Economics of Education Review 62 (2018) 113–128

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

A classroom experiment on effort allocation under relative grading[☆]

Andy Brownback

Department of Economics, University of Arkansas, AR 72701, Fayetteville, United States

ABSTRACT

Grading on the curve is a form of relative evaluation similar to an all-pay auction or rank-order tournament. When students are drawn from a population distribution into a class, their realized distribution of abilities 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.[1] Since Texas high schools vary in size by orders of magnitude,[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

☆ I would like to thank James Andreoni for his guidance and support. Additional thanks are due to Prashant Bharadwaj, Simone Galperti, Laura Gee, Uri Gneezy, Michael Kuhn, Paul Niehaus, Marta Serra-Garcia, Joel Sobel, Charles Sprenger, Isabel Trevino, and Joel Watson. This study was approved by the UC San Diego IRB. Project #131456. The author declares that he has no relevant or material financial interests that relate to the research described in this paper.

E-mail address: abrownback@walton.uark.edu.

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.

2 For example, Valentine HS has fewer than 20 students, while Skyline HS has over 4,500.

https://doi.org/10.1016/j.econedurev.2017.11.005
Received 15 June 2017; Received in revised form 3 October 2017; Accepted 16 November 2017
Available online 21 November 2017

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 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.[3]

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 (Kessler and Vesterlund (2015); Levitt and List (2007)). In addition, the classroom environment allows me to explore the theory without relying on induced values for student ability. The 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). Equilibrium predictions and equilibrium behavior require common knowledge of the distribution of ability. In a natural environment, I cannot control the distribution nor can I make as strong of a case for common beliefs about the distribution, all of this may lead to weaker results with respect to heterogeneity.

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 Angrist and Lavy (1999); Glass and Smith (1979); Hoxby (2000); Mosteller (1995), and Krueger (2003) explore the *full effect* of class size reductions and generally agree on their benefits.[4] 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, Schotter, and Weigelt (2004) find mixed results about the impact of small changes in the cohort size. Both use a laboratory setting, focus their attention on mean effort, and have only small-scale contests with small-scale changes. In contrast, I test our hypotheses in a natural setting, explore heterogeneity, and change the contest size by an order of magnitude. Gill, Kissová, Lee, and Prowse (Forthcoming) 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 example) or in cases where the influence of relative ranking is perceived by the agents (in a cohort of graduate students, for example).[5]

[3] This result requires complete ignorance of other students' realized types. My results may not hold if types are observed or if knowledge of types is correlated with cohort size (e.g. students in small cohorts are more familiar with other students' types). Our experiment intentionally eliminates this possibility in order to reflect the conditions in the majority of college courses where there is insufficient student interaction to determine types regardless of cohort size.

[4] Angrist, Lavy, Leder-Luis, and Shany (2017) attempt to replicate this result with a larger and more recent dataset and find conflicting evidence.

[5] See Paredes (2016) and Czibor, Onderstal, Sloof, and van Praag (2014) for