A Classroom Experiment on Effort Allocation under Relative Grading

Andy Brownback

September 27, 2014

Abstract

Grading on the curve, or relative grading is one of the most commonly used grading schemes in education. The law of large numbers implies that as the size of a class grows, the percentile ranks of its students draw closer to their percentile ranks in the population, which changes the students’ incentives. I model this environment in order to predict how changes in the class size differentially affect students with different abilities as measured by their GPAs. I test this model in a large-scale classroom experiment measuring effort in terms of time spent on online quizzes. I randomly assign students to “cohorts” of 10 or 100 students, where they receive high grades if they score in the top 70 percent of their cohort. I observe quizzes with both cohort sizes for all students every Quiz Week, so my design controls for student- and week-specific effects as it tests the causal relationship between the cohort size and effort. My results show that the lower variance of the larger cohorts elicits greater mean effort and greater effort from all but the lowest ability students. The greater variance of the smaller cohort elicits more effort from the lowest ability students. Many low ability students fail to take advantage of the randomness of the smaller cohort, an allocation failure consistent with “cursed” beliefs about their classmates and other behavioral biases. I discuss student welfare and make policy recommendations for classroom designers.

Preliminary draft. Comments greatly appreciated.
1 Introduction

Under a relative or “curved” grading scheme, the evaluation of a student’s academic performance is based on her percentile rank within her comparison group. Importantly, these grades are determined independently from any absolute measures of performance. Eschewing cardinal measures of academic performance has proved desirable enough to establish these grading mechanisms as a fixture in many university classrooms and law schools.\textsuperscript{1,2} Indeed, mechanism designers across many areas of education have employed relative awarding schemes in competitions for scholarships, college admissions, and even teacher pay.\textsuperscript{3,4,5}

In its simplest form, relative grading assigns grades to students based only on their percentile rank within the class. For example, a professor may award “A” grades to students who are at or above the 80th percentile of performers, “B” grades to students between the 80th and 60th percentiles, and so on. I will refer to the quantiles that serve to distinguish different grade levels as “cutoffs.”

Consider an example, Texas HB 588, which grants automatic admission to any Texas state university—including the University of Texas at Austin—to all Texas high school seniors who graduate in the top 10 percent of their high school class.\textsuperscript{6} The incentive structure of this policy mirrors that of a curved grading scheme with a single cutoff at the 10th percentile. Like many other curved grading schemes, this policy is invariant to the size of a given graduating class despite the fact that graduating classes vary by orders of magnitude.\textsuperscript{7}

\footnotesize
\textsuperscript{1}Though frequently discussed, good statistics on the percentage of undergraduate courses graded on the curve are hard to come by.
\textsuperscript{2}Mroch (2005) estimates that 79\% of law schools standardize scores according to a grading curve.
\textsuperscript{3}Missouri’s Bright Flight scholarship program awards scholarships to the top 3\% of high school seniors based on ACT or SAT scores.
\textsuperscript{4}In California, the top 9\% of graduating seniors are guaranteed admission to one of the University of California campuses. In Kansas, the top 33\% are guaranteed access to the state college of their choice. In Texas, the top 10\% are offered similar incentives.
\textsuperscript{5}North Carolina Senate Bill 402 section 9.6(g) grants favorable contracts to the top 25\% of teachers at each school as evaluated by the school’s administration.
\textsuperscript{6}The bill was modified in 2009 to stipulate that the University of Texas at Austin may cap the number of students admitted under this measure to 75\% of in-state freshman students.
\textsuperscript{7}Plano East High School in Plano, TX has an enrollment of 6,015 students, while Valentine High School in Valentine, TX has an enrollment of 9 students.
Under curved grading, a student’s comparison group is critical in determining her outcome. I will refer to the draw of students in a comparison group as a “cohort.” The law of large numbers implies that the larger is a draw of students from a distribution, the more that draw resembles the distribution itself. Since this draw of students determines the incentives for effort, the size of a cohort must therefore affect each student’s incentives. To put this chain of events in context, the larger is a Texas high school graduating class, the closer does the draw of students in it resemble the population distribution of Texas high school seniors. Small schools are more likely to draw a senior class of outlying students than large schools, resulting in cohorts with higher variance. This variance then affects the incentives for effort among students. A similar intuition holds for university courses, grant applications, and tenure-track positions, to list a few examples.

This paper presents a theoretical and experimental treatment of a simplified grading curve where students either receive a high or low grade. The theoretical model finds the equilibria for expected-grade maximizing students under two identical grading curves with differently sized cohorts. This formalizes the ways in which the incentive structure of relative grading schemes are sensitive to changes in the number of competitors and provides closed form and falsifiable predictions for the effects of those changes. The existence of different incentives, however, does not ensure that participants’ behavior will respond predictably. Prior research has demonstrated that economic agents often ignore sample size when drawing inference, causing systematic biases in their beliefs (Tversky and Kahneman, 1971; Kahneman and Tversky, 1973; Rabin, 2002; Benjamin, Rabin and Raymond, 2014). In this paper, I conduct a controlled test of this question in a natural environment to determine the sensitivity of student effort choices to incentives that change with sample size.

I conduct an experiment on relative grading in a large, upper-division economics course at UC San Diego. I present students with a pair of quizzes graded according to identical curves, that is, with identical quantiles set as cutoffs. For each student, I randomly determine which quiz will be graded based on the 10-Student cohort and which will be graded based on
the 100-Student cohort. I observe the amount of time each student spends taking each quiz and use this as a measure of effort. My design allows me to control for the amount of time studied in preparation for each quiz by analyzing a pair of quizzes taken in quick succession. Under these controls, I can use the time spent on a given quiz to directly analyze the causal relationship between the effort exerted on quizzes and the size of the cohort. Similar to existing models, my model predicts that effort will increase with the size of the cohort. My results confirm this prediction, showing that larger cohorts elicit a statistically significantly 27 second increase in effort on each quiz. This increase in effort is more than 3 percent over the mean, an increase that may appear small, but would lead to multiple hours of additional study time each quarter.

I enrich my analysis by including GPA to control for student ability. The model makes clear predictions about the heterogeneity of incentives across students of different ability levels. First, the model predicts that students whose ability level lies below the cutoff will exert more effort in the smaller cohorts, while students with abilities above the cutoff will do the opposite. Second, the model predicts that the maximum and minimum differences in effort between cohort sizes should occur near the cutoff, above and below, respectively.

My empirical results deviate from these predictions in important ways. First, while the lowest ability students exert more effort on quizzes with the smaller cohort, as a whole, students below the cutoff exert more effort in the larger cohorts on average. Second, the maximum difference in favor of the large cohort actually occurs for students with abilities below the cutoff, not above it. For students near but below the cutoff, this effort allocation fails to take advantage of the greater variance of the smaller cohort.

My model’s predictions, however, rely on each student holding accurate beliefs about his ability relative to the classroom distribution. These accurate beliefs must take into account the fact that higher ability students are more likely to enroll in upper-division

---

8 Most notably Moldovanu and Sela (2006)
9 Students reported spending an average of 6.6 hours per week studying for the course on their course evaluations. A 3% increase, therefore results in approximately 2.2 additional hours of studying in a quarter.
economics courses. “Cursed” beliefs (Eyster and Rabin, 2005), on the other hand, suggest that students may wrongly perceive that the distribution of students in their class is identical to the distribution of students in their previous lower-division courses and best respond to these inaccurate beliefs. I test the fit of my model under cursed beliefs by using GPA data from all lower-division courses to construct a hypothetical distribution of abilities in those courses and show that observed behavior is closer to the optimal strategic behavior when students’ fail to account for the selection of students into the course. But, even under fully cursed beliefs there is a residual misallocation of effort. This misallocation is consistent with several possible behavioral biases, among them, overconfidence, updating failures, and myopic reference dependence.

The tension between mean effort and the distribution of effort is central to the debate over optimal classroom design. I propose ways in which the effect of contest size can be addressed within the curved grading mechanism, depending on the designer’s objectives. I also propose different policies targeted at the cursed beliefs or behavioral biases potentially causing students to misallocate effort across quizzes.

To fix ideas, I refer to “relative” or “curved” grading schemes throughout this paper, but this should not distract from the generality of the results. While my discussion and experiment all exist within the classroom setting, relative awarding schemes are found throughout the modern economy, in job promotion contests, performance bonuses, and lobbying contests, to name a few. I believe that my results generalize to any of these settings for several reasons. To begin, the costs of effort in my experiment, the means of exerting it, and the norms surrounding its exertion are all similar to a professional setting. Additionally, the means of evaluating performance in my experiment closely reflects those of many professional settings. Finally, the heterogeneity of ability in a classroom likely follows similar patterns as the heterogeneity of professional competencies, since intelligence, motivation,

---

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

The outline of this paper is as follows. In the following section, I discuss the existing literature on contests in experimental and non-experimental settings, as well as the education literature on performance evaluations. In Section 3 I outline a simple model of the type of relative grading that my experimental subjects face and solve it under a specific set of production and cost functions. This model yields the qualitative predictions that I will test in the experiment. The experiment itself is formally presented in Section 4. Section 5 presents the results of the experiment and tests the predictions of the model under accurate and cursed beliefs. Section 6 discusses the implications of my results for the optimal design of relative evaluation methods. Section 7 concludes the paper.

2 Literature

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

2.1 Experimental Economics

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

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

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

### 2.2 Microeconomic Theory

The contest theory literature began in order to capture the incentives for rent-seeking (Tullock, 1967; Krueger, 1974), and it has since evolved into a more general branch of research
that considers various environments with costly effort and uncertain payoffs. The three models most often employed are the all-pay auction (Hirshleifer and Riley, 1978; Hillman and Samet, 1987; Hillman and Riley, 1989), the Tullock contest (Tullock, 1980), and the rank-order tournament (Lazear and Rosen, 1981).

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

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

2.3 Education Literature

Grading mechanisms are frequently studied in the economics of education. Costrell (1994) explores the endogenous selection of grading standards by policy makers seeking to maximize social welfare, subject to students who best respond to those standards. Betts (1998) expands this framework to include heterogeneous students. Betts and Grogger (2003) then look at the impact of grading standards on the distribution of students. Notably, these paper do not consider strategic interaction between students, a distinction between my paper and theirs.

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

Kokkelenberg, Dillon, and Christy (2008) find a negative effect of class size on the grades awarded to individual students in a non-strategic environment. The independent effect of class size on strategically interacting students, however, remains unstudied. Contest size is often taken as given or assumed to be determined exogenously. In this paper, I demonstrate that cohort size plays a significant role in a student’s selection of effort when grading on the curve. Thus, a classroom designer optimizing student outcomes needs to take into consideration the heterogeneous effects of cohort size on students with different abilities.

3 A Simple Model of Academic Effort

In this section, I develop a model of strategic interaction in a classroom graded on the curve. To do this, I make certain assumptions about the preferences and production functions of the students. I include a more general treatment of the model in the appendix. I use this simplified model to display generic properties that convey intuition for how an experimental intervention in the size of a cohort will affect a student’s effort. My model borrows heavily from the independent private value auction literature (Vickrey, 1961), and the all-pay auction literature (Baye, Kovenock, and de Vries, 1993).

Suppose there are \( N \) students exerting costly effort in order to increase their chances of winning one of \( M \equiv P \times N \) prizes in the form of high grades. These high grades are awarded on a strictly relative basis to the highest \( M \) performers.

Suppose each student has an ability, \( a_i \), distributed uniformly from 0 to 1. That is \( a_i \sim U[0, 1] \), meaning \( F(a_i) = a_i \). Students are evaluated at each period, \( t \), based on their academic output, or “score,” \( s_{i,t} \) from that period. Score is determined as a function of effort, \( e_{i,t} \), and ability. Suppose this production function is Cobb-Douglas and marginal costs are
constant and homogeneous according to,

\[ s_{i,t} = e_{i,t} \times a_i \quad (1) \]

\[ C(e_{i,t}; a_i) = c \times e_{i,t} \quad (2) \]

This production function is appealing for its simplicity and its ability to capture an intuitive complementarity between ability and effort. Constant marginal costs are appealing in that they allow the costs of a given score to be decreasing in ability for all positive ability levels. Assuming homogeneous marginal costs will not affect the predictions for my experiment, since my within-subjects design controls for student-specific costs of effort. Under these assumptions, any change in the marginal costs that does not result in a corner solution simply rescales the equilibrium scores, making the actual value of \( c \) irrelevant.

### 3.1 Student’s Utility

The expected utility of a student is determined by both the likelihood of receiving a high grade at a given score and the cost of scores. Normalize the value of receiving a high grade to one. Heterogeneity across students with different ability levels is now captured by the effort required to generate a given level of output. Thus, a student’s utility can be represented as

\[ U(e_{i,t}; a_i) = \Pr(e_{i,t} \times a_i \geq \bar{S}) - c \times e_{i,t} \quad (3) \]

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

I restrict my attention to the set of functions, \( S : (a_i; N, P) \mapsto s_{i,t} \), that take parameters, \( N \) and \( P \), and map abilities to scores and constitute symmetric equilibria of the model. In the appendix, I prove that any such function must be monotonic in a symmetric equilibrium. In addition to monotonicity, it is straightforward to show that scores must also be continuous in ability.\(^{11}\) All continuous, monotonic functions are invertible, so there must exist a function

\(^{11}\)Suppose not. With discontinuities within the support \( S(a_i) \), some students would be failing to best
that maps a given student’s score back onto the student’s ability implied by that score. Given that the equilibrium scores depend on the parameters, $N$ and $P$, this inverse function, too, depends on these parameters. This defines the function $A(s_{i,t}; N, P) \equiv S^{-1}(s_{i,t}; N, P)$.

With monotonicity and invertibility established, the probability of receiving a high grade is equivalent to the probability that the ability level implied by a student’s score is higher than the ability levels implied by the scores of $N - M$ other students in her cohort. Given that $F(a_i) = a_i$, this probability is characterized by the order statistic,

$$\Pr\left(s_{i,t} \geq S\right) = \sum_{j=0}^{N-1} \frac{(N-1)!}{j!(N-1-j)!} A(s_{i,t}; N, P)^j \times (1 - A(s_{i,t}; N, P))^{N-1-j}. \quad (4)$$

Substituting (4), (1), and (2) into (3) completes the student’s utility function.

$$U(e_{i,t}; a_i, N, P) = \sum_{j=0}^{N-1} \frac{(N-1)!}{j!(N-1-j)!} A(e_{i,t} \times a_i; N, P)^j \times (1 - A(e_{i,t} \times a_i; N, P))^{N-1-j} \times c \times e_{i,t}. \quad (5)$$

Maximizing (5) with respect to the choice variable, $e_{i,t}$, yields the first-order condition,

$$\frac{\partial U(e_{i,t}; a_i, N, P)}{\partial e_{i,t}} = \sum_{j=0}^{N-1} \frac{(N-1)!}{j!(N-1-j)!} \left\{ j \times A(e_{i,t} \times a_i; N, P)^{j-1} \right. \times A'(e_{i,t} \times a_i) \times a_i \times (1 - A(e_{i,t} \times a_i; N, P))^{N-1-j} \right. \\
\left. \times A'(e_{i,t} \times a_i) \times a_i \left[ A(e_{i,t} \times a_i; N, P)^j (1 - A(e_{i,t} \times a_i; N, P))^{N-2-j} \right. \right. \\
\left. \left. \times A'(e_{i,t} \times a_i) \times a_i \right] \right\} \equiv c. \quad (6)$$

At equilibrium, the ability implied by a student’s score must equal that student’s ability. A student scoring just above the discontinuity would be able to increase his expected utility by lowering his effort, which would lower his score and his costs, up until the discontinuity in score has vanished.
That is, \( A(s_{i,t}) = a_i \). Substituting this into (6) and solving for \( \frac{1}{A'(s_{i,t})} \) results in

\[
\sum_{j=N-1}^{N-1} \left( \frac{(N-1)!}{j!(N-1-j)!} \right) \left\{ \frac{1}{c} a_i^j (1 - a_i)^{N-1-j} - \left( \frac{N-1-j}{c} \right) \right\} \left\{ a_i^{j+1} (1 - a_i)^{N-2-j} \right\} \equiv \frac{1}{A'(s_{i,t}; N, P)} .
\] (7)

By the definition of \( A(s_{i,t}) \),

\[
S(\overbrace{A(s_{i,t})}) = S\left( S^{-1}(s_{i,t}) \right) = s_{i,t}
\]

\[\Rightarrow S'(A(s_{i,t})) A'(s_{i,t}) = 1\]

\[\Rightarrow S'(a_i) = \frac{1}{A'(s_{i,t})} .
\]

Substituting this into (7) yields the differential equation defining the relationship between ability and equilibrium score given by,

\[
S'(a_i; N, P) \equiv \sum_{j=(1-P)\times N}^{N-1} \left( \frac{(N-1)!}{j!(N-1-j)!} \right) \left\{ \frac{1}{c} a_i^j (1 - F(a_i; N, P)))^{N-1-j} - \left( \frac{N-1-j}{c} \right) \right\} \left\{ a_i^{j+1} (1 - a_i)^{N-2-j} \right\} .
\] (8)

The solution to (8) takes the parameters \( N \) and \( P \), and calculates the equilibrium score as a function of student ability. Dividing the equilibrium score by the corresponding ability returns the student’s choice of effort at equilibrium.

Now consider two different sizes of contest, \( N \) and \( N' \), each with the same proportion of winners, \( P \). Evaluating (8) at the different parameter values reveals the effect that changing the contest size from \( N \) to \( N' \) has on the equilibrium scores. From here, it is straightforward to deduce the change in equilibrium effort levels. I will refer to this change in effort as the “treatment effect.”
3.2 Experimental Model

Solving for the optimal score and effort requires parameter values for \( N \) and \( P \). This subsection restricts attention to the experimental parameters, \( N = 10 \) and \( N = 100 \), and explores the effect that moving from one cohort size to another has on the effort exerted by students at equilibrium.

Since each of the two cohort sizes are randomly assigned to quizzes, I will refer to the two treatments as the “10-Student Quiz” and the “100-Student Quiz.” The proportion of winners on each quiz is held constant at \( P = 0.7 \). That is, the top 70 percent of scores in a cohort receive high grades on a given quiz. Figure 1 maps a student’s ability to the probability they draw a cohort in which their ability falls among the top 70 percent. Figure 2 plots the equilibrium score functions that solve (8) for \( P = 0.7 \) and \( N = \{10, 100\} \).

![Probability that a Given Ability is among the Top 70% of Abilities in a Cohort](image-url)

Figure 1: Probability that a Given Ability is among the Top 70% of Abilities in a Cohort
The intuition behind these equilibrium score functions can be gathered through a thought experiment. Consider the symmetric best response function in an environment where the proportion of winners remains constant at $P = 0.7$, but the number of students in a cohort approaches infinity. While a cohort of this size will still have some variability in the draw of students, the law of large numbers ensures that the distribution of students in the cohort approaches a perfect reflection of the probability distribution from which they are drawn. Thus, common knowledge of the probability distribution is sufficient for a student’s belief about her relative position in her cohort to approach certainty.

In this infinitely large cohort, a student whose ability is greater than the 30th percentile in the probability distribution will best respond by choosing a score that no student below the 30th percentile can match and receive non-negative expected surplus. That score, given
the production and cost functions, is approximately $s_{i,t} = 0.3$. Students below the 30th percentile best respond by exerting no effort, producing a score of zero. Thus, the equilibrium score function in this setting approaches a step function that starts at $s_{i,t} = 0$ until $a_t \geq 0.3$, at which point the equilibrium score jumps $s_{i,t} = 0.3$.

Keeping in mind the limiting case, consider the equilibrium scores in Figure 2. For the 100-Student Quiz, the scores more closely reflect the infinitely large cohort, while the 10-Student Quiz scores are more affected by the randomness of the smaller cohort. This randomness manifests itself in the marginal costs and benefits of changing scores. Consider students who choose $s_{i,t} > 0$ in the limiting case. Their marginal benefit of lowering a score is constant and identical across treatments, since effort has a constant marginal cost, so foregone effort has a constant marginal benefit. The marginal cost of lowering a score is paid through reductions in the probability of receiving a high grade. In the 10-Student Quiz, that probability changes more gradually, so, at the margin, reducing a score is less costly, and the 10-Student Quiz scores drop below the 100-Student Quiz scores.

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

Equilibrium effort is deduced by inverting the equilibrium score function while holding constant student ability. My experiment pairs 10- and 100-Student Quizzes each week, and my analysis takes the difference in effort between the two quizzes as its dependent variable, so I will focus on the model’s predictions for this variable. Analyzing the difference within an individual controls for student-specific heterogeneity, and will provide a cleaner test of the treatment effect. To see the model’s predictions for how the difference in effort will evolve with ability consider Figure 3, which plots the equilibrium effort in the 100-Student Quiz minus the equilibrium effort in the 10-Student Quiz as a function of ability.

---

12 This value itself means little except as an ordinal measure of academic output.
13 $a_t = 0.3$ occurs with zero probability, so it can be included in either side of the step function.
While the function mapping the ability of a student to his or her effort at equilibrium is clearly quite complicated, the intuition behind it is rather simple. Students with lower abilities benefit from the introduction of randomness into the draw of their cohort, and exert more effort under this randomness. Conversely, students with higher abilities benefit from decreases in randomness brought about by larger cohorts. Therefore, high ability students exert more effort under the less uncertain regime.

### 3.3 Predictions From the Model

The model provides 3 primary predictions about the treatment effects displayed in Figure 3. While the numeric estimates of the model depend on too many assumptions to be informative, the generic qualities provide clarity about the ways in which an expected-grade maximizing
student reacts to changes in the grading environment.\textsuperscript{14}

**Hypothesis 1: Mean effort is increasing in cohort size**

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

**Hypothesis 2: The treatment effect crosses the axis from below at \( a_i = 0.3 \)**

My model predicts that the treatment effect crosses the horizontal axis exactly once. Call this single-crossing point \( a^* \). Figure 3 shows that the treatment effect is negative for \( a_i < a^* \) and positive for \( a_i \geq a^* \). In my experiment, \( a^* \) closely corresponds to the cutoff, \( a_i = 0.3 \), so my predictions focus on this value as \( a^* \).\textsuperscript{15}

**Hypothesis 3: The local minimum of the treatment effect is located near and before \( a^* \), and the local maximum is located near and after \( a^* \)**

Figure 3 predicts that the minimum treatment effect should exist near the cutoff and below while the maximum treatment effect should exist near the cutoff and above. The extrema identify the students for whom the relative returns to effort in one cohort size is maximally different from the corresponding returns in the other cohort size. Students with abilities near and below \( a^* \) have the greatest relative gains from the randomness of the 10-Student Quiz, while the opposite is true for students with abilities near and above \( a^* \).

\textsuperscript{14}Moldovanu and Sela (2006) prove these properties for a general class of cost functions. Each of the three predictions arise out of a single-crossing property that they demonstrate must exist for symmetric equilibria in contests with different values of \( N \), holding constant the proportion of winners.

\textsuperscript{15}While the single-crossing point in Figure 3 is not exactly equal to 0.3, the combination of the salience of the 30th percentile in my experiment and the proximity of the single-crossing point to this value make it a natural candidate.
4 Experimental Design

My experiment takes the experimental design used in Andreoni and Brownback (2014) and adapts it for a classroom context. My design simultaneously presents students with a pair of quizzes, one treated as the 10-Student Quiz and the other as the 100-Student Quiz, and records student behavior on each. The paired quiz design borrows from the paired auction design of Kagel and Levin (1993) and Andreoni, Che, and Kim (2007). I analyze the difference in behavior between the two quizzes in order to control for student-specific and week-specific effects.

4.1 Recruitment and Participation

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

4.2 Quiz Design, Scoring, and Randomization

There were 5 Quiz Weeks in the quarter. At noon on Thursday of each Quiz Week, 2 different quizzes covering material from the previous week were posted on TED, the online course website used by UC San Diego. Both quizzes were due by 5pm the following day. Each quiz had a time limit of 30 minutes, and students could take the quizzes in any order. The time limits will ensure that the time recorded for students is reflective of their effort on the quiz and includes little time spent idle.

I refer to the two quizzes each Quiz Week as Quiz A and Quiz B. All students saw the same 2 quizzes, but it was randomly determined which of the two quizzes would receive
the 10-Student Quiz treatment and which would receive the 100-Student Quiz treatment. Therefore, while every student was assigned both Quiz A and B, approximately half of them had Quiz A graded as the 10-Student Quiz and half had it graded as the 100-Student Quiz. The opposite treatment was assigned to Quiz B in each case. The questions on Quizzes A and B were designed to have as little overlap as possible to eliminate order effects in effort and scores. Before beginning the quiz, students only observed the grading treatment, and not the content of the quiz. No student was aware of the treatment received by any other student. Table 1 shows the balance across treatment.

<table>
<thead>
<tr>
<th>Quiz Version</th>
<th>Treatment</th>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 4</th>
<th>Week 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10-Student</td>
<td>201</td>
<td>282</td>
<td>258</td>
<td>253</td>
<td>249</td>
</tr>
<tr>
<td></td>
<td>100-Student</td>
<td>262</td>
<td>282</td>
<td>259</td>
<td>253</td>
<td>243</td>
</tr>
<tr>
<td>B</td>
<td>10-Student</td>
<td>267</td>
<td>280</td>
<td>262</td>
<td>251</td>
<td>243</td>
</tr>
<tr>
<td></td>
<td>100-Student</td>
<td>200</td>
<td>282</td>
<td>259</td>
<td>256</td>
<td>252</td>
</tr>
</tbody>
</table>

Note: Asymmetries across treatments may arise out of chance, failed submissions, or withdrawals.

The number of correctly answered questions determined the score for each student. The top 7 scores received high grades in each 10-Student Quiz cohort, and the top 70 scores received high grades in each 100-Student Quiz cohort. Students were randomly assigned to new cohorts each week and all students in a cohort had been assigned the same quiz and the same grading treatment.

All quizzes were graded out of 3 possible points. Students receiving high grades were awarded 3 points, while students receiving low grades were awarded 1. Non-participants received 0 points. In total, the quizzes counted for approximately 13 percent of the grade in the class. Students whose scores were tied at the 70th percentile of a cohort were all awarded 3 points unless the tied students all failed to participate, in which case the students all received 0 points.
4.3 Effort

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

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

4.4 Ability

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

Figure 4 shows the cumulative distribution function of all student GPAs in the class. I use the value of the cumulative distribution function at a given GPA to represent the ability level of that student in my predictions. Importantly, the 30 percentile GPA is 2.72. The median and mean GPA are 3 and 2.99, respectively. 7 students have GPAs of 4.0, while only 1 student has the minimum GPA of 1.0.

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

I begin this section by describing my data and their basic statistics. Then, I specify the dependent variable I will use in the analysis and test its aggregate characteristics. Next, I demonstrate heterogeneous treatment effects across students of different ability levels and test where the model does and does not hold predictive power. Finally, I explore explanations for the model’s failures.

5.1 Data and Descriptive Statistics

In total, 579 students submitted 5,094 online quizzes in this experiment. Of those, 2,546 were assigned the 10-Student Quiz treatment and 2,548 were assigned the 100-Student Quiz treatment. The duration of each quiz was recorded, and my analysis will include every recorded time. Table 2 reports the means and standard deviations of the unpaired quiz duration for both cohort sizes.

Since I did not receive GPA data until the end of the quarter, I was not able to observe
the GPAs of the students who dropped during the quarter. There were only 20 submitted pairs of quizzes from students who dropped the course. The mean treatment effect for these quizzes is approximately -36.6 seconds with a standard deviation of 432 seconds. I will include these data into the analysis of the mean treatment effect, but will exclude it from the analysis of heterogeneity in treatment effects, since I have no measure of their ability. Each of these decisions biases my results away from the model’s predictions. For the means, it lowers the average effect, diminishing my results. With respect to heterogeneity, their inclusion only strengthens my results, because they are more likely to be low GPA students, and their treatment effect is negative, on average.

5.2 Dependent Variable

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

5.3 Endogenous Selection and Controls

One confound in these results is that the order of quiz completion is endogenous. This is a limitation of the online environment. Fortunately, the order in which the quizzes are

\[^{17}\]In order to force students to take quizzes in a specific order the second quiz must be hidden from view until the completion of the first quiz. In order to ensure that students knew they were assigned two quizzes,
presented is randomly assigned and provides a relevant instrument for the order of completion that is mechanically designed to be valid as well. The effect of this endogenous selection is not large—51.4 percent of 100-Student Quizzes were presented first online, while 55.3 percent of them were completed first—but is statistically significant.

The first column of Table 3 captures the relevance of the instrument. Column 2 shows that the instrument has a problematic correlation with student GPAs. Despite being randomly assigned before each Quiz Week, the order in which the quizzes were presented happened to correlate to the GPA of the students. This is unfortunate but was unavoidable, since I did not have access to student GPAs until the end of the quarter. Additionally, the explanatory power is negligible, with an $R^2$ value below 0.002. Column 3 demonstrates that the residual effect of GPA on the order in which a student completes the quizzes is not significant once I control for the order in which the quizzes were presented. To demonstrate that none of these selection issues drive any results, all tables will feature results with and without direct controls for quiz order as well as the instrumented version of that control.

### Table 3: Testing the Relevance and Validity of the Instrument

<table>
<thead>
<tr>
<th></th>
<th>Pr{100-St. Quiz presented first}</th>
<th>Pr{100-St. Quiz presented first}</th>
<th>Pr{100-St. Quiz presented first}</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-St. Quiz presented first</td>
<td>0.759*** (0.05)</td>
<td>0.757*** (0.05)</td>
<td></td>
</tr>
<tr>
<td>GPA</td>
<td>0.105** (0.05)</td>
<td>0.060 (0.05)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-0.243*** (0.04)</td>
<td>-0.290** (0.14)</td>
<td>-0.422*** (0.15)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.064</td>
<td>0.002</td>
<td>0.065</td>
</tr>
<tr>
<td>N</td>
<td>2486</td>
<td>2486</td>
<td>2486</td>
</tr>
</tbody>
</table>

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

**Hypothesis 1: Mean effort is increasing in cohort size**

My model provides a straightforward prediction that mean effort is increasing in the cohort size. The data confirm this prediction. On average, students spend approximately 27 more

I posted them both simultaneously.
seconds on the 100-Student Quiz than on the 10-Student Quiz ($t = 2.39 \ P = 0.017$). Instrumenting for the order of completion only strengthens these results ($t = 4.28 \ P = 0.000$). When viewed relative to the means from Table 2, this difference is economically meaningful as well. A 27 second increase in time is more than a 3 percent increase over the mean.

5.4 Heterogeneity in Treatment Effects

Both Hypotheses 2 and 3 predict a systematic heterogeneity in the treatment effect on students with different abilities. This section presents the results of a semi-parametric test of heterogeneity in the treatment effect, and, in doing so, a test of the model’s ability to predict that heterogeneity. For this analysis, I first divide the data by the cutoff at the 30th percentile (corresponding to a GPA of 2.72) into a top and bottom portion. I then split each of the two portions in such a way that half of the students in each portion land in each bin. The specification of the bins are given by,

$$\text{LowestBin}_i = \mathbb{1}_{\{\text{GPA}_i \in [1, 2.433]\}}$$
$$\text{LowBin}_i = \mathbb{1}_{\{\text{GPA}_i \in (2.433, 2.72]\}}$$
$$\text{HighBin}_i = \mathbb{1}_{\{\text{GPA}_i \in (2.72, 3.246]\}}$$
$$\text{HighestBin}_i = \mathbb{1}_{\{\text{GPA}_i \in (3.246, 4]\}} .$$

Regressing the treatment effect on these 4 bins provides a semi-parametric characterization of the general patterns of student effort allocation that I will use to test Hypotheses 2 and 3. The results of this uncontrolled regression are displayed in column 1 of Table 4. Column 2 presents the same regression after including a dummy variable indicating whether the 100-Student Quiz was taken first. Column 3 repeats this, but instruments for the completion

$^{18}$Both estimates come from the constant term of a regression of the difference in duration with standard errors clustered at the student level.

$^{19}$Inserting direct controls for the order of completion is not a valid measure, since the endogenous order of completion is collinear with the treatment effect.
order using the presentation order. Notice that the coefficient of the instrumental variables regression is dramatically different than the non-instrumented coefficient, while this affects the ordering of \textit{LowestBin} and \textit{HighestBin}, they are never statistically distinguishable, so none of the following results depend on the specification chosen.

<table>
<thead>
<tr>
<th>Table 4: 100-Student Quiz Duration Minus 10-Student Quiz Duration</th>
<th>OLS</th>
<th>OLS IV Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>LowestBin $GPA_i \in [1, 2.433]$</td>
<td>-0.031</td>
<td>-2.618***</td>
</tr>
<tr>
<td>LowBin $GPA_i \in (2.433, 2.72]$</td>
<td>1.454***</td>
<td>-1.454***</td>
</tr>
<tr>
<td>HighBin $GPA_i \in (2.72, 3.246]$</td>
<td>0.515</td>
<td>-2.568***</td>
</tr>
<tr>
<td>HighestBin $GPA_i \in (3.246, 4]$</td>
<td>0.151</td>
<td>-2.724***</td>
</tr>
<tr>
<td>100-Std. Quiz Taken First</td>
<td>5.266***</td>
<td>-6.158***</td>
</tr>
<tr>
<td>Instrumented</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>N</td>
<td>2506</td>
<td>2506</td>
</tr>
</tbody>
</table>

* p<0.10, ** p<0.05, *** p<0.01
All standard errors clustered at the individual level.

Hypothesis 2: The treatment effect crosses the axis from below at $a_i = 0.3$

The 30 percentile in my experiment corresponds to a GPA of approximately 2.72. Therefore, Hypothesis 2 predicts that the mean treatment effects for students with below and above this cutoff should be negative and positive, respectively. Table 4 offers little support of this hypothesis, as the treatment effect appears to cross the axis between \textit{LowestBin} and \textit{LowBin}. A simple comparison of means between the treatment effect for $GPA \leq 2.72$ and its complement fails to reject the null hypothesis that the treatment effect is identical in the two regions ($t = -1.03 \ P = 0.304$).\footnote{I test this with a regression of the treatment effect on indicators for the two regions with standard errors clustered at the student level.} In fact, the mean difference between the 100- and 10-Student Quizzes below the cutoff, 44.4 seconds, is greater than the mean difference above
the cutoff, 20.0 seconds.

Importantly, this test rejects the location of the single crossing point, not the existence of a single crossing point. As I impose more structure on the evolution of the treatment effect across different abilities, the prediction about the existence of a single crossing point is increasingly supported by the data.

**Hypothesis 3:** The local minimum of the treatment effect is located near and before $a_i = 0.3$, and the local maximum is located near and after $a_i = 0.3$

The coefficients from Table 4 indicate that the treatment effect is maximized over $LowBin$ and minimized over $LowestBin$. The difference between the mean treatment effect found in $LowBin$ and the mean treatment effect found in its complement is approximately 71 seconds and is statistically significant ($t = 2.42 \; P = 0.016$). The difference between the mean treatment effect from $LowestBin$ and its complement is approximately 33 seconds, but is not statistically significant ($t = -1.17 \; P = 0.241$).²¹ These results contradict Hypothesis 3, which predicts that the treatment effect should be minimized in $LowBin$ and maximized in $HighBin$.

### 5.5 Treatment Effect as a Continuous Function of Ability

Imposing continuity on the evolution of the treatment effect across abilities provides an empirical counterpart to Figure 3 and allows me to test the heterogeneity of the treatment effect across students of different abilities in a more structured way. Figure 5 plots a linear polynomial smoothing function along with a 95 percent confidence interval for the treatment effect.

Figure 5 confirms the patterns indicated by the previous tests, and provides additional clarity about the evolution of the treatment effect across students with different abilities.²¹

---

²¹For each of the previous two tests, I regressed the difference in time spent on the 100-Student Quiz and time spent on the 10-Student Quiz onto an indicator variable for the relevant bin. Standard errors were clustered at the student level.
The peak of the predicted treatment effect occurs before the cutoff, where it is significant and positive, and trends downward towards 0 as a student’s ability increases. This fit also captures the negative treatment effect for low-ability students. Under the continuity restriction, the treatment effect is now statistically distinguishable from zero.

### 5.6 Perception vs Reality

Before concluding that the model holds no predictive power over the heterogeneity in the treatment effects, I will first test the model’s critical dependence on the specification of students’ beliefs. Specifically, I will test how the model’s predictions change when agents are no longer required to correctly infer their ability relative to their classmates. Classical assumptions require that students make accurate inferences based on their past academic
experiences and the selection of students expected to enroll in each class. Prior research suggests, however, that students may be “cursed” to believe that enrollment decisions are independent of ability. Without accounting for the selection of classmates, students will best respond to the previous distributions of classmates.

Using data on the distribution of grades from each UC San Diego undergraduate course over the last 5 years will allow me to estimate the extent to which students best respond to their current environment rather than their past environments. To construct a perceived distribution of classmates based on students’ past academic experiences, I aggregate the grade distributions from all lower-division courses in all departments. This generates a sort of composite grade distribution representing the likely GPAs associated with students at each percentile rank in introductory courses at UC San Diego. Since my experiment takes place in what is often the first upper-division economics course that students take, this perceived distribution is a good proxy for the level of competition students have faced thus far in their academic careers. Figure 6 plots the CDF of GPAs in this perceived grade distribution along with the CDF of GPAs from Figure 4.

Figure 6 shows how low-ability students appear to select out of this upper-division economics course. This selection effect moves the 30th percentile cutoff from a GPA of approximately 2.72 to a GPA of approximately 2.24. Cursed students fail to account for the fact the low GPA students select out of intermediate economics courses and believe their percentile rank is based on this composite distribution. Selection effects draw the realized percentile rank of each GPA in the experiment downward for low GPAs and upward for high GPAs. Thus, a cursed student with a GPA of 2.24 perceives his percentile rank to be at the cutoff of 30, while his realized percentile rank is approximately 10.

Figure 7 imposes this cursed belief structure on the strategic effort predictions displayed in Figure 3. For these predictions, students are aware of neither their own cursed beliefs, nor the cursed beliefs of others, so the predictions simply shift from one ability level to another. Notice that the data now capture all three qualitative hypotheses from the model.
The maximum of the predicted treatment effect occurs over the region that was represented by LowBin, and the minimum occurs over the region that was represented by LowestBin. Since less than 10 percent of students in my experiment possess GPAs below the modified single crossing point, the mean treatment effect is still predicted to be positive.

Hypothesis 2 predicts that the difference between the time spent on the 100- and 10-Student Quizzes is positive for students with $GPA > 2.72$ and negative for students with $GPA \leq 2.72$. One measure of the fit of this prediction is simply the number of differences with accurately predicted signs. An alternative measure would be the sum of differences with the predicted sign minus the sum of the differences without the predicted sign. Using both of these measures, I estimate the cutoff that maximizes the fit of the data and find that
Figure 7: Predicted Difference in Effort Based on Perceived Rank

The maximum occurs at a cutoff GPA of 1.81.\textsuperscript{22} This implies that, while cursed beliefs move the predicted single crossing point much closer to the observed point, there are still residual deviations that need to be accounted for.

While this is not proof of cursedness among students in my experiment, it is suggestive that, with a better understanding of the belief structure of students, models of strategic interaction in the classroom can generate useful predictions for the allocation of effort by students. Further experimentation will be required to understand the complete process of belief formation and belief updating that students undergo in a classroom setting.

\textsuperscript{22}In both cases, there is a flat maximum. The former case is maximized at GPAs of 1.81, 1.57, 1.54, 1.51, and 1.45. The latter is maximized at GPAs of 1.81 and 1.78.
5.7 Alternative Explanations

Cursedness is not the only possible explanation for the deviations from the model. In this section I address 2 competing explanations for the inconsistencies, rejecting their predictive ability.

**GPA correlates with ability to understand the incentives in the environment**

The complexity of the equilibrium effort prediction raises the concern that the lower ability students will be less able to intuit the benefits to randomness, while the higher ability students will be more able to understand the benefits to decreases in randomness. In fact, students below the cutoff are most sensitive to the treatments, and the lowest ability students respond as predicted to the increases in randomness by exerting additional effort. This shows that comprehension of the equilibrium incentives is independent on the student’s GPA.

**GPA correlates with intrinsic motivation**

Using GPA as a proxy for ability raises the concern that the students labeled high ability may be more intrinsically motivated to exert effort on quizzes than the students labeled as low ability. If higher ability students are more intrinsically motivated, they will exert similar effort on each of the two quizzes, while lower ability students may respond more to grading incentives. Since my model and a model of pure intrinsic motivation both generate the prediction that high ability students will be unaffected by the cohort size, I will need to look at lower ability students to distinguish between the two theories. A model of intrinsic motivation suggests that differences in effort are globally decreasing as intrinsic motivation becomes more important in the selection of effort. The data clearly refute this by showing that the strongest treatment effect exists for students in the middle region, meaning that sensitivity to treatments cannot be globally decreasing. Therefore, while I cannot reject that the highest types may base decisions in part on intrinsic motivation, the general patterns of behavior cannot be driven by this effect.
6 Discussion

Any discussion of effort exertion in classrooms first needs to address the social benefit of inducing greater effort on classroom quizzes. While it is not certain that additional time spent on quizzes has any return in the form of performance or learning, what can be said is that additional time spent on quizzes is a measure of costly effort exertion.\textsuperscript{23} If students do not perceive it to increase performance, then positive effort would be strictly dominated for any student with a non-zero value of time. Accordingly, it serves as an appropriate proxy for the relative amount of effort a student would assign to each type of quiz in the absence of the paired quiz design. That is, if a student is willing to spend significantly more time on one quiz than the other when they are presented simultaneously, that same student would likely be willing to study longer or attend more class in preparation for a quiz graded based on the revealed preferred cohort size in the absence of this experiment.\textsuperscript{24}

The connection between greater effort in studying or attendance and greater academic performance is well-documented (Romer, 1993; Stinebrickner and Stinebrickner, 2008; De Fraja, Oliveira, and Zanchi, 2010; Arulampalam, Naylor, and Smith, 2012), implying that classroom motivation generates positive returns. Understanding the shifting of student’s strategic effort resulting from subtle changes within a commonly applied grading scheme therefore serves to further the social goal of increasing academic output, and deserves attention as such.

Towards that end, my experiment uncovers systematic trends among students of different abilities to react to changes in the randomness of their grading environment that arise from changes in the number of students in their cohort. My experimental design addresses the effect that changing the enrollment of a course graded on the curve has on the strategic effort of students while holding constant all other characteristics of the classroom, such as

\textsuperscript{23}Indeed, statistical tests fail to reject the null hypothesis that there is no relationship between the treatment and quiz scores of students.

\textsuperscript{24}Ultimately, however, this experiment was not designed to test the connection between the motivation to spend more time taking a given quiz and the motivation to study harder before quizzes, so I must leave that connection as well-founded speculation.
teaching quality, student observability, or access to resources. With these clean controls, I can contribute to the discussion of optimal classroom size by demonstrating that changes in a student’s strategic setting causally affect effort exertion on that course.

6.1 Patterns of Effort Allocation

Three prominent stylized facts manifest themselves in my data.

First, effort is increasing in the cohort size. Thus, on average, students increase their effort as the variance of their grading environment decreases. This result confirms the prediction from my model that when the variance of a grading environment decreases, the increases in effort from high ability students more than offset the decreases in effort from low ability students.

Second, there is a marginally significant, negative effect of increasing cohort size for students with abilities farthest below the cutoff. This indicates a preference among low-ability students to exert greater effort in higher variance environments.

Third, the students most motivated by increasing the cohort size have abilities near but below the cutoff. This is contrary to the model’s prediction that these students understand that the returns to effort are higher in the 10-student cohort than in the 100-student cohort.

6.2 Misallocation of Effort

The positive treatment effect for students with abilities below the cutoff represents a misallocation of effort between quizzes. For students in the bottom 30th percentile, the higher variance environment offers a greater opportunity to receive a favorable draw for their cohort, increasing the returns to effort on the 10-Student Quizzes relative to the 100-Student quizzes for low ability students. Misallocating effort to the 100-Student Quiz increases the aggregate effort for that quiz, at the cost of foregone potential grades for lower-ability students.

While the mechanisms for this misallocation of effort are not clear from the data, it appears that students are “cursed” to believe that their classmates’ enrollment decisions
for the upper-division economics course were unrelated to their abilities. The fit of the model is greatly improved by supposing that students believe that their draw of classmates is simply an average draw from the population they experienced in lower-division courses. This adjustment, however, cannot fully explain the deviations from the equilibrium predictions. Three behavioral biases are likely candidates to explain this residual misallocation of effort.

First, a failure of students to properly update about their their relative abilities would cause students to favor one cohort size over the other. Upon receiving signals about their abilities from the results of the quizzes, students should reallocate effort based on their posterior belief about their relative ability, but it may be the case that students cling to their prior beliefs in this domain. A second possibility is that students possess a general overconfidence about their abilities. If students respond to their strategic grading environment according to an inflated estimate of their relative ability, then some students below the cutoff will favor the 100-Student Quiz, while students far enough below the cutoff will still favor the 10-Student Quiz. Both of these biases are consistent with existing results about updating about perceived intelligence (Eil and Rao, 2011; Mobius et al., 2011).

A third explanation is reference dependent utility, specifically myopic loss aversion (Benartzi and Thaler, 1995). If students consider each quiz independently, then they will overexert on quizzes with lower returns to their effort. Recall that increasing effort only increases the probability of high grades and does not alter those grades themselves. As Sprenger (2010) points out, loss aversion over probability can only arise under expected value based reference points as in disappointment aversion (Bell, 1985; Loomes and Sugden, 1986; Gul, 1991), because under stochastic reference points in the style of Kőszegi and Rabin (2006, 2007) students are risk neutral over changes in probability within the support of the original gamble. If a student is risk neutral over changes in probability, then she would exert more effort on the quiz that led to the greatest increases in her probability of gains. Under expected value based reference points, a student would exert additional effort on the quiz that was considered to be “losing” relative to her reference point, in the case of students near but
below the cutoff, this would be the 100-Student Quiz.

6.3 Policy Prescriptions

Based on the positive average treatment effects, a mechanism designer with preferences over aggregate effort should implement a grading environment with the lowest possible variance to maximize the total effort exerted. This result could be accomplished through combining multiple classes into one grading unit and compensating for classroom-level differences.

While decreasing the variance increases aggregate effort, it is not without certain costs. For the lowest ability students, the higher variance environment induces more effort. The intuition for this result relates back to Figure 1, which shows how increases in the number of students in a cohort make it increasingly unlikely that a low-ability student receives a high grade. From a policy-perspective, this result suggests that low-ability students can become discouraged by relative grading when the cohort size becomes large enough. If motivating low-ability students is part of an educator’s objective, then smaller cohorts can accomplish this. This could be achieved by splitting large classes up into smaller sections and grading each individually.

The results also show that lower variance environments induce greater effort from many students with abilities below the cutoff. While the data do not explain the origin of this allocation failure, it still generates several policy prescriptions. If the designer hopes to motivate students below the cutoff on average, then decreasing the variance of the grading environment appears to achieve that goal. This change, however, comes at the cost of students at the lowest end of that spectrum, so any mechanism designer will need to balance gains from the former group with losses from the latter.

My results identify several failures of the model to capture all of the relevant phenomena. Students appear to hold systematically inaccurate beliefs about their classmates, themselves, or both, causing them to misallocate effort to tasks or choose their courses sub-optimally. In these cases, feedback about relative ability could increase student utility, but may reduce
aggregate effort, since the misallocation of effort by students below the cutoff added to the aggregate effort in the 100-Student Quiz. This may give rise to circumstances where a student over-exerts effort as a result of a bias, and the instructor actually wishes to encourage this over-exertion.

7 Conclusion

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

In order to ground the intuition for why class size may affect student effort choices, I first develop a theoretical model of the situation to generate an understanding of the ways in which students respond to shifts in the class size. My experiment tests these qualitative predictions and measures the causal impact of cohort size on effort.

My results highlight an important tension between mean effort and the distribution of effort. I confirm that students do indeed identify the changes in the environment that result from shifting the number of students in a cohort. The mean effort exerted increased significantly with the cohort size. Effort among low ability students, however, showed marginally significant decreases with the cohort size. Meanwhile, students with abilities near but beneath the cutoff exert significantly more effort under the larger cohort. This suggests that smaller classes promote effort from low-ability students as a result of the increase in variance, and larger classes promote effort from students who believe themselves to be within reach of the cutoff.

The misallocation of effort by students may not be completely atheoretical, as it is consistent with well-documented behavioral biases, such as cursedness, overconfidence, non-
Bayesian updating, and reference dependence. Further experimentation is needed to confirm or reject these theories, however.

My results make it clear that the relative grading mechanisms currently in place generate many unintended consequences as class size changes. This information can serve to identify the different demographics who are put at risk by the grading mechanisms and provide the basis for exploration of optimal grading mechanisms that find the desired balance between greater mean effort and the distribution of effort among students.

References


