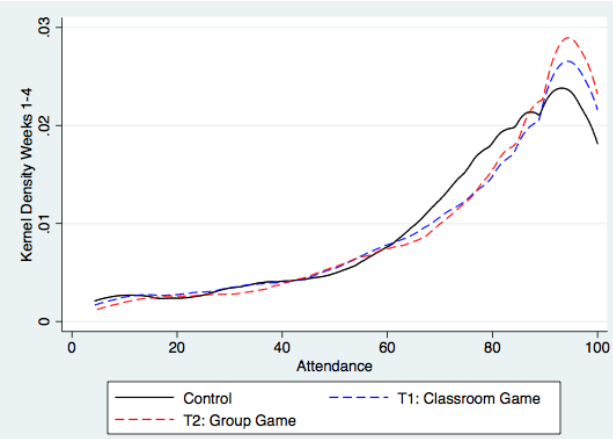


Figure 2.3: Kernel Density Plot: Week 1 to 4

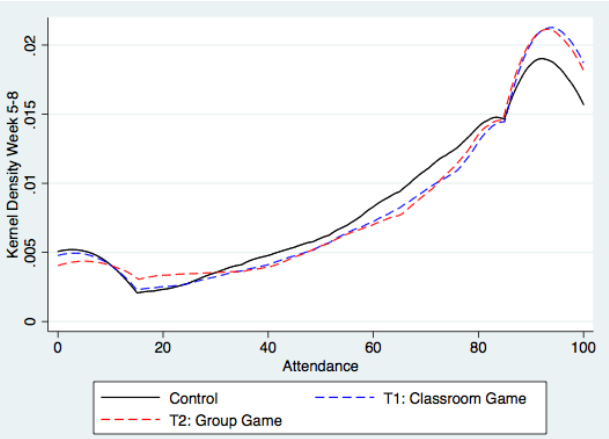


Figure 2.4: Kernel Density Plot: Week 5 to 8

Note: Week1-4 variable represents aggregate attendance from weeks 1 to 4. Week5-8 represent aggregate attendance from weeks 5 to 8. Week1-8 represents aggregate attendance over all weeks from week 1 to week 8.

2.2.5 Econometric Model for Analysis

The intervention used in this project allows us to answer two following important questions:

1. Can attendance be improved by developing simple games and using symbolic incentives

	(1) Control	(2) T1: Classroom Game	(3) T2: Group Game	(4) p-value	(5) T1 or T2	(6) p-value
Student-level variables						
Baseline Listening Score	33.20	32.89	32.16	0.38	32.56	0.32
	(19.70)	(19.25)	(18.75)		(19.02)	
Baseline Reading Score	52.56	52.51	52.35	0.95	52.44	0.84
	(19.97)	(18.14)	(18.95)		(18.51)	
Baseline Writing Score	47.49	48.37	47.21	0.45	47.84	0.67
	(24.98)	(24.72)	(24.81)		(24.76)	
Female	0.46	0.46	0.46	0.97	0.46	0.92
	(0.498)	(0.498)	(0.499)		(0.498)	
Age	19.12	19.09	19.01	0.73	19.06	0.60
	(3.777)	(3.260)	(3.112)		(3.193)	
Grade Level	1.93	1.89	1.85	0.27	1.87	0.17
	(1.183)	(1.154)	(1.111)		(1.135)	
Other Education	1.15	1.27	1.22	0.01	1.25	0.00
	(0.815)	(0.818)	(0.804)		(0.812)	
Employment Status	0.13	0.11	0.10	0.19	0.11	0.08
	(0.456)	(0.402)	(0.405)		(0.404)	
Married	0.03	0.01	0.02	0.01	0.02	0.00
	(0.181)	(0.121)	(0.136)		(0.128)	
Family-level variables						
Mother's Education	2.41	2.36	2.26	0.14	2.32	0.14
	(1.834)	(1.771)	(1.789)		(1.780)	
Father's Education	3.35	3.39	3.25	0.11	3.33	0.66
	(1.659)	(1.636)	(1.656)		(1.646)	
Family Size	3.98	4.06	4.58	0.00	4.30	0.01
	(3.665)	(3.562)	(3.575)		(3.577)	
N	1328	1405	1165		2570	

Note: Standard error in parenthesis. p-value in column 4 tests the null hypothesis that mean values in column 1, 2 and 3 are indifferent. p-value in column 6 tests the null hypothesis that values in column 1 and 5 are indifferent. Baseline Listening, Reading and Writing scores are scores on the standardized test that every student has to take before joining the institute. Female takes value 1 if girl and 0 if boy. Grade level takes value from 1 to 4 (we did not work with grade level 5 which is the highest grade level at the school). Other education represents their education level outside FEA. Student employment status takes 1 if unemployed; 2 if employed full time and 3 if employed part-time. Parental (mother's and father's) education is a categorical variable including options like no education, primary education, class 10, completed high school, graduate, etc. Family Size represents the number of members that live within their household.

Table 2.1: Pre-intervention Descriptive Statistics and Balance Check

	(1) Control	(2) T1: Classroom Game	(3) T2: Group Game	(4) p-value	(5) T1 or T2	(6) p-value
Week 1	79.28	79.48	81.82	0.03	80.54	0.17
	(27.26)	(27.76)	(25.34)		(26.71)	
Week 2	74.89	77.93	79.56	0.00	78.66	0.00
	(31.14)	(29.97)	(28.50)		(29.32)	
Week 3	73.60	73.72	73.48	0.98	73.61	0.99
	(31.56)	(33.07)	(33.24)		(33.14)	
Week 4	70.24	70.75	72.92	0.11	71.73	0.19
	(33.74)	(34.21)	(33.69)		(33.98)	
Week 5	69.49	72.22	72.10	0.07	72.16	0.02
	(34.89)	(34.29)	(34.75)		(34.49)	
Week 6	67.68	70.01	70.17	0.15	70.09	0.05
	(36.51)	(35.53)	(35.92)		(35.70)	
Week 7	67.85	71.14	69.56	0.24	70.37	0.13
	(37.79)	(36.39)	(36.07)		(36.22)	
Week 8	62.12	60.89	63.68	0.72	62.40	0.92
	(37.16)	(39.57)	(37.36)		(38.38)	
Week 1-4	74.80	75.86	77.45	0.02	76.58	0.03
	(24.23)	(24.72)	(23.42)		(24.15)	
Week 5-8	68.14	70.45	70.80	0.06	70.61	0.02
	(31.33)	(31.54)	(31.08)		(31.33)	
Week 1-8	71.86	73.81	74.48	0.02	74.12	0.01
	(24.80)	(25.27)	(24.36)		(24.86)	
Dropout	0.05	0.05	0.04	0.58	0.05	0.75
	(0.217)	(0.221)	(0.203)		(0.213)	
N	1328	1405	1165		2570	

Note: Standard error in parenthesis. p-value in column 4 tests the null hypothesis that mean values in column 1, 2 and 3 are indifferent. p-value in column 6 tests the null hypothesis that values in column 1 and 5 are indifferent. Week 1, 2, 3, 4, 5, 6, 7 and 8 represent the week-wise percentage attendance. Week1-4 variable represents aggregate attendance from weeks 1 to 4. Week5-8 represent aggregate attendance from weeks 5 to 8. Week1-8 represents aggregate attendance over all weeks from week 1 to week 8. Dropout represents the proportion of students dropped-out of the school.

Table 2.2: Post-intervention Summary Statistics

to extrinsically motivate students? This result can be obtained using a simple OLS regression measuring the change in average student attendance as we switch from control to treatment classrooms.

$$a_{ic} = \beta_0 + X'_{ic}\beta_1 + \beta_2 D_{ic} + \epsilon_{ic} \tag{2.1}$$

where,

$$D_{ic} = \begin{cases} 1, & \text{if student } i \text{ is from control classroom} \\ 0, & \text{if student } i \text{ is from one of the treatment classrooms} \end{cases}$$

a_{ic} is the attendance of student i studying in classroom c. X_{ic} is the set of student, parents or school related variables which are collected by the FEA administration. After clustering the standard error at the classroom level, the estimate of β_2 obtained will inform us of the causal effect that an extrinsic motivator to attend classes in the form of our attendance-game has on student attendance.

2. Does the impact of these games vary based on competing group size? This result can also be obtained using econometric model mentioned above. We will, however, consider the observations from only those treatment classrooms which we will be interested in studying. We can compare Control with T1 classrooms; Control with T2 classrooms and T1 with T2 classrooms. During the comparison of attendance for T1 with T2 classrooms, the estimate of the treatment dummy will inform us about whether a smaller group game has different effect on attendance than a game played within an entire classroom.

	(1) Attendance1-8	(2) Attendance1-8	(3) Attendance1-4	(4) Attendance5-8	(5) Dropout
T1 (CG)	1.60				
	(1.35)				
T2 (GG)	1.41				
	(1.51)				
T1 or T2		1.51	1.17	1.76	-0.003
		(1.22)	(1.11)	(1.65)	(0.01)
Constant	65.43	65.42	68.40	61.62	0.08
School F.E.	Yes	Yes	Yes	Yes	Yes
Observations	3710	3710	3710	3710	3710
R-square	0.09	0.09	0.08	0.09	0.02

Note: Standard error in parentheses and clustered at classroom level. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$. Classrooms were selected from 28 schools. Regression analysis is conducted over all students who participated and were not late admissions, i.e., who did not join the school after the classes and the intervention had already begun at least a week back.

Table 2.3: Regression - Attendance

	(1) Attendance1-8	(2) Attendance1-8	(3) Attendance1-4	(4) Attendance5-8	(5) Dropout
T1 (CG)	1.41***				
	(0.49)				
T2 (GG)	1.18**				
	(0.50)				
T1 or T2		1.30***	1.29***	1.70**	-0.006
		(0.43)	(0.46)	(0.83)	(0.006)
Constant	87.30	87.29	87.71	86.60	0.02
School F.E.	Yes	Yes	Yes	Yes	Yes
Observations	2215	2215	2215	2189	2215
R-square	0.05	0.05	0.04	0.03	0.02

Standard error in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Standard error in parentheses and clustered at classroom level. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$. Classrooms were selected from 28 schools. Regression analysis is conducted over all students who attended more than 75% of their classes and were not late admissions, i.e., those who did not join the school after the classes and the intervention had already begun at least a week back.

Table 2.4: Regression - Attendance if greater than 75%

	(1) Attendance1-8	(2) Attendance1-8	(3) Attendance1-4	(4) Attendance5-8	(5) Dropout
T1 (CG)	-0.65				
	(1.41)				
T2 (GG)	0.05				
	(1.43)				
T1 or T2		-0.32	-0.81	-0.67	0.009
		(1.22)	(1.27)	(2.08)	(0.02)
Constant	46.79	46.87	52.06	39.53	0.13
School F.E.	Yes	Yes	Yes	Yes	Yes
Observations	1495	1495	1495	1421	1495
R-square	0.03	0.03	0.03	0.04	0.05

Note: Standard error in parentheses and clustered at classroom level. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$. Classrooms were selected from 28 schools. Regression analysis is conducted over all students who attended lesser than 75% of their classes and were not late admissions, i.e., those who did not join the school after the classes and the intervention had already begun at least a week back.

Table 2.5: Regression - Attendance if lesser than 75%

2.3 Results

Table 2.3 presents the estimation results, controlling for school fixed effects and clustering the standard error at classroom level. We find that, on average, students in both the treatment groups attend higher percentage of classes than students in the control group. T1 leads to an average increase in attendance by 1.6%, while T2 does that by 1.41%. The estimates, however, are statistically insignificant.

Based on the preliminary suggestive results from the Kernel Density plots presented earlier, we run two further regressions that evaluate the effectiveness of our intervention on two sets of students differentially - students whose attendance is less than 75%, and those whose attendance is greater than 75%[8].

This bifurcation leads to striking results. When the model is used to ascertain the impact of the intervention on students whose attendance is greater than 75% (Table B.1), we get significant impact of the intervention. The introduction of the symbolic reward increases the

[8]From kernel density plot, the distinction appears around 80% attendance group. We ran the estimation at both 80% and 75%, the estimated effect size seems more distinct at 75% than at 80%, even though the effects at both cut-off are significant. Results at 80% cut-off presented in the appendix.

attendance, on average, by 1.3%[9]. This effect is stronger in weeks 5 to 8 (1.7%) as compared to weeks 1 to 4 (1.3%). This indicates that the weighting strategy adopted in the design of the competitive game worked well in providing greater incentive to attend classes in the latter half of the intervention period.

However, the intervention does not have any impact on students whose attendance is less than 75% as can be seen in Table B.2. The results presented in Table 2.3, B.1 and B.2 also do not show any significant change in the dropout rate when the intervention was introduced.

Thus, we find that the extrinsic motivation is most effective with students with relatively higher attendance. This could be because the effectiveness of the intervention, which is a competitive game, relies on the reward's ability to induce competition. One could argue the greater attendance leads to greater involvement, and thus, a stronger competitive spirit. This could lead to greater intervention effectiveness in students with higher attendance while not so much in students with lower attendance. This is also evident from no change observed in the drop-out rate since it is typically the students with lower attendance levels who dropout and this game, therefore, did not either influence their attendance nor did it influence their dropout rate.

Another potential reason for these results comes from the relatively small number of students whose attendance is less than 75% to begin with. This group had a total of 1405 observations, which in turn could have led to a low power for the regression we ran. On the contrary, the number of students in the group with attendance greater than 75% was over 50% greater, at 2295.

2.4 Conclusion

Education literature in the past has identified, and attempted to solve the access problem, and the learning problem. However, the same can not be said about attendance, or its lack thereof leading to gaps in learning. Ensuring more schools and learning spaces, running

[9]We are presenting results by pooling both treatment groups together to increase the statistical power

campaigns to encourage enrolment in schools, and legislating education as a fundamental right, are examples of initiatives that have led to increased student enrolment. However, there is sufficient evidence to indicate that higher enrolment does not necessarily lead to enhanced learning outcomes. Solutions to the learning problem like teacher training, greater school regulation, etc. are being explored. These measures presume a students ability to be motivated by the long term benefits of education. There is, however, an empirical case to be made for present-biased or myopic perspectives of young learners, making them unable to be driven by potential future gains, and thus, the effectiveness of these measures often falls short of expected results.

This study looks at more immediate, even if seemingly less consequential, rewards as a means to increase student attendance. It further evaluates if such immediate rewards produce significantly different results based on the size of the competing group, and confirms the effectiveness of such measures. Straightforward extrapolation of these results nudges policymakers in the direction of symbolic, but timely reward systems, and makes a case for their potential success. The behavioral changes brought by this low-cost symbolic reward offered through a competitive attendance based game can have vast policy implications[10].

The study also provides a template to tweak game designs in different contexts to ensure that the observed effectiveness is maintained. A meaningful next step could be an evaluation of real time symbolic rewards - what if a student was rewarded for their attendance after every single class, while ensuring additional rewards for consistency? Would this lead to even greater attendance?

Lastly, the study calls attention to a missing piece in education policy - attendance. It goes beyond solving the access problem and the learning problem, and tries to emphasise the need to include interventions that positively impact attendance in discussions in education policy in contemporary times.

[10]In the future version of this paper, we intend to include the effect of this intervention on students' academic performance, the data on which will soon be available.

Chapter 3

Government Effectiveness in the Provision of Public Goods: The Role of Institutional Quality

3.1 Introduction

High taxation and a large public sector can potentially distort choices and also lead to political corruption and rent seeking, thereby afflicting government's effectiveness in the delivery of public goods and services[1]. Higher taxes also incentivize firms to move their investments from the formal to the informal sector and thus, impeding economic growth. One of the most striking differences between the economies in advanced countries and in developing countries is in the role of the public sector, the former typically having a relatively large public sector, with a substantial commitment to public health, public education, infrastructure, and social security, whereas in developing countries these programs either do not exist, or do not entail broad population coverage[2]. Consequently, the tax burden is sub-

[1]Se e.g. Olken, 2006, and references therein for work that documents, using micro data, how taxation gives rise to corruption.

[2]For example, the average for central government spending as a share of the GDP between the years 1996-2000 was almost 40% in the high-income group of countries and less than 15% in the low-income group

stantially larger in developed than in developing countries.[3] Yet, despite the overall lighter tax burden in developing countries, there has been remarkably little, if at all, convergence in incomes with the developed world and scarce evidence that growth in the latter has been impeded by a large public sector (see Lindert, 2004, for historical analysis, and Easterly and Rebelo, 1993, for contemporaneous evidence).

One of this paper's goals is to reconcile these observations in light of the role that institutional quality plays in mitigating the detrimental effects of a large public sector and the consequential, high tax burden. We assume that law enforcement, bureaucratic efficiency (or political stability) and absence of political corruption constitute institutional quality for an economy. Public good provision with its corresponding tax burden, on the other hand, constitutes government effectiveness in an economy. It is argued by means of a simple model that, where the institutional quality is high, size of informal sector is smaller and taxation to finance public spending is much less detrimental than with a lax institution. This implies that the formal sector is bigger and optimal tax rates are higher, i.e., taxation is more affordable for an economy with better institutional quality, ceteris paribus. These results feed into the main theme of this paper which talks of improved provision of public goods due to better institutional quality. Adding this aspect of institutional quality to a relatively standard framework helps explain some of the empirical regularities related to public sector's effectiveness.

We then test some of the implications of the theoretical framework. The focus of our empirical analysis constitutes firm-level perceptions on the quality of public services (which parallels public good provision in our theoretical model) in general and in specific areas such as infrastructure, health, and education; and on the severity of the tax burden (which

of countries (authors' calculations based on the World Bank Development Reports).

[3]Thus, the share of the GDP collected in tax revenues in recent years was about 30% in high-income countries, but only some 10% in low-income countries data on which are available. A strong robust relationship between the GDP and tax revenues across countries can be easily discerned from glancing at the data with some high-income countries such as Belgium, Italy, and the Netherlands collecting almost 50% of the GDP in tax revenues, whereas many low-income countries collect 10 percent and even less (World Development Reports, recent years).

parallels the public good maximizing tax rate in our theoretical model). It suggests that, consistent with the model's implications, a better institutional quality reinforces the perceived effectiveness of the public sector and, therefore, lowers the perception of the tax burden as an obstacle to firms' business activity.

This paper is related to several literatures. One is the relatively small but evolving literature on the determinants and the growth effects of informality pioneered in De Soto, 2000, see also Loayza, 1996, and Sarte, 2000, for some analytical approaches. Friedman, Johnson, Kaufman, and Zoido-Lobaton, 2000, Johnson et al., 2000, and Dabla-Norris et al., 2008, provide evidence that enforcement quality is a more important determinant of informality than fiscal policies. More recent papers (Manolas et al., 2013, Remeikiene and Gaspareniene, 2015, Shahab et al., 2015, Bayar, 2016, and Goel and Nelson, 2016) also provide empirical evidence on the negative and significant role that institutional quality plays in determining the size of the shadow economy.

Other related work emphasizes the role of public investment in development. Barro, 1990, is a seminal contribution in this regard, which however disregards the informal sector in its model. There is also work on the determinants of the size and the capacity of the public sector, see Boix, 2001, that also contains a careful literature review. The more directly relevant literature on the effective capacity of the public sector is much more limited. La Porta et al., 1999, is the only contribution we are aware of in this regard, and we will comment on this paper more in detail below; our paper can be viewed as complementary to it in providing additional pieces of evidence on the determinants of government quality.

There is some recent work exploring the effect of specific institutional quality measures in various contexts. Desai et al., 2007, shows how the effect of corporate taxes is mediated via the quality of corporate governance. Rajkumar and Swaroop, 2008, shows that differences in the efficacy of public spending in health and education can be largely explained by corruption and bureaucratic inefficiency. Lledó and Poplawski-Ribeiro, 2013, investigates political and institutional constraints to fiscal policy implementation in Sub-Saharan Africa and finds

that planned fiscal adjustments or expansions are less likely to be implemented with weaker institutions framing. Hauner and Kyobe, 2010, finds that increased accountability of government institutions has an effect on improving efficiency from government expenditures on health and education. Abiad et al., 2016, shows that public investment inefficiency (such as poor project selection, implementation, and monitoring) affects the output growth in an economy. Our work can be viewed as an extension for a broader measure of institutional quality while focusing on more elements of public good expenditures, other than health and education; and the corresponding tax burden on firms.

We now proceed as follows. Section 2 presents the basic analytical framework, followed by the empirical analysis of some of the theoretical implications in Section 3, and Section 4 concludes with brief remarks.

3.2 Conceptual framework

Our theoretical analysis models the interaction between government and firm of an economy, where the former decides about how much tax to charge the firm which is then used towards the provision of public goods, while the latter responds to government's tax choice by choosing the proportion of investment to be hidden in the informal economy. In this standard framework, we introduce two parameters, one for bureaucratic inefficiency and political corruption and another for law enforcement, both representing institutional quality.

The comparative statics analysis predicts that the detrimental effect of high taxation on informality is weakened in the presence of higher institutional quality, ceteris paribus. This result seems to be well consistent with various recent findings. While early work found that tax burden and government regulations lead to a larger informal sector (see Schneider and Enste, 2000), more recent research suggests that when institutional variables are included in the regression specification, they trump the tax and regulation variables (Chong and Gradstein, 2007). Further, using firm-level data, Friedman et al., 2000, and Johnson et al.,

2000, in their analysis of transition economies find that firms' trust in the rule of law explains their tendency to go informal much better than measures of the tax burden. Dabla-Norris et al., 2008, using firm-level data, find that, while both taxes and regulations tend to be associated with higher levels of informality, the rule of law emerges as its dominant predictor. Regression analysis indicates that the adverse effect of taxes in this regard is moderated by a high level of the rule of law as perceived by the firms, which is again consistent with our analytical findings; it also indicates that stronger rule of law is associated with more efficient government, which in turn also decreases the propensity to go informal.

The main focus of our theoretical model and our paper, however, is on how institutional environment is associated with the provision of public good and corresponding optimal tax rate (or perception of tax as an obstacle). Some preliminary insights here may be derived from La Porta et al., 1999, which exhibits highly significant correlations across countries between measures of institutional quality such as the political rights index on one hand and measures of the size of the public sector (the fraction of the labor force employed in the public sector) and its outcomes (such as in health, education, and infrastructure) on the other hand.[4] Their cross-country regressions also reveal that institutional proxies are associated with the size of the public sector. These empirical findings are in alignment with the predictions of our theoretical model below.

3.2.1 The model

The illustrative framework presented is relatively standard. For the simplicity of our analysis, we consider an economy populated by only one economic agent, the firm, which has to make an investment, k. The production out of this investment will be subject to a statutory tax at the rate of T, decided by government. The firm can, however, evade paying their tax dues by hiding their endowment or by moving their activity into the informal sector. Thus, we assume that the production out of a declared part of investment, $1- h$, is taxed at

[4]For example, the correlation of the political rights index with the infant mortality variable is -.57; with school attainment is .67; with the infrastructure index is .67

the rate of T, and the proceeds are used by government to provide the public good. The complementary part, h, is hidden from the tax authority and shifted to the informal sector.

In case of an audit, however, the agent is subject to a penalty. It is assumed that the penalty results in a net loss. This is presumably because of the outlays to cover the costs of monitoring and auditing, which increase the probability of detection of informal activities. These aspects are not explicitly modeled here as our interest is more with the implications of this interaction between the state and the agent rather than its microeconomic foundations.

Without specifying the details of the auditing procedure, we let $P(h;\ \phi) = \frac{\phi h^2}{2}$ denote the penalty – as a fraction of investment - imposed on an agent hiding h, where $0<\phi<1$ is interpreted as the law enforcement quality.[5] The seminal paper by Allingham and Sandmo (1972) and the subsequent work provide useful framework for microeconomic analyses of tax evasion and auditing; this literature enables an endogenous derivation of the penalty and the evasion activity. As our interest here is less with these aspects and more with their macroeconomic implications, a reduced form specification as above is adopted. The share of hidden resources hk is interpreted as the size of the informal sector.

Our model also assumes that there is rent seeking behavior in the economy in the form of bureaucratic inefficiency (or policy instability) and political corruption to the extent of parameter.[6] This is analogous to another tax that a firm has to pay on its production out of the declared investment in the formal economy. We assume $0<\gamma<1$. While some rent seeking would be present in almost every economy, McGuire and Olson (1996) explains why it would not be optimal for a government (even an autocrat) to expropriate all the income generated in the economy as tax or rent. Different countries will have different rent seeking behavior depending on their internal political and economic dynamics, however, it is usually expected to be higher in developing countries. We consider only one period decision making in our model where we don't derive the optimal for an economy. We assume to be a constant

[5]The particular quadratic formulation is mainly for tractability purposes.

[6]Correspondingly, $1-\gamma$ is a measure of "political stability and absence of political corruption" used in the empirical analysis.

and known to firms and government at the beginning of the period before they choose h and T, respectively. Measures of law enforcement, ϕ, and bureaucratic efficiency (or political instability) and absence of political corruption, $1-\gamma$, determine the institutional quality in our economy.

We assume firm's production function (or the generated income) in this economy to be:

$$z = (1-T-\gamma)(1-h)kA + (h - \phi\frac{h^2}{2})k \tag{3.1}$$

In this function, $A > 1$, is a parameter representing government investment in infrastructure that complements private investment, $(1-h)k$, made in formal economy by firm. Although parameter A would increase in magnitude with greater tax collection and public good investment by government, it is assumed to be constant (by firms) in any given period. Parameter A is expected to be higher for developed countries. In a given period, $(1-h)kA$ represents the income generated out of firm's investment in formal sector and $(1-T-\gamma)(1-h)kA$ represents the share that firm gets to keep after paying tax T and rent. Additionally, we assume that money invested in informal sector, $(h_{\mathrm{i}} - \phi h^2/2)k$, neither increases nor decreases in monetary value by the end of the period. This implies that informal sector does not get to enjoy the complementarities with government investment in infrastructure, given by parameter A. For simplicity, we assume firm consumes all its earnings from formal and informal sector at the end of this one-period model leaving nothing behind.

We assume that the government budget is balanced in each period, i.e., government spending on public goods equals its tax collection.

$$G = T(1-h)kA \tag{3.2}$$

For simplicity, we assume rent, kA, does not add either to production in the economy or to public good spending.

In this period, the government acting as a welfare maximizer, selects a tax rate, upon which the firm makes its decision, determining the fraction of unreported income or the size of informal economy.[7] In equilibrium, these are mutually consistent.

3.2.2 Analysis

The government's end goal in this model is to maximize public good spending G which is a function of variables h and T. Since government is aware of the tendency of firm to react to a higher tax, T, by hiding a greater share, h, of investment in informal economy, they would find an optimal tax keeping firm's response function in mind.

The firm's decision about optimal h as a function of T is determined as below:

$$max_h(1-T-\gamma)(1-h)kA + (h-\phi\frac{h^2}{2})k$$

Maximizing this expression gives:

$$h^*\{T|\phi,\gamma\,A\} = min\{max\{0,(\frac{1}{\phi})[1-(1-T-\gamma)A]\},1\} \tag{3.3}$$

Given $A > 1$ and $0 < \phi < 1$, if both $T = 0$ and $\gamma = 0$ hold for an economy, we can check that $h^* = 0$. Intuitively, if firm faces no tax or rent seeking behavior, it will have no incentive to resort to informal economy. And if $T + \gamma = 1$, as would be expected of an autocrat when it is the "end of the world" period for him, $h^* = 1$. Intuitively, if firm has to pay all its income as tax or rent, it will have no incentive to invest in the formal economy. This suggests that, for any given period, a firm's investment share, h, in informal economy increases with tax T for given γ, ϕ and A.

Let's assume $(1-T-\gamma)A < 1$, this yields an interior solution such that:

[7] A previous version also contained analysis of a political equilibrium, whereby the majority voting determines the tax rate; the analysis yields similar insights.

$$h^*\{T|\phi,\gamma\,A\} = (\frac{1}{\phi})[1-(1-T-\gamma)A] \tag{3.4}$$

We can see that size of informal economy is an increasing function of the tax rate ($\frac{\partial h^*}{\partial T} = \frac{A}{\phi} > 0$) , more so when enforcement quality is lax ($\frac{\partial^2 h^*}{\partial T \partial \phi} = -\frac{A}{\phi^2} < 0$) ; a decreasing function of both enforcement quality ($\frac{\partial h^*}{\partial \phi} = -\left(\frac{1}{\phi^2}\right)[1-(1-T-)\,A] < 0$) and political stability and absence of corruption ($\frac{\partial h^*}{\partial(1-\gamma)} = -\frac{A}{\phi} < 0$). All these results are intuitively appealing wherein one would expect a developing country to have a lower ϕ and $1-\gamma$, i.e., lower institutional quality, both causing a higher h, in comparison to a developed country. This also suggests that for two identical countries except for institutional quality, the country with better institutional quality has a bigger formal sector (Remeikiene and Gaspareniene, 2015; Shahab et al., 2015; Bayar, 2016; Goel and Nelson, 2016).

Given a firm's choice of informality as a function of tax, h^*, government will find an optimal tax rate, T^* :

$$\max_{\mathrm{T}} \; G = T\,(1-h^*)\,kA$$

$$\text{s.t. } h* = \left(\tfrac{1}{\phi}\right)[1-(1-T-)\,A]$$

Consider $\frac{\partial G}{\partial \phi} = -TkA\frac{\partial h^*}{\partial \phi} = \left(\frac{TkA}{\phi^2}\right)[1-(1-T-\gamma)\,A] > 0$. This inequality is obtained using our previous assumption $(1-T-\gamma)\,A < 1$ for an interior $h^* solution. Additionally,\ consider \frac{\partial G}{\partial(1-\gamma)} = -TkA\frac{\partial h^*}{\partial(1-\gamma)} = \left(\frac{TkA^2}{\phi}\right) > 0$. This shows that enforcement quality, and bureaucratic efficiency and absence of political corruption enhances the public good provision. In other words, higher institutional quality bolsters public good provision, ceteris paribus. These results make intuitive sense, since for developing countries where we generally observe lower law enforcement, ϕ, and higher rent seeking behavior, γ, we usually find lower public good provision in comparison to developed countries.

On substituting for $h^* from (4) in the public good function G, we get$:

$$\underset{T}{max} G = aT - bT^2$$

where $a = \left(\frac{kA}{\phi}\right)[\phi - 1 + (1-\gamma)A]$

and $b = \left(\frac{kA}{\phi^2}\right)$.

This reveals that the relationship between tax rate, T, and public good, G, is a non-monotonic one, increasing initially and decreasing afterwards. This is not surprising as, when the tax rate is high, the agent reacts by hiding a larger portion of the bequeathed resources, generating a decreasing portion of the Laffer curve.

Solving the above expression with respect to tax, T, yields the optimal tax rate as:

$$T^* = \frac{(\phi - 1) + (1-\gamma)A}{2A} \tag{3.5}$$

This public good maximizing tax rate, T^*, is an increasing function of enforcement quality ($\frac{\partial T^*}{\partial \phi} = 1/2A > 0$), and of bureaucratic efficiency and absence of political corruption ($\frac{\partial T^*}{\partial (1-)} = \frac{1}{2} > 0$). This shows that the public good maximizing tax rate increases as law enforcement, or as bureaucratic efficiency and absence of political corruption, $1-\gamma$, improve. In other words, higher institutional quality mediates some of the effects that higher taxes have on firms, ceteris paribus, consequently making them appear as less of an obstacle to investing in formal economy.

Collecting the results, we obtain:

Proposition 1. The effect of taxation on informality ($\frac{\partial h^*}{\partial T} > 0$) works through enforcement quality and is stronger when the latter is lax ($\frac{\partial^2 h^*}{\partial T \partial \phi} < 0$). Also, informality is reduced with bureaucratic efficiency and absence of political corruption ($\frac{\partial h^*}{\partial (1-\gamma)} < 0$) .

Proposition 2. Better enforcement quality implies a higher public good maximizing tax rate ($\frac{\partial T^*}{\partial \phi} > 0$); bureaucratic efficiency and absence of political corruption has the same effect

on public good maximizing tax ($\frac{\partial T^*}{\partial(1-\gamma)} > 0$). This suggests that public good maximizing tax rate increases in institutional quality or, in other words, the perception of tax as an obstacle decreases in institutional quality.

Proposition 3. Public good provision increases in institutional quality due to a smaller optimal size of informal economy, h^*, and a larger optimal tax rate, T^* . Given optimal firm behavior, h^*, public good provision is a non-monotonic function of the tax rate, increasing first and then decreasing ($G = aT - bT^2$) .

Proposition 2 and 3 together imply that government effectiveness improves with institutional quality, where government effectiveness is given by public good provision and the corresponding tax as a burden on firms, and institutional quality is given by measures of law enforcement, bureaucratic efficiency and absence of political corruption.

It must be noted that the static model considered above was simplified with assumptions of one firm; exogenously determined enforcement rate, ϕ, and rent, γ; convex penalty function $P(h;\, \phi) = \phi h^2/2$; and a constant complementarity measure, A, between firm's formal sector investment and government investment on infrastructure. In this one-period context where firm decides, h, and government decides, T, at the beginning of the period, it makes sense to assume constant ϕ, γ and A which would be the values firm and government perceive to be present in the economy. However, in the real world multi-period and many firms setting, one would expect ϕ to be an increasing function of government's tax collection; rent, γ, to depend on the long-run motive of the social planner or the political system of the economy; penalty function, $P(h;\phi)$, to be linear rather than convex; and complementarity parameter, A, to be an increasing function of government's spending on intermediate investment goods. While these parameter values will keep varying between periods in the general multi-period and many firms model in response to changes in government tax revenues and subsequent government spending, its analysis will give similar results as our one-period and one firm

model.

The static model provides an important insight into the role that institutional quality may have played in the remarkably little convergence in incomes of developing countries with the developed world (Lindert, 2004). Government's tax revenue is majorly used towards the provision of final consumption goods (such as health and education) and intermediate investment goods (such as infrastructure), where former adds directly to an economy's GDP while latter affects GDP through its positive impact on the complementarity parameter, A, making private investments in formal sector more productive. Better institutional quality (higher ϕ and $1-\gamma$), therefore, increases GDP through increased government spending (on consumption and investment goods) out of higher tax collection, and through increased production and productivity in the formal sector. Improved institutional quality reduces the size of not so productive informal sector. In other words, better institutional quality increases both private and public sector production, thus, contributing to the income gap between nations with varying institutional quality, ceteris paribus.

3.3 Empirical evidence

Our theoretical analysis generates several implications. Proposition 1 talks about how taxation affects informality through the intermediation of institutional quality ($\frac{\partial^2 h^*}{\partial T \partial \phi} < 0$) . There is overwhelming evidence in favor of this result (Friedman et al., 2000; Johnson et al., 2000; Chong and Gradstein, 2007; Dabla-Norris et al., 2008). The focus of our paper, on the other hand, is on providing more disaggregated evidence to further enhance the preliminary insights derived from La Porta et al., 1999, consistent with the Proposition 3 of our model, namely, that better institutional environment is associated with better functioning public sector. The dataset generated through the World Business Environment Survey (WBES) by the World Bank allows us to provide such evidence. In addition, we also use this dataset to test our Proposition 2 in the model which suggests that better institutional environment

reduces the perception of tax as an obstacle to growth. We now proceed by describing this dataset.

3.3.1 Data and empirical strategy

The survey was taken as an initiative of the World Bank Group, in partnership with many other institutions seeking to provide feedback from enterprises on the state of the private sector in client countries; to measure the quality of governance and public services including the extent of corruption; to provide better information on constraints to private sector growth, from the enterprise perspective; to establish the basis for internationally comparable indicators which can track changes in the business environment over time thus allowing both for competitive assessment and impact assessments of market-oriented reforms; and to stimulate systematic public-private dialogue on business perceptions and the agenda for reform. The field work was done between 1999 and 2000 by private polling of each country's firms that fulfilled the basic requirements. The survey was targeted to a representative sample of firms filling criteria as sector, size, location, and ownership characteristics[8]. The objective was to gather information on a sizeable number of firms in several countries around the world, which was accomplished for most of the sample.[9]

The sample consists of firm level survey responses of thousands of firms in more than eighty countries, many of them developing and in transition. The survey asked each business to rank the constraints or problems impacting their operations. This process involved an

[8]The particular requirements that had to be filled by the sample selected were as follows. Sector: In each country, the sectoral composition in terms of Manufacturing (including agro-processing) versus Services (including commerce) will be determined by relative contribution to GDP, subject to a 15% minimum for each category. Size: At least 15% of the sample shall be in the small and 15% in the large size categories. Ownership: At least 15% of the firms will have foreign control. Exporters: At least 15% of firms will be exporters, meaning that some significant share of their output is exported. Location: At least 15% of firms will be in the category "small city or countryside" .

[9]The countries and number of firms (in parentheses) included in the survey are: Argentina (76), Bangladesh (38), Belarus (101), Bolivia (72), Brazil (148), Bulgaria (84), Canada (87), Chile (80), Colombia (88), Costa Rica (51), Czech Republic (81), Dominican Republic (68), Ecuador (52), El Salvador (63), France (72), Germany (75), Guatemala (51), Haiti (71), Honduras (50), Hungary (102), India (123), Indonesia (70), Italy (67), Malaysia (43), Mexico (43), Nicaragua (62), Pakistan (72), Panama (49), Peru (77), Philippines (90), Poland (175) , Portugal (78), Romania (114), Slovakia (23), Spain (82), Sweden (76), Thailand (71), Turkey (113), United Kingdom (59), United States (86), Ukraine (158), and Uruguay (57).

extensive questionnaire undertaken via a face-to-face interview with either the firm managers or firm owners of each company. As a result, the survey reports comparative measurements based on firms' perceptions about their business environment as shaped by a variety of economic and policy factors[10].

For testing the theoretical model's implications about public goods provision in Proposition 3, we use answers to questions regarding the quality of public services such as infrastructure, health and education, security, etc., and the efficiency of the government on delivering those services as proxies. A corresponding World Bank question in the survey is as follows: "how would you generally rate the efficiency of central and local government in delivering services?" with responses ranging from "1=very inefficient" to "6=very efficient". Also, as proxies to our main explanatory variable of interest, "institutional quality", we use answers to questions related to firm's perception of the quality of the judicial system and its functioning, as well as the main institutional stimulants for firm's growth, such as policy stability and absence of political corruption[11]. To test Proposition 2, we use answers to questions related to taxes and their regulation as obstacles posed to business' growth as proxies to the optimal tax or the tax burden.

Additionally, we also include country wide variables, in particular, institutional quality, the logarithm of the GDP, and the tax rate. The former is taken from International Country Risk Guide (2006), a well-known comprehensive index including the assessment of corruption within the political system, the strength and impartiality of the judicial system, the assessment of the popular observance of laws; and the institutional strength and quality of the bureaucracy. This index is taken as an average for the period 1998 and 2002, in order to assess the long term quality of the institutional framework. As for the tax rate and the GDP, we use the VAT rate as of August 2004 which is taken from the International Monetary Fund

[10]In recent years, several researchers have employed these data, including Misch et al., (2014), Hallward-Driemier and Pritchett (2015), among many others.

[11]These institutional measures are highly correlated with other standard institutional measures employed in the literature, such as BERI, ICRG, and the measures originally collected by by Kaufmann et al (1999) A summary of such measures can be found at this World Bank site: http://info.worldbank.org/governance/wgi/# home

	Obs	Mean	S.D	Min	Max
Firm's characteristics					
Company is owned by a foreign investor	9673	0.19	0.39	0	1
Government owns the company	9645	0.12	0.33	0	1
Size: Medium	10007	0.40	0.49	0	1
Size: Large	10007	0.19	0.39	0	1
Manufacturing	9141	0.36	0.48	0	1
Service	9141	0.43	0.50	0	1
Agriculture	9141	0.07	0.26	0	1
Construction	9141	0.10	0.29	0	1
Firm's perception about institutional quality					
Political stability	9034	2.21	1.08	1	4
Absence of corruption	8376	2.47	1.15	1	4
Confidence in judicial system	9539	3.76	1.43	1	6
Courts-enforceability	8902	3.42	1.47	1	6
Courts-consistent	8614	3.13	1.41	1	6
Courts-affordable	8875	3.18	1.46	1	6
Courts-quick	9067	2.35	1.28	1	6
Courts-honest	8814	3.35	1.50	1	6
Courts–fair & impartial	9012	3.44	1.44	1	6
Firm's Perception about Quality of public services					
Efficiency of government in delivering services	7786	3.16	1.20	1	6
Quality of education	8874	3.59	1.27	1	6
Quality of public health	9227	3.23	1.35	1	6
Quality of water	9390	4.00	1.29	1	6
Quality of power	9485	4.11	1.28	1	6
Quality of telephones	9518	4.17	1.24	1	6
Quality of public works	9035	3.35	1.36	1	6
Country level institutional quality					
Quality of Institutions index	8935	8.55	2.78	0	15.88
Log(GDP)	10032	24.14	1.98	20.32	29.79
Taxes					
General constraint-taxes and regulations	9382	2.86	1.01	1	4
Current VAT rate	9467	16.20	4.63	5	25

Table 3.1: Summary Statistics

(2006), and from the World Development Indicators, respectively. Finally, as basic controls, we base our specification on existing literature and, in particular, include basic firm characteristics, such as ownership, size, and industrial sector. Table in the appendix provides detailed definitions of all the variables used in this paper, and Table 1 provides corresponding summary statistics.

3.3.2 Specification and results

Our analysis concentrates on testing some of the implications of the theoretical model above[12]. Table 2 presents our benchmark specification for determinants of government effectiveness in delivering public services. As our dependent variable is categorical, we run ordered probit regression and show the coefficients obtained.[13]

We find that, on average, government-owned firms perceive the government as relatively efficient in delivering public services. Also, the size of the firm is positively linked to the perception of the effectiveness of the government. In contrast, we do not find any significant relationship between the sector where the firm operates and the opinion on the efficiency of the government. Also, we do not find any robust evidence that size of the economy, as measured by the gross domestic product, is associated with the perception of government effectiveness in public goods provision.

Consistent with the model's predictions and similar to previous country level evidence (La Porta, et. al., 1999), we find a significant association between the quality of institutions and the efficiency in provision of public services at the firm level. Furthermore, in order

[12] In particular, we do not provide empirical results on the link between taxes and government efficiency as in this specific case endogeneity issues can be particularly problematic. When applying an IV approach similar to the one used in the paper we find results consistent with the predictions of the model. Also, La Porta et al. (1999) provide some empirical tests on this link at the country level.

[13] Since our GDP term is not statistically significant, we also tested the same specification with (i) one and two-lagged terms of GDP, (ii) a quadratic term in GDP and (iii) Interactive terms of GDP and other controls. In all cases, the single GDP term remains statistically insignificant. One must bear in mind that while the coefficients obtained from ordered probit cannot be interpreted directly, as we need to calculate marginal coefficients, the significance and sign of such coefficients are normally reported. We provide marginal coefficients for benchmark results in the appendix. We would be happy to provide the additional marginal calculations upon request.

Dependent variable: Efficiency of government in delivering services (1=very inefficient, 6=very efficient)			
Company is owned by a foreign investor	0.033	0.032	0.002
	(0.70)	(0.68)	(0.040)
Government owns the company	0.136**	0.094	0.102
	(2.13)	(1.39)	(1.60)
Size: Medium	0.107**	0.084	0.078
	(2.00)	(1.45)	(1.42)
Size: Large	0.207***	0.181***	0.118*
	(3.48)	(2.80)	(1.94)
Manufacturing	-0.148	-0.127	-0.156
	(-0.63)	(-0.51)	(-0.65)
Service	-0.149	-0.150	-0.150
	(-0.63)	(-0.61)	(-0.62)
Agriculture	-0.287	-0.321	-0.270
	(-1.15)	(-1.22)	(-1.04)
Construction	-0.249	-0.209	-0.245
	(-1.06)	(-0.83)	(-1.00)
Log(GDP)	0.008	-0.006	0.005
	(0.26)	(-0.19)	(0.15)
Quality of Institutions Index	0.039*	0.048**	0.040*
	(1.88)	(1.99)	(1.88)
Political stability	0.215***		
	(6.96)		
Absence of corruption		0.162***	
		(5.48)	
Confidence in judicial System			0.277***
			(11.8)
Observations	6039	5721	6107
Num. of countries	55	55	55
Log pseudo likelihood	-9264	-8827	-9138
Pseudo R-sq	0.0294	0.0247	0.0526
Chi-sq	216.4	132.7	284.9

Robust z-statistics in parentheses. Standard errors clustered at the country level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.2: Institutional quality and public services (ordered probit)

to exploit the between and within country variation that our data allows, we include both country-level and firm-level variables that takes into account the quality of institutions. As described above, at the domestic level, we use the institutional quality index from ICRG (2006) and at the firm level, we use question on perceptions of institutions as growth obstacles, in particular, those related with policy stability, absence of corruption, and the overall assessment of quality of the judiciary. The evidence presented in Table 2 shows that there is a highly significant association between the quality of institutions and the effectiveness of government in providing public services. Particularly, our results show that firm perception of a lesser corrupt political system, a more stable and predictable policy environment, and a more reliable judiciary implies firm perception of a more efficient government in delivering public services[14]. This result concurs with Proposition 3 of our model.

According to these findings, an increase of one standard deviation in the quality of institutions index - equivalent to moving from the institutional quality level of Mexico (7.8) to the one in Spain (11.7) - is associated with a 0.3 percent increase in the probability of ranking the performance of government as "very efficient". Similarly, at the firm-level, an improvement in political stability represented by a move from a response that policy instability poses a "minor obstacle to growth" to "no obstacle at all" is associated with an increase of about 0.7 percent in the probability of ranking the government as "very effective"[15].

Further, we present evidence on firm's perception of the tax rates and regulation as obstacles for growth as determined by firm characteristics, overall institutional quality, current tax rates, and the quality of public goods provided by the government, see Table 3. As expected, higher tax rates, measured by the value added tax (VAT) rate, are positively related

[14]When we add Barro and Lee's measure of education (years of secondary school) and a political rights measure (Freedom House) the statistical significance and signs of our (i) index of Quality of Institutions, (ii) the General constraint-political stability variable, (iii) the General constraint- absence of corruption variable, (iv) and the Confidence in judicial System variable do not change. However, since the number of observations is reduced drastically in relation to our core results (to 3500 observations approximately) we do not report these findings but they are available upon request.

[15]Table in the appendix shows the marginal coefficients of our variables of interest based on our benchmark regression on the first column of Table 2.

General constraint-taxes and regulations (1=no obstacle and 4=major obstacle)							
Quality of Institutions Index	-0.065	-0.070	-0.065	-0.068	-0.076	-0.068	-0.072
	(3.87)***	(3.81)***	(3.30)***	(3.48)***	(4.09)***	(3.84)***	(4.19)***
Current VAT rate	0.044	0.049	0.047	0.046	0.049	0.045	0.048
	(2.95)***	(3.37)***	(2.99)***	(2.98)***	(3.20)***	(3.11)***	(3.46)***
Quality of education							-0.110
							(5.27)***
Quality of public health						-0.115	
						(5.46)***	
Quality of water					-0.048		
					(1.88)*		
Quality of power				-0.059			
				(2.18)**			
Quality of telephones			-0.089				
			(3.21)***				
Quality of postal system		-0.075					
		(3.54)***					
Quality of public works	-0.110						
	(5.57)***						
Observations	6604	6733	6760	6782	6803	6436	6349
Num. of countries	70.00	70.00	69.00	69.00	70.00	70.00	70.00
Log pseudolikelihood	-8083.57	-8278.03	-8386.86	-8435.30	-8387.58	-7857.42	-7758.01
Pseudo R-sq	0.06	0.05	0.05	0.05	0.05	0.06	0.06
Chi-sq	319.03	313.35	240.28	245.66	267.76	367.13	327.24

Same controls as in Table 2. Robust z-statistics in parentheses. Standard errors clustered at the country level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.3: Taxation as an obstacle (ordered probits)

with the perception of taxes as an obstacle for growth while the quality of public services are negatively related with it. Most specific to our interest is the result that the institutional quality index is negatively associated with the perception of taxes as an obstacle for growth [16]. This result concurs with Proposition 2 of our model.

3.3.3 Robustness

Table 4 presents some further evidence on the impact of institutional quality on the quality of public goods serviced by government. We use various dependent variables that capture quality in the delivery of public goods, in particular, education, public health, water service, electric power, postal system, and the overall quality of public works. We find that there is a robust, positive, and statistically significant link between the measures of institutional quality and the quality of government services.

Since the survey does a detailed coverage of firms' perceptions of the legal system, it further enables us to do the analyses of its various features, such as its speed, fairness and impartiality, enforceability and others as the determinants of government effectiveness in public good provision. As can be seen in Table 5, each of the aspects of legal system is positively related to the government perception as an effective provider of services. As the country level institutional proxy remains highly significant and with a positive effect, also does our measures of the effectiveness of the courts.

To address the potential bias generated by endogeneity in the perceptions data, we employ an instrumental variables approach. In particular, we use a two-stage procedure that includes both country and firm level instruments for our two variables of interest, namely, our index of quality of institutions and our political stability variable. In the case of the former, a country-level variable, we use continental dummies and legal origin as country-level instruments. As has been shown in the literature (e.g., La Porta et al., 1997, 1999) legal origin is a very strong determinant of the current institutional quality of a country. Furthermore, it is reasonable to assume that the legal origin of a country may be minimally related to the effectiveness

	Edu.	Pub Health	Water	Power	Tele.	Pub Works
Quality of Institutions index	0.052	0.075	0.090	0.083	0.074	0.041
	(1.82)*	(2.51)**	(5.52)***	(3.64)***	(3.46)***	(1.78)*
Confidence in judicial system	0.174	0.173	0.136	0.152	0.144	0.128
	(9.85)***	(9.08)***	(7.97)***	(8.47)***	(7.88)***	(7.49)***
Observations	6786	7055	7169	7206	7222	7052
Pseudo R-sq	0.03	0.04	0.05	0.05	0.04	0.02
Quality of Institutions index	0.048	0.072	0.087	0.080	0.077	0.042
	(1.65)∗	(2.32)**	(5.81)***	(3.69)***	(3.63)***	(1.82)∗
Absence of corruption	0.152	0.159	0.104	0.104	0.057	0.048
	(6.04)***	(5.81)***	(4.13)***	(3.48)***	(2.40)**	(1.84)∗
Observations	6442	6689	6817	6842	6861	6714
Pseudo R-sq	0.02	0.03	0.04	0.04	0.03	0.01
Quality of Institutions index	0.054	0.074	0.085	0.080	0.075	0.034
	(1.76)∗	(2.25)**	(5.70)***	(3.73)***	(3.45)***	(1.44)
Political stability	0.129	0.142	0.123	0.120	0.080	0.105
	(4.66)***	(4.84)***	(5.59)***	(4.97)***	(3.11)***	(4.05)***
Observations	6214	6451	6562	6577	6591	6465
Pseudo R-sq	0.02	0.03	0.04	0.04	0.03	0.02

Robust z-statistics in parentheses. Standard errors clustered at the country level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.4: Institutional quality and public services. Robustness checks

Efficiency of government in delivering services (1=very inefficient 6=very efficient)						
Company is owned by a foreign investor	0.014	0.006	0.039	0.023	0.017	0.005
	(0.36)	(0.17)	(0.98)	(0.52)	(0.44)	(0.13)
Government owns the company	0.089	0.099	0.057	0.076	0.061	0.086
	(1.44)	(1.53)	(0.86)	(1.10)	(0.96)	(1.28)
Size: Medium	0.050	0.045	0.093	0.059	0.063	0.080
	(0.99)	(0.84)	(1.93)*	(1.09)	(1.33)	(1.42)
Size: Large	0.124	0.129	0.184	0.136	0.142	0.157
	(1.99)**	(1.97)**	(2.92)***	(1.94)*	(2.37)**	(2.29)**
Manufacturing	0.010	0.082	-0.065	-0.060	0.032	-0.048
	(0.04)	(0.30)	(0.29)	(0.19)	(0.11)	(0.16)
Service	0.010	0.089	-0.079	-0.063	0.046	-0.056
	(0.04)	(0.33)	(0.35)	(0.20)	(0.16)	(0.19)
Agriculture	-0.177	-0.057	-0.318	-0.244	-0.122	-0.200
	(0.64)	(0.20)	(1.33)	(0.72)	(0.40)	(0.63)
Construction	-0.097	-0.012	-0.211	-0.185	-0.063	-0.158
	(0.37)	(0.04)	(0.93)	(0.58)	(0.21)	(0.52)
Log(GDP)	-0.002	-0.003	0.009	0.007	-0.002	0.001
	(0.07)	(0.10)	(0.28)	(0.21)	(0.06)	(0.03)
Quality of Institutions index	0.044	0.044	0.059	0.071	0.049	0.055
	(2.20)**	(2.17)**	(3.10)***	(2.96)***	(2.37)**	(2.44)**
Courts-enforceability						0.154
						(6.13)***
Courts-consistent					0.231	
					(9.40)***	
Courts-affordable				0.128		
				(5.31)***		
Courts-quick			0.279			
			(9.01)***			
Courts-honest		0.199				
		(8.50)***				
Courts–fair & impartial	0.219					
	(9.17)***					
Observations	5949	5814	5997	5882	5897	5886
Num. of countries	55.00	55.00	55.00	55.00	55.00	55.00
Log pseudolikelihood	-8993	-8812	-8986	-9013	-8881	-8987
Pseudo R-sq	0.04	0.04	0.05	0.03	0.04	0.03
Chi-sq	205.98	185.19	183.17	128.36	214.04	152.07

Robust z-statistics in parentheses. Standard errors clustered at the country level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.5: Institutional quality, courts, and public services (ordered probits)

Efficiency of government in delivering services (1=very inefficient 6=very efficient)	
Company is owned by a foreign investor	-0.055
	(-0.97)
Government owns the company	-0.000
	(-0.00091)
Size: Medium	0.167***
	(2.83)
Size: Large	0.244***
	(3.60)
Manufacturing	-0.245
	(-1.60)
Service	-0.306*
	(-1.83)
Agriculture	-0.085
	(-0.42)
Construction	-0.348**
	(-2.15)
Log(GDP)	-0.048
	(-1.04)
Quality of Institutions index	0.082**
	(2.02)
General constraint-political stability	0.987***
	(2.83)
Observations	5343
Log pseudo likelihood	-8213.6397
Pseudo R-sq	0.0139
Chi-sq	97.23

Table 3.6: Institutional Quality and Public Services: Ordered probits with instrumental variables (Benchmark regression)

of the government in their delivery of public services as, unlike the overall quality of the institutions of a country, it is more likely that effectiveness in the delivery of services may be determined by short-run conditions rather than those that originated in the legal framework of the country some time ago (La Porta et al., 1999)[17].

In the case of Political stability, a firm-level variable, we also use the ownership and legal organization of the firm obtained from the WBES dataset. The first instrument reflects whether the owner is public or private, and if the latter then whether it is an individual, a family and whether or not it has supervisory board members. We believe that this is a good instrument because political stability may be directly correlated with the behavior of a firm given the potential influence of the State[18]. In fact, even non-State companies may be subject to political stability via influence of board members with specific interests, or direct links between top management and government officials.[19] On the other hand, the other instrument, legal organization of the firm, reflects whether it is formed as a partnership, a cooperative, or a privately-held corporation, and is the analogous to legal origin at the country level. The manner in which the firm is legally organized may be prone to having more links with the political system. It is believed that some types of arrangements may better shield for such external influence (Sokolov and Solanko, 2016).

Table in the appendix provides detailed definitions of these variables. It is reasonable to expect that such firm-level instruments may have an impact on our firm-level variables of interest; it seems also unlikely that such variables have any bearing on the perception of the quality of provision of public services. The results are shown in Table 6[20]. Overall, we find that institutions quality index at country level when instrumented by legal origin of the country and political stability at firm level when instrumented by ownership and legal

[17]In fact, the pairwise correlation between legal origin and provision of public services is below 0.15 and it is not statistically significant at conventional levels.

[18]A current example would be the case of the State oil company PDVSA in Venezuela.

[19]An example is Fisman (2001) who shows how the stock value of several firms changed dramatically once the dictator Suharto died.

[20]We also instrumented the other regressions obtaining similar results. For space reasons we do not present these results, but will be happy to provide them upon request.

organization of the firm still yield significant effects on the quality of public goods provided by the government.

3.4 Conclusion

This paper's starting point is the observation that neither the size of government nor the tax burden in themselves seem to impede economic performance in a cross section of countries. It then provides a theoretical model whereby the effect of taxes is mediated through institutional quality of the economy. The results then indicate that the optimal tax rate, hence the size of the public sector, increases with the institutional quality.

We then test these results using firm level data that contain information about satisfaction with public services and the extent to which taxation is viewed as an obstacle to growth. It turns out that institutional quality affects both: the better it is, the better public services are perceived and the less detrimental taxation seems to be. All this lends support to the, analytically derived and commonly observed across countries, positive association between institutional quality and government effectiveness.

An important direction for additional work would deal with the endogenization of institutional quality, possibly by studying how it interacts with the determination of the tax rate. Another, empirical direction would be an examination of the effect of both on firms' growth. We plan to address these aspects in future research.

Appendix A

Appendix Chapter 1

A.1 Theorems and Proofs

Theorem 1: Under idiosyncratic error assumption and risk-neutrality, $\mu_P^{N*} = \mu_I^{N*}$. In other words, optimal student effort under NGS will be same whether measurement is perfect or imperfect.

Proof 1: Under idiosyncratic error assumption, we know that:

$$\int_{\epsilon}(\underline{q}_a + \mu + \epsilon)g(\epsilon)d\epsilon = \underline{q}_a + \mu \tag{A.1}$$

Under perfect measurement, student's optimal effort is:

$$\mu_P^{N*} = argmax_{\mu}\ [W_a^N(\underline{q}_a + \mu) - C_a(\mu)]$$

Under imperfect measurement, student's optimal effort is:

$$\mu_I^{N*} = argmax_{\mu}\ [\int_{\epsilon}[W_a^N(\underline{q}_a + \mu + \epsilon)g(\epsilon)]d\epsilon - C_a(\mu)]$$

Consider a risk-neutral Expected Utility maximizing student, i.e., one for whom $W_a^N(.)$ is linear in test scores. From Expected Utility Theory, we know:

$$[\int_{\epsilon} W_a^N(\underline{q}_a + \mu + \epsilon)g(\epsilon)d\epsilon = W_a^N(\int_{\epsilon}(\underline{q}_a + \mu + \epsilon)g(\epsilon)d\epsilon)]_I = [W_a^N(\underline{q}_a + \mu)]_P \tag{A.2}$$

where I and P subscripts indicate unbiased imprecision and precision in measurement,respectively, and the latter equality follows from (A.1).

Thus, any unbiased imprecision in measurement does not change the reward function of an expected utility maximizing risk-neutral student from the one under perfect measurement scenario. Therefore, given an unchanged cost function under the perfect and imperfect measurement cases, optimal efforts will be identical under the two.

Theorem 2: Under idiosyncratic error assumption and risk-aversion, $\mu_P^{N*} > \mu_I^{N*}$. In other words, optimal student effort under NGS will decrease when measurement is imperfect. (For a risk-loving student, optimal effort under NGS will increase when measurement is imperfect.)

Proof 2:

Consider a risk-averse expected utility maximizing student, i.e., one for whom reward function $W_a^N(.)$ is concave in test scores. From Expected Utility Theory, we know:

$$[\int_{\epsilon} W_a^N(\underline{q}_a + \mu + \epsilon)g(\epsilon)d\epsilon < W_a^N(\int_{\epsilon}(\underline{q}_a + \mu + \epsilon)g(\epsilon)d\epsilon)]_I = [W_a^N(\underline{q}_a + \mu)]_P$$

where the latter equality follows from (A.1). This equation can be re-written as:

$$[\int_{\epsilon} W_a^N(\underline{q}_a + \mu + \epsilon)g(\epsilon)d\epsilon]_I < [W_a^N(\underline{q}_a + \mu)]_P$$

This implies that each effort level under imperfect measurement gives lesser expected reward than when same effort is exerted under perfect measurement case (see Figure 1.2). Therefore, imprecision pivots the expected rewards curve downward while keeping the cost curve same. Therefore, given an unchanged cost function under the perfect and imperfect measurement cases and using standard marginal analysis, it is trivial to see that such a pivot will lead to

a lower optimal effort level under imperfect measurement case.

For a risk-loving student ($W_a^N(.)$ is convex in test scores), however, imperfect measurement will pivot the expected rewards curve upward, still staying convex though. Following the same line of argument as under the case of a risk-averse student, it is trivial to see that imperfect measurement increases effort level, however, I cannot find that using marginal conditions due to both cost and expected reward functions being convex.

Corollary 2: Effort level under NGS is a decreasing function of risk-aversion, ceteris paribus.

Proof Corollary 2: With an increase in risk aversion of a given student, expected rewards curve will pivot down further while cost function does not change. This will lead to the choice of a lower optimal effort level.

Theorem 3: Under ideosyncratic error assumption, ${\mu_P^L}^* \leq {\mu_I^L}^*$. In other words, optimal student effort under LGS with imperfect measurement of effort will be at least as much as that with perfect measurement of effort.

Theorem 4: Under assumption A1, optimal effort levels across students in LGS system will bunch right above the lower threshold, and at or right below the higher threshold, corresponding to their optimizing Precautionary or Anticipatory efforts, respectively.

Proof 3 & 4: Suppose student with ability a scores $q_{it} < \underline{q}_a$ with probability P_1 and rewards $W_a^L(a-1)$; $q_{it} > \bar{q}_a$ with probability P_2 and rewards $W_a^L(a+1)$; and $\underline{q}_a < q_{it} < \bar{q}_a$ with probability $1 - P_1 - P_2$ and and rewards $W_a^L(a)$.

From assumption A1, $W_a^L(a-1) = W_a^N(\bar{q}_{a-1})$; $W_a^L(a) = W_a^N(\bar{q}_a)$; and $W_a^L(a+1) = W_a^N(\bar{q}_{a+1})$.

Also, $P_1 = Pr(q_{it} < \underline{q}_a) = Pr(\underline{q}_a + \mu + \epsilon < \underline{q}_a) = Pr(\epsilon < -\mu) = G(-\mu)$. Similarly, $P_2 = Pr(q_{it} \geq \bar{q}_a) = 1 - G(\bar{q}_a - \underline{q}_a - \mu) = 1 - G(\alpha - \mu)$, where α is a constant exogenous to the student.

The expected rewards function of student with ability a under LGS will be:

$$EB_a^L(\mu) = P_1 W_a^L(a-1) + (1-P_1-P_2)W_a^L(a) + P_2 W_a^L(a+1)$$

$$= \text{G(-}\mu)W_a^N(\bar{q}_{a-1}) + [G(\alpha-\mu) - G(-\mu)]W_a^N(\bar{q}_a) + [1-G(\alpha-\mu)]W_a^N(\bar{q}_{a+1})$$

Therefore,

$$EB_a^L(\mu) = W_a^N(\bar{q}_{a+1}) - G(\alpha-\mu)[W_a^N(\bar{q}_{a+1}) - W_a^N(\bar{q}_a)] - G(-\mu)[W_a^N(\bar{q}_a) - W_a^N(\bar{q}_{a-1})] \quad \text{(A.3)}$$

First order condition:

$$\frac{\partial EB_a^L(\mu)}{\partial \mu} = \underbrace{g(\alpha-\mu)}_{>0}\underbrace{[W_a^N(\bar{q}_{a+1}) - W_a^N(\bar{q}_a)]}_{>0} + \underbrace{g(-\mu)}_{>0}\underbrace{[W_a^N(\bar{q}_a) - W_a^N(\bar{q}_{a-1})]}_{>0} > 0$$

Second order condition:

$$\frac{\partial^2 EB_a^L(\mu)}{\partial \mu^2} = -\underbrace{g'(\alpha-\mu)}_{A}\underbrace{[W_a^N(\bar{q}_{a+1}) - W_a^N(\bar{q}_a)]}_{>0} - \underbrace{g'(-\mu)}_{B}\underbrace{[W_a^N(\bar{q}_a) - W_a^N(\bar{q}_{a-1})]}_{>0}$$

For $\mu \in [0, \alpha]$, term $-A$ is positive and term $-B$ is negative. As effort μ increases from 0, $-A$ decreases initially and then increases and $-B$ increases initially and then decreases. This opposite movement of $-A$ and $-B$ causes function $\frac{\partial^2 EB_a^L(\mu)}{\partial \mu^2}$ to keep switching signs as effort increases. From the shape of normal density function $g(.)$, we can extrapolate the curvature of expected rewards function which is increasing in effort, concave above the lower threshold and convex below the upper thresholds, and fairly flat in between the thresholds (see Figure A.2).

Due to missing concavity of the expected rewards function, we cannot use marginal conditions to find the effort level that optimizes expected rewards net of cost of effort:

$$E(W_a|\mu) = \int_\epsilon [W_a^L(f(\mu+\epsilon))g(\epsilon)]d\epsilon - C_a(\mu) \quad \text{(A.4)}$$

However, it is fairly evident from the explored functional form of expected rewards curve

and the typically assumed convex cost curve (and as showed in the Figure A.4) that the optimal effort level will be either a precautionary effort right above the lower threshold or an anticipatory effort right at or below the upper threshold, depending on the curvature of the cost curve.

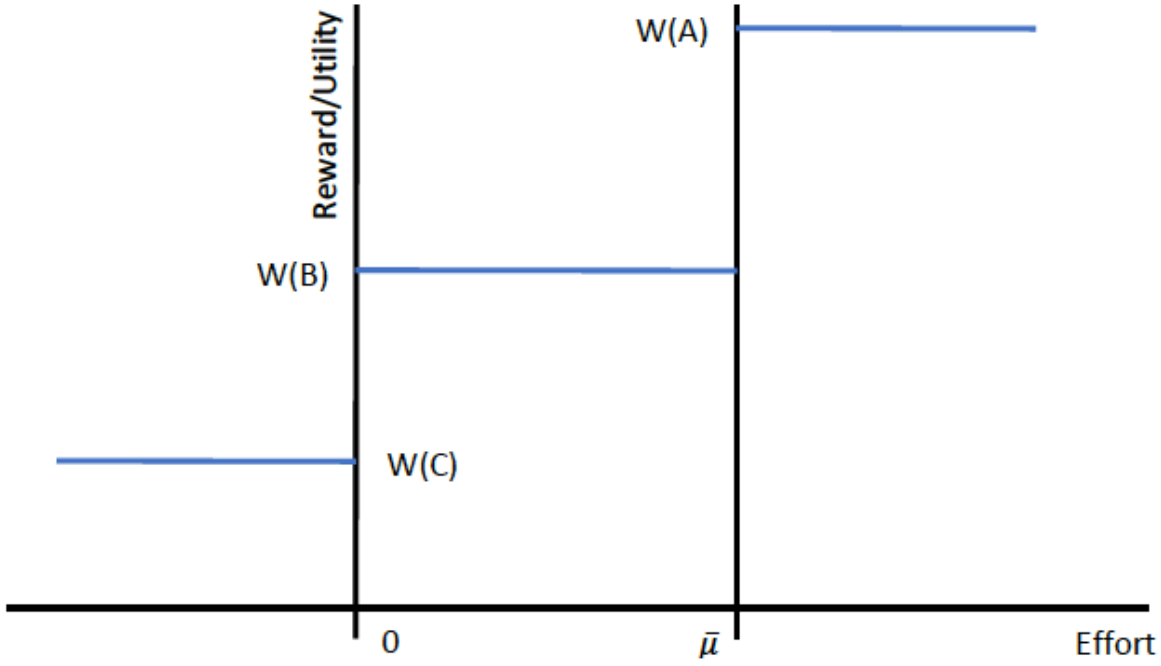
In either cases, the effort level will be greater than 0, which is the optimal effort under perfect measurement. Therefore, $0 = \mu_P^{L^*} \leq \mu_I^{L^*}$.

This result is robust to risk-attitudes of a student since changes in risk-attitude will only change $W_a^N(\bar{q}_{a-1}), W_a^N(\bar{q}_a)$ and $W_a^N(\bar{q}_{a+1})$ values. In (A.3), this only affects the constant terms, $W_a^N(\bar{q}_{a+1}), [W_a^N(\bar{q}_{a+1}) - W_a^N(\bar{q}_a)]$ and $[W_a^N(\bar{q}_a) - W_a^N(\bar{q}_{a-1})]$. Such a change does not change the curvature of Expected Rewards curve as a function of effort μ under LGS. Thus, while this may cause a difference between the optimal efforts exerted by risk-neutral and risk-averse students of ability a, it won't change our general results given in Theorem 3, i.e., $\mu_P^{L^*} \leq \mu_I^{L^*}$, and in Theorem 4 of precautionary and anticipatory efforts.

A.2 Figures

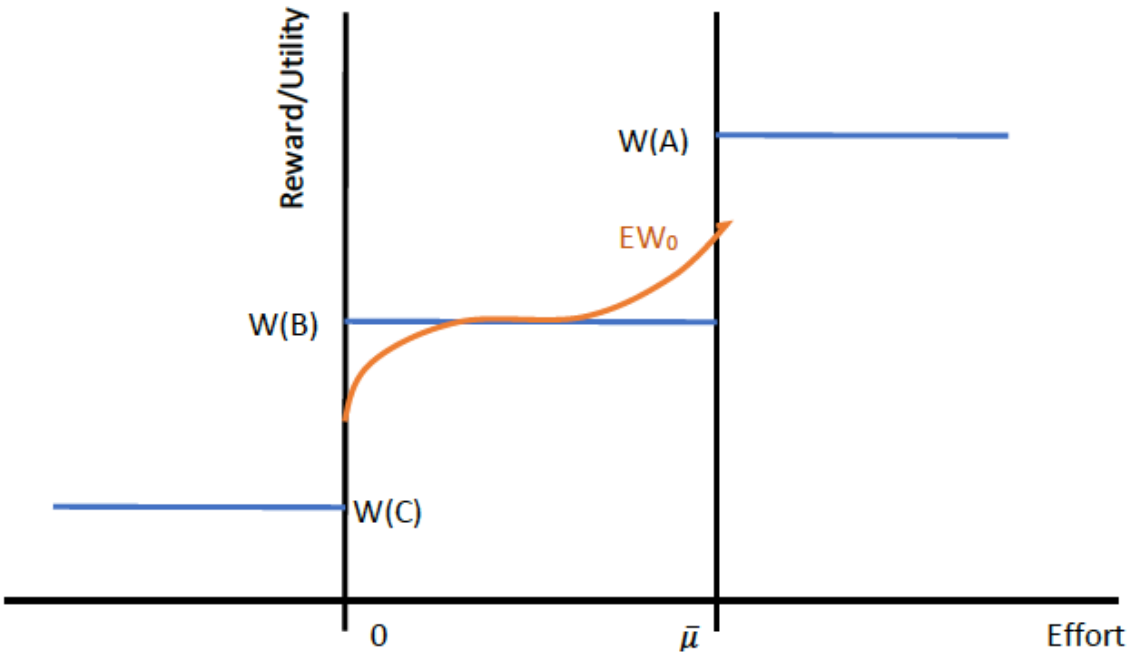
Figures

Figure A.1: Reward function under LGS



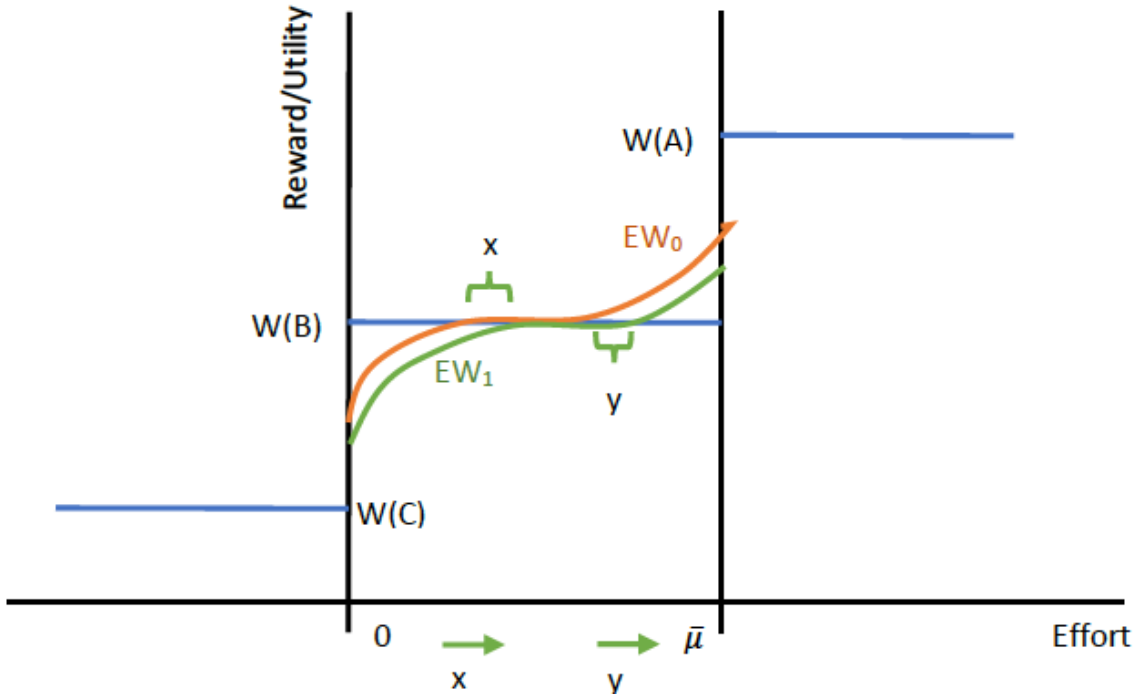
Rewards is equivalent to utility in our theoretical framework. This figure represents the student whose effort level cannot go below 0 and above $\bar{\mu}$, i.e., $\mu \in [0, \bar{\mu}]$. In a perfect measurement world of our theoretical model, this student will get reward W(B) for any exerted effort level. In an imperfect measurement world, W(A) and W(C) represent the possible rewards corresponding to letter grades A and C, respectively.

Figure A.2: Expected Reward function under LGS



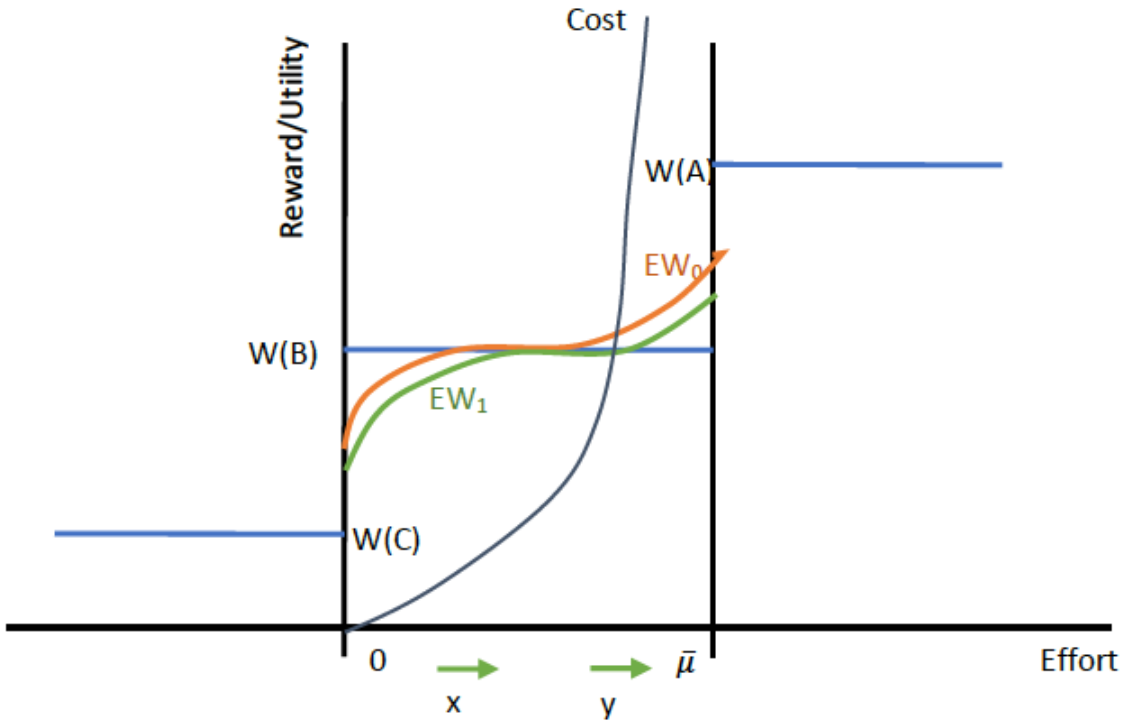
Rewards is equivalent to utility and expected rewards is equivalent to expected utility in our theoretical framework. In an imperfect measurement world, EW_0 represents the expected rewards to a student who will otherwise receive W(B) in a perfect measurement world, for all his efforts, $\mu \in [0, \bar{\mu}]$. Under assumption that $\epsilon_{it} \sim N(0, \sigma^2)$, EW_0 increases as effort increases from 0, minimizing the possibility of dropping to letter grade C. As effort increases further, the expected rewards EW_0 become steady around W(B) and as effort approaches $\bar{\mu}$, the expected rewards EW_0 rise again in expectation of letter grade A. The EW_0 curve corresponding to effort below 0 and above $\bar{\mu}$ is never realized since $\mu \in [0, \bar{\mu}]$.

Figure A.3: Expected Reward functions with varying risk-attitudes under LGS



Rewards is equivalent to utility and expected rewards is equivalent to expected utility in our theoretical framework. In an imperfect measurement world, EW_0 represents the expected rewards to a student who will otherwise receive W(B) in a perfect measurement world for all his efforts, $\mu \in [0, \bar{\mu}]$. The expected rewards curve shifts to EW_1 as risk-aversion of this student increases, everything else staying constant. x and y represent the risk-premium to be paid in terms of extra effort because of increased risk-aversion. x represents the extra effort required to minimize the possibility of letter grade C and corresponding reward W(C). y represents the extra effort required to start anticipating letter grade A and corresponding reward W(A). EW_0 and EW_1 curves corresponding to effort below 0 and above $\bar{\mu}$ are never realized since $\mu \in [0, \bar{\mu}]$.

Figure A.4: Optimal effort level under LGS



Rewards is equivalent to utility and expected rewards is equivalent to expected utility in our theoretical framework. EW_0 and EW_1 are the expected rewards functions of two identical students, except in risk-attitudes. EW_1 represents a more risk-averse student than EW_0. Cost function is convex in effort. For the shown cost function in this figure, we can see that with EW_0, student may (or may not) exert anticipatory effort close to $\bar{\mu}$ but with EW_1, student will definitely not exert anticipatory effort and will exert a precautionary effort closer to 0. This figure shows that increasing risk-aversion raises the possibility of exerting precautionary effort closer to 0 and diminishes the possibility of anticipatory effort closer to $\bar{\mu}$, among otherwise identical students.

A.3 Grading History

A.3.1 Grading history of the (US) K-12 education system

In the early days of 19th century when the number of schools or students were low, grading system was about teachers' visit to students' homes presenting an oral progress report with little standardization of content which later switched to a written narrative description of student performance (Guskey & Bailey, 2001). The increasing complexity, diversity and demand of education in late 19th and early 20th century could not keep up with the earlier subjective, time-consuming and cost-ineffective method of narrative description (Farr, 2000). This resulted in the movement towards a 0-100 grading system in schools.

Early 20th century observed a discussion about the imprecise measurement of student performance across instructors and schools. Starch (1913) showed that three major factors, (a) differences due to the pure inability to distinguish between closely allied degrees of merit", (b) "differences among the standards of different teachers", and (c) "differences in the relative values placed by different teachers upon various elements in a paper, including content and form" together produced an average probable measurement error of 5.4 on a 100-point scale across instructors. He did not find significant differences among the standards of different schools. Several other studies found this between-teacher error in measurement of student performance (Ashbaugh, 1924; Brimi, 2011; Eells, 1930; Healy, 1935; Silberstein, 1922). This piling evidence about the variability in grades across teachers led to the movement of grading system away from a 100-point scale and towards use of a 9-point scale (A+, A, B+, B, C+, C-, D+, D and F) initially and then to the current more commonly used 5-point (A-F) scale. By 1940s, more than 80% of U.S. schools had adopted the 5-point A-F grading scale.

Until the mid 20th century, grading in most schools, especially the high schools, used to be relative (or Norm-referenced) to ensure student ranking for college admissions. However, in 1950s and 1960s, a debate had started around the ideology of education with learning being the main objective and not comparison with peers (Crooks (1933), A. Z. Smith and

Dobbin (1960)). This led to the later movement in schools towards an absolute grading system with grades being based on student's own mastery on the subject matter [1].

A.3.2 Grading history of the (US) college education system

The history of grading systems in the US starting 18th century has been very experimental in nature. Grading systems ranging from descriptive adjectives to various numerical systems to several new scales of merit and demerit were tested during this period. Mary Lovett Smallwood (1935) gives a detailed history of evolution of grading in American university system. She reports that Yale was the first to use marking or grading in 18th century to differentiate students learning. Quoting from Ezra Stiles' (American academic, educator, seventh president of Yale College, and one of the founders of Brown University) 1785 diary footnotes, she mentions that the scale was made up of descriptive adjectives.

The college of William and Mary used scales of No. 1. (for the first in their respective classes), No. 2. (for Orderly, correct, and attentive), No. 3. (for the ones who made very little improvement), and No. 4. (for those who learnt little or nothing) in 1817. Harvard in 1830 used a 20-point scale, and mathematical and philosophical professors at Harvard started using 100-point scale in 1837. The University of Michigan used the numerical system in early 19th century, pass-no pass system in 1851, plus sign for passing student in 1852, grade "condit" (abbreviation for conditional) in addition to the plus sign in 1860s and then moving back to the 100 scale numerical system in late 19th century. However, it was Mount Holyoke College in 1898 that adopted the letter grading system (A (95-100); B (90-94); C (85-89); D (80-84); E (75-79); F (Failed)) that is closest in spirit to the one known today. Despite this increasing usage of 5-point A-F scale, grading systems differed vastly across US universities in 19th century.

At the beginning of 20th century, akin to similar movements in K-12, a consensus had developed among educators and researchers about the unjustified precision of the 100-point

[1] Reference: Brookhart et al, 2016

scale. This caused a major shift towards the normal curved A-F grading system (Relative or Norm referenced grading) or to similar other systems (Rugg, 1918). Meyer (1908) proposed the following grade categories along the normal curve of grades: Excellent (3% of students), Superior (22%), Medium (50%), Inferior (22%), and Failure (3%). The normal curve further became the underlying criterion through 1960s for the grading systems (majorly A-F) adopted by various American universities in alignment with the developing notion that grades are a method of ranking students.

In 1960s, people criticized grading systems for grading students' performance against the performance of their peers and proposed grading students' own level of mastery on the subject matter being assessed (Glaser (1963), Bloom (1971)). In other words, the proponents of Absolute (also called Criterion Referenced) grading systems had become more vocal during this time period, thus, making people to rethink about the proper aims behind an education system. While the discussions in the education arena towards this switch were still going on, it was the Vietnam war of 1960s that instantly pushed this movement from relative to absolute letter grading system. American need for higher numbers of eligible people for the draft created pressure on professors to not to fail students, thereby, dramatically increasing the proportions of A and B grades and decreasing that of F (Rojstaczer & Healy, 2012).

Thus, grading systems have evolved from grading students on the numerical 100-point scale (NGS hereafter) to grading on the norm-referenced 5-point A-F scale (Relative LGS hereafter) to eventually grading on the criterion-referenced 5-point A-F scale (Absolute LGS hereafter). While several other countries (Canada, Kenya, Hong Kong, South Korea, United Kingdom, Sweden, etc.) follow the same absolute 5-point letter grading system as the US, there still are countries (Costa Rica, Nicaragua, Chile, Venezuela, India, Pakistan, China, Israel, Indonesia, Poland, etc.) that have stood by the old 100-point numerical grading system. Through this historical journey, the move between different grading regimes was motivated by teacher-related factors and lesser so by the factors concerning the students. One such important factor that was missed out from that entire discussion about the evolution

of grading systems was the incentive embedded in each of those grading systems.

In recent years, economists and educators have caught up to that missed incentive story on two (Relative LGS and Absolute LGS) of these three grading systems and have explored (both theoretically and empirically) the possible differences that could exist between these two systems. This strand of research comparing Relative with Absolute grading answers the question of "How to evaluate student performance?", where Relative grading evaluates performance of each student in comparison to his peers while Absolute grading evaluates performance of each student based on his own mastery at the subject matter. This comparison between Relative and Absolute LGS is of most interest to institutions of higher education, i.e., universities where this debate between relative and absolute LGS still continues and several professors still choose between Absolute and Relative LGS systems for their respective courses. However, response to this question of how to evaluate student performance (Relative LGS vs Absolute LGS) is not much relevant at K-12 educational institutions. Instead, the debate at K-12 schools has always revolved around the choice between NGS and Absolute LGS. These are both absolute systems of grading where they both evaluate student performance on his mastery at the subject matter. These two grading systems differ only in how they report student performance (fine 100-point scale in NGS and coarse 5-point A-F scale in Absolute LGS), with both systems judging each student independently of others. There is currently no existing research, to the best of my knowledge, on "How to report student performance - using NGS or using Absolute LGS?".

Appendix B

Appendix Chapter 2

B.1 Additional tables

	(1) Attendance1-8	(2) Attendance1-8	(3) Attendance1-4	(4) Attendance5-8	(5) Dropout
T1 (CG)	1.09***				
	(0.42)				
T2 (GG)	0.75*				
	(0.44)				
T1 or T2		0.94**	1.05**	0.94**	-0.009
		(0.37)	(0.42)	(0.71)	(0.007)
Constant	90.77	90.75	91.11	86.60	0.02
School F.E.	Yes	Yes	Yes	Yes	Yes
Observations	1958	1958	1958	1935	1958
R-square	0.04	0.03	0.03	0.02	0.02

Standard error in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Standard error in parentheses and clustered at classroom level. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$. Classrooms were selected from 28 schools. Regression analysis is conducted over all students who attended more than 80% of their classes and were not late admissions, i.e., those who did not join the school after the classes and the intervention had already begun at least a week back.

Table B.1: Regression - Attendance if greater than 80%

	(1) Attendance1-8	(2) Attendance1-8	(3) Attendance1-4	(4) Attendance5-8	(5) Dropout
T1 (CG)	-0.72				
	(0.59)				
T2 (GG)	0.81				
	(0.59)				
T1 or T2		-0.77	-1.20	-0.95	0.01
		(1.21)	(1.26)	(1.87)	(0.02)
Constant	52.24	52.23	56.61	46.38	0.11
School F.E.	Yes	Yes	Yes	Yes	Yes
Observations	1760	1760	1760	1682	1760
R-square	0.04	0.04	0.05	0.04	0.04

Note: Standard error in parentheses and clustered at classroom level. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$. Classrooms were selected from 28 schools. Regression analysis is conducted over all students who attended lesser than 80% of their classes and were not late admissions, i.e., those who did not join the school after the classes and the intervention had already begun at least a week back.

Table B.2: Regression - Attendance if lesser than 80%

	(1) Attendance	(2) Attendance	(3) Attendance
T1 or T2	1.521	1.231	1.186
	(1.221)	(1.172)	(1.153)
Baseline Listening Score		0.0216	0.0252
		(0.0242)	(0.0242)
Baseline Reading Score		0.0525**	0.0518**
		(0.0245)	(0.0242)
Baseline Writing Score		0.0866***	0.0871***
		(0.0213)	(0.0211)
Female		4.265***	4.327***
		(0.921)	(0.930)
Age		-0.372**	-0.316**
		(0.144)	(0.151)
Grade Level		-1.915***	-2.026***
		(0.531)	(0.521)
Employment Status		-1.548	-1.643
		(1.173)	(1.179)
Mother's Education		-0.0641	-0.0209
		(0.281)	(0.282)
Father's Education		-0.554**	-0.499*
		(0.263)	(0.262)
Other Education			-0.892*
			(0.513)
Marital Status			-5.207*
			(3.116)
Family Size			0.365**
			(0.148)
Schhol F.E.	Yes	Yes	Yes
Constant	65.42***	69.76***	69.06***
	(3.591)	(4.906)	(5.028)
N	3718	3456	3456

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Standard error in parentheses and clustered at classroom level. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$. Classrooms were selected from 28 schools. Regression analysis is conducted over all students were not late admissions, i.e., those who did not join the school after the classes and the intervention had already begun at least a week back. Both the treatment groups are pooled together for improved statistical power. Column 1 presents results without any controls, Column 2 presents results with controls that were balanced in the summary statistics table 2.1 and Column 3 presents results with controls that also include variables that were not balanced in the summary statistics table 2.1

Table B.3: Regression - Attendance

	(1) Attendance	(2) Attendance	(3) Attendance
T1 or T2	1.325***	1.412***	1.404***
	(0.432)	(0.420)	(0.418)
Baseline Listening Score		-0.0152	-0.0147
		(0.00921)	(0.00924)
Baseline Reading Score		0.0119	0.0120
		(0.0104)	(0.0104)
Baseline Writing Score		0.00485	0.00498
		(0.00780)	(0.00780)
Female		0.646*	0.669*
		(0.380)	(0.382)
Age		-0.193***	-0.186***
		(0.0596)	(0.0632)
Grade Level		-0.0327	-0.0482
		(0.180)	(0.180)
Employment Status		0.410	0.378
		(0.514)	(0.514)
Mother's Education		-0.103	-0.105
		(0.112)	(0.112)
Father's Education		-0.187	-0.180
		(0.120)	(0.120)
Other Education			-0.189
			(0.211)
Marital Status			-0.940
			(1.046)
Family Size			0.0149
			(0.0565)
School F.E.	Yes	Yes	Yes
Constant	87.28***	91.65***	91.73***
	(0.793)	(1.515)	(1.632)
N	2219	2076	2076

Note: Standard error in parentheses and clustered at classroom level. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$. Classrooms were selected from 28 schools. Regression analysis is conducted over all students who attended more than 75% of their classes and were not late admissions, i.e., those who did not join the school after the classes and the intervention had already begun at least a week back. Both the treatment groups are pooled together for improved statistical power. Column 1 presents results without any controls, Column 2 presents results with controls that were balanced in the summary statistics table 2.1 and Column 3 presents results with controls that also include variables that were not balanced in the summary statistics table 2.1

Table B.4: Regression - Attendance if greater than 75%

	(1) Attendance	(2) Attendance	(3) Attendance
T1 or T2	-0.304	-0.751	-0.754
	(1.208)	(1.198)	(1.198)
Baseline Listening Score		0.0223	0.0236
		(0.0362)	(0.0363)
Baseline Reading Score		-0.0125	-0.0125
		(0.0308)	(0.0307)
Baseline Writing Score		0.0334	0.0346
		(0.0276)	(0.0276)
Female		2.581**	2.520**
		(1.211)	(1.198)
Age		0.0797	0.0341
		(0.173)	(0.182)
Grade Level		-0.786	-0.806
		(0.696)	(0.695)
Employment Status		-2.107	-2.120
		(1.358)	(1.373)
Mother's Education		-0.459	-0.436
		(0.368)	(0.373)
Father's Education		0.269	0.288
		(0.392)	(0.393)
Other Education			-0.352
			(0.692)
Marital Status			1.674
			(3.426)
Family Size			0.318*
			(0.181)
School F.E.	Yes	Yes	Yes
Constant	46.86***	44.68***	45.11***
	(2.588)	(4.879)	(4.944)
N	1499	1380	1380

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Standard error in parentheses and clustered at classroom level. $p^* < 0.10, p^{**} < 0.05, p^{***} < 0.01$. Classrooms were selected from 28 schools. Regression analysis is conducted over all students who attended lesser than 75% of their classes and were not late admissions, i.e., those who did not join the school after the classes and the intervention had already begun at least a week back. Both the treatment groups are pooled together for improved statistical power. Column 1 presents results without any controls, Column 2 presents results with controls that were balanced in the summary statistics table 2.1 and Column 3 presents results with controls that also include variables that were not balanced in the summary statistics table 2.1

Table B.5: Regression - Attendance if lesser than 75%

Appendix C

Appendix Chapter 3

C.1 Additional tables

Variables Description

Variables	Definition	Source
Firm characteristics		
Company is owned by a foreign investor	Answer to the question on the nationality of the owners. The variable takes the value of 1 if the company is owned by a foreign investor, and zero otherwise.	
Government owns the company	Answer to the question on the ownership of the firm. The variable takes the value of 1 if the company is owned by the government, and zero otherwise.	
Size: Medium	A firm is defined as medium size if it has between 51 and 500 employees.	
Size: Large	A firm is defined large size if it has more than 500 employees.	WBES
Manufacturing	Firm belongs to the manufacturing sector.	
Service	Firm belongs to the service sector.	
Agriculture	Firm belongs to the agriculture sector.	
Construction	Firm belongs to the construction sector.	

Firm's perception about institutional quality

Political stability	Answer to the question: Please judge on a four point scale how problematic are the following factors for the operation and growth of your business: Policy instability/uncertainty. (1) Major obstacle; (2) Moderate obstacle; (3) Minor obstacle; (4) No Obstacle.	
Absence of corruption	Answer to the question: Please judge on a four point scale how problematic are the following factors for the operation and growth of your business: Corruption. (1) Major obstacle; (2) Moderate obstacle; (3) Minor obstacle; (4) No Obstacle.	
Confidence in judicial system	Answer to the statement: "I am confident that the legal system will uphold my contract and property rights in business disputes" . The answer ranges from 1 to 6, where 1=fully disagree, and 6=fully agree.	WBES
Courts-enforceability	Answer to the question: In resolving business disputes, do you believe your country's court system to be: Decisions Enforced. The answer ranges from 1 to 6, where, 1=never, and 6=always.	
Courts-consistent	Answer to the question: In resolving business disputes, do you believe your country's court system to be: Consistent. The answer ranges from 1 to 6, where, 1=never, and 6=always.	
Courts-affordable	Answer to the question: In resolving business disputes, do you believe your country's court system to be: Affordable. The answer ranges from 1 to 6, where, 1=never, and 6=always.	
Courts-quick	Answer to the question: In resolving business disputes, do you believe your country's court system to be: Quick. The answer ranges from 1 to 6, where, 1=never, and 6=always.	WBES
Courts-honest	Answer to the question: In resolving business disputes, do you believe your country's court system to be: Honest/Uncorrupt. The answer ranges from 1 to 6, where, 1=never, and 6=always.	
Courts–fair & impartial	Answer to the question: In resolving business disputes, do you believe your country's court system to be: Fair and Impartial. The answer ranges from 1 to 6, where, 1=never, and 6=always.	

Firm's Perception about Quality of public services

Efficiency of government in delivering services	Answer to the question: How would you generally rate the efficiency of central and local government in delivering services? The answer ranges from 1 to 6, where, 1=very inefficient, and 6=very efficient.	
Quality of education	Rating of the overall quality and efficiency of services delivered by the following public agencies or services: Education services/Schools. Answer ranges from 1=Very bad, to 6=Very good.	
Quality of public health	Rating of the overall quality and efficiency of services delivered by the following public agencies or services: Public Health Care Service/Hospitals. Answer ranges from 1=Very bad, to 6=Very good.	
Quality of water	Rating of the overall quality and efficiency of services delivered by the following public agencies or services: The Water/Sewerage Service/Agency. Answer ranges from 1=Very bad, to 6=Very good.	WBES
Quality of power	Rating of the overall quality and efficiency of services delivered by the following public agencies or services: The Electric Power Company/Agency. Answer ranges from 1=Very bad, to 6=Very good.	
Quality of telephones	Rating of the overall quality and efficiency of services delivered by the following public agencies or services: The Telephone Service/Agency. Answer ranges from 1=Very bad, to 6=Very good.	
Quality of public works	Rating of the overall quality and efficiency of services delivered by the following public agencies or services: Roads Department/Public Works. Answer ranges from 1=Very bad, to 6=Very good.	

Country level institutional quality		
Quality of Institutions index	Average of the index in the period 1998-2002. The aggregated index comprises: (a) Corruption - Assessment of the corruption within the political system. The most common form of corruption met directly by business is financial corruption in the form of demands for special payments and bribes connected with import and export licenses, exchange controls, tax assessments, police protection, or loans. It is also more concerned with actual or potential corruption in the form of excessive patronage, nepotism, job reservations, 'favor-for-favor', secret party funding, and suspiciously close ties between politics and business, (b) Law and Order - Law and Order are assessed separately, with each sub-component comprising zero to three points. The Law sub-component is an assessment of the strength and impartiality of the legal system, while the Order sub-component is an assessment of popular observance of the law. A country can enjoy a high rating – 3 – in terms of its judicial system, but a low rating - 1 – if it suffers from a very high crime rate or if the law is routinely ignored without effective sanction (for example, widespread illegal strikes), and (c) Bureaucratic Quality - The institutional strength and quality of the bureaucracy is another shock absorber that tends to minimize revisions of policy when governments change. High points are given to countries where the bureaucracy has the strength and expertise to govern without drastic changes in policy or interruptions in government services. Countries that lack the cushioning effect of a strong bureaucracy receive low points because a change in government tends to be traumatic in terms of policy formulation and day-to-day administrative functions. The index takes values between 0 and 18.	ICRG
Log(GDP)	Log of the Gross Domestic Product for the year when the interview was done.	WDI
Taxes		
General constraint-taxes and regulations	Answer to the question: Please judge on a four point scale how problematic are the following factors for the operation and growth of your business: Policy instability/uncertainty. (1) Major obstacle; (2) Moderate obstacle; (3) Minor obstacle; (4) No Obstacle.	WBES
Current VAT rate	Data correspond to the current standard VAT rate as of August 2004. The information was comprised by the IMF "VAT Database: VAT Rates for Fund Member Countries" , which in turn was based on calculations by the International Bureau of Fiscal Documentation; and Corporate Taxes 2003-2004, Worldwide Summaries (PricewaterhouseCoopers).	IMF

Instruments

Legal origin	Dummies related to the origin of the commercial law of a country: British, French, Scandinavian, Socialist or German.	La Porta, et al (1997)
Region	Dummies of the regions that are covered in the sample: Transition, East Asia, South Asia, Latin America, and OECD.	WDI
Legal organization of the company	Answer to the question: What is the legal organization of this company: (1) Single proprietorship, (2) Partnership, (3) Cooperative, (4) Corporation, privately-held, (5) Corporation listed on stock exchange.	WBES
Firm's ownership	Answer to the question: Which of the following best describes the overall control of your firm, where control means making major decisions concerning the enterprise's direction? (1) Individual owner(s), (2) A family, (3) A company group, (4) A bank, (5) Its board of directors/supervisory board, (6) Its managers, (7) Its workers, (8) Others.	WBES

All the data are retrievable at this site: http://www.enterprisesurveys.org

Efficiency of government in delivering services (1=very inefficient 6=very efficient)						
	Pr[Y=1]	Pr[Y=2]	Pr[Y=3]	Pr[Y=4]	Pr[Y=5]	Pr[Y=6]
Quality of Institutions index	-0.007	-0.007	-0.001	0.008	0.006	0.001
	(-1.87)*	(-1.88)*	(-1.42)	(1.94)**	(1.87)*	(1.48)
Political stability	-0.036	-0.038	-0.008	0.043	0.032	0.007
	(-7.02)***	(-6.53)***	(-1.94)**	(6.94)***	(5.99)***	(3.03)***

Note: The number of observations is 6039, the Log-likelihood is -9264.29, the Pseudo-R-squared is 0.03, and the corresponding Chi-Squared is 193.15. Robust z-statistics in parentheses. Standard errors clustered at the country level. * significant at 10%; ** significant at 5%; *** significant at 1%

Table C.1: Institutional quality and public services

Bibliography

ADB, Abdul Abiad, Davide Furceri, and Petia Topalova IMF. "The macroeconomic effects of public investment: Evidence from advanced economies." Journal of Macroeconomics 50 (2016): 224-240.

Allingham, Michael G., and Agnar Sandmo. "Income tax evasion: A theoretical analysis." Journal of public economics 1, no. 3-4 (1972): 323-338.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. "Report Cards: The Impact of Providing Test-score information on Educational Markets." (2008).

Angrist, Joshua, and Victor Lavy. "The effects of high stakes high school achievement awards: Evidence from a randomized trial." American economic review 99, no. 4 (2009): 1384-1414.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. "Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment." American economic review 92, no. 5 (2002): 1535-1558.

Aschauer, David Alan. "Is public expenditure productive?." Journal of monetary economics 23, no. 2 (1989): 177-200.

Attanasio, Orazio P., Costas Meghir, and Ana Santiago. "Education choices in Mexico: using a structural model and a randomized experiment to evaluate Progresa." The Review of Economic Studies 79, no. 1 (2012): 37-66.

Aucejo, Esteban M., and Teresa Foy Romano. "Assessing the effect of school days and absences on test score performance." Economics of Education Review 55 (2016): 70-87.

Baker, H. E., and Homer L. Bates. "Student and faculty perceptions of the impact of plus/minus grading: a management department perspective." Journal on Excellence in College Teaching 10, no. 1 (1999): 23-33.

Baez, Javier E., and Adriana Camacho. Assessing the long-term effects of conditional cash transfers on human capital: evidence from Colombia. The World Bank, 2011.

Baird, Sarah, Craig McIntosh, and Berk Özler. "Cash or condition? Evidence from a cash transfer experiment." The Quarterly journal of economics 126, no. 4 (2011): 1709-1753.

Banerji, Rukmini, James Berry, and Marc Shotland. "The impact of maternal literacy and participation programs: Evidence from a randomized evaluation in India." American Economic Journal: Applied Economics 9, no. 4 (2017): 303-37.

Barham, Tania, Karen Macours, and John A. Maluccio. More schooling and more learning? Effects of a three-year conditional cash transfer program in Nicaragua after 10 years. No. IDB-WP-432. IDB Working Paper Series, 2013.

Bayar, Yilmaz. "Public governance and shadow economy in Central and Eastern European countries." Revista Administratie si Management Public (RAMP) 27 (2016): 62-73.

Becker, William E. and Sherwin Rosen. "The Learning Effect of Assessment Evaluation in High School," Economics of Education Review, 1992, 11(2), pp. 107-118.

Benhassine, Najy, Florencia Devoto, Esther Duflo, Pascaline Dupas, and Victor Pouliquen. "Turning a shove into a nudge? A" labeled cash transfer" for education." American Economic Journal: Economic Policy 7, no. 3 (2015): 86-125.

Bettinger, Eric P. "Paying to learn: The effect of financial incentives on elementary school test scores." Review of Economics and Statistics 94, no. 3 (2012): 686-698.

Blimpo, Moussa P. "Team incentives for education in developing countries: A randomized field experiment in Benin." American Economic Journal: Applied Economics 6, no. 4 (2014): 90-109.

Bobba, Matteo, and Veronica Frisancho. "Learning about oneself: The effects of signaling academic ability on school choice." Unpublished working paper (2014).

Boix, Carles. "Democracy, development, and the public sector." American Journal of Political Science (2001): 1-17.

Bressette, Andrew. "Arguments for plus/minus grading: A case study." Educational Research Quarterly 25, no. 3 (2002): 29.

Celik Katreniak, Dagmara. "Dark Side of Incentives: Evidence From a Randomized Control Trial in Uganda." Available at SSRN 3288474 (2018).

Charness, Gary, and Uri Gneezy. "Strong evidence for gender differences in risk taking." Journal of Economic Behavior Organization 83, no. 1 (2012): 50-58.

Chong, Alberto, and Mark Gradstein. "Inequality and informality." Journal of public Economics 91, no. 1-2 (2007): 159-179.

Czibor, Eszter, Sander Onderstal, Randolph Sloof, and Mirjam Van Praag. "Does relative grading help male students? Evidence from a field experiment in the classroom." (2014).

Dabla-Norris, Era, Mark Gradstein, and Gabriela Inchauste. "What causes firms to hide output? The determinants of informality." Journal of development economics 85, no. 1-2 (2008): 1-27.

De Soto, Hernando, and Harry P. Diaz. "The mystery of capital. Why capitalism triumphs in the West and fails everywhere else." Canadian Journal of Latin American Caribbean Studies 27, no. 53 (2002): 172.

del Pero, Angelica Salvi, and Alexandra Bytchkova. "A bird's eye view of gender differences in education in OECD countries." (2013).

Desai, M., Dyck, A., and L. Zingales. "Theft and Taxes." Journal of Financial Economics 84: 591-623.

Dobkin, Carlos, Ricard Gil, and Justin Marion. "Skipping class in college and exam performance: Evidence from a regression discontinuity classroom experiment." Economics of Education Review 29, no. 4 (2010): 566-575.

Dubey, Pradeep, and John Geanakoplos. "Grading exams: 100, 99, 98,... or a, b, c?." Games and Economic Behavior 69, no. 1 (2010): 72-94.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. "School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools." Journal of Public Economics 123 (2015): 92-110.

Dynarski, Susan. "Building the stock of college-educated labor." Journal of human resources 43, no. 3 (2008): 576-610.

Easterly William, R., and T. Rebelo Sergio. "Fiscal Policy and Economic Growth: An Empirical Investigation." Journal of Monetary Economics 32, no. 3 (1993): 417-458.

Eckel, Catherine C., and Philip J. Grossman. "Differences in the economic decisions of men and women: Experimental evidence." Handbook of experimental economics results 1 (2008): 509-519.

Else-Quest, Nicole M., and Oana Peterca. "Academic attitudes and achievement in students of urban public single-sex and mixed-sex high schools." American Educational Research Journal 52, no. 4 (2015): 693-718.

Fisman, Raymond. "Estimating the value of political connections." American economic review 91, no. 4 (2001): 1095-1102.

Friedman, Eric, Simon Johnson, Daniel Kaufmann, and Pablo Zoido-Lobaton. "Dodging the grabbing hand: the determinants of unofficial activity in 69 countries." Journal of public economics 76, no. 3 (2000): 459-493.

Fries, Ryan N., Jamie Conklin, Jessica S. Krim, and Deborah A. Smith. "Student and Faculty Perceptions on Plus-Minus Grading: A Case Study." Educational Research Quarterly 36, no. 4 (2013): 49-68.

Fryer Jr, Roland G. "Financial incentives and student achievement: Evidence from randomized trials." The Quarterly Journal of Economics 126, no. 4 (2011): 1755-1798.

Furceri, Davide, and Marcos Poplawski Ribeiro "Government consumption volatility and the size of nations." (2009).

Gershenson, Seth, Alison Jacknowitz, and Andrew Brannegan. "Are student absences worth the worry in US primary schools?." Education Finance and Policy 12, no. 2 (2017): 137-165.

Goel, Rajeev K., Ummad Mazhar, Michael A. Nelson, and Rati Ram. "Different forms of decentralization and their impact on government performance: Micro-level evidence from 113 countries." Economic Modelling 62 (2017): 171-183.

Goel, Rajeev K., and Michael A. Nelson. "Shining a light on the shadows: Identifying robust determinants of the shadow economy." Economic Modelling 58 (2016): 351-364.

Glewwe, Paul, and Karthik Muralidharan. "Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications." In Handbook of the Economics of Education, vol. 5, pp. 653-743. Elsevier, 2016.

Gong, Binglin, and Chun-Lei Yang. "Gender differences in risk attitudes: Field experiments on the matrilineal Mosuo and the patriarchal Yi." Journal of economic behavior organization 83, no. 1 (2012): 59-65.

Goodman, Joshua. "Flaking Out: Snowfall, Disruptions of Instructional Time, and Student Achievement." Applied Statistics (2012).

Gramlich, Edward M. "Infrastructure investment: A review essay." Journal of economic literature 32, no. 3 (1994): 1176-1196.

Grant, Darren. "The essential economics of threshold-based incentives: Theory, estimation, and evidence from the Western States 100." Journal of Economic Behavior Organization 130 (2016): 180-197.

Grant, Darren, and William B. Green. "Grades as incentives." Empirical Economics 44, no. 3 (2013): 1563-1592.

Hallward-Driemeier, Mary, and Lant Pritchett. "How business is done in the developing world: Deals versus rules." Journal of Economic Perspectives 29, no. 3 (2015): 121-40.

Handa, Sudhanshu. "Raising primary school enrolment in developing countries: The relative importance of supply and demand." Journal of development Economics 69, no. 1 (2002): 103-128.

Hanushek, Eric A. "Why quality matters in education." Finance and development 42, no. 2 (2005): 15-19.

Hanushek, Eric A., and Lei Zhang. Quality-consistent estimates of international returns to skill. No. w12664. National Bureau of Economic Research, 2006.

Hauner, David, and Annette Kyobe. "Determinants of government efficiency." World Development 38, no. 11 (2010): 1527-1542.

Heckman, James J., Jora Stixrud, and Sergio Urzua. "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior." Journal of Labor economics 24, no. 3 (2006): 411-482.

Hirshleifer, Sarojini. "Incentives for effort or outputs? A field experiment to improve student performance." Unpublished manuscript. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (J-PAL) (2015).

Jalava, Nina, Juanna Schrøter Joensen, and Elin Pellas. "Grades and rank: Impacts of non-financial incentives on test performance." Journal of Economic Behavior Organization 115 (2015): 161-196.

Jensen, Robert. "Do labor market opportunities affect young women's work and family decisions? Experimental evidence from India." The Quarterly Journal of Economics 127, no. 2 (2012): 753-792.

Johnson, Simon, Daniel Kaufmann, John McMillan, and Christopher Woodruff. "Why do firms hide? Bribes and unofficial activity after communism." Journal of Public Economics 76, no. 3 (2000): 495-520.

Kaufmann, Daniel, Aart Kraay, and Massimo Mastruzzi. Governance matters V: aggregate and individual governance indicators for 1996-2005. The World Bank, 2006.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. "Incentives to learn." The Review of Economics and Statistics 91, no. 3 (2009): 437-456.

Li, Tao, Li Han, Linxiu Zhang, and Scott Rozelle. "Encouraging classroom peer interactions: Evidence from Chinese migrant schools." Journal of Public Economics 111 (2014): 29-45.

Lindert, Peter H. "Social spending and economic growth since the Eighteenth Century." (2004).

Lledó, Victor, and Marcos Poplawski-Ribeiro. "Fiscal policy implementation in sub-Saharan Africa." World Development 46 (2013): 79-91.

Main, Joyce B., and Ben Ost. "The impact of letter grades on student effort, course selection, and major choice: A regression-discontinuity analysis." The Journal of Economic Education 45, no. 1 (2014): 1-10.

Manolas, George, Kostas Rontos, George Sfakianakis, and Ioannis Vavouras. "The determinants of the shadow economy: The case of Greece." International Journal of Criminology and Sociological Theory 6, no. 1 (2013).

McClure, James E., and Lee C. Spector. "Plus/minus grading and motivation: an empirical study of student choice and performance." Assessment Evaluation in Higher Education 30, no. 6 (2005): 571-579.

McEwan, Patrick J. "Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments." Review of Educational Research 85, no. 3 (2015): 353-394.

McGuire, Martin C., and Mancur Olson. "The economics of autocracy and majority rule: The invisible hand and the use of force." Journal of economic literature 34, no. 1 (1996): 72-96.

Misch, Florian, Norman Gemmell, and Richard Kneller. "Using surveys of business perceptions as a guide to growth-enhancing fiscal reforms." Economics of Transition 22, no. 4 (2014): 683-725.

Mukherjee, Shagata. "Essays on Gender and Microfinance." (2017).

Muralidharan, Karthik. "Field experiments in education in developing countries." In Handbook of economic field experiments, vol. 2, pp. 323-385. North-Holland, 2017.

Nores, Milagros, Clive R. Belfield, W. Steven Barnett, and Lawrence Schweinhart. "Updating the economic impacts of the High/Scope Perry Preschool program." Educational Evaluation and Policy Analysis 27, no. 3 (2005): 245-261.

Oettinger, Gerald S. "The effect of nonlinear incentives on performance: evidence from "Econ 101"." Review of Economics and Statistics 84, no. 3 (2002): 509-517.

Olken, Benjamin A. "Corruption and the costs of redistribution: Micro evidence from Indonesia." Journal of public economics 90, no. 4-5 (2006): 853-870.

Ors, Evren, Frédéric Palomino, and Eloic Peyrache. "Performance gender gap: does competition matter?." Journal of Labor Economics 31, no. 3 (2013): 443-499.

Paredes, Valentina. "Grading system and student effort." Education Finance and Policy 12.1 (2017): 107-128.

Porta, Rafael La, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert W. Vishny. "Law and finance." Journal of political economy 106, no. 6 (1998): 1113-1155.

Porta, Rafael La, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny. "The quality of government." The Journal of Law, Economics, and Organization 15, no. 1 (1999): 222-279.

Rajkumar, Andrew Sunil, and Vinaya Swaroop. "Public spending and outcomes: Does governance matter?." Journal of development economics 86, no. 1 (2008): 96-111.

Remeikiene, R. I. T. A., and L. I. G. I. T. A. Gaspareniene. "Evaluation of the shadow economy influencing factors: Lithuanian case." In 9th International Conference on Business Administration, Dubai, United Arab Emirates, pp. 148-154. 2015.

Robinson, Carly D., Jana Gallus, Monica G. Lee, and Todd Rogers. "The Demotivating Effect (and Unintended Message) of Retrospective Awards." No. rwp18-020. 2018.

Sarte, Pierre-Daniel G. "Informality and rent-seeking bureaucracies in a model of long-run growth." Journal of Monetary Economics 46, no. 1 (2000): 173-197.

Schneider, Friedrich, and Dominik H. Enste. "Shadow economies: size, causes, and consequences." Journal of economic literature 38, no. 1 (2000): 77-114.

Shahab, Mohammad Reza, Jamshid Pajooyan, and Farhad Ghaffari. "The effect of corruption on shadow economy: an empirical analysis based on panel data." International Journal of Business and Development Studies 7, no. 1 (2015): 85-100.

Sohn, Kitae. "The role of cognitive and noncognitive skills in overeducation." Journal of Labor Research 31, no. 2 (2010): 124-145.

Sokolov, Vladimir, and Laura Solanko. "Political influence, firm performance and survival." Higher School of Economics Research Paper No. WP BRP 60 (2017).

Springer, Matthew G., Brooks A. Rosenquist, and Walker A. Swain. "Monetary and nonmonetary student incentives for tutoring services: A randomized controlled trial." Journal of Research on Educational Effectiveness 8, no. 4 (2015): 453-474.

Urquiola, Miguel, and Eric Verhoogen. "Class-size caps, sorting, and the regression-discontinuity design." American Economic Review 99, no. 1 (2009): 179-215.

World Bank. World Development Report 2004 (Overview): Making Services Work for Poor People. World Bank, 2003.

Younger, Michael Robert, and Molly Warrington. "Would Harry and Hermione have done better in single-sex classes? A review of single-sex teaching in coeducational secondary

schools in the United Kingdom." American Educational Research Journal 43, no. 4 (2006): 579-620.

Vita

Puneet Arora was born on December 3, 1988 in Palwal (Haryana), India. He pursued his studies in Economics from University of Delhi where he earned his Bachelor's (2009) and Master's degree (2013). He obtained his Ph.D. degree in Economics from Georgia State University (GSU) in 2020 under the supervision of Dr. Alberto Chong.

Puneet's research utilizes both experimental (lab and field) and quasi-experimental methods to study issues in developing countries, with a primary focus on issues in the education sector. In his most recent research work, Puneet conducted field experiments in India testing behavioral interventions to improve student learning and attendance outcomes. In addition, he has also worked on a lab experiment studying different audit mechanisms to improve tax compliance; and a natural experiment studying the nature of petty corruption among notary publics. He has presented his work at the research seminars at University of Chicago, Georgia State University, Ashoka University, Delhi School of Economics and several others.

Puneet has taught courses in Principles of Microeconomics, Intermediate Microeconomics and Mathematics for Economists. He has served as a research assistant to professors at Delhi School of Economics, Monash University and Georgia State University. He has also been awarded with Center for the Economic Analysis of Risk Scholarship, Andrew Young School Dissertation Fellowship, Russell Sage Foundation Research Grant and the Outstanding Graduate Research Assistant Award from Georgia State University.

Puneet will join the faculty at the Amrut Mody School of Managament, Ahmedabad University as an Assistant Professor in Economics in fall 2020.