Abstract

Grading on the curve is a form of relative evaluation similar to an all-pay auction or rank-order tournament. The distribution of students drawn into the class from the population is predictably linked to the size of the class. Increasing the class size draws students’ percentile ranks closer to their population percentiles. Since grades are awarded based on percentile ranks in the class, this reallocates incentives for effort between students with different abilities. The predicted aggregate effort and the predicted effort from high-ability students increases while the predicted effort from low-ability students decreases. Andreoni and Brownback (2017) find that the size of a contest has a causal impact on the aggregate effort from participants and the distribution of effort among heterogeneous agents. In this paper, I randomly assign “class sizes” to quizzes in an economics course to test these predictions in a real-stakes environment. My within-subjects design controls for student, classroom, and time confounds and finds that the lower variance of larger classes elicits greater effort from all but the lowest-ability students, significantly increasing aggregate effort.

“You embrace the top 20 [percent of employees], deal with the middle 70, and you face into the bottom 10, and you do what’s right for them and for you.” –Jack Welch

1 Introduction and Background

Relative evaluation is often employed to mitigate the effects of asymmetric information between a mechanism designer and economic agents. Teachers, for example, may grade on the curve because they do not have perfect information about exactly how difficult their exams are for their students. Similarly, management philosophies such as the one embodied by the quote from Jack Welch assert that the relative ranking of employees is a sufficient metric for their evaluation. Job promotion or bonuses, tenure decisions, and lobbying contests all rely to a degree on evaluating options relative to their peers.

Under relative evaluation, an agent’s incentives for effort are dependent on the composition of their comparison group or “cohort.” According to the law of large numbers, the larger this cohort becomes, the
more closely its composition resembles the population distribution from which it is drawn. The size of a cohort affects the incentives for effort by influencing the cohort composition. Any agent with knowledge of the population distribution can draw inference about her cohort distribution from her cohort size. Since cohort size has a predictable effect on cohort composition, it also has a predictable effect on the distribution of effort incentives among agents. In this paper, I demonstrate that the theoretical connection between cohort size and effort bears itself out in a large-scale classroom experiment on relative grading.

Consider an example, Texas HB 588 grants automatic admission to any Texas state university to all Texas high school seniors who graduate in the top 10 percent of their high school class.\textsuperscript{1} Since Texas high schools vary in size by orders of magnitude,\textsuperscript{2} a student of a given ability may face dramatically different incentives for effort under this program, depending on her class size. In smaller high schools, the student would be more likely to face a class full of outliers—high or low—causing the returns to effort to vary wildly. In larger schools, the composition of students is more likely to reflect the characteristics of the population, reducing the uncertainty around her returns to effort. While this is only one of many factors at play in an environment as complex as a classroom, my paper uses a within-student design to hold all other factors constant and cleanly identify the predictable effect of class size on student effort.

I follow the model of Andreoni and Brownback (2017), which connects the size of a cohort and the incentives for effort in an all-pay auction. This model builds on the work of Lazear and Rosen (1981) and Becker and Rosen (1992), which explore the use of tournaments as labor contracts and grading mechanisms, respectively. In focusing exclusively on contest size, the theory borrows from Moldovanu and Sela (2001) and Moldovanu and Sela (2006), which provide in-depth theoretical treatments of real-effort contest “architecture.”

The model operates as follows: students draw private valuations for “pass” grades independently from a known distribution before being assigned to cohorts and competing for grades. Valuation can be interpreted as ability. Grades will be awarded based on whether or not a student’s effort exceeds a certain percentile rank or “cutoff.” The distribution of valuations in a larger cohort is more reflective of the population distribution. Thus, a student above the cutoff percentile in the population is more likely to also be above the cutoff in her cohort if the cohort is large. On the other hand, a student below the cutoff in the population is more likely to luckily find herself above the cohort cutoff when her cohort is small. As a result, the model predicts that larger cohorts diminish incentives for effort among low-ability

\textsuperscript{1}The bill was modified in 2009 to stipulate that the University of Texas at Austin may cap the number of students admitted under this measure to 75% of in-state freshman students.
\textsuperscript{2}For example, Valentine HS has fewer than 20 students, while Skyline HS has over 4,500.
students and increase incentives for effort among high-ability students. Additionally, the model predicts that the decrease in uncertainty from larger cohorts will increase aggregate effort.

My experimental results confirm that mean effort increases in the cohort size. However, I reject the distributional consequences associated with the larger cohort sizes. Both low- and high-ability students appear to increase their effort in larger cohorts. My results on aggregate effort confirm what Andreoni and Brownback (2017) found in their laboratory study. However, Andreoni and Brownback (2017) found that larger cohorts had a small, negative effect on effort from low-types while I find positive effects on the majority of low-ability students. The importance of replicating the main results of Andreoni and Brownback (2017) in a natural environment should not be understated as external validity is often a concern in laboratory experiments (Levitt and List (2007); Kessler and Vesterlund (2015)). In addition, the classroom environment allows me to explore the theory without relying on induced values for student ability. Indeed, this lack of salience surrounding what constitutes “low-ability” in the classroom may drive the difference in results between this study and Andreoni and Brownback (2017).

To arrive at these conclusions, I conducted a classroom experiment on relative grading in a large, upper-division economics course at the University of California, San Diego (UCSD). Five times during the semester, students completed a pair of online quizzes that count towards their course grade. One of the two quizzes was assigned to a cohort of 10 students and the other was assigned to a cohort of 100 students. The top 70 percent of scores in a cohort—the top 7 out of 10 and 70 out of 100, respectively—received “pass” grades. I 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. My experimental design allows me to use the within-student difference in effort between the two quizzes of a given week to measure the causal impact of a change in the cohort size on effort. This measure eliminates any potential classroom-specific, student-specific, or time-specific confounds that often plague classroom studies. The 100-Student Quiz elicits over 3 percent more effort than the 10-Student Quiz, and this difference is statistically significant. Thus, a costless change in grading policy adds up to considerable gains in student effort over a semester.

Next, I use student GPA data as a proxy for ability in order to test the heterogeneous impact of cohort size on students with different abilities. I confirm that students above the cutoff in the population distribution exert more effort on the 100-Student Quiz, but I also find that students below the cutoff exert more effort on the 100-Student Quiz. The lowest-ability students, however, exert significantly less effort on the 100-Student Quiz. Even though the smaller cohort provides an opportunity for low-ability students to take advantage of the uncertainty in their draw of opponents, only the lowest ability students
do so. In my analysis, I test possible explanations of this allocation failure.

My results indicate that the randomness of the smaller cohort has a negative effect on aggregate student effort. This unintended consequence of reducing class sizes seems to conflict with a large literature on the benefits of class size reductions. This is a false dichotomy. Studies such as Mosteller (1995), Angrist and Lavy (1999), Glass and Smith (1979), Hoxby (2000), and Krueger (2003) explore the full effect of class size reductions and generally agree on their benefits. My paper, on the other hand, focuses on one partial effect—the effort response under relative grading. Indeed, addressing the unintended consequences that I identify could make class size reductions even more effective.

In addition to the positive aggregate effect of increasing cohort size on student effort, my results uncover meaningful differences in the intensity of this effect. Heterogeneity between low- and high-types is consistent with both the theoretical literature (Amann and Leininger (1996); Krishna and Morgan (1997); Olszewski and Siegel (2016)) and experimental literature (Müller and Schotter (2010); Noussair and Silver (2006)) on private value all-pay auctions and tournaments—environments designed to simulate real-effort tasks. Andreoni and Brownback (2017) explore the effect of increasing the cohort size on all-pay auction bidding and find, along with a significant increase in aggregate bidding, that the effect is much stronger on high-types than low-types. Harbring and Irlenbusch (2005) and Orrison et al. (2004) both focus their attention on mean effort in laboratory tournaments, finding mixed results about the impact of small changes in the cohort size. Gill et al. (2015) explore real-effort responses to rank in a given contest finding that people exert a disproportionate amount of effort to avoid last place or earn first place. While students in my experiment do not learn their rank, this motivation could interact with cohort sizes by varying the proportion of students receiving the highest rank.

To fix ideas, I refer to grading mechanisms throughout this paper, but this should not distract from the generality of my results. Relative awarding mechanisms are found in job promotion contests, performance bonuses, and lobbying contests, among others. Since the costs of effort, the means of exerting it, and the ways in which heterogeneous abilities manifest themselves are similar across academic and professional settings, my results provide a framework for predicting how agents will respond to changes in their environment when their performance is evaluated relative to their peers.

It is important to note that I am not comparing relative evaluation to other awarding mechanisms. I only consider the effect of changes in cohort size conditional on relative evaluation. This analysis is important for any environment where relative evaluation is unavoidable (because of legislation, for

---

3Angrist et al. (2017) attempt to replicate this result with a larger and more recent dataset and find conflicting evidence.
example) or in cases where the influence of relative ranking is perceived by the agents (in a cohort of graduate students, for example).\footnote{See Paredes (2016) for a comparison of behavior under different evaluation mechanisms.}

\section{A Model of Academic Effort}

Andreoni and Brownback (2017) develop a simple framework outlining the incentives for effort when grades are awarded on a relative basis. This model supposes there are $N$ students exerting costly effort in order to increase their chances of winning one of $M \equiv P \times N$ prizes, where $1 - P$ is the cutoff percentile that distinguishes winners from losers. Prizes in this environment take the form of “pass” grades. Students compete by exerting effort, $e_i$, which has a constant marginal cost of 1. “Pass” grades are awarded to the $M$ students with the highest effort.

Further, each student draws an independent private value for a “pass” grade from a uniform distribution from 0 to 1. That is, $v_i \sim U[0, 1]$. This value for a “pass” can be thought of as the student’s ability or her percentile rank. High-ability students gain more surplus from a “pass” grade through lower costs, greater motivation, or other factors.\footnote{A framework of independent private effort costs reduces to an affine transformation of the utility function. Thus, behavior is identical under these isomorphic utility specifications.}

I consider class sizes of 10 and 100 where 70 percent of students receive passing grades. That is, $N \in \{10, 100\}$ and $P = 0.7$. A naïve approach to the student effort-selection problem is instructive for understanding differences in these two class sizes. Figure 1 captures the probability that a student is randomly drawn into a cohort where they are in the top 70 percent of students. This reveals the key intuition of the paper: The uncertainty of the 10-Student Quiz increases the likelihood that a low-ability student encounters a cohort 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. These changes in probability correspond to changes in the returns to effort by students, meaning that the cohort size should have a predictable but heterogeneous effect on student effort.

The incentives that the model identifies are only a few of many often unobservable incentives such as intrinsic motivation, instructor attention, and peer-learning. This limits the ability of a formal model to predict the level of effort a student chooses. Therefore, I rely on this formalization to develop predictions about the difference in student effort between cohort sizes, holding other incentives for effort constant. Andreoni and Brownback (2017) provide the following generic predictions about relative effort in contests
of different sizes but with a fixed proportion of prizes:

- **Prediction 1**: Aggregate effort increases in the size of the contest increases.

- **Prediction 2**: There exists a single-crossing point for student abilities, \( v^* \), above which students increase their effort as the contest size increases and below which students decrease their effort as the contest size increases.

The underlying intuition for these predictions is simple. On average, larger cohorts have less uncertainty in the returns to effort. Since the costs of effort are certain, decreasing the uncertainty around the marginal benefit of effort induces higher effort.\(^6\) However, this increase in effort is not uniform across students. Low-ability students benefit from greater randomness in the draw of their cohort, and therefore put forth more effort in the small cohorts. Conversely, randomness is detrimental to high-ability students, so their effort decreases. This generates a single-crossing point in the effort functions for large and small classes.

\(^6\)This is generically true for any proportion of winners when considering changes in the size of contest.
2.1 Predictions From the Model

My experiment pairs the 10- and 100-Student Quizzes, so my analysis takes their difference—specifically, the 100-Student Quiz duration minus the 10-Student Quiz duration—as the dependent variable. I refer to this difference as the “treatment effect.” This within-subjects design cleanly measures the treatment effect by controlling for student-specific heterogeneity. The model predicts the treatment effect to be heterogeneous across student abilities. Below, I translate the model’s two primary predictions about the treatment effect into potential objectives of a mechanism designer.\(^7\)

**Objective 1: Maximize aggregate effort**

The 100-student cohort has a more predictable distribution of abilities. This decrease in uncertainty raises the marginal benefit of effort for high-ability students. Andreoni and Brownback (2017) find that the increased effort from high-ability students outweighs the decreased effort from low-ability students as the cohort grows.

**Objective 2: Balance gains from low- and high-ability students**

The model predicts that heterogeneity in the treatment effect will allocate gains from larger cohorts disproportionately to high-ability students. The treatment effect is negative for low-ability students up until a single-crossing point between the equilibrium effort functions. I will refer to this single-crossing point as \(v^*\). The specific location of \(v^*\) will depend on the distribution of abilities and individual utility functions, but based on Figure 1, it is natural to think of \(v^*\) as corresponding with the cutoff percentile of \(P = 0.3\).\(^8\)

A designer may want to balance the gains in effort across students with different abilities. To do so, it is critical to understand the way in which class size affects students heterogeneously. Accordingly, I test the model’s ability to inform designers about the tension between the aggregate gains of Objective 1 and the distributional consequences of Objective 2.

---

\(^7\)While effort is a continuous variable, scores are discrete, complicating the application of the model to our setting. This may generate small deviations from our predictions at the margins of different scores, but the average effect of the contest size should be consistent with our model. See Cohen and Sela (2007) for a theoretical treatment of how discrete outcomes and ties affect contests.

\(^8\)While the single-crossing point in the equilibrium effort functions 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.
3 Experimental Design

My experiment takes the paired-auction design used in Andreoni and Brownback (2017) and adapts it for a classroom context, replacing auctions with grading cohorts. Each week for five weeks of the semester, students simultaneously receive a pair of online quizzes, a 10-Student Quiz and a 100-Student Quiz. The online environment then captures student behavior on each quiz.

I analyze the difference in behavior between the two simultaneous quizzes in order to control for class-specific, student-specific, and time-specific effects. This paired design provides a powerful test of student responses to cohort size while holding constant all classroom characteristics. The majority of studies on class size reduction focus on the full effect, which is generally large and positive but combines the effects of changes in student effort, instructional quality, student resources, assignment frequency and difficulty, and others. Since my experiment operates within-student and within-week, all factors associated with the classroom experience, instruction, or resources are held constant to isolate the effect of cohort size on student effort.

3.1 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 online announcement at the beginning of the course. All students were given the opportunity to withdraw their information from the study, though none exercised this option. The announcement can be found in appendix Section A.1.

3.2 Quiz Design, Scoring, and Randomization

There were five Quiz Weeks in the quarter. At noon on Thursday of each Quiz Week, two different quizzes covering material from the previous week were posted to the course website. Both quizzes were due by 5 pm the following day. Each quiz consisted of four multiple choice or numeric questions. The quizzes had a time limit of 30 minutes to ensure that the quiz was given focused attention with little time spent idle and that the time recorded for students reflects the actual effort they exerted on the quiz.

---

9This draws from the paired auction designs of Kagel and Levin (1993) and Andreoni et al. (2007).
10Economics is the second largest major at UCSD and has “capped” enrollment. Intermediate microeconomics is typically taken in the second year after all calculus prerequisites have been completed. The enrollment limits and prerequisites generate a positive selection of students. Section 3.4 describes this in more detail.
Students could take the quizzes in any order.

I refer to the content of the two quizzes of a pair as “Quiz A” and “Quiz B.” The two quizzes were presented to students in the same order with Quiz A always first. I then randomly assigned cohort sizes to each quiz. One of the quizzes was assigned to be the 10-Student Quiz and the other the 100-Student Quiz. Thus, for half of the students, Quiz A was graded relative to a cohort of 10 and for the other half, it was graded relative to a cohort of 100. In each case, Quiz B was graded relative to the other cohort size. The questions on Quizzes A and B covered the same material but were designed to have as little overlap as possible to eliminate order effects. Prior to starting a quiz, the students observed the grading treatments, but not the quiz content, so selection effects could only be based on grading treatment. No student was told what treatments any other student received. Table 1 shows the balance across treatments.

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

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

Students were informed that all quizzes would be graded relative to a cohort of students randomly selected each Quiz Week who took the same quiz (A or B) under the same grading treatment (10 or 100). The number of questions correct determined the score for each student. The top 7 scores received 3 points in each 10 student cohort, and the top 70 scores received 3 points in each 100 student cohort. Students who completed the quiz but were not in the top 70 percent received 1 point and non-participants received 0 points. The quizzes amounted to approximately 13 percent of the total grade for the class, providing strong incentives for effort.

3.3 Effort

The time at which every quiz was started and completed was recorded to the millisecond. My analysis takes the difference in the amount of time a student spent on the two quizzes as a measure of the relative...
effort spent on each quiz. This measure of relative effort will reveal which quiz the student believed to hold the greater relative returns to her effort. Random assignment leaves student ability and quiz difficulty independent of the cohort size, thus if a student spends statistically more time on quizzes of one cohort size, it indicates that she believes her marginal product is higher for that cohort size.

It may seem that longer times spent on a quiz could be an indication that the student was actually exerting less effort on the quiz by, for instance, distractedly answering questions while surfing the Internet, but it is important to remember that the level effect of each cohort size is unimportant for my analysis, as I only consider the difference in time. It is therefore the marginal difference between the two quizzes that matters. I assert that—conditional on the average level of attention (or inattention) a student gives towards quizzes this week—the quiz on which the student spends more time is the quiz on which the student exerted greater effort.\footnote{At very least, the student is expending more of a costly resource (time) to complete the quiz.}

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, with 86 percent of students waiting less than an hour between the two quizzes and a median interval of 32 minutes.

3.4 Ability

I use a student’s grade point average (GPA) as a proxy for her ability.\footnote{Administrative delays prevented me from receiving this information until after the quarter. Therefore, the measure is not entirely exogenous from quiz performance, but is arguably exogenous from relative quiz performance. Regardless, the quizzes amounted to only 13 percent of a student’s grade in one of dozens of classes she would have taken, so I do not see this as a major problem.} GPA provides a clean instrument because there is no sense in which quiz effort and GPA are substitutable. Using exam performance or course grades would introduce endogeneity into the analysis because of the trade-off between allocating effort to exams or quizzes. There may be a correlation between student effort levels and GPA, but my analysis eliminates these level effects by taking the difference in effort between two quizzes.

Figure 2 shows a histogram of student GPAs in the class. To proxy ability, I use the percentile rank of a given GPA. This maps back to the model by giving a uniform distribution of abilities from 0 to 1. The GPA at the 30th percentile cutoff is 2.72, while the median and mean GPA are 3.0 and 2.99, respectively. Seven students have a GPA of 4.0, and only one student has the minimum GPA of 1.0.
4 Results

579 students submitted 5,094 online quizzes in this experiment. Of those, 2,546 were 10-Student Quizzes, and 2,548 were 100-Student Quizzes. My analysis includes all recorded times from all completed pairs of quizzes. Table 2 reports the means and standard deviations of the quiz durations for each cohort size in the first two columns. The unpaired difference in means presented in Column 3 provides a descriptive but weak evaluation of the treatment effect.

Table 2: Duration of Quizzes in Minutes

<table>
<thead>
<tr>
<th>Quiz100</th>
<th>Quiz10</th>
<th>Difference (Unpaired)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Duration</td>
<td>14.97</td>
<td>14.58</td>
</tr>
<tr>
<td>Standard Deviation/Error</td>
<td>(9.21)</td>
<td>(9.00)</td>
</tr>
</tbody>
</table>

A potential complication in these results is that the order in which students completed the quizzes is endogenous. The effect of this endogenous selection is small—51.4% of 100-Student Quizzes were pre-
sented first, while 55.3% were completed first—but is statistically significant. Fortunately, the randomly
assigned presentation order provides a randomly assigned, relevant instrument for completion order.

Table 3 demonstrates that presentation order is a relevant instrument in column 1. Column 2 shows
that the mechanically random presentation order has a spurious correlation with student GPA despite
being re-randomized each week. I could not block-randomize on GPA because I did not gain access to
student GPAs until after the quarter had ended. The explanatory power of this relationship 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, I will present all results with and without the instrument for quiz completion order.

Table 3: Testing the Relevance and Validity of the Instrument

<table>
<thead>
<tr>
<th></th>
<th>Pr[Quiz100 completed first]</th>
<th>Pr[Quiz100 presented first]</th>
<th>Pr[Quiz100 completed first]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiz100 presented first</td>
<td>0.293***</td>
<td>0.292***</td>
<td>0.292***</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>GPA</td>
<td>0.042**</td>
<td>0.022</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.404***</td>
<td>0.384***</td>
<td>0.337***</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.05)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.087</td>
<td>0.002</td>
<td>0.088</td>
</tr>
<tr>
<td>N</td>
<td>2,486</td>
<td>2,486</td>
<td>2,486</td>
</tr>
</tbody>
</table>

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

Since no probabilities are near the endpoints of the [0, 1] interval, I employ a linear probability model.
Standard errors clustered at the subject level.

4.1 Objective 1: Maximize aggregate effort

Table 4 tests the straightforward prediction that mean effort increases with the cohort size. Column 1
shows that the mean difference between the 100- and 10-Student quizzes is 0.45 minutes, representing
an increase in effort of 3.1% from the 10-Student Quiz mean. Column 2 instruments for the endogenous
ordering of quizzes and estimates the independent effect of cohort size on effort. Using this instrument,
the mean difference rises to 3.83 minutes, an increase of 26.2% over the 10-Student Quiz mean. This
dramatic increase should be interpreted with caution because of the added noise from the relatively coarse
instrument.\(^{16}\) Nonetheless, it suggests that the larger quiz may garner up to an extra minute of effort
per question, an effect with large economic significance.

\(^{16}\)Inserting direct controls for completion order is a biased measure, since the endogenous order of completion is collinear
with the treatment effect. I have included this analysis in appendix Section C.1 to show that the sign of the treatment effect
is unaffected, though the magnitude and significance are attenuated.
Table 4: Mean Difference between 100- and 10-Student Quiz Duration

<table>
<thead>
<tr>
<th>OLS Instrumental Variables</th>
<th>Quiz&lt;sub&gt;100&lt;/sub&gt; Taken First</th>
<th>Constant</th>
<th>( v_i &lt; v^* \implies GPA &lt; 2.72 )</th>
<th>( v_i \geq v^* \implies GPA \geq 2.72 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6.113***</td>
<td>0.450**</td>
<td>0.706**</td>
<td>0.348</td>
<td></td>
</tr>
<tr>
<td>(1.52)</td>
<td>(0.19)</td>
<td>(0.31)</td>
<td>(0.23)</td>
<td></td>
</tr>
<tr>
<td>N 2,507</td>
<td>N 2,507</td>
<td>0.954***</td>
<td>0.699***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.32)</td>
<td>(0.25)</td>
<td></td>
</tr>
</tbody>
</table>

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

Values reported in minutes. Standard errors clustered at the student level.

### 4.2 Objective 2: Balance gains from low- and high-ability students

A classroom designer may want to offer students similar incentives for effort, but the model predicts that cohort size will necessarily influence the distribution of incentives across ability levels. To test for this heterogeneity, I estimate the treatment effect above and below the predicted single crossing point, \( v^* = 0.3 \), or a GPA of 2.72. Recall that the model predicts a negative treatment effect for students below \( v^* \) and a positive treatment effect for students above \( v^* \). As a second test of heterogeneity, I impose continuity on the treatment effect to see how it evolves with students’ abilities. I use this approximation to fit an estimated single crossing point, \( \hat{v}^* \), to the data and compare it to the model’s prediction.

To estimate the treatment effect above and below \( v^* \), I regress the treatment effect on indicator variables for a students above and below \( v^* = 0.3 \). Table 5 presents these results. The first column shows a significant and positive treatment effect for low-ability students and a positive but insignificant treatment effect for high-ability students. In the second column, I repeat the analysis instrumenting for the endogenous completion order. The treatment effect is positive for students above and below \( v^* \), and both are significant.

Table 5: Difference between 100- and 10-Student Quiz Duration by Ability

<table>
<thead>
<tr>
<th>OLS Instrumental Variables</th>
<th>Quiz&lt;sub&gt;100&lt;/sub&gt; Taken First</th>
<th>( v_i &lt; v^* \implies GPA &lt; 2.72 )</th>
<th>( v_i \geq v^* \implies GPA \geq 2.72 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6.043***</td>
<td>0.706**</td>
<td>0.348</td>
<td></td>
</tr>
<tr>
<td>(1.26)</td>
<td>(0.31)</td>
<td>(0.23)</td>
<td></td>
</tr>
<tr>
<td>N 2,507</td>
<td>N 2,507</td>
<td>0.954***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.32)</td>
<td></td>
</tr>
</tbody>
</table>

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

Values reported in minutes. Standard errors clustered at the student level.

The results of Table 5 reject the prediction that changes in the cohort size unevenly affect students
whose abilities lie on either side of the cutoff. We do not find evidence of a trade-off between increasing aggregate effort and increasing effort from low-ability students. Indeed, the largest gains in effort come from low-ability students. Below, I uncover heterogeneity within low-ability students.

Imposing continuity on the treatment effect allows us to analyze the heterogeneity at a finer resolution. I fit an estimated single crossing point, $v^*$, to the data in order to provide a second test of the model’s predictions about heterogeneous effects of changes in the cohort size. Figure 3 estimates the treatment effect using a locally linear polynomial smoothing function.\textsuperscript{17}

![Local Polynomial Fit of the Treatment Effect (in Minutes)](image)

Figure 3: Local Polynomial Fit of the Treatment Effect (in Minutes)

In contrast to Table 5, Figure 3 finds some support for a conflict between increasing aggregate effort and increasing effort from low-ability students but only among the lowest-ability students. The point at which Figure 3 estimates the treatment effect crosses from negative to positive is approximately a GPA of 1.74, or $v^* = 0.02$. Thus, only the bottom 2 percent of students are negatively affected by the increased cohort size.\textsuperscript{18} This tail represents a small percent of the population and should be interpreted with caution because of its small sample size. Nonetheless, this population is economically important,

\textsuperscript{17}I include a less restrictive test of evolution of the treatment effect in appendix Section C.2, showing that the results are not driven by the continuity restriction.

\textsuperscript{18}Estimating the single-crossing point by maximizing the number of differences with the correctly predicted sign results in a similar estimate of $v^* = 0.03$ or a GPA of 1.81
since policy makers are often motivated by the interests of the most at-risk students. \(^{19}\)

This estimated single-crossing point highlights strong heterogeneity among students below the theoretical cutoff. While the majority of these low-ability students see positive and sometimes large treatment effects, the lowest-ability students significantly decrease their effort on the 100-Student Quiz. The estimated treatment effect is minimized at the lowest GPA, where the predicted treatment effect is \(-2.93\) minutes or a 19.3% reduction.

For students above the cutoff, the treatment effect is everywhere positive but decreasing in GPA, nearing zero at the upper limit. This pattern accords with Figure 1, where the probability that a high-ability student is drawn into a cohort in which they are above the cutoff is near 1 for both cohort sizes.

### 4.3 Robustness: Stated Preference

From the effort exerted on each quiz, I can measure the revealed preferences of students towards the larger quiz, on average. To corroborate this, I measured the stated preferences of students at the end of the quarter with the question:

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

Table 6 shows the likelihood that a student stated a preference for the 10-Student Quiz. Column 2 breaks down these preferences by a student GPA. In general, students prefer to be graded relative to a cohort of 100 students, but students above the cutoff are significantly more likely to hold that preference. This offers support for the prediction that higher ability students benefit more from larger cohorts. \(^{20}\)

### Table 6: Stated Preference for the 10-Student Quiz

<table>
<thead>
<tr>
<th>GPA &lt; 2.72</th>
<th>Pr(Prefer Quiz(_{10}))</th>
<th>0.360**</th>
<th>(\text{(0.15)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.081***</td>
<td>-1.196***</td>
<td>(\text{(0.07)})</td>
</tr>
<tr>
<td>N</td>
<td>493</td>
<td>493</td>
<td></td>
</tr>
</tbody>
</table>

\(^{19}\)This suggests that confusion is not driving the deviations of the lower-ability student from the model’s predictions, as the lowest-ability students seem to understand where their returns to effort are highest.

\(^{20}\)Alternatively, it could be the case that students below the cutoff, who exerted significantly more effort on the 100-Student Quiz, are more likely to prefer the 10-Student Quiz because they recalled not working as hard on it.
4.4 Robustness: Stationary Behavior

My predictions assume that behavior remains stationary across weeks. To test this assumption, I plot the mean treatment effect by week in Figure 4. The treatment effect does not show any clear pattern across the 5 weeks of the experiment. Clearly, each individual week will have a larger confidence interval as a result of the decreased sample size, though no week shows more than a slightly negative estimate of the treatment effect. The confidence intervals of every week include the overall mean treatment effect.

![Figure 4: Mean Treatment Effect by Week (in Minutes)](image)

4.5 Robustness: Alternative Motivations for Effort

My within-subjects design controls for individual-specific variation in motivation by differencing out the level effects from a given student. Implicitly, however, this assumes that these motivations for effort do not correlate with cohort size. It is important that my results address these alternative motivations for effort that may drive differences between cohort sizes. Specifically, I address the possibility that intrinsic motivation, risk aversion, gender, or beliefs may be responsible for the observed treatment effects.

Figure 5 addresses the role of intrinsic motivation in the treatment effect. It shows that average time spent on a quiz does indeed increase in a student’s GPA. However, this is driven almost exclusively by the behavior of the lowest-ability students. The correlation between GPA and aggregate effort is near-zero above GPA = 2.0, making it unlikely that level-effects in effort are falsely indicating that the treatment effect is growing while the proportion of effort is remaining constant.
Risk preferences are also a natural explanation for the treatment effect since the two treatments bear different levels of uncertainty.\textsuperscript{21} To understand the effect of risk preferences, I surveyed students at the end of the quarter with a question from Dohmen et al. (2011) about their propensity for risk-taking.\textsuperscript{22}

Figure 6 plots the effort curves divided by risk preferences. There is a level-effect in effort for low-ability, risk-averse students, but this will be eliminated when I take the difference in quiz effort. There is little difference between risk-averse and risk-tolerant students in the differential effort allocated to each quiz. Moreover, the data reject any relationship between GPA and risk aversion ($t = -0.71$, $P = 0.478$), casting doubt on its ability to explain any heterogeneity in the treatment effect.\textsuperscript{23}

As a final motivation for effort, I explore the relationship between gender and the treatment effect. Gender differences in risk preferences and competitiveness are well-documented, often finding large effects.\textsuperscript{24} However, gender and GPA are not significantly related ($t = -1.44$, $P = 0.150$), making it unlikely that the heterogeneity in the treatment effect is driven by gender-differences.\textsuperscript{25} Additionally, Figure 7 plots the treatment effect by gender. Men appear to be slightly more sensitive to changes in cohort size but the qualitative features of the treatment effect are similar across genders.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{quiz_effort_by_gpa}
\caption{Quiz Effort by GPA (Min)}
\end{figure}

\textsuperscript{21}Standard models of risk aversion assume preferences are linear in probabilities. Since the two treatments manipulate the probability of being awarded a given grade, they would not identify a difference at face value. Nonetheless, risk preferences may affect effort choices because of how effort influences the variance in possible outcomes. Alternatively, additional effort could create disappointment-averse tendencies (Gul (1991); Gill and Prowse (2012)).

\textsuperscript{22}Dohmen et al. (2011) show that the question, “How likely are you to take risks, in general.” correlates closely with revealed risk preferences.

\textsuperscript{23}This regression can be found in appendix Section C.3.

\textsuperscript{24}See Niederle and Vesterlund (2011) for a review of the literature on gender and competitive preferences.

\textsuperscript{25}This regression can be found in appendix Section C.4.
4.6 “Cursed” Beliefs or Overconfidence

All “rational” models depend on students holding accurate beliefs by drawing correct inferences about the abilities of their classmates. Deviations from predicted behavior may be a result of mistaken beliefs. Simple overconfidence may cause students to believe—and thus behave like—their ability to be higher than it is. This could explain the large percentage of low-ability students increasing their effort in the larger cohort.

This overconfidence could arise from a psychological predisposition towards over-optimistic self-evaluation or from a failure to draw accurate inference from available information. In order to draw accurate inference about their relative ability, students must account for the positive selection of classmates who enroll in upper-division economics courses. Students with “Cursed” beliefs (Eyster and Rabin (2005)), on the other hand, may believe enrollment decisions to be independent of ability and assume that the distribution of students is similar to the distribution from lower-division courses. Figure 8 plots the CDF for the actual GPA distribution from the course along with a “cursed” distribution from all lower-division UCSD course grades from the 5 years prior to the experiment.\(^\text{26}\)

Cursed students who fail to account for the fact the low GPA students select out of upper-division economics courses believe their percentile rank to be based on a relatively weaker distribution. For example, a cursed student with a GPA of 2.24 perceives his percentile rank to be 30, while his actual percentile rank is approximately 10. While this may be a compelling story for the origins of overconfidence, in

\(^{26}\)To construct the GPA distribution, I aggregate the grade distributions from all lower-division courses in all departments to find a composite GPA for each percentile rank. I then suppose there are students at the 1st percentile, 2nd percentile, \ldots, 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.
data it is observationally indistinguishable from a simpler model of over confidence.

![Graph](image)

**Figure 8: Percentile Ranks of Students**

5 Policy Prescriptions

Before addressing the policy implications of my experiment, I must first address the social benefit of inducing greater effort on classroom quizzes. One approach is to consider the effect of the treatments on the student scores. This analysis shows no significant treatment effects ($t = -0.28, P = 0.781$).27 There are three primary reasons why I believe time spent on the quizzes is a more valuable measure of treatment effects. First, time is the costly resource that students exert in order to achieve higher scores and the decision of how much time to allocate to any task is based on an ex-ante perception of the relative returns to effort while scores are an ex-post measure of outcomes. For this reason, I believe the amount of time allocated to a quiz better captures the motivation behind a student’s studying decision. Second, quiz scores are a much coarser measure than quiz duration, meaning that my experiment is underpowered to identify a treatment effect in scores. Finally, random fluctuations in quiz difficulty drive much more variation in scores than quiz effort does. A sufficiently long panel could overcome this, but each student only saw five pairs of quizzes, so the effect of changes in effort may not be perceptible.

The connection between greater effort in studying or attendance and greater academic performance is well-documented (Romer (1993); Stinebrickner and Stinebrickner (2008); De Fraja et al. (2010); Arulampalam et al. (2012)). Therefore, understanding the effect of changes in the grading environment on the allocation of effort by students can help increase their academic performance. This is of particular

---

27 The full specification of this regression is found in Table 11 in appendix Section C.5.
relevance in an education setting, since this policy change is potentially costless. Thus, even if the gains are small, the costs are even smaller.

Similar to the laboratory results of Andreoni and Brownback (2017), I find that aggregate effort increases as the size of the contest increases. Thus, mechanism designers with preferences over aggregate effort should implement an evaluation mechanism with the lowest possible variance in order to maximize the total effort exerted. This could be accomplished through, for instance, combining multiple classes into one grading unit, pooling all employees together when making promotions or awarding bonuses, or combining multiple job openings into one search process.

Andreoni and Brownback (2017) found that, while decreasing the variance increases aggregate effort, it is unevenly distributed between high- and low-types. In contrast, I find little evidence that low-types decrease effort in larger cohorts. Indeed, low-ability students see the largest increases when moving to larger cohorts. Only among the lowest ability students did smaller cohorts with higher variance 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 agent is awarded a prize. If motivating the lowest-ability agents is part of the mechanism designer’s objective, then smaller cohorts may accomplish this, though it would come at the cost of other low-ability students. This could be achieved by, for example, breaking up large classes and grading students based on their performance relative to their discussion section or by choosing promotions or awarding bonuses within working groups.

While I conducted my experiment under a Pass-Fail grading structure, the results extend naturally to more granular grading scales, too.28 As class sizes grow, students will again separate to the extremes on either side of a cutoff. In this setting, students simply have more cutoffs to divide them.

Feedback in this environment plays an interesting role. The failure of many low-ability students to take advantage of the randomness of smaller cohorts suggests that they may hold biased beliefs about their relative ability. Better feedback about relative ability may allow students to make better decisions about effort allocation, but may also reduce their aggregate effort, since their mis-allocated effort contributed to the greater aggregate effort of the 100-Student Quiz. These misaligned incentives may put classroom designers and students at odds with respect to optimal feedback.

28See Dubey and Geanakoplos (2010) for a discussion of optimal granularity in grading.
6 Conclusion

In this paper, I cleanly isolate the effect of class size on effort induced by relative grading mechanisms. Understanding these student responses adds to the discussion of optimal class size by identifying one unintended consequence of class size reductions.

The probabilities plotted in Figure 1 suggest, and Andreoni and Brownback (2017) theoretically demonstrate that increasing the cohort size in order to increase aggregate effort will decrease effort from the weakest students. I found no evidence of this trade-off for the majority of low-ability students. In fact, most low-ability students exerted greater effort in the larger and less random cohorts. The inability or unwillingness of low-ability students to take advantage of the randomness of the smaller cohorts suggest that low-ability students may be subject to certain behavioral biases that affect their beliefs about their relative ability. Further experimentation is needed to confirm this.

The trade-off between mean effort and effort from low-ability students may have been overstated, but it was not non-existent. The weakest students exerted significantly more effort in the smaller cohorts, making a compelling case for these more random environments when motivating the most at-risk population is the key objective.

My results provide a basis for exploring how grading mechanisms interact with class size. I hope that my investigation of the potential unintended consequences of reducing class size contributes to the discussion of optimal classroom design.
References


A Appendix A: Experimental Procedures

A.1 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: 4,4,4,3,3,1,1,1,1,1. 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.
A.2 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 (99) of your classmates. If your Score is among the top 7 (70), you will receive a Grade of 3/3 for this quiz. If your score is among the bottom 3, you will receive a Grade of 1/3 simply for participating.

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

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

All answers will be in WHOLE NUMBERS.”

A.3 Online Environment
A.4 Post-Experiment Survey

Survey on TED Quizzes for Econ 100A

PID: _______________

1. If we were to offer quizzes in your next econ class, but graded all quizzes in one way, which would you prefer?
   □ 100-Student Quiz  □ 10-Student Quiz

2. Which Quiz did you work harder on?
   1 2 3 4 5 6 7
   Worked Harder on 10-student Quiz
   The Same on both
   Worked Harder on 100-student Quiz

3. How hard did you work on the Quizzes, in general?
   1 2 3 4 5 6 7
   (Not Hard)
   (Very Hard)

4. Which Quiz made you learn more?
   □ 10-Student Quiz  □ 100-Student Quiz  □ They were the same

5. Would you prefer less challenging or more challenging Quizzes?
   1 2 3 4 5 6 7
   (Less Challenging)
   (More Challenging)

6. Would you prefer Quizzes with less or more predictable “Cutoffs”?
   1 2 3 4 5 6 7
   (Less Predictable)
   (More Predictable)

7. How likely are you to take risks, in general?
   1 2 3 4 5 6 7
   (Not Likely)
   (Very Likely)

Thank you so much for your feedback and for your patience this quarter!
B Appendix B: Theory

B.1 Equilibrium Effort

Andreoni and Brownback (2017) find that the following first order differential equation characterizes the equilibrium effort function for $N$ players, $M$ prizes, a uniform distribution of abilities, and constant cost of effort:

$$E'(v_i) = v_i \times \sum_{k=N-M}^{N-1} \left\{ \frac{(N-1)!}{(N-1-k)!k!} \times \left[ k \times f(v_i) F(v_i)^{k-1} (1-F(v_i))^{N-1-k} - (N-1-k) \times f(v_i) F(v_i)^k (1-F(v_i))^{N-2-k} \right] \right\}.$$  \hspace{1cm} (1)

To solve this, I set the initial condition to $E(0) = 0$. This is simply a dominance argument. Any positive effort would exceed the valuation of the student, so is strictly dominated by $e = 0$. Negative effort is not possible, so the optimal effort for a student with $v_i = 0$ must be $e = 0$.

After setting the initial condition, I used Mathematica to solve the equation numerically for both $(N, M) = (10, 7)$, $(100, 70)$. While closed-form solutions exist, numerical solutions, which are much less computationally costly, are sufficient to graph the equilibrium effort in Figure 9 and generate predictions about the treatment effects.

![Figure 9: Effort at Equilibrium in the 10- and 100-Student Quizzes](image)

C Appendix C: Empirical Results

C.1 Direct Controls for Order

To directly control for quiz order, I will use fixed-effects to capture the student-week combination and insert indicator variables for the quiz completed first. As mentioned, this measure introduces multi-
collinearity between the treatment effect and the quiz order. It is important to note that the sign of
the effect does not change, though its magnitude and significance are diminished.

Table 7: Quiz Duration (in minutes)

<table>
<thead>
<tr>
<th></th>
<th>Quiz Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Quiz</td>
<td>2.634***</td>
</tr>
<tr>
<td></td>
<td>(0.19)</td>
</tr>
<tr>
<td>100-Student Quiz</td>
<td>0.167</td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
</tr>
<tr>
<td>Constant</td>
<td>13.368***</td>
</tr>
<tr>
<td></td>
<td>(0.12)</td>
</tr>
<tr>
<td>Fixed Effects</td>
<td>Student-Week</td>
</tr>
<tr>
<td>N</td>
<td>5012</td>
</tr>
</tbody>
</table>

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

C.2 Semi-Parametric Tests

For robustness, I use semi-parametric tests to confirm the heterogeneity shown in Figure 3. For this analysis, I first divide the data by the cutoff at the 30th percentile, corresponding to a GPA of 2.72, then split each of the two portions by their conditional median student’s GPA. The specification of the bins are given by,

\[
\begin{align*}
LowestBin_i &= 1_{\{GPA_i \in [1.2.433]\}} \\
LowBin_i &= 1_{\{GPA_i \in (2.433, 2.72]\}} \\
HighBin_i &= 1_{\{GPA_i \in (2.72, 3.246]\}} \\
HighestBin_i &= 1_{\{GPA_i \in (3.246, 4]\}}
\end{align*}
\]

Regressing the treatment effect on these 4 bins provides a semi-parametric test of the distribution of incentives. This regression is displayed in column 1 of Table 8. Column 2 repeats the regression, instrumenting for the completion order using the presentation order.

Under a less parametric specification, low-ability students still show positive treatment effects, on average. The lowest-ability students, however, appear to show near-zero treatment effects with the better low-ability students showing large and positive treatment effects. Similarly, high-ability students still show positive treatment effects, though the treatment effect for the highest ability students is not statistically significant. Thus, in general, the evolution of the treatment effect does not appear to depend critically on the specification of the analysis.
Table 8: Duration of the 100-Student Quiz minus the 10-Student Quiz (in minutes)

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>IV Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>LowestBin</td>
<td>-0.031</td>
<td>0.190</td>
</tr>
<tr>
<td>GPA_i ∈ [1, 2.433]</td>
<td>(0.43)</td>
<td>(0.43)</td>
</tr>
<tr>
<td>LowBin</td>
<td>1.454***</td>
<td>1.737***</td>
</tr>
<tr>
<td>GPA_i ∈ (2.433, 2.72]</td>
<td>(0.44)</td>
<td>(0.44)</td>
</tr>
<tr>
<td>HighBin</td>
<td>0.515</td>
<td>0.878***</td>
</tr>
<tr>
<td>GPA_i ∈ (2.72, 3.246]</td>
<td>(0.33)</td>
<td>(0.34)</td>
</tr>
<tr>
<td>HighestBin</td>
<td>0.151</td>
<td>0.496</td>
</tr>
<tr>
<td>GPA_i ∈ (3.246, 4]</td>
<td>(0.33)</td>
<td>(0.34)</td>
</tr>
<tr>
<td>Quiz100 First</td>
<td>-6.125***</td>
<td>(1.26)</td>
</tr>
<tr>
<td>Instrumented</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>2,506</td>
<td>2,506</td>
</tr>
</tbody>
</table>

*p < 0.10, **p < 0.05, ***p < 0.01
All standard errors clustered at the student level.
All times in minutes.

C.3 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

<table>
<thead>
<tr>
<th></th>
<th>Risk Measure (From 1-7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td>-0.094</td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
</tr>
<tr>
<td>Constant</td>
<td>4.202***</td>
</tr>
<tr>
<td></td>
<td>(0.41)</td>
</tr>
<tr>
<td>N</td>
<td>504</td>
</tr>
</tbody>
</table>

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

C.4 Gender and GPA

The gender and GPA of the subjects were recorded. Male GPAs are slightly lower, on average, but the relationship between GPA and gender is not significant.
Table 10: Regression of GPA on gender dummy

<table>
<thead>
<tr>
<th></th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>-0.073</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
</tr>
<tr>
<td>Constant</td>
<td>3.048</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
</tr>
</tbody>
</table>

N 512

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

C.5 Treatment Effect on Scores

All quizzes were scored out of 4 points. Column 2 of Table 11 tests for an endogeneity concern by measuring the effect of quiz order on scores. The effect of quiz ordering on relative score is small and not significant, rejecting the endogeneity concern. Thus, our estimate is derived from the random-effects linear regression represented in Column 1. The constant terms in both Columns 1 and 2 confirm the limited effect of the quiz size on the scores of students.

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

<table>
<thead>
<tr>
<th>Quiz100 First</th>
<th>Treatment Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.033</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.010</td>
</tr>
<tr>
<td></td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
</tr>
</tbody>
</table>

N 2,491 2,419

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

Note: Grades were 3 for “passing” students and 1 for others.