

# A Field Experiment on Effort Allocation under Relative Grading

Andy Brownback

November 27, 2014

## **Abstract**

Grading on the curve is one of the most commonly used evaluation mechanisms in education. Under this mechanism, grades are assigned based on each student's percentile rank in the class. Considering classes to be samples from the population implies that as the class size grows, the percentile ranks of the students in it draw closer to their percentile ranks in the population, which changes the students' incentives. I model this environment to predict how changes in the class size reallocate incentives for effort between students with different abilities, an effect that holds true in any relative evaluation method. I use a field experiment in an intermediate economics course to test the model's predictions about effort exertion in a real-stakes environment. My results show that the lower variance of larger classes elicits greater mean effort and greater effort from all but the lowest-ability students. The greater variance of smaller classes elicits more effort from only the lowest-ability students. Many low-ability students fail to take advantage of the randomness of the smaller class size, an allocation failure consistent with "cursed" beliefs about their classmates and other behavioral biases. My results shed new light on the debate over the effects of class size reductions.

# 1 Introduction

Relative evaluation mechanisms are often used to mitigate the effects of asymmetric information between the mechanism designer and the agents. Under complete information about agents' cost functions and their valuations for prizes, a mechanism designer could set absolute thresholds that extract the full surplus from all agents who receive a prize. Under incomplete information, however, designers may prefer relative evaluation mechanisms, since a poorly calibrated absolute mechanism offers little incentive for effort.

Consider, for example, a teacher who wishes to maximize effort from his students. Setting an absolute threshold for each grade that is too high or too low could cause students to exert minimal effort. This teacher may remedy his incomplete information about his students' abilities by grading on the curve. Under a relative or "curved" grading mechanism, the evaluation of a student's performance is based on her percentile rank within her comparison group independent of any absolute measure of performance. These mechanisms have become a fixture in many university classrooms and law schools.<sup>1</sup> Indeed, mechanism designers across many areas of education employ relative evaluation in competitions for scholarships, college admissions, and even teacher pay.<sup>2,3,4</sup>

Under relative evaluation, the composition of an agent's comparison group is critical in determining the agent's outcome at a given level of effort. I will refer to the sample drawn from the population into a comparison group as a "cohort." The law of large numbers implies that the larger this cohort becomes, the more it comes to resemble the distribution from which it is drawn. Since the composition of a cohort determines the incentives for effort by an agent, the size of a cohort affects those incentives by bringing its expected

---

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

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

<sup>3</sup>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 admission to the state college of their choice. In Texas, the top 10% are offered similar incentives.

<sup>4</sup>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.

composition closer to the population distribution. Therefore, in theory, the incentives for effort under any relative evaluation mechanism can be modeled and understood as a function of cohort size. A mechanism designer concerned about the allocation of incentives to agents could then equalize incentives across any two cohort sizes by manipulating the number or value of prizes awarded. In practice, however, mechanism designers typically operate under constraints restricting them to a given number or percentage of prizes awarded regardless of cohort size. This paper will use the framework provided by the law of large numbers to evaluate the theoretical and empirical effect of cohort size on incentives for effort in a relative evaluation mechanism under such a constraint.

Consider an example, Texas HB 588 grants automatic admission to any Texas state university—including the University of Texas at Austin—to all Texas high school seniors who graduate in the top 10 percent of their high school class.<sup>5</sup> Each graduating class is therefore subject to a distinct relative evaluation mechanism with the restriction that each mechanism must award 10 percent of the students automatic admission, regardless of the class size. Since Texas high schools vary in size by orders of magnitude,<sup>6</sup> the law of large numbers implies that the incentives for effort under this policy will vary dramatically. Smaller high schools are more likely to draw graduating classes full of outliers where the returns to effort are more uncertain, while larger high schools are more likely to have cohorts of students whose characteristics more closely reflect the characteristics of the population, reducing the uncertainty around the returns to effort.

While it may seem obvious that incentives change with the size of a cohort, prior research casts doubt on the ability of economic agents to draw accurate inference about information that depends critically on sample size (Tversky and Kahneman, 1971; Kahneman and Tversky, 1973; Rabin, 2002; Benjamin, Rabin, and Raymond, 2014). Therefore, research is necessary to uncover to what degree agents identify and respond to these shifts in incentives

---

<sup>5</sup>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.

<sup>6</sup>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.

for effort.

This paper uses an experimental intervention in the size of a cohort to test its causal effect on effort by students who are graded on the curve. Specifically, I measure how student effort changes as a result of changes in their strategic environment that originate from the size of their cohort. In order to provide the most powerful test of this effect, I use treatments that only vary the size of the cohort while holding constant all other classroom factors. Additionally, I provide a measure of the importance of strategic incentives relative to other motivations for effort from students.

I use a field experiment on relative grading in a large, upper-division economics course at the University of California, San Diego (UCSD) to test the sensitivity of student effort to cohort size. I present students with a pair of quizzes, each graded such that the top 70 percent of scores in a cohort receive high grades. For each student, I randomly determine which quiz in the pair will be graded relative to a 10 student cohort and which will be graded relative to a 100 student cohort. Since cohort size is randomly assigned to quizzes, I will refer to these as the “10-Student Quiz” and the “100-Student Quiz.” I measure effort as the time a student spends on a given quiz. By comparing a pair of quizzes taken in quick succession but with different cohort sizes, I can use the difference in the amount of time spent on the two quizzes to directly measure the causal impact of cohort size on effort. This test cleanly identifies the influence of cohort size independent from any other motivations for effort. Additionally, my experiment determines which of the assumptions of the model may need to be adjusted before it can reliably predict effort allocation patterns in the classroom.

In order to generate predictions about the ways in which an expected-grade maximizing student under curved grading would respond to changes in the cohort size, I develop a theoretical model of strategic effort exertion in the classroom. Strategic concerns are just one of countless motivations a student may have to exert effort in the classroom. My results address the relevance of these other motivations per se and allay concerns about their potential to interact with the strategic incentives. I therefore abstract away from these other

motivations and develop predictions based on expected-grade maximizing responses to the underlying strategic incentives of the curved grading environment. In this simplified model, I make several assumptions about students' utility functions, their beliefs about their own ability and the abilities of their classmates, their responses to those beliefs, and the separability of their cost functions. As such, its value is in generating qualitative predictions about the allocation of effort by strategic students responding to the changes in the size of their grading cohorts. Indeed, from a policy perspective, it is the empirical effect of the cohort size on effort allocation that is paramount. Confirming or rejecting a given model of strategic effort will always be a secondary concern.

My model predicts that mean effort will increase with the size of the cohort. This results from decreases in the uncertainty about the returns to effort. That is, as the cohort size grows and the characteristics of the cohort draw closer to the population characteristics, the uncertainty surrounding the returns to any given level of effort decrease, drawing up the mean effort.

My model also provides structure for how incentives may differ by ability level. In particular, incentives diverge on either side of the the 30th percentile of scores, where high and low grades are separated. I refer to the quantile that distinguishes the two outcomes as the "cutoff." My model predicts that students with ability levels below the cutoff will exert more effort on the 10-Student Quiz, where the increase in uncertainty actually benefits their expected returns to effort. Students with abilities above the cutoff are predicted to exert more effort on the 100-Student Quiz, since decreases in uncertainty benefit their expected returns to effort. These two predictions imply that the effort functions cross each other once, and near the cutoff. The model makes a related prediction that the maximum and minimum differences between effort in the 100- and 10-Student Quizzes occur above and below the cutoff, respectively. These heterogeneous effects highlight the tradeoffs between mean effort and the distribution of effort, a key tension for optimizing classroom design.

I first confirm the prediction that the mean effort will increase in the cohort size. The 100-

Student Quizzes elicit 5 percent more effort than the 10-Student Quizzes, and this difference is significant at the 1 percent level. I go on to test the model's predictions about the heterogeneous impact of cohort size on students with different abilities by using GPA as a proxy for ability. My experimental results deviate from the model's predictions in important ways. First, while the *lowest* ability students exert more effort on the 10-Student Quizzes, this effect is not found for all students below the cutoff. In fact, *mean* effort by students below the cutoff is higher on 100-Student Quizzes. Second, the maximum difference in effort in favor of the 100-Student Quiz occurs at ability levels below the cutoff, not above it. For these students, allocating greater effort to the 100-Student Quiz fails to take advantage of the higher variance of the smaller cohort. These deviations from the theoretical predictions highlight the limits to the predictive ability of a purely strategic model of classroom effort. Specifically, a strategic model under classical assumptions about beliefs and best responses will overstate the distributional costs of increasing mean effort. That is, effort from low ability students is less negatively affected by increasing cohort sizes than theory predicts.

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

for selection into the course.

My results suggest that cursed beliefs are not the only potential complication in evaluating strategic effort exertion. Allowing for “fully cursed” beliefs<sup>7</sup> shifts the predicted single crossing point towards lower abilities and increases the number of students predicted to exert greater effort on the 100-Student Quiz but still cannot explain the behavior of many low ability students. The misallocation of effort by these students is consistent with several possible behavioral biases, among them, overconfidence, updating failures, and reference dependence. My experiment cannot independently identify these possible behavioral responses, but I discuss the impacts of each and outline experiments that could identify the extent to which each is responsible for deviations from the theoretical predictions.

The data can cleanly reject several alternative explanations for the results. A model where high-ability students have greater intrinsic motivation provides some similar predictions about how incentives for effort evolve with student ability. Intrinsic motivation may be present as an incentive for effort, but without allowing for sensitivity to the strategic incentives present, this model cannot generate the patterns of effort observed in my experiment. In a similar way, my data can reject the notion that risk preference or demographic characteristics can explain the results entirely. I can also reject the hypothesis that effort allocation is driven by a correlation between GPA and the ability to intuit the strategic incentives of the environment. Indeed, my paper provides clean evidence that students’ responses to changes in their strategic environment are largely predictable by models of purely strategic effort, despite many other observable and unobservable motivations for effort.

To fix ideas, I refer to grading mechanisms throughout this paper, but this should not distract from the generality of the results. Relative awarding mechanisms are found throughout the modern economy, in job promotion contests, performance bonuses, and lobbying contests, among others.<sup>8</sup> The qualitative predictions of models of strategic effort, therefore,

---

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

<sup>8</sup>For example, in his book *Straight from the Gut*, former GE CEO Jack Welch recommends that managers rank employees according to a 20-70-10 model of employee vitality where 20% of employees are labeled “A”

provide useful intuition in these settings. Since the costs of effort, the means of exerting it, and the ways in which heterogeneous abilities manifest themselves are all similar in academic and professional settings, my results may be useful in predicting how certain mechanisms will affect the allocation of effort across employees.

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

## 2 Literature

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

Experiments testing effort exertion in different laboratory settings date back to Bull, Schotter and Weigelt (1987), who test bidding in laboratory rank-order tournaments. They find that bidders approach equilibrium after several rounds of learning. Equilibrium behavior in laboratory all-pay auctions is more elusive with the majority of studies demonstrating overbidding (Potters, de Vries, and van Winden, 1998; Davis and Reilly, 1998; Gneezy and players, 70% “B” players, and 10% “C” players. “A” players are rewarded, and “C” players are eliminated.

Smorodinsky, 2006; Barut, Kovenock, and Noussair, 2002). Müller and Schotter (2010) and Noussair and Silver (2006) confirm the overbidding result, but also uncover heterogeneous effects for different types of bidders. For an exhaustive survey of the experimental literature on contests and auctions, refer to Dechenaux, Kovenock, and Sheremeta (2012).

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

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

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

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

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

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

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

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

The education literature traditionally studies class size like an input to the production function. The aforementioned Mosteller (1995) paper operates in this vein, finding that

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

### **3 A Model of Academic Effort**

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

Strategic incentives will be present amidst myriad other incentives for effort. As such, the contribution of this model is in the structure that it gives to the relative incentives for effort, not the point estimates it provides for the amount of effort each student chooses. Additionally, the heterogeneity this model predicts relies on several assumptions about the beliefs and information that each student has. I test some of these assumptions alongside the predictions model in order to evaluate the model's predictive ability with and without the assumptions. This model also provides a characterization for the way in which strategic incentives interact with alternative motivations for effort. With this structure, my experiment can test these motivations individually.

In solving this model, I borrow heavily from the independent private value auction (Vickrey, 1961) and the all-pay auction literatures (Baye, Kovenock, and de Vries, 1993).

Suppose there are  $N$  students exerting costly effort in order to increase their chances of winning one of  $M \equiv P \times N$  prizes in the form of high grades. Effort appears as scores on quizzes, and high grades are awarded to the students with the highest  $M$  scores.

Suppose each student has an ability,  $a_i$ , distributed uniformly from 0 to 1. That is  $a_i \sim U[0, 1]$ , meaning  $F(a_i) = a_i$ . Students are evaluated at each period,  $t$ , based on their academic output, or “score,”  $s_{i,t}$  from that period. Suppose that scores have a constant marginal cost that is inversely related to ability,<sup>9</sup>

$$C(s_{i,t}; a_i) = \frac{s_{i,t}}{a_i}. \quad (1)$$

### 3.1 Student’s Utility

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

$$U(s_{i,t}; a_i) = \Pr(s_{i,t} \geq \bar{S}) - \frac{s_{i,t}}{a_i}, \quad (2)$$

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

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

---

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

in ability.<sup>10</sup> All continuous, monotonic functions are invertible, so there exists a function that maps a given score back onto the ability implied by that score. Given that the equilibrium scores depend on the parameters,  $N$  and  $P$ , this inverse function, too, depends on these parameters. 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. This probability is represented as an order statistic involving the CDF of  $a_i$ . Substituting in the experimental parameters,  $N = \{10, 100\}$  and  $P = 0.7$ , yields the order statistics presented in Figure 1.

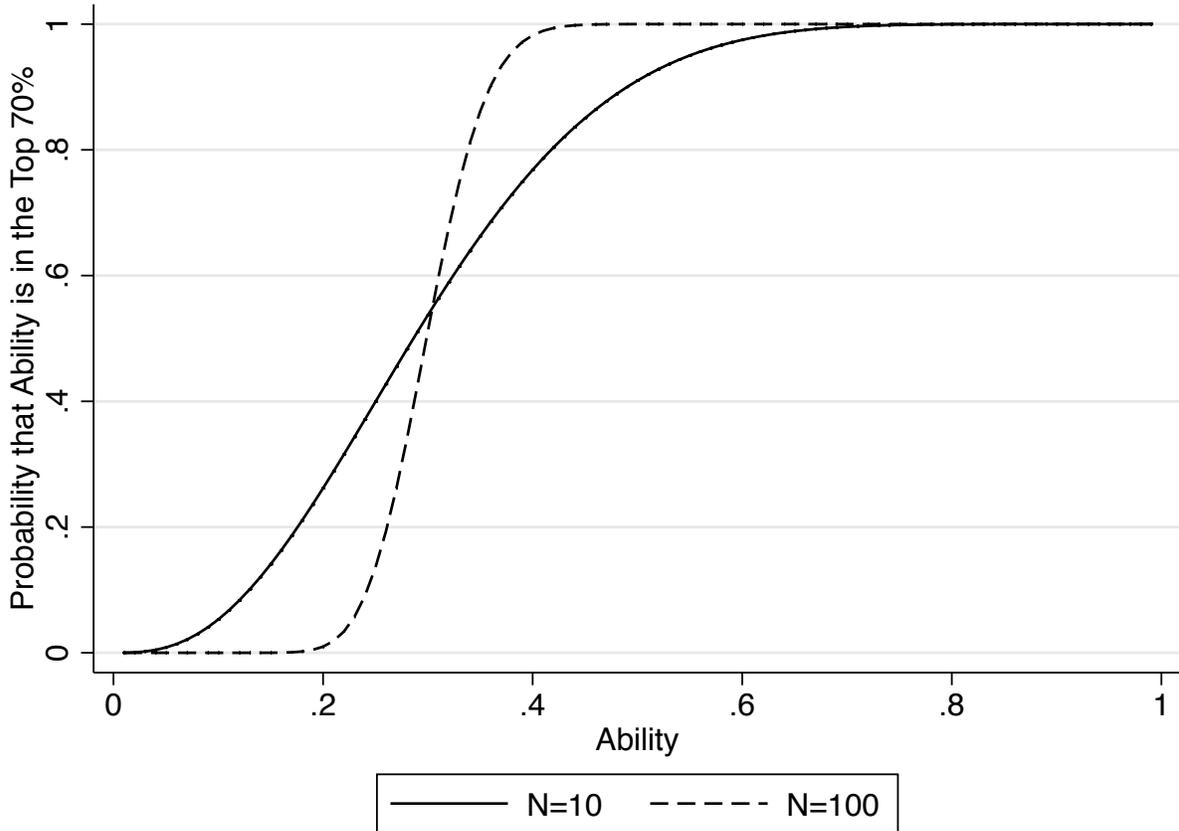


Figure 1: Probability a Given Ability is in the Top 70% of Abilities in a Cohort of Size  $N$

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

Figure 1 reveals the key intuition of this paper. For low-ability students, the uncertainty of the 10-Student Quiz increases their likelihood of encountering cohorts in which they are among the top 70 percent. For high-ability students, that same uncertainty decreases the likelihood that they are among the top 70 percent of their cohort. These probabilities represent variation in the expected returns to student effort. Importantly, this variation heterogeneously affects students based on their ability.

Plugging the order statistics into (2) completes the student's utility function.

$$U(s_{i,t}; a_i, N, P) = \sum_{j=N-NP}^{N-1} \left( \frac{(N-1)!}{j!(N-1-j)!} \right) A(s_{i,t}; N, P)^j \times (1 - A(s_{i,t}; N, P))^{N-1-j} - \frac{s_{i,t}}{a_i}. \quad (3)$$

In the appendix, I solve for the equilibrium scores as a function of ability. Figure 2 plots the equilibrium score functions at the experimental parameter values,  $N = \{10, 100\}$ . It is worth noting that the point estimates represented are not valuable per se, but only for their representation of the relative scores in each treatment.

While the function mapping the ability of a student to his or her score at equilibrium is clearly quite complicated, the intuition behind it is rather simple. Students with lower abilities benefit from the introduction of randomness into the draw of their cohort, and put forth greater 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.

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

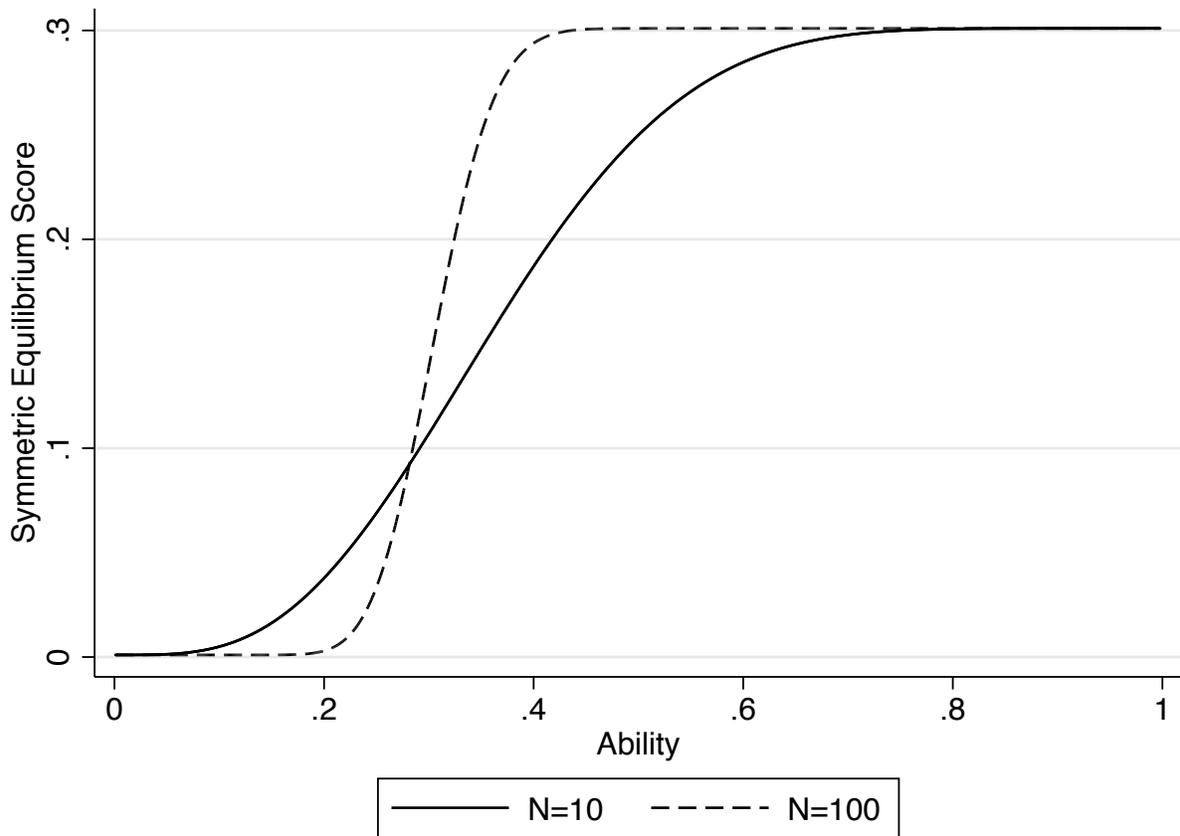


Figure 2: Scores at Equilibrium in the 10- and 100-Student Quizzes

Thus, common knowledge of the probability distribution is sufficient for a student’s belief about her relative position in her cohort to approach certainty.

In this infinitely large cohort, a student whose ability is greater than the 30th percentile in the probability distribution will best respond by choosing a score that no student below the 30th percentile can match and receive non-negative expected surplus. That score, given the assumed cost function, is approximately  $s_{i,t} = 0.3$ .<sup>11</sup> Students below the 30th percentile best respond by 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  $a_i \geq 0.3$ , at which point the equilibrium score jumps to  $s_{i,t} = 0.3$ .<sup>12</sup>

Keeping in mind the limiting case, consider the equilibrium scores in Figure 2. For the

<sup>11</sup>This value itself means little except as an ordinal measure of academic output.

<sup>12</sup> $a_i = 0.3$  occurs with zero probability, so it can be included in either side of the step function.

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. Consider the marginal costs and benefits of adjusting scores from their infinite-cohort equilibrium. For students who choose  $s_{i,t} > 0$  in the limiting case, the marginal benefit of lowering a score is constant and identical across treatments, since scores have a constant marginal cost, so foregone scores have a constant marginal benefit. The marginal cost of lowering a score is paid through reductions in the probability of receiving a high grade. In the 10-Student Quiz, that probability changes more gradually, 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 scores bears a constant marginal cost for both the 10- and 100-Student Quizzes, but holds a higher marginal benefit in the 10-Student Quiz because the randomness increases the likelihood of states of the world in which low scores receive high grades. So, the 10-Student Quiz scores rise above the 100-Student Quiz scores.

### 3.2 Predictions From the Model

My experiment pairs 10- and 100-Student Quizzes each week, and my analysis takes the difference in effort between the two quizzes—specifically, the 100-Student Quiz duration minus the 10-Student Quiz duration—as its dependent variable. I refer to this difference as the “treatment effect.” I use the model’s predictions for the difference in scores as a proxy for its predictions about the difference in effort. This allows me to remain agnostic about the production function for scores, only assuming that higher predicted scores imply higher predicted effort. My within-subjects analysis of this difference in effort controls for student-specific heterogeneity, and will provide a cleaner test of the treatment effect. To see the model’s predictions for how the treatment effect will evolve with ability consider Figure 3, which plots the equilibrium score in the 100-Student Quiz minus the equilibrium score in the 10-Student Quiz as a function of ability.

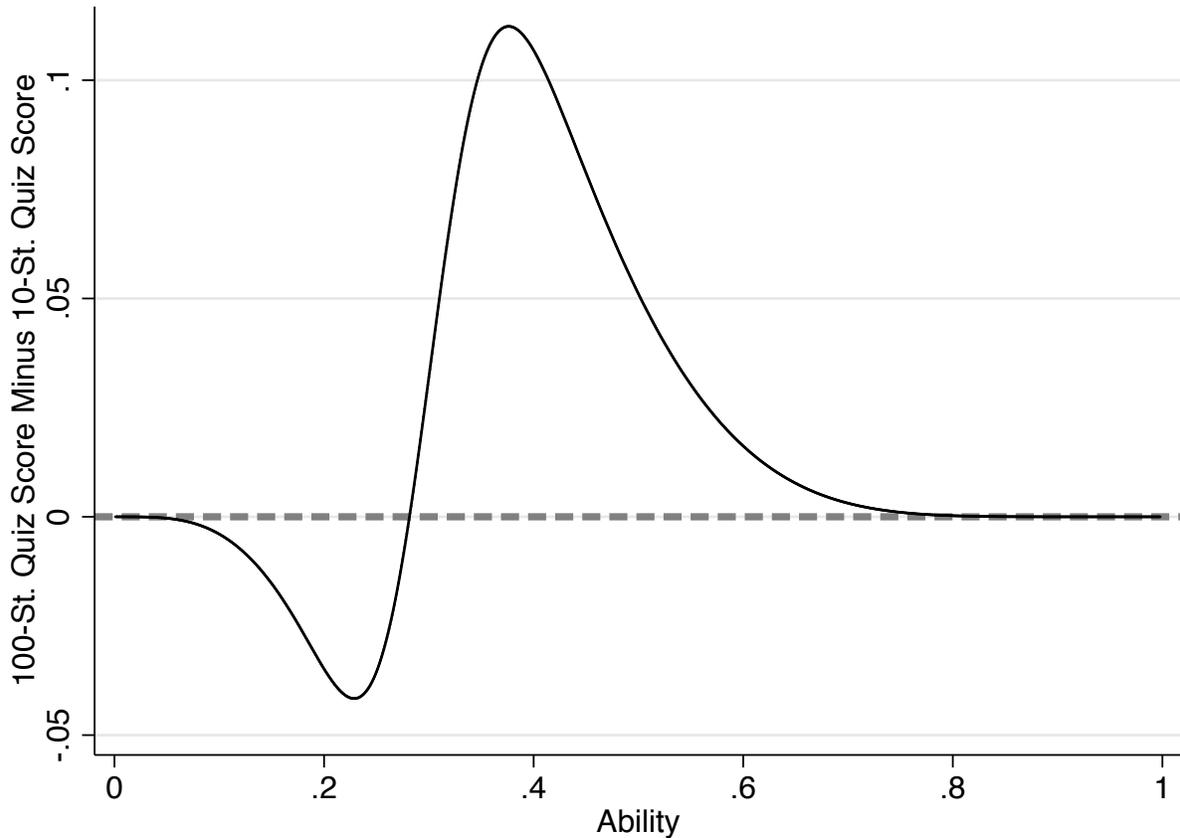


Figure 3: Difference in Score between 100- and 10-Student Quizzes at Equilibrium

The model provides three primary predictions about the treatment effect displayed in Figure 3. While the point estimates of the model are based on specific assumptions, the following predictions represent generic qualities that provide clarity about the ways in which an expected-grade maximizing student may react to changes in the grading environment.<sup>13</sup>

**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.

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

**Hypothesis 2: The treatment effect crosses the axis once and from below.**

My model predicts that the treatment effect begins negative and moves positive, crossing the horizontal axis exactly once. Call this single-crossing point  $a^*$ . Based on Figure 1, it is natural to think of  $a^*$  as corresponding with the cutoff. This is generally true, but the specific location of  $a^*$  will depend on the distribution of abilities, and the cost function for scores. My predictions focus on the cutoff of  $a_i = 0.3$  as  $a^*$ .<sup>14</sup>

**Hypothesis 3: The local minimum of the treatment effect is located below the cutoff, and the local maximum is located above the cutoff.**

Figure 3 shows the extrema of the treatment effect. These 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. These returns are closely tied to Figure 1, where it is clear that the difference between the order statistics is minimized below the cutoff and maximized above it. Accordingly, students with abilities below the cutoff have the greatest relative gains from the randomness of the 10-Student Quiz, while the opposite is true for students with abilities above the cutoff.

## 4 Experimental Design

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

I analyze the difference in behavior between the two quizzes in order to control for student-specific and week-specific effects. Importantly, this paired design will provide a powerful test

---

<sup>14</sup>While the single-crossing point in Figure 3 is not precisely 0.3, the salience of the 30th percentile in my experiment and the proximity of the single-crossing point to this value make it a natural candidate.

of responses to cohort size that occur in an environment riddled with alternative motivations for effort. Indeed, the design is powerful enough to independently test the interaction between many of these motivations and the treatments.

## **4.1 Recruitment and Participation**

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

## **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 to the course website. 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 30-minute limit ensures that the quiz was given focused attention with little time spent idle, meaning that the time recorded for students is reflective of their effort on the quizzes.

I refer to the content of the two quizzes in a Quiz Week as “Quiz A” and “Quiz B.” All students saw these 2 quizzes in the same order, but I randomly assignment grading treatments to the quizzes. One of the quizzes received the 10-Student Quiz treatment and one received the 100-Student Quiz treatment. Therefore, while every student was assigned both Quizzes A and B, approximately half of them had Quiz A graded as the 10-Student Quiz and half had it graded as the 100-Student Quiz. The opposite treatment was assigned to Quiz B in each case. The questions on Quizzes A and B were designed to have as little overlap as possible to eliminate order effects in effort and scores. Before beginning the quiz, students only observed the grading treatment, and not the quiz content. No student was informed of

the treatments received other students. Table 1 shows the balance across treatments.

Table 1: Submitted Quizzes from Each Week and Each Treatment

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

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

The number of questions correct determined the score for each student. The top 7 scores received high grades in each 10-Student Quiz cohort, and the top 70 scores received high grades in each 100-Student Quiz cohort. Students were anonymously re-randomized into cohorts each week. All students in a cohort had taken the same quiz under the same grading treatment. Students receiving high grades were awarded 3 points, while students receiving low grades were awarded 1. Non-participants received 0 points. The quizzes counted for approximately 13 percent of the total grade in the class, providing strong incentives for students to participate seriously. 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 time at which every quiz was started and completed was recorded to the millisecond. My analysis will take the amount of time that a student spent taking a quiz to be the measure of effort that the student exerted on that quiz. This measure of effort will reveal which quiz the student believed to hold the greater returns to her effort.

Both quizzes were posted simultaneously, meaning that the amount of time a student could spend studying prior to starting either quiz was roughly constant between the two

quizzes. Student behavior supports this assertion. 86 percent of students waited less than an hour between the two quizzes, with a median interval of 32 minutes between quizzes.

## 4.4 Ability

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

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

## 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 investigate possible explanations for the model's failures.

---

<sup>15</sup>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.

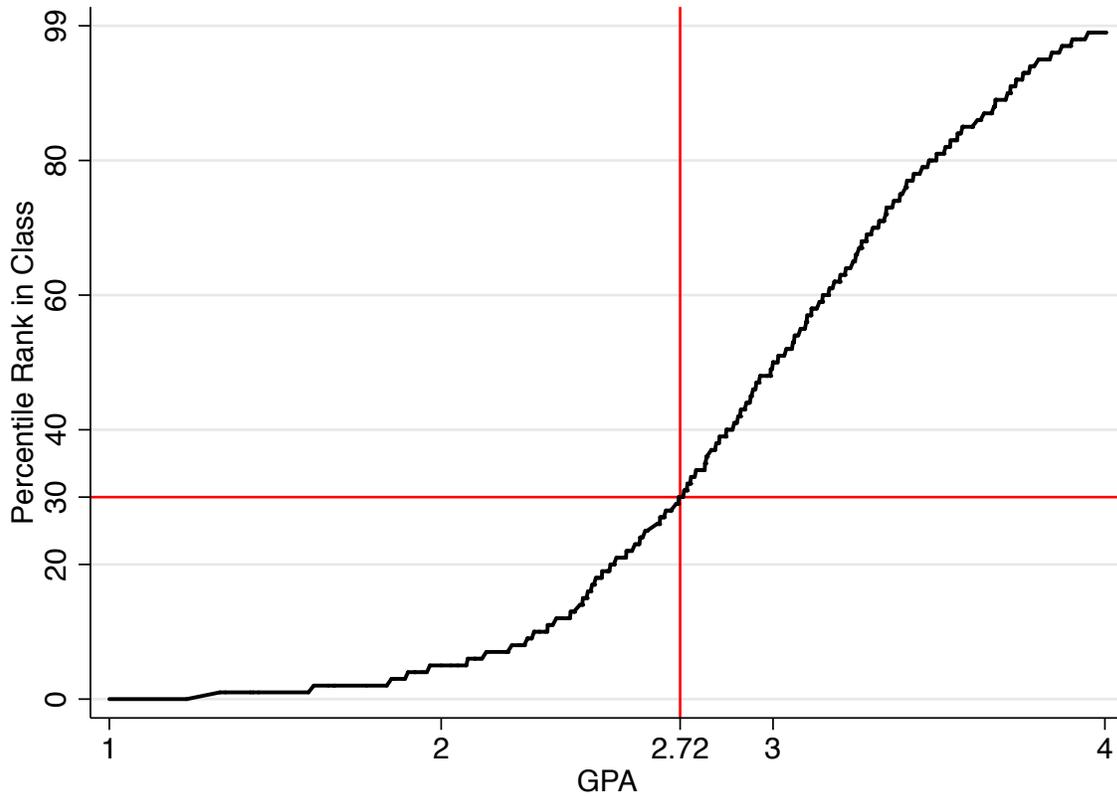


Figure 4: Cumulative Distribution Function for GPAs of All Students in the Experiment

## 5.1 Data and Descriptive Statistics

In total, 579 students submitted 5,094 online quizzes in this experiment. Of those, 2,546 had been assigned to a 10-Student Quiz, and 2,548 had been assigned to a 100-Student Quiz. The duration of each quiz was recorded, and my analysis will include every recorded time appearing in a completed pair of quizzes. Table 2 reports the means and standard deviations of the unpaired quiz duration for both cohort sizes.<sup>16</sup>

<sup>16</sup>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 -36.7 seconds with a standard deviation of 432 seconds. I include these data in the analysis of the mean treatment effect, but exclude them from the analysis of heterogeneity in treatment effects, since I have no measure of ability. Each of these decisions biases my results away from the model's predictions. For the means, it lowers the average effect, diminishing my results. With respect to heterogeneity, their inclusion only strengthens my results, because they are more likely to be low GPA students, and their average treatment effect is negative.

Table 2: Descriptive Statistics on the Duration of the Quizzes

	100-Std. Quiz	10-Std. Quiz
Mean Duration ( <i>in minutes</i> )	14.97	14.58
Standard Deviation/Error	(9.21)	(9.00)

Note: In this table, the quiz means are unpaired, so will provide a much weaker test of significant differences.

## 5.2 Dependent Variable

My analysis uses the difference between the time allocated to the 100- and 10-Student Quizzes as the dependent variable. Recall that I refer to this difference as the treatment effect. This dependent variable is appealing because it reveals a student’s beliefs about which quiz will yield higher returns to her effort. Since random assignment leaves effort costs and quiz difficulty independent of the cohort size, if a student shows a general trend toward spending more time on the 10- or 100-Student Quiz, then the student must believe that her marginal product is higher on that quiz.<sup>17</sup> 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 potential complication in these results is that the order of quiz completion could not be controlled and thus is endogenous. This is a limitation of the online environment.<sup>18</sup> Fortunately, even though quizzes are presented simultaneously, one quiz is positioned above the other vertically. This presentation order is randomly assigned and provides a relevant instrument for the order of completion that is mechanically designed to be valid. The effect of this endogenous selection is not large—51.4 percent of 100-Student Quizzes were presented first online, while 55.3 percent were completed first—but is statistically significant.

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

<sup>18</sup>In order to force students to take quizzes in a specific order the second quiz must be hidden from view until the completion of the first quiz. I posted both quizzes simultaneously in order to ensure that students knew they were assigned two quizzes.

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

Table 3: Testing the Relevance and Validity of the Instrument

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

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

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

### Hypothesis 1: Mean effort is increasing in cohort size.

Table 4 tests my model's straightforward prediction that mean effort is increasing in the cohort size. Column 1 shows that, on average, students spend approximately 27 more seconds on the 100-Student Quiz than on the 10-Student Quiz, an increase of 3 percent over the mean. With endogenous selection of ordering, however, this number is confounded by the effect of quiz order on quiz duration. Instrumenting for the order of completion removes

this endogeneity and provides a clearer picture of the treatment effect, showing that the 100-Student Quiz elicits 46.4 seconds more effort from students, an increase of more than 5 percent over the mean.<sup>19</sup>

Table 4: Mean Difference between 100- and 10-Student Quiz Duration

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

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

All values are reported in minutes.

All standard errors clustered at the student level.

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

## 5.4 Heterogeneity in the Treatment Effect

My model predicts heterogeneity in the treatment effect across ability levels. In this section, I impose continuity on the treatment effect in order to understand how it evolves across different abilities. This continuous evolution of the treatment effect across abilities provides an empirical counterpart to Figure 3 and tests if the observed heterogeneity of the treatment effect coincides with the predictions of the model.<sup>20</sup> Figure 5 plots a locally linear polynomial smoothing function along with 95 percent confidence intervals for the treatment effect.

It is important to note that the complicating influence of other incentives for effort will be most apparent when considering the heterogeneity of the treatment effect. After testing the specific predictions of the model, I test for several of these effects. Deviations from the theory may manifest themselves through biased beliefs, correlations between intrinsic motivation

<sup>19</sup>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.

<sup>20</sup>I include a less parametric test of heterogeneity in the appendix, showing that the results are not driven by the continuity restriction.

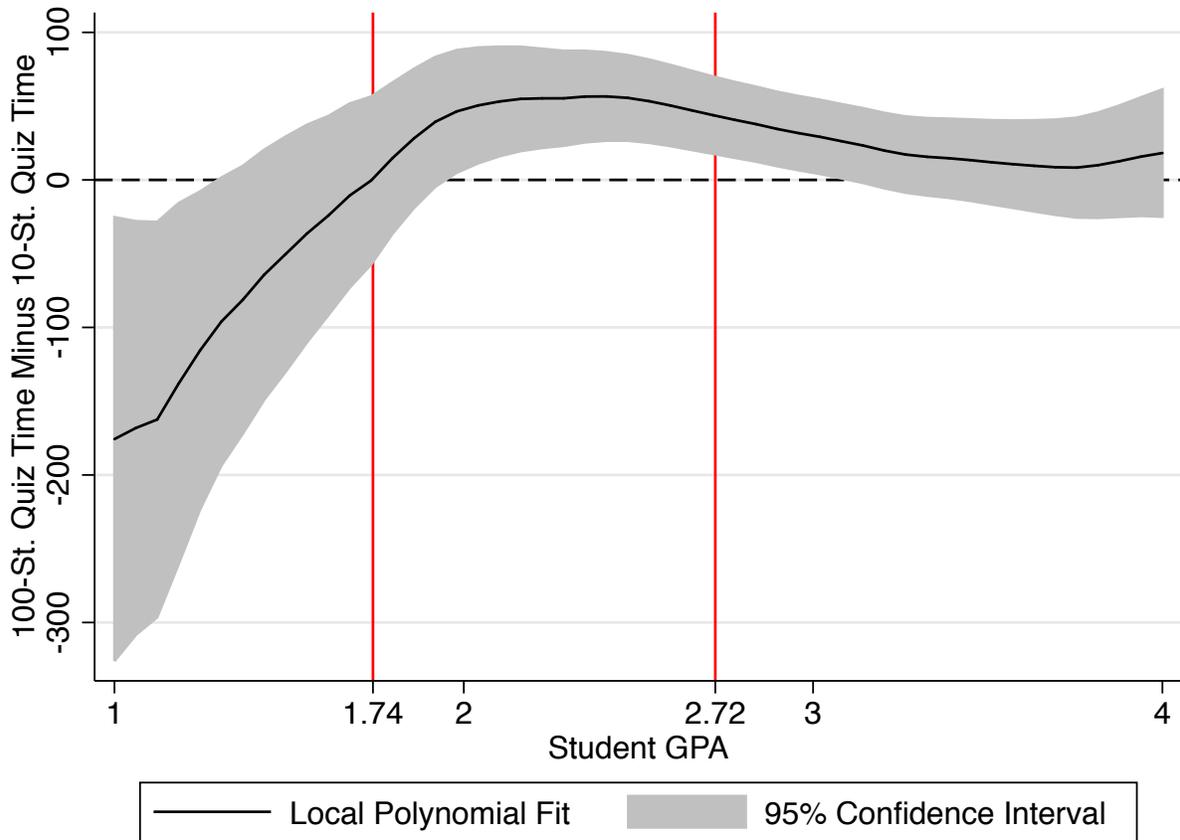


Figure 5: Local Polynomial Fit of the Difference in Effort between 100- and 10-Student Quizzes (Times Reported in Seconds)

and GPA, heterogeneous comprehension of the strategic incentives, or correlations between demographic characteristics and GPA. I address each of these in turn.

**Hypothesis 2: The treatment effect crosses the axis once and from below.**

To consider the single-crossing nature of the treatment effect, I examine its evolution beginning with the lowest-ability students. For the left tail of the distribution of GPAs, there is a strong negative treatment effect that shows statistical significance. This tail represents a small percentage of the population, approximately 2 percent, but an economically significant one, as policy makers regularly design policies around the interests of the weakest students. The treatment effect then crosses the axis and becomes significant and positive for low- and

middle-ability students. There is a much larger population that shows significant and positive treatment effects, approximately 45 percent of students, than significant and negative treatment effects. Finally, the treatment effect returns to near-zero for high-ability students. Qualitatively, this pattern is exactly what my model predicted—that the treatment effect would cross the axis once and from below.

While these tests support the existence of a single point where the treatment effect crosses the axis from below, they also suggest that the model misses the mark with respect to its location. Specifically, my model makes a focused prediction that this single-crossing point occurs near the cutoff at the 30th percentile, or a GPA of approximately 2.72. A simple comparison of means easily rejects this location for the single-crossing point. Testing the mean treatment effects for  $GPA \leq 2.72$  and  $GPA > 2.72$  fails to reject the null hypothesis that the treatment effect is identical in the two regions ( $t = -1.03$   $P = 0.304$ ).<sup>21</sup> In fact, the mean difference between the 100- and 10-Student Quizzes below the cutoff is 44.4 seconds, a greater value than the mean difference above the cutoff, 20.0 seconds.

Figure 5 provides a more specific test of the single-crossing point. The point at which the smoothing estimate crosses the horizontal axis is approximately a GPA of 1.74, or the second percentile of student GPAs. For a more non-parametric estimate of the single-crossing point, I first note that all differences in effort above the single-crossing point should be positive, while all differences in effort below it should be negative. I then use two measures to assess the fit of different possible single-crossing points. First, I find the point that maximizes the absolute number of positive differences in effort above it and negative differences in effort below it. My second measure finds the point that maximizes the sum of the differences with the predicted sign minus the sum of the differences without the predicted sign. Using both of these measures, the single-crossing point that best fits the data occurs at a GPA of 1.81 or approximately the 3rd percentile.<sup>22</sup>

---

<sup>21</sup>I test this with a regression of the treatment effect on indicators for the two regions with standard errors clustered at the student level. The full regression results are found in the appendix.

<sup>22</sup>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.

**Result 2:** The treatment effect does cross the axis once and from below, but the location of the single crossing point deviates substantially from the model’s prediction.

**Hypothesis 3: The local minimum of the treatment effect is located below the cutoff, and the local maximum is located above the cutoff.**

Using Figure 5 to identify the extrema of the treatment effect rejects the claim that the treatment effect is maximized above the cutoff. In fact, the treatment effect is maximized for GPAs between 2.0 and the cutoff of 2.72. Figure 5 predicts that the maximum treatment effect is 56.7 seconds and occurs at a GPA of 2.39.<sup>23</sup> This is a direct rejection of the prediction that students below the cutoff understand and take advantage of the benefits from the randomness of the 10-Student Quiz.

Figure 5 also identifies the minimum of the treatment effect. The model predicts that the lowest ability students understand the futility of their efforts and show no sensitivity to the treatment effects. This assertion is rejected by the data, where the lowest-ability students demonstrate the strongest tendency to exert effort under the randomness of the 10-Student Quiz. This minimum occurs at the boundary, with the lowest-ability student, one with a GPA of 1.0, showing the smallest predicted treatment effect, -175.7 seconds.<sup>24</sup> This pattern is not consistent with the specific predictions of the model but is consistent with the general intuition that weaker students benefit more from the randomness of the 10-Student Quiz.

**Result 3:** The maximum of the treatment effect does not occur above the cutoff, but below it. The minimum of the treatment effect also occurs below the cutoff, but is only statistically distinguishable from zero for the lowest-ability students.

---

<sup>23</sup>In the appendix, I use a semi-parametric test regressing the treatment effect on bins of different abilities to reject the null hypothesis that the mean treatment effect is equal across all bins. The bin containing the maximum is below the cutoff and has a treatment effect 68 seconds higher than its complement ( $t = 2.34$   $P = 0.019$ ). Standard errors were clustered at the student level.

<sup>24</sup>In the appendix, I also use a semi-parametric test showing that the bin containing the minimum of the treatment effect does cover the region occupied by the weakest students, but does not display significant differences from its complement. This is when it is broadly defined as the lowest 15 percent of abilities ( $t = -1.42$   $P = 0.156$ ). A finer specification of bins would capture the significant negative effect seen in Figure 5.

## 5.5 Alternatives to the Neo-Classical Assumptions of the Model

Using my model as the backbone for understanding the patterns of effort allocation when students are graded on the curve provides several qualitatively accurate predictions, but understandably fails with regards to more focused predictions. Interactions between the treatment effect and alternative motivations for effort could cause any number of shifts in the locations of specific phenomena but will nonetheless preserve the shape and single-crossing nature of the treatment effect that the data show.

To the extent that my experiment can identify the effects of these alternative motivations for effort, I address them below. Additional behavioral phenomena are likely to be present, but may not be identified by my data, I explore the implications of several of these possibilities in the following section.

### 5.5.1 “Cursed” Beliefs

My model depends critically on the specification of students’ beliefs. Characterizing how the predictions of the model change when students are no longer required to make accurate inferences about their relative ability will allow me to test if deviations from the model can be explained by beliefs alone. Classical assumptions require that students make accurate inferences based on their past academic experiences and the selection of students expected to enroll in each class. Prior research suggests, however, that students may be “cursed” to believe that enrollment decisions are independent of ability.<sup>25</sup> Without accounting for the selection of classmates, students will best respond as if they are facing a distribution of classmates similar to the ones they have previously faced.

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

---

<sup>25</sup>Eyster and Rabin (2005) provide an extensive review of empirical and experimental phenomena that can be attributed to cursedness.

from all lower-division UCSD courses over the last 5 years can provide suggestive evidence that student behavior is consistent with cursedness. To construct the distribution of classmates that a cursed student would perceive based on experiences in lower-division courses, I aggregate the grade distributions from all lower-division courses in all departments. From this composite grade distribution, I find the likely GPAs associated with students at each percentile rank in lower-division courses at UCSD.<sup>26</sup> Figure 6 plots the cumulative distribution function of GPAs in this perceived grade distribution alongside the GPAs from Figure 4.

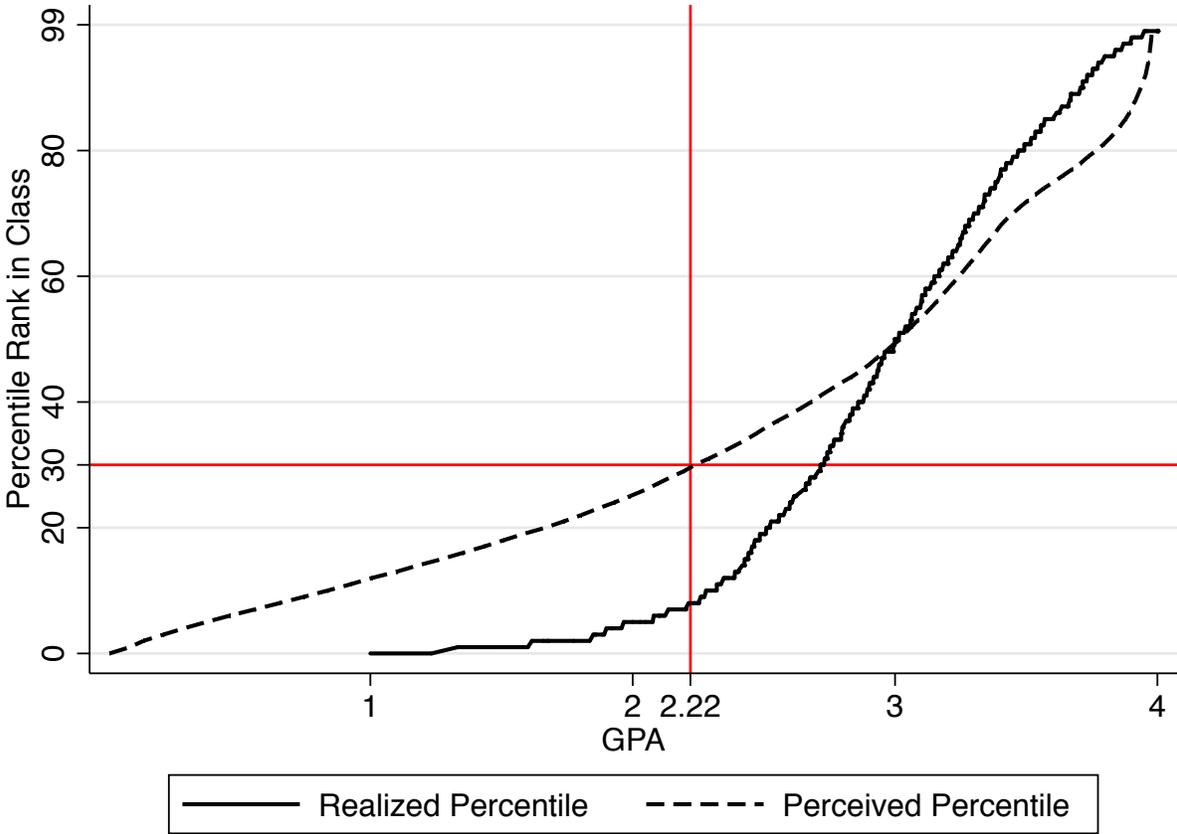


Figure 6: Perceived and Actual Percentile Ranks of Students

<sup>26</sup>Specifically I suppose there are students at the 1st percentile, 2nd percentile, . . . , 99th percentile. The 99th percentile students receive the grades associated with the top 1 percent of grades awarded in each class, the 98th percentile students receive the grades associated with the top 2 percent, and so on. This generates a function mapping a student’s percentile to their lower-division GPA. The measure is imperfect, but does capture the selection effects of upper-division economics courses.

Figure 6 shows how low-ability students appear to select out of this upper-division economics course. Ignoring this selection effect moves the perception of 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 relatively weaker distribution. Selection effects draw the realized percentile ranks 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 30, while his actual percentile rank is approximately 10.

Figure 7 imposes a fully-cursed belief structure on the equilibrium predictions displayed in Figure 3.<sup>27</sup> Referring back to Figure 5, it is clear that the model can now more accurately predict the locations of the single-crossing point and the extrema of the treatment effect.

Under cursed beliefs, my model predicts that the single-crossing point occurs at a GPA near 2.14, a shift that places it much closer to the estimated single-crossing point of 1.81 from the previous section. Less than 10 percent of students in my experiment possess GPAs below this adjusted single-crossing point, so the vast majority are now predicted to have positive treatment effects, a phenomenon confirmed by the data. The adjusted maximum of the treatment effect now occurs for abilities below the cutoff. This coincides much closer with the data, as does the adjusted location of the minimum of the treatment effect, which now is predicted to occur for the lowest 5 percent of student abilities.

Cursedness adjusts the locations of the relevant phenomena closer to their respective locations in the data, but even the adjustment of full cursedness was not sufficient for them to coincide exactly with the data. This implies that, while cursed beliefs may have a dramatic impact on the fit of the model, there are still residual deviations that need to be accounted for. It is also important to note that this should not be taken as proof of cursedness among students in my experiment, rather, it is suggestive that with a better understanding of the

---

<sup>27</sup>These predictions assume that students are aware of neither their own cursed beliefs, nor the potential for cursed beliefs in others. This allows for a simple mapping from the actual GPA to the equilibrium action at the perceived GPA percentile.

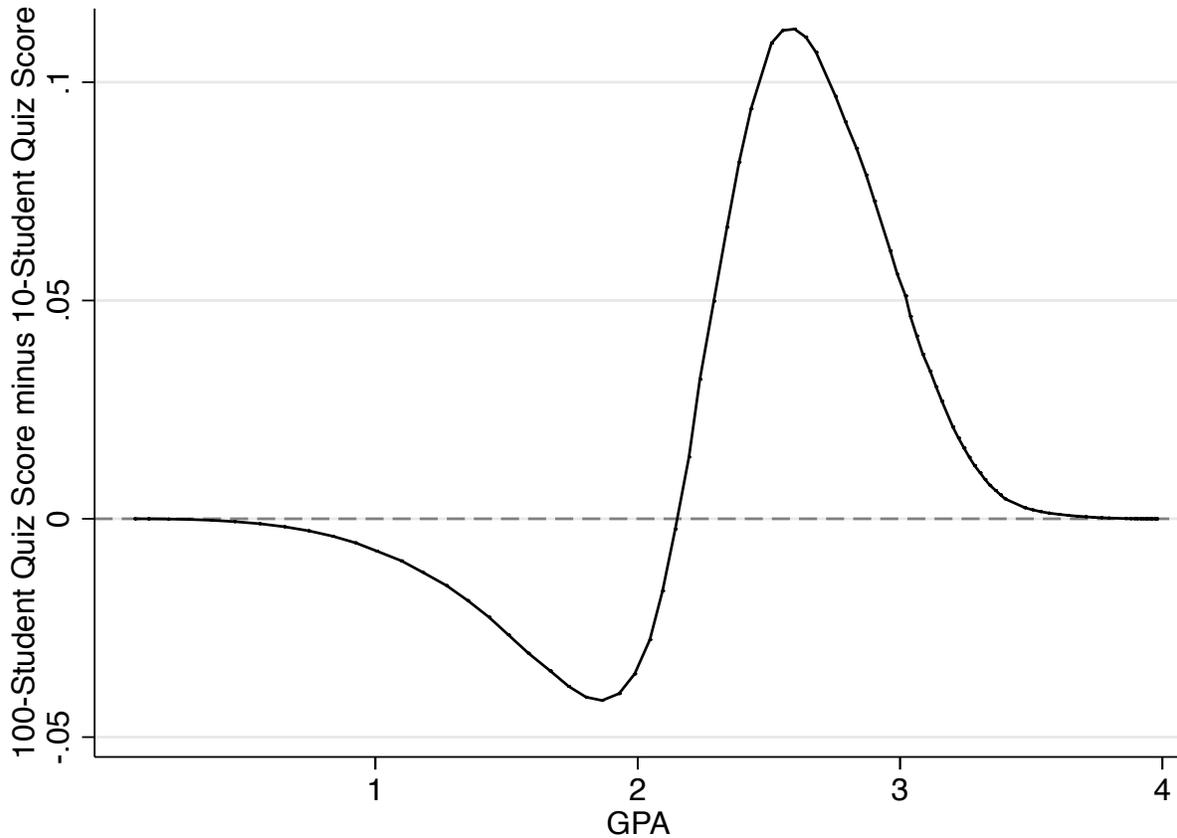


Figure 7: Predicted Difference in Effort Based on Perceived Rank

belief structures of students, models of strategic interaction in the classroom can generate useful predictions for the allocation of effort by students. Further experimentation will be required to understand the complete process of belief formation and updating that students undergo in a classroom setting. In particular, careful experimentation will be needed to distinguish the effects of cursedness from the effects of general overconfidence.

### 5.5.2 Comprehension of Strategic Incentives

The complexity of the equilibrium effort prediction raises the concern that lower ability students will be less able to intuit the benefits to randomness, while higher ability students will be more able to understand the benefits to decreases in randomness. In fact, the weakest students appear to be the most sensitive to the treatments and respond as predicted to the

increases in randomness by exerting more effort. Additionally, students below the cutoff are significantly more likely to state a preference for the 10-Student Quiz on post-experiment questionnaires ( $z = 2.41, P < 0.02$ ).<sup>28</sup> This shows that comprehension of incentives cannot be globally increasing in GPA and cannot drive the result.

### 5.5.3 Intrinsic Motivation

Using GPA as a proxy for ability may raise the concern that the students labeled high ability may be more intrinsically motivated to exert effort on quizzes than the students labeled low ability. Figure 8 uses a locally linear polynomial smoothing function to plot the amount of time allocated to each quiz on the left-hand scale and the aggregate amount of time allocated to quizzes on the right-hand scale and allows me to address the question of intrinsic motivation directly.

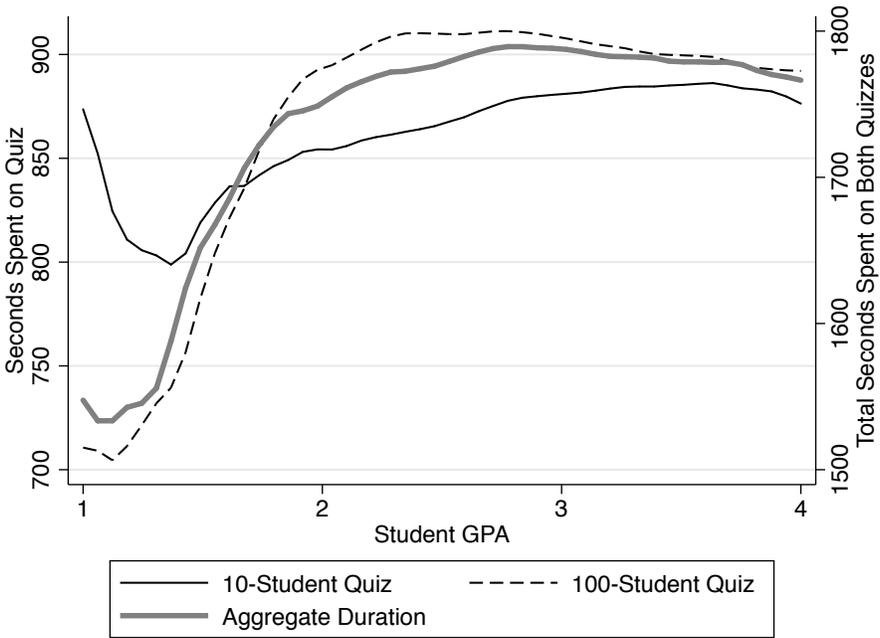


Figure 8: Seconds spent on each quiz and aggregate seconds spent on quizzes by GPA

Figure 8 clearly shows that aggregate quiz duration is generally higher for students with

<sup>28</sup>The post-experiment questionnaires as well as the full specification of this regression can be found in the appendix.

higher GPAs, but reaches a maximum at a fairly low GPA of 2.81, approximately the 37th percentile. Level effects of effort will be removed by differencing the time on the 100- and 10-Student Quizzes, so this phenomenon would only pose a threat to the experimental results if the level effect interacted with the differences between the two quizzes. The data reject this interaction effect. While differences are very large for students with the lowest aggregate duration, they are also very large for students with the highest aggregate duration. Indeed, the two points where the differences are the smallest have very different levels of aggregate duration. This shows that, while level effects in effort that trend with GPA are a reality in my experiment, they cannot drive the observed pattern of differences in effort.

#### 5.5.4 Risk Aversion

Risk aversion could provide an alternative explanation to the observed behavior, since the two quizzes inherently bear different levels of uncertainty. In general, as risk aversion increases, low-ability students will decrease their effort, and high-ability students will increase their effort. To control for risk preferences, at the end of the course students answered a survey question about their likelihood of taking risks. Dohmen et al. (2011) show that this survey question is a strong predictor of revealed risk preferences.<sup>29</sup>

In general, the patterns of effort predicted under different risk preferences will simply increase the standard errors of the estimates of the treatment effect unless risk aversion is correlated with GPA. The data refute this correlation ( $t = -0.71$ ,  $P = 0.478$ ) and show that positive treatment effects are common across different risk preferences.<sup>30</sup> Figure 9 splits students into risk averse and risk loving based on their responses to the survey question and plots the absolute duration of each quiz treatment to test if the impact of risk aversion is different across quizzes. The locally linear polynomial fit of the data show that the level effect of risk aversion is much more pronounced than any differential effect across the two quizzes. In general, low-ability, risk-loving students exert more effort on all quizzes than their

---

<sup>29</sup>The question specifically asks, “How likely are you to take risks, in general.”

<sup>30</sup>The full specification of this regression can be found in the appendix.

risk-averse counterparts, but this level effect attenuates as ability increases. Importantly, the qualitative features of the treatment effect—that is, the difference between the two curves—and, indeed, the point estimates for the values of the treatment effect are similar across students with different risk preferences but equal abilities.

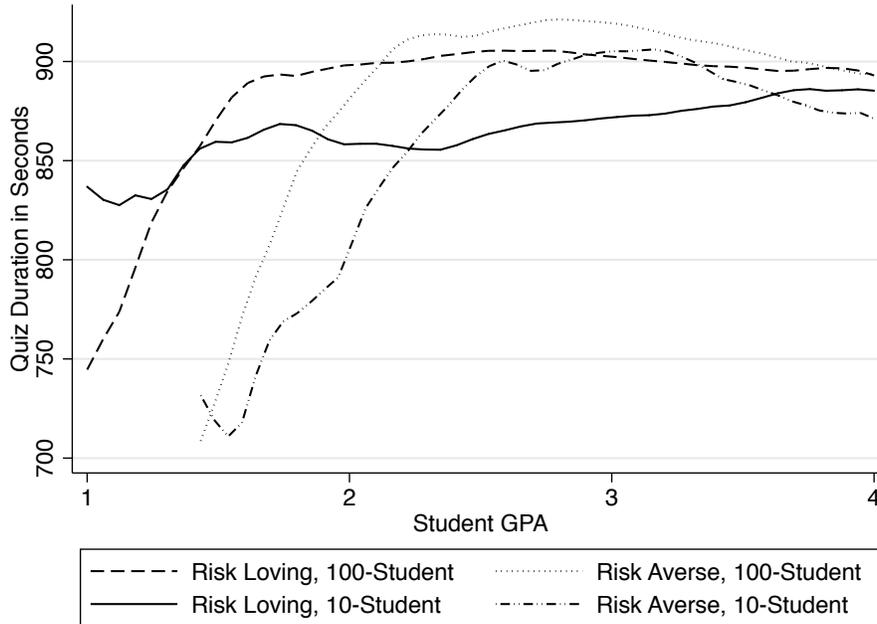


Figure 9: Duration of each quiz separated by risk preference (in seconds)

Figure 9 confirms the prediction that low-ability, risk-averse students diminish their effort, but this impulse does not differentially affect the 10- and 100-Student Quizzes, even though one is inherently riskier than the other. Additionally, as the ability increases, there are no meaningful differences between students of equal ability but different risk preference. Thus, the general patterns of the treatment effect cannot be driven by risk preferences alone.

### 5.5.5 Gender and Competitiveness

The role of gender in determining risk and competitive preferences has been widely studied, and deserves attention in this paper, as my model supposes tournament-style competition between students.<sup>31</sup> The effect of gender does not present a likely confound in the identifica-

<sup>31</sup>See Niederle and Vesterlund (2011) for a review of the literature on gender and competitive preferences.

tion of the heterogeneity of the treatment effect across abilities, as GPA does not correlate with gender ( $t = -1.44, P = 0.150$ ).<sup>32</sup> Nonetheless, understanding the ways that students of different genders react to the treatments will be instructive in designing optimal grading mechanisms. Figure 10 plots the treatment effects by gender and shows that the qualitative features of the treatment effect are similar across genders, making it an unlikely driver of the observed patterns of effort allocation between quizzes.

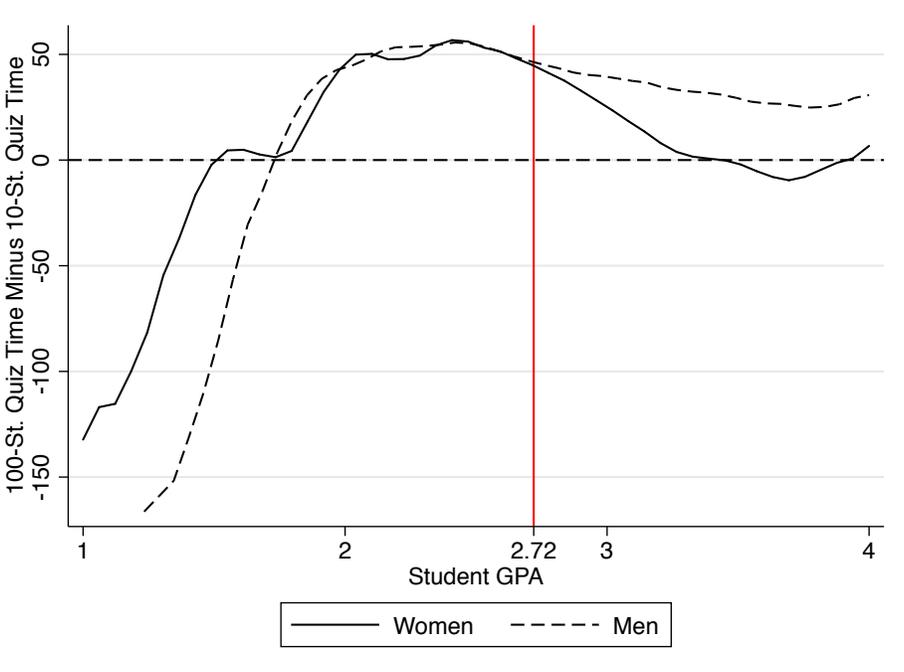


Figure 10: Local Polynomial Fit of Difference in Effort between 100- and 10-Student Quizzes by Gender (Times Reported in Seconds)

## 6 Discussion

A discussion of the policy implications of my experiment must first address the social benefit of inducing greater effort on classroom quizzes. One approach this is to consider the effect of the treatments on the scores of the students. This analysis shows no significant effect of the treatments ( $t = -0.37, P = 0.710$ ).<sup>33</sup> There are two primary explanations for why

<sup>32</sup>This regression can be found in the appendix.

<sup>33</sup>The full specification of this regression is found in the appendix.

time spent on the quizzes is a superior measure to scores on quizzes for my purposes. First, quiz scores are much coarser, with all quizzes being graded based on 4 questions. More variation allows me to observe changes in effort at a smaller margin than changes in scores, overcoming the problem of low-powered tests. Second, the differences in the difficulty of the quizzes assigned to each treatment are amplified in the analysis of quiz scores, driving relatively much more variation in scores than in durations. A sufficiently large sample of quizzes could overcome this problem, but each subject saw only 5 pairs of quizzes, so the variation in effort is lost amidst the variation in quiz difficulty.

It seems, therefore, that in this setting, it is not certain that additional time spent on quizzes has a positive return in the form of scores, but it is certain that additional time spent on quizzes is a measure of costly effort exertion. If students do not perceive that additional time increases performance, then any positive amount of time spent on quizzes 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 absent the paired quiz design. That is, if a students are willing to spend significantly more time on the 100-Student Quiz than the 10-Student Quiz when they are posted simultaneously, then students would likely study longer or attend more class in preparation for a 100-Student Quiz in the absence of this experiment.<sup>34</sup>

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

Towards that end, my experiment uncovers systematic trends among students of different

---

<sup>34</sup>Ultimately, 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.

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 experiment isolates the strategic uncertainty of small classes by manipulating a grading cohort size while holding constant all other characteristics of the classroom, such as teaching quality, student observability, or access to resources. With these clean controls, I can contribute to the discussion of optimal classroom size by demonstrating that even small changes in a students strategic setting causally affect student effort exertion.

## 6.1 Patterns of Effort Allocation

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

Second, the students most motivated by increasing the cohort size have abilities near but below the cutoff. This contradicts the model's prediction that these students understand that their returns to effort are higher on the 10-Student Quiz than on the 100-Student Quiz. This misallocation of effort may reflect biased beliefs about relative ability.

Third, there is a negative effect of increasing cohort size for the lowest-ability students, indicating that these students identify the higher returns to their effort in the higher variance environment. Stated preference confirms that students below the cutoff are more likely to prefer the higher variance environment than students above it, though both groups state a general preference for the lower variance environment.

## 6.2 Misallocation of Effort

Despite the number of competing explanations for the results addressed in the previous section, none were able to independently generate the patterns of the treatment effect demonstrated in Figure 5. Indeed, the effects of several alternative motivations for effort are clear but never consistent with the observed behavior and thus not a viable alternative to a strate-

gic effort model. The data indicate that the strategic incentives present in the environment do cause shifts in effort that trend systematically with student ability in ways qualitatively similar to the predictions of the model. Nonetheless, the accuracy of the predictions of the model can still be improved by incorporating some of these effects, most importantly, biased beliefs. The data provide suggestive evidence that beliefs are partially responsible for effort allocations that are ex-ante suboptimal.

The positive treatment effect for students with abilities below the cutoff represents the most robust misallocation of effort by students. While the mechanisms for this misallocation of effort are not clear from the data, there is suggestive evidence that students may be “cursed” to believe that their classmates’ enrollment decisions for upper-division economics courses were unrelated to their abilities. The fit of the model is greatly improved by supposing that students believe that their classmates represent an average draw from the population of students. Adjusting for cursed beliefs, however, cannot fully explain the deviations from the model’s predictions. The following three behavioral biases are potential causes of this residual misallocation of effort.

First, students may possess a general overconfidence about their abilities. If students respond to their strategic grading environment according to an inflated estimate of their relative ability, then many students below the cutoff may favor the 100-Student Quiz, but students far enough below the cutoff will still favor the 10-Student Quiz. Second, students may fail to properly update their beliefs about their relative abilities. Upon receiving their results from the quizzes, students should reallocate effort based on their posterior beliefs about their relative ability, but updating failures allow students to ignore the informational content of quiz results and cling to their potentially biased prior beliefs. Both of these biases are consistent with existing results showing biased updating about self-perceived intelligence (Eil and Rao, 2011; Mobius et al., 2011).

Third, students may possess reference dependent utility, specifically myopic loss aver-

sion (Benartzi and Thaler, 1995).<sup>35</sup> If students consider each quiz independently, they will overexert on quizzes with lower returns to their effort. Recall that increasing effort only increases the probability of high grades and does not alter the grades themselves, since grades are fixed at 0, 1, or 3 points. As Sprenger (2010) points out, loss aversion over probabilities can only arise under expected value based reference points as in disappointment aversion (Bell, 1985; Loomes and Sugden, 1986; Gul, 1991), because under stochastic reference points (Kőszegi and Rabin, 2006, 2007), students are risk neutral over changes in probability within the support of the original gamble. A student who is risk neutral over changes in probability would exert more effort on the quiz that led to the greatest increases in her probability of high grades. Under loss aversion, a student would exert additional effort on the quiz that was considered to be “losing” relative to her reference probability, and in the case of students below the cutoff, this would be the 100-Student Quiz.

### 6.3 Policy Prescriptions

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

While decreasing the variance increases aggregate effort, it incurs certain costs. For the lowest ability students, higher variance environments induce more effort. The intuition for this result relates back to Figure 1, which shows how increases in the size of a cohort make it increasingly unlikely that a low-ability student receives a high grade. From a policy-perspective, this result suggests that low-ability students can become discouraged by relative grading when the cohort size becomes large enough. If motivating low-ability students is part of an educator’s objective, then smaller cohorts can accomplish this. This could be achieved by splitting large classes up into smaller sections and grading each individually.

---

<sup>35</sup>Here, “myopic” does not refer to time horizons, rather it refers to students evaluating outcomes of each quiz individually instead of evaluating the impact of each quiz on the total course grade.

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

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

## 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 mechanisms to achieve their objectives with respect to student effort.

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

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

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

My results make it clear that 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 different grading mechanisms. It additionally provides the basis for exploration of grading mechanisms that find the desired balance between increasing mean effort and promoting a more desirable distribution of effort among students.

## References

- Amann, Erwin and Wolfgang Leininger**, “Asymmetric all-pay auctions with incomplete information: the two-player case,” *Games and Economic Behavior*, 1996, *14* (1), 1–18.
- Andreoni, James and Andy Brownback**, “Grading on a Curve, and other Effects of Group Size on All-Pay Auctions,” Technical Report, National Bureau of Economic Research 2014.
- , **Yeon-Koo Che, and Jinwoo Kim**, “Asymmetric information about rivals’ types in standard auctions: An experiment,” *Games and Economic Behavior*, 2007, *59* (2), 240–259.

- Arulampalam, Wiji, Robin A Naylor, and Jeremy Smith**, “Am I missing something? The effects of absence from class on student performance,” *Economics of Education Review*, 2012, *31* (4), 363–375.
- Barut, Yasar, Dan Kovenock, and Charles N Noussair**, “A Comparison of Multiple-Unit All-Pay and Winner-Pay Auctions Under Incomplete Information\*,” *International Economic Review*, 2002, *43* (3), 675–708.
- Baye, Michael R, Dan Kovenock, and Casper G de Vries**, “Rigging the Lobbying Process: An Application of the All-Pay Auction,” *The American Economic Review*, 1993, *83* (1), 289–294.
- , – , **and** – , “The all-pay auction with complete information,” *Economic Theory*, 1996, *8* (2), 291–305.
- Becker, William E and Sherwin Rosen**, “The learning effect of assessment and evaluation in high school,” *Economics of Education Review*, 1992, *11* (2), 107–118.
- Bell, David E**, “Disappointment in decision making under uncertainty,” *Operations research*, 1985, *33* (1), 1–27.
- Benartzi, Shlomo and Richard H Thaler**, “MYOPIC LOSS AVERSION AND THE EQUITY PREMIUM PUZZLE,” *The Quarterly Journal of Economics*, 1995, *110* (1), 73–92.
- Benjamin, Daniel J, Matthew Rabin, and Collin Raymond**, “A Model of Non-Belief in the Law of Large Numbers,” 2014.
- Betts, Julian R**, “The impact of educational standards on the level and distribution of earnings,” *American Economic Review*, 1998, pp. 266–275.
- **and Jeff Grogger**, “The impact of grading standards on student achievement, educational attainment, and entry-level earnings,” *Economics of Education Review*, 2003, *22* (4), 343–352.
- Bull, Clive, Andrew Schotter, and Keith Weigelt**, “Tournaments and piece rates: An experimental study,” *The Journal of Political Economy*, 1987, pp. 1–33.
- Chen, Jennjou and Tsui-Fang Lin**, “Class attendance and exam performance: A randomized experiment,” *The Journal of Economic Education*, 2008, *39* (3), 213–227.

- Costrell, Robert M**, “A simple model of educational standards,” *The American Economic Review*, 1994, pp. 956–971.
- 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.
- Davis, Douglas D and Robert J Reilly**, “Do too many cooks always spoil the stew? An experimental analysis of rent-seeking and the role of a strategic buyer,” *Public Choice*, 1998, 95 (1-2), 89–115.
- Dechenaux, Emmanuel, Dan Kovenock, and Roman M Sheremeta**, “A survey of experimental research on contests, all-pay auctions and tournaments,” Technical Report, Discussion Paper, Social Science Research Center Berlin (WZB), Research Area ‘Markets and Politics’, Research Professorship & Project ‘The Future of Fiscal Federalism’ 2012.
- Dobkin, Carlos, Ricard Gil, and Justin Marion**, “Skipping class in college and exam performance: Evidence from a regression discontinuity classroom experiment,” *Economics of Education Review*, 2010, 29 (4), 566–575.
- Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner**, “Individual risk attitudes: Measurement, determinants, and behavioral consequences,” *Journal of the European Economic Association*, 2011, 9 (3), 522–550.
- Dubey, Pradeep and John Geanakoplos**, “Grading exams: 100, 99, 98, or A, B, C?,” *Games and Economic Behavior*, 2010, 69 (1), 72–94.
- Eil, David and Justin M Rao**, “The good news-bad news effect: asymmetric processing of objective information about yourself,” *American Economic Journal: Microeconomics*, 2011, pp. 114–138.
- Eyster, Erik and Matthew Rabin**, “Cursed equilibrium,” *Econometrica*, 2005, 73 (5), 1623–1672.
- Fraja, Gianni De, Tania Oliveira, and Luisa Zanchi**, “Must try harder: Evaluating the role of effort in educational attainment,” *The Review of Economics and Statistics*, 2010, 92 (3), 577–597.
- Gneezy, Uri and Rann Smorodinsky**, “All-pay auctions—An experimental study,” *Journal of Economic Behavior & Organization*, 2006, 61 (2), 255–275.

- Gul, Faruk**, “A theory of disappointment aversion,” *Econometrica: Journal of the Econometric Society*, 1991, pp. 667–686.
- Harbring, Christine and Bernd Irlenbusch**, “Incentives in tournaments with endogenous prize selection,” *Journal of Institutional and Theoretical Economics JITE*, 2005, 161 (4), 636–663.
- Hillman, Arye L and Dov Samet**, “Dissipation of contestable rents by small numbers of contenders,” *Public Choice*, 1987, 54 (1), 63–82.
- **and John G Riley**, “Politically Contestable Rents and Transfers\*,” *Economics & Politics*, 1989, 1 (1), 17–39.
- Hirshleifer, Jack and John G Riley**, “Elements of the Theory of Auctions and Contests,” Technical Report, UCLA Department of Economics 1978.
- Kagel, John H and Dan Levin**, “Independent private value auctions: Bidder behaviour in first-, second- and third-price auctions with varying numbers of bidders,” *The Economic Journal*, 1993, pp. 868–879.
- Kahneman, Daniel and Amos Tversky**, “On the psychology of prediction.,” *Psychological review*, 1973, 80 (4), 237.
- Kokkelenberg, Edward C, Michael Dillon, and Sean M Christy**, “The effects of class size on student grades at a public university,” *Economics of Education Review*, 2008, 27 (2), 221–233.
- Kőszegi, Botond and Matthew Rabin**, “A model of reference-dependent preferences,” *The Quarterly Journal of Economics*, 2006, pp. 1133–1165.
- **and –**, “Reference-dependent risk attitudes,” *The American Economic Review*, 2007, pp. 1047–1073.
- Krueger, Anne O**, “The political economy of the rent-seeking society,” *The American economic review*, 1974, pp. 291–303.
- Lazear, Edward P and Sherwin Rosen**, “Rank-Order Tournaments as Optimum Labor Contracts,” *The Journal of Political Economy*, 1981, 89 (5), 841–864.
- List, John A, Daan Van Soest, Jan Stoop, and Haiwen Zhou**, “On the role of group size in tournaments: Theory and evidence from lab and field experiments,” Technical Report, National Bureau of Economic Research 2014.

- Loomes, Graham and Robert Sugden**, “Disappointment and dynamic consistency in choice under uncertainty,” *The Review of Economic Studies*, 1986, 53 (2), 271–282.
- Mobius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat**, “Managing self-confidence: Theory and experimental evidence,” Technical Report, National Bureau of Economic Research 2011.
- Moldovanu, Benny and Aner Sela**, “The optimal allocation of prizes in contests,” *American Economic Review*, 2001, pp. 542–558.
- and –, “Contest architecture,” *Journal of Economic Theory*, 2006, 126 (1), 70–96.
- Mosteller, Frederick**, “The Tennessee study of class size in the early school grades,” *Future of children*, 1995, 5, 113–127.
- Mroch, Andy**, “Law School Grading Curves,” 2005.
- Müller, Wieland and Andrew Schotter**, “Workaholics and dropouts in organizations,” *Journal of the European Economic Association*, 2010, 8 (4), 717–743.
- Niederle, Muriel and Lise Vesterlund**, “Gender and competition,” *Annu. Rev. Econ.*, 2011, 3 (1), 601–630.
- Noussair, Charles and Jonathon Silver**, “Behavior in all-pay auctions with incomplete information,” *Games and Economic Behavior*, 2006, 55 (1), 189–206.
- Olszewski, Wojciech and Ron Siegel**, “Large Contests,” 2013.
- Paredes, Valentina**, “Grading System and Student Effort.”
- Potters, Jan, Casper G de Vries, and Frans Van Winden**, “An experimental examination of rational rent-seeking,” *European Journal of Political Economy*, 1998, 14 (4), 783–800.
- Rabin, Matthew**, “Inference by Believers in the Law of Small Numbers,” *Quarterly Journal of Economics*, 2002, pp. 775–816.
- Romer, David**, “Do students go to class? Should they?,” *The Journal of Economic Perspectives*, 1993, pp. 167–174.
- Siegel, Ron**, “All-Pay Contests,” *Econometrica*, 2009, 77 (1), 71–92.

**Sprenger, Charles**, “An endowment effect for risk: Experimental tests of stochastic reference points,” Technical Report, working paper 2010.

**Stinebrickner, Ralph and Todd R Stinebrickner**, “The Causal Effect of Studying on Academic Performance,” *The B.E. Journal of Economic Analysis & Policy*, 2008, 8 (1).

**Tullock, Gordon**, “The welfare costs of tariffs, monopolies, and theft,” *Economic Inquiry*, 1967, 5 (3), 224–232.

**Tversky, Amos and Daniel Kahneman**, “Belief in the law of small numbers.,” *Psychological bulletin*, 1971, 76 (2), 105.

**Vickrey, William**, “Counterspeculation, auctions, and competitive sealed tenders,” *The Journal of finance*, 1961, 16 (1), 8–37.

**Welch, Jack and John A Byrne**, “Straight from the Gut,” *New York*, 2001.

# Appendix A: Experimental Procedures

## Syllabus Instructions for Quizzes

### Economics 100A Quizzes

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

#### Overview:

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

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

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

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

## Online Instructions for Quizzes

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

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

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

All answers will be in WHOLE NUMBERS.”

## Online Environment

The screenshot shows the Canvas LMS interface for the course 'ECON 100A - Andreoni [W114]'. The top header displays the UC San Diego logo and the URL 'TED.UCSD.EDU'. The user's name 'Andrew Brownback' and a score of '97' are visible in the top right corner. The 'Content' tab is active, and 'Edit Mode is: ON' is indicated. The main content area shows two quiz items: '100-Student Quiz (Week 10)' and '10--Student Quiz (Week 10)'. The left navigation menu includes options like 'Home Page', 'Information', 'Content', 'Discussions', 'Groups', 'Tools', and 'Help'.



## Appendix B: Theory

### Score is Monotonic in Ability

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

Incentive compatibility dictates that for  $i$  and  $j$ , respectively,

$$\begin{aligned} U(s_{i,t}, a_i, N, P) &\geq U(s_{j,t}, a_i, N, P) \\ U(s_{j,t}, a_j, N, P) &\geq U(s_{i,t}, a_j, N, P). \end{aligned}$$

Expanding these equations yields

$$P_{N,P}(s_{i,t}) - \frac{C(s_{i,t})}{a_i} \geq P_{N,P}(s_{j,t}) - \frac{C(s_{j,t})}{a_i} \quad (4)$$

$$P_{N,P}(s_{j,t}) - \frac{C(s_{j,t})}{a_j} \geq P_{N,P}(s_{i,t}) - \frac{C(s_{i,t})}{a_j}, \quad (5)$$

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

Solve for common terms and combine (4) and (5) to get

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

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

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

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

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

which is a contradiction. QED.

## 7.1 Equilibrium Effort Functions

Maximizing (3) with respect to score yields the first-order condition,

$$\begin{aligned} \frac{\partial U(s_{i,t}; a_i, N, P)}{\partial s_{i,t}} &= \sum_{j=N-NP}^{N-1} \left( \frac{(N-1)!}{j!(N-1-j)!} \right) \left\{ \left[ j \times A(s_{i,t}; N, P)^{j-1} \times A'(s_{i,t}) \right] \right. \\ &\quad \left. (1 - A(s_{i,t}; N, P))^{N-1-j} - (N-1-j) \right. \\ &\quad \left. \times \left[ A(s_{i,t}; N, P)^j (1 - A(s_{i,t}; N, P))^{N-2-j} \times A'(s_{i,t}) \right] \right\} \equiv \frac{1}{a_i}. \end{aligned} \quad (6)$$

At equilibrium, the ability implied by a student's score must equal that student's ability. That is,  $A(s_{i,t}) = a_i$ . Substituting this into (6) and solving for  $\frac{1}{A'(s_{i,t})}$  results in

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

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

$$\begin{aligned} S(A(s_{i,t})) &= S(S^{-1}(s_{i,t})) = s_{i,t} \\ \Leftrightarrow S'(A(s_{i,t})) A'(s_{i,t}) &= 1 \\ \Rightarrow S'(a_i) &= \frac{1}{A'(s_{i,t})}. \end{aligned}$$

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

$$\begin{aligned} S'(a_i; N, P) &\equiv \\ \sum_{j=N-NP}^{N-1} \left( \frac{(N-1)!}{j!(N-1-j)!} \right) \left\{ j \times a_i^j (1 - a_i)^{N-1-j} - (N-1-j) \left[ a_i^{j+1} (1 - a_i)^{N-2-j} \right] \right\}. \end{aligned} \quad (8)$$

Solving this under the initial condition that  $S(0; N, P) = 0$  yields the equilibrium score as a function of ability.<sup>36</sup>

<sup>36</sup>This initial condition states that, at equilibrium, the weakest student understands the futility of effort regardless of cohort size, and chooses a score of zero.

# Appendix C: Empirical Results

## Semi-Parametric Tests

For robustness, I use semi-parametric tests to confirm the heterogeneity shown in Figure 5. 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,

$$\begin{aligned} \text{LowestBin}_i &= \mathbb{1}_{\{GPA_i \in [1, 2.433]\}} \\ \text{LowBin}_i &= \mathbb{1}_{\{GPA_i \in (2.433, 2.72]\}} \\ \text{HighBin}_i &= \mathbb{1}_{\{GPA_i \in (2.72, 3.246]\}} \\ \text{HighestBin}_i &= \mathbb{1}_{\{GPA_i \in (3.246, 4]\}} \end{aligned}$$

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. This regression is displayed in column 1 of Table 5. Column 2 repeats the regression, instrumenting for the completion order using the presentation order.

Table 5: Duration of the 100-Student Quiz minus the 10-Student Quiz (in minutes)

	OLS	IV Regression
<i>LowestBin</i>	-0.031	0.190
<i>GPA<sub>i</sub> ∈ [1, 2.433]</i>	(0.43)	(0.43)
<i>LowBin</i>	1.454***	1.737***
<i>GPA<sub>i</sub> ∈ (2.433, 2.72]</i>	(0.44)	(0.44)
<i>HighBin</i>	0.515	0.878***
<i>GPA<sub>i</sub> ∈ (2.72, 3.246]</i>	(0.33)	(0.34)
<i>HighestBin</i>	0.151	0.496
<i>GPA<sub>i</sub> ∈ (3.246, 4]</i>	(0.33)	(0.34)
<i>100-Std. Quiz</i>		-6.125***
<i>Taken First</i>		(1.26)
Instrumented	No	Yes
N	2,506	2,506

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

All standard errors clustered at the student level.

All times in minutes.

The maximum of the treatment effect is identified as occurring over *LowBin*. The minimum of the treatment effect appears to be less robust, but is consistently located over *LowestBin*.

### Hypothesis 2: The treatment effect crosses the axis once and from below.

To consider the single-crossing nature of the treatment effect, I examine its evolution beginning with the students in *LowestBin*. In both the OLS regression and the IV regression, the

coefficient on the treatment effect is lowest in this first bin. The coefficient on the treatment effect in *LowBin* moves into the positive domain for the OLS coefficients, and remains positive for *HighBin* and *HighestBin*. This implies that the treatment effect does cross the axis once, from below, and that the crossing occurs between *LowestBin* and *LowBin*.

**Result 2:** Under a less parametric specification, the treatment effect still crosses the axis a single time and from below, but the location of the single crossing point deviates substantially from the prediction.

**Hypothesis 3: The local minimum of the treatment effect is located below the cutoff, and the local maximum is located above the cutoff.**

The coefficients from Table 5 identify the regions over which the treatment effect is maximized and minimized. While this coarse measure is imperfect, it is sufficient to reject the claim that the treatment effect is maximized above the cutoff. In fact, the treatment effect is clearly maximized in *LowBin*, where the coefficient is nearly double the value of the other coefficients. While a pairwise comparison fails to reject the equality of coefficients between *LowBin* and *HighBin* ( $F = 2.49$   $P = 0.115$ ), the difference between the mean treatment effect found in *LowBin* and the mean treatment effect found in its complement is approximately 68 seconds and is statistically significant ( $t = 2.34$   $P = 0.019$ ).<sup>37</sup>

Table 5 is less successful in identifying the minimum of the treatment effect. The coefficient for *LowestBin* is the smallest, but not significantly different from the coefficients in *HighBin* or *HighestBin*. Additionally, the difference between the mean treatment effect in *LowestBin* and its complement is approximately 41 seconds, but is not statistically significant ( $t = -1.42$   $P = 0.156$ ).<sup>38</sup>

## Means on either side of the cutoff

.

## Pairwise tests

.

---

<sup>37</sup>To test the difference in means, I regressed the treatment effect onto an indicator variable for the relevant bin. Standard errors were clustered at the student level. The full specification of the regression is in the appendix.

<sup>38</sup>To test the difference in means, I regressed the treatment effect onto an indicator variable for the relevant bin. Standard errors were clustered at the student level. The full specification of the regression is in the appendix.

Table 6: 100-Student Quiz Duration Minus 10-Student Quiz Duration

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

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

Table 7: 100-Student Quiz Duration Minus 10-Student Quiz Duration

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

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

## Stated Preference

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

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

Table 8: Stated Preference for the 10-Student Quiz

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

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

## Risk Preference

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

“How likely are you to take risks, in general?”

Table 9: Stated Risk Preference

Risk Measure (From 1-7)	
$GPA$	-0.094 (0.13)
Constant	4.202*** (0.41)
N	504

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

## Gender and GPA

The gender and GPA of the subjects were recorded. GPA is does not significantly correlate with gender.

Table 10: Probit regression of likelihood of being male on GPA

	Pr(Male)
<i>GPA</i>	-0.151 (0.11)
Constant	0.723** (0.32)
N	512

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

### Treatment Effect on Scores

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

Table 11: Scores on 100-St Quiz minus Scores on 10-St Quiz

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

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