Discouraged or Motivated at University? Goals and Reference Points in Academic Performance

Jaimie W. Lien
Department of Decision Sciences and Managerial Economics
The Chinese University of Hong Kong

Jie Zheng
Department of Economics
School of Economics and Management
Tsinghua University

Current Version: June 13th, 2015
Initial Version: August 15th, 2014

Abstract:[1]

Recent evidence has shown that individuals have reference points for performance which shape their effort provision on various tasks (ex. Berger and Pope, 2011; Pope and Simonsohn, 2011). We analyze the academic performance of about 700 university students enrolled in a large core course, hypothesizing that students' initial performances may induce reference-dependent effort provision around common or individual-specific reference points. We do not find strong evidence of the announced class average as a reference point, nor other plausible salient scores such as round number cutoffs (ie. 70, 80, 90). Instead, we find that effort behavior of students appears more consistent with a reference point near 100. That is, personal improvement tends to be relatively greater in the marginal gains frame than in the marginal loss frame, for almost any reference point candidate - except reference point candidates arbitrarily close to 100. Consideration of possible individual-based reference points, such as students' previous grades in other courses, supports the general finding that students seem to become discouraged in the loss frame rather than being motivated to 'catch up'. This closely resembles the *discouragement effect* found in Gill and Prowse's (2012) experimental real effort tournaments. An implication is that the final grade distribution tends to be regressive in students' initial performances.

Keywords: Reference-dependence, effort provision, education economics, academic achievement
JEL Codes: D03, I21

[1] jaimie.lien.tsinghuasem@gmail.com ; zhengjie@sem.tsinghua.edu.cn ; We are grateful to David Gill, Peter Hinrichs, Juanjuan Meng, Se Yan, Xi Weng, Guozhong Zhu, Li-An Zhou and seminar participants at Guanghua School of Management, for helpful comments and insights. We gratefully acknowledge research funding from the National Science Foundation of China, Ministry of Education of China, and Tsinghua University. All errors are our own.

There is a growing body of evidence pointing to the importance of goals as effort motivators. One important context where the influence of goal setting has obvious social benefits is in education. Educators have traditionally been using academic evaluations to motivate students to learn, from primary school all the way through the university level. According to the literature on reference-dependence, the structure of such evaluation schemes combined with natural psychological goal-setting tendencies, could have substantial predictive power in students' effort provision and their ultimate performance success.

In this paper we examine the performance of nearly 700 undergraduate students over three years, enrolled in a core required course at a large, top-ranked university in China. We test whether students show signs of reference-dependent effort provision around a publicly announced reference point: the class average on the midterm exam. If students use the class average as a reference point, reference-dependent loss aversion predicts that students performing slightly below the class average will experience a sensation of being 'behind' (ie. in the loss frame), and subsequently exert higher effort than those who performed slightly above this benchmark (ie. those in the gain frame). This hypothesis about effort provision being substantially affected by one's marginal position relative to a salient reference point, has been tested and confirmed in Berger and Pope (2011) for NBA and NCAA basketball teams, and Pope and Schweitzer (2011) in PGA golf tournaments. Thus, given that the class average performance is publicly announced each year in our field setting, it is reasonable to think that the average could serve as a motivator for students in our sample.

Our main empirical approach follows Berger and Pope (2011) which uses a local regression discontinuity design to test for reference point effects. To summarize our results briefly, we do not find evidence supporting the announced class average as an effective reference point for students, at least in our particular field setting. We also check for other plausible salient reference points such as round number scores (ie. 60, 70, 80, 90), given that evidence has been found for the influence of such salient numbers as goals in both sports and standardized university admissions exams (Pope and Simonsohn, 2011). We do not find supporting evidence for these reference point candidates either. Instead, we find that improvement tends to be (monotonically) increasing in students' initial performance in the midterm exam, consistent with a reference point close to a perfect score. This is striking given that for some of our performance improvement measures, the difficulty of improving one's performance is arguably higher when one's initial score is higher. Another measure of effort we consider is participation in an optional extra credit essay. Similarly, we do not find many significant reference points on this dimension either, except that the most likely students to participate, tend to be those who initially perform very well.

This effort pattern has important consequences for students' motivation. If effort is increasing in initial performance, it implies that students with lower initial scores are less incentivized to improve, even though they may be most in need of improvement. By contrast, those students who are already doing very well, expend the most effort in trying to improve, yet are least in "need" of improvement in their performance. The implied incentive structure is reminiscent of a winner-take-all tournament, in which individuals who have relatively low chances of obtaining a top score may reduce their effort.

We discuss the grading system at the university, and how it externally and internally motivates the effort and improvement patterns we detect. In particular, the university assigns final grades on students' transcripts as a numerical score, 0 through 100, but does not assign categories to the numerical values, such as A, B, C, D or F. This is a possible reason why we do not find reference-dependent effects around 'round' number scores such as 70, 80, 90, etc.; that is, the labels of A, B, C, D and F may serve as a key encouragement for students in using these numbers as goals. The reference candidate of

60 does hold a special meaning in our context, since 60 is the threshold for passing the course. Thus students in danger of not meeting this threshold have a clear external incentive to improve their performance. While we do find that the most improved students tend to be from the below 60 category, we do not find that those performing slightly worse than 60 are significantly more incentivized on average compared to those performing slightly better than 60, the key finding needed to identify 60 as a reference point.

Our analysis contributes to the literature on education economics, which addresses students' responses to evaluation systems. Costrell (1994) provides a model of socially optimal educational standards for a heterogeneous population. Betts and Grogger (2003) find that higher grading standards tend to increase test scores the most at the top of the score distribution, however, tending to reduce educational attainment. Grant and Green (2013) test the idea that university students may be motivated by the explicit discrete rewards provided by the A through F grading system. They find that across courses, at the university which is the subject of their study, the score thresholds for each letter grade do not seem to be effective effort motivators. Oettinger (2002) tests a similar question, also under a grading system which has discrete thresholds, and finds evidence that students' effort does respond to the discrete grading system. Our study builds upon this literature by testing the motivation of students in the absence of explicit grade thresholds, but under a hypothesis of loss aversion around various salient scores on the 0 to 100 scale.

To our knowledge, there are just a few studies which explicitly address the role of reference-dependence in academic performance.[2] Consistent with reference-dependence, Pope and Simonsohn (2011) find that falling short of a round number score on the SAT, makes it more likely that a student will retake the exam. They further find in baseball and lab experimental contexts, individuals also tend to have an aversion to performing worse than a salient round number, highlighting the importance of round numbers as motivators. Wang (2013) examines the effect of high-stakes testing on suicidal ideation in a reference-dependent loss-averse framework, finding evidence that high school students in South Korea are loss-averse around their practice college entrance exam score. Galdon and Gonzalbez (2013) conduct an experiment which tests the effect of academically performing below or above one's expectations, as well as the interaction effect with musical stimuli, on students' satisfaction levels.

Since our field setting involves curved scores relative to the performance of others, our study also relates to the branch of literature analyzing behavioral responses of students to relative performance feedback. Azmat and Iriberri (2010) examine the effect of a natural experiment which provided high school students with information on the average grade point average and the student's own performance. They find that the temporary policy change in fact increased the raw performance of all students. Specifically, the improvement was on average 5%, with the strongest effects among students at the top and bottom of the grading distribution. Tran and Zeckhauser (2012) found that students informed of their ranking on practice English exams, performed better than students who were not informed of their rankings. These studies focus on the effect of relative performance information, but do not look specifically for differential impacts of initial scores on effort in the grade distribution.

The findings of our paper are closely related to Gill and Prowse (2012), which shows that subjects in a real effort experiment are discouraged by rivals' efforts in a sequential move tournament.

[2] Our study is also related to a broader literature on reference dependent labor supply, such as Camerer, Babcock, Loewenstein and Thaler (1997), Farber (2005, 2008), Fehr and Goette (2007), and Crawford and Meng (2011). A difference between our setting and these, is that our students are not facing a labor supply problem in which their efforts are monetarily rewarded with a well-specified payment for each unit of effort or productivity.

They develop a model of disappointment aversion, based on loss aversion around an endogenously-determined, expectations-based reference point. The effort of the second player tends to be inversely associated with the observed effort of the first player, in accordance with this model. Our findings roughly replicate their findings in a field setting. Since we focus on students' performances after the announcement of the midterm score, our setting also has a sequential feature. However, in contrast to their experiment, our students may not have as clear a reference point as in the two player setting, due to the large number of students potentially 'competing' for grades. As in their study, we find that when students observe tough competition in the field through a first round of results, they tend to become discouraged rather than becoming further motivated.

There are serious distributive consequences to our finding that the main motivating reference point in our data is a near-perfect grade, and that for most of the candidate reference points, effort and improvement is lower on the loss side in a regressive manner. It implies that students who are not already performing near the top of the class, seem to have little motivation for self-improvement. Students already doing well will thus reap more benefits from the later part of the course (in terms of knowledge as well as score benefits resulting from higher effort), while students not doing well initially, will subsequently obtain fewer benefits. On the other hand, if students were reference-dependent around a class average or other near-median benchmark, the result would tend to be progressive in nature. We discuss the external validity of our finding to different grading schemes and institutional structures. In particular, we speculate that using lettered grading systems (ex. A, B, C, D, F) may give students at the university studied, more concrete reference-points near their initial performances to use as goals. There may be a behaviorally optimal partition of scores which induces a socially optimal effort provision across students. Recent theoretical work in this direction, without addressing loss aversion specifically, includes Dubey and Geanakoplos (2010).

The remainder of the paper is structured as follows: Section 2 describes the field setting and data; Section 3 discusses hypotheses and empirical strategy; Section 4 reports on results for the common reference points and individually-based reference points; Section 5 discusses external validity and interpretation of the results; Section 6 concludes.

2. Our Context and Data

The education system in China from primary school to university level focuses heavily on test scores and grading, with students' performance on these measures serving as the overwhelming admission criteria to students' next possible level of study, as well as post-graduation employment. China has a long-standing tradition of the use of academic testing as a means for social achievement and mobility. Relative comparisons among students are common, and at pre-university levels such social comparisons may even be sanctioned by teachers by making some students' performances known to the class.

A famous example of the intense emphasis on academic testing in China is the *gaokao* (College Entrance Examination), a three-day-long high-stakes test which determines high school students' eligibility for university study.[3] Another popular example is the National Civil Service exam, which all university graduates wishing to enter government service must take. Hence, students generally pay quite close attention to their test scores and relative academic performance, and a near-universal mentality is that the higher the score, the better. The education system's focus on scores as a measure of achievement

[3] See Liu and Wu (2006) for a discussion of the main issues regarding the Chinese College Entrance Exam.

makes students' preferences over scores much more homogenous than in other country contexts previously studied in the literature.

The specific setting for our analysis is the core Intermediate Microeconomics course in the business school at a large, top Chinese university. The course lasts for one semester, and consists of 16 weekly lectures. The course is a requirement for "economics and finance" majors and accounting majors, and enrolls over 200 students each semester, divided into two lecture sessions. Business school disciplines are among the most popular of major choices in recent years, and the students who have gained admission to this program can be fairly considered among the best performing students in the country. Due to China's large population, the intense competition for slots at this university and the popularity of economics-related majors, students successfully entering the program are nearly universally of high academic ability, particularly in exam performance. We thus have good reason to believe that students have the academic ability to convert their effort into visible score improvements.

Intermediate Microeconomics is an appropriate setting for testing for reference-effects in academic performance for a few key reasons. First, students in the course have already been enrolled at the university for at least one year prior to their enrollment in the course, allowing time for adjustment of their expectations for university level work difficulty and grading schemes. This is potentially quite important, in light of the heterogeneous educational backgrounds which may define performance in the first year's Principles of Economics sequence. Second, Intermediate Microeconomics is a quantitative problem-based course, with clear correct and incorrect answers, so that scoring of exams can be achieved objectively. Any grade adjustments such as score curving are conducted after calculating students' raw scores. Third, the course is a requirement for the degree, so that there is no selection bias, and each cohort of Intermediate Microeconomics is essentially equivalent to a cohort in the entire program. Finally since it is a core course for the degree, students are motivated to do well in it, as their performance in this course is likely to influence their chances for future opportunities (study abroad, graduate school, top jobs, recommendation letters, etc.).

The grading in the course consists of a midterm exam (worth 30%), a series of quiz sessions (worth 20%), and a final exam (worth 50%). The midterm exam is given at the midpoint of the course, and is potentially graded on a monotonic curve. After the midterm, students are informed of their own curved score and the averaged curved score. The quiz sessions are conducted periodically throughout the semester, and the grading counts only the best 3 scores for each student (equally weighted). The final exam is also potentially graded on a curve, but is not reported to students. After the semester is over, students observe their total score for the course on their transcript, which may also have a monotonic curve applied to it. Students have a chance to increase their total score by writing an extra credit essay at the end of the semester, which can add up to 3 percentage points to their total score, depending on their performance on the essay. These grading rules are made clear to students from the beginning of the course, and should be considered common knowledge among students. A summary of the course grading structure is given in Table 1.

Our data consist of all students who enrolled in the course between years 2011 and 2013, in which the course was always offered in the Fall semester. The same two instructors co-taught the course for all three years, and the course structure remained identical across years.

All course grades at the university are reported on students' transcripts as integers between 0 and 100, with 60 as the passing mark. Students who do not pass this required course for their majors must retake the course in some form, or they are unable to graduate. Students' transcripts at the end of the semester also inform them of their score rank within the course. Non-major students also occasionally enroll in the course, but comprise a small minority of students.

Table 1: Course grading structure

Midterm Exam	30%
Quizzes (Best 3)	20%
Final Exam	50%
Short Essay (Optional)	Adds up to 3 percentage points to total score

Since our analysis relies on within-student performance evaluation between the midterm exam and the end of the course, we can only include those students who sat for both the midterm and the final exam. Thus, our sample for analysis consists of 696 students over the three years of the course.

2.1 Summary Statistics

Table 2 shows summary statistics for our key raw variables of interest, by year. The average raw scores for the midterm and final exams tend of fall in the 70's range. One interesting observation is that the yearly essay participation rate seems to be positively related to the difficulty of the final exam. This can be explained by the fact that the optional essay assignment is due *after* students have taken the final exam.

Table 2: Summary Statistics

Year	Mean raw midterm score (out of 100)	Mean raw final exam score (out of 100)	Essay participation rate	Number of Students
2011	78.4	77.1	35.8%	240
2012	75	67.1	52.5%	236
2013	74.2	70.1	42.9%	220

Figure 1 shows scatter plots of the raw midterm score and improvement in raw score between the midterm and final exam, by year. In each year the class distribution across these two metrics shared several features: 1. Improvement between the midterm and the final exam was far from being universal – declining performance was almost as common as improved performance, and both improvement and decline occurred across the distribution of original midterm raw scores; 2. Students were concentrated in the upper half of possible raw scores, but there was substantial variety in score outcomes; 3. The variance of improvement (or decline) is wider, the lower the original score – this is in part an artifact of bounds on improvement at the top of the distribution, and in part due to student heterogeneity (initially top performing students being less likely to perform poorly on the final).

Figure 2 displays histograms of the raw score improvement between the midterm exam and the final exam, by year and whether students were in the gain frame or the loss frame relative to the average class performance. Consistent with Figure 1, improvement levels are more dispersed for the loss frame than for the gains frame, with neither frame showing obvious greater improvements proportionally.

Figure 1: Raw midterm scores and raw improvement, by year

(horizontal axis: raw midterm score, vertical axis: raw improvement)

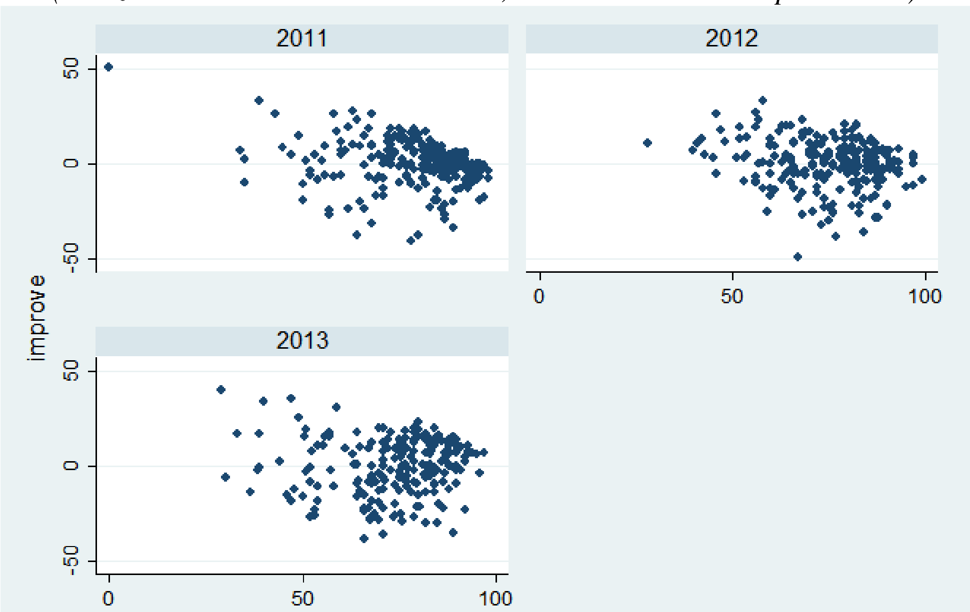
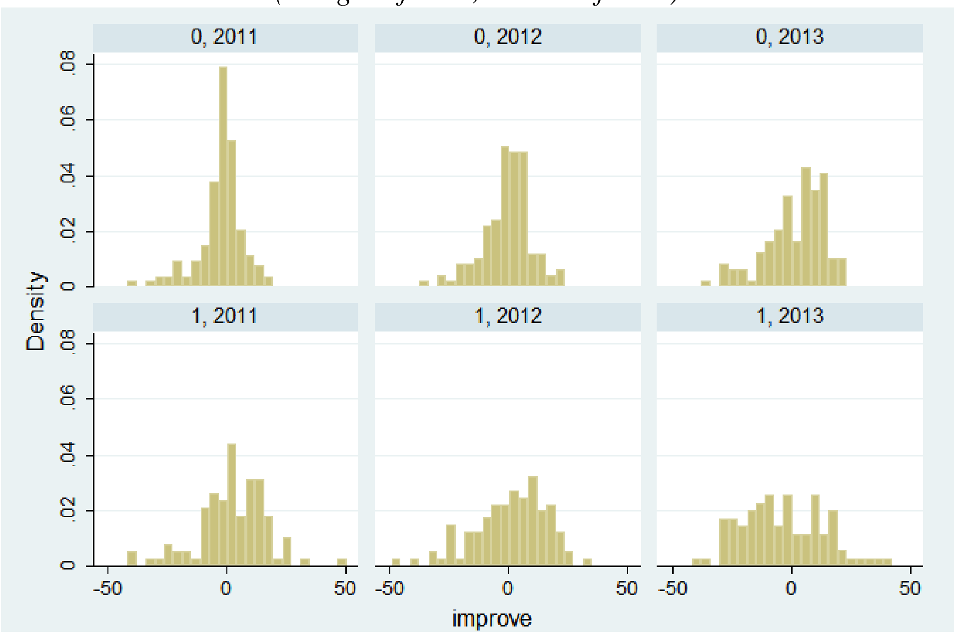


Figure 2: Distribution of raw improvements by loss and gain frames relative to average score

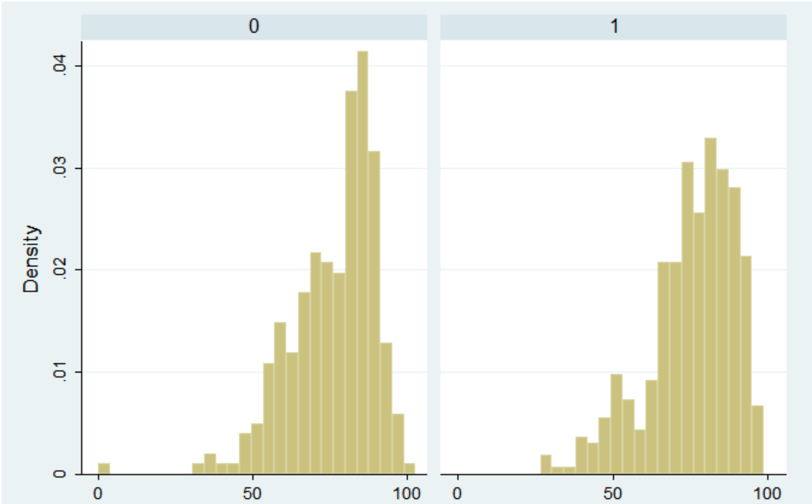
(0 = gain frame, 1 = loss frame)



We use the raw midterm exam and final exam scores as our main measures of students' performance. In Figures 1 and 2, raw improvement of student i is calculated relative to the difficulty of that year's exam, $rawimprove_i = (finalexam_i - \overline{finalexam_t}) - (midterm_i - \overline{midterm_t})$ where the average final exam and midterm exam raw scores of that year t represent difficulty. Although the difficulty of the exams is intended to remain about constant over the years, the actual difficulty as perceived by students may depend on the specific questions asked. Thus, it is important to control for some measure of how difficult students found the exam in general, which we capture using the average raw score.

Besides score improvements and variants of score improvement measure, we also consider students' participation in the optional essay assignment. Figure 3 shows histograms of the raw midterm scores separated by participation in the essay (1) or not (0). A pattern is not necessarily clear from the histogram alone.

Figure 3: Raw midterm scores and participation in optional essay
(0 = no participation in essay, 1 = participation in essay)



3. Hypotheses and empirical approach

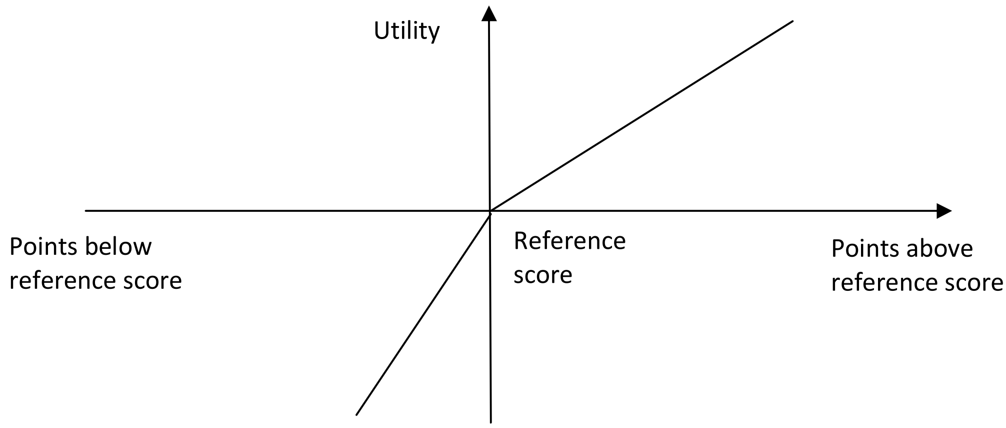
Our main hypothesis follows directly from models of reference-dependent loss aversion (Kahneman and Tversky, 1979; Koszegi and Rabin, 2006). Since individuals tend to find losses more painful than gains of the same magnitude are enjoyable to them, falling short of a specific goal Consequences of loss aversion around a reference point on individuals' effort provision have been tested in several field settings: Professional golf (Pope and Schweitzer, 2011), NBA and NCAA basketball (Berger and Pope, 2011), professional baseball and SAT taking behavior of high school students (Pope and Simonsohn, 2011). We follow the local regression discontinuity approach of Berger and Pope (2011) closely.[4]

The intuition is simple: Students who hold a particular score value as a reference point or goal, will experience disutility from outcomes below the goal, and if they are within reach of the goal, will work harder to attain it than students who obtained outcomes which exceed the goal by the same amount. The key is that a reference point serves as a break point around which we expect to see differences in behavior above compared to below: specifically, we expect students marginally below the reference point to be more motivated than students marginally above it. The basic reference-dependent loss averse utility function with the commonly used linear gain and loss segments is shown in Figure 4. The key feature of loss aversion is that students have a steeper marginal disutility of an outcome below the reference point, compared to the marginal utility of an outcome above it. Thus, students who receive a

[4] See Imbens and Lemieux (2008) for details on regression discontinuity design practice.

midterm score slightly below a salient reference score, are predicted to exert more effort than those slightly above, in order to avoid the realization of a permanent score which is below the reference score.

Figure 4: Loss Aversion with Academic Score as Reference Point



A noticeable difference between our approach and the related papers in the reference-dependence literature is that in contrast to the previous papers using the regression discontinuity method, we do not have a single salient reference-point which we can strongly anticipate that students will react to (ie. Score ties between teams in Berger and Pope, 2011 and "par" in Pope and Schweitzer, 2011). Instead, we test among several candidate reference points, by checking whether behavior in the vicinity of *each* potential reference point is consistent with loss aversion when that particular reference point is used as a goal. This is implemented by conducting a separate regression for each candidate reference point. This means that different subsamples of students will be used in each regression in the case of narrow, non-overlapping bands on the score window, while for larger windows, the same students may appear in more than one regression. Students appearing in the regression for more than one candidate reference point is not a problem for the interpretation of our results, since each regression has a self-contained hypothesis in the loss aversion framework. Each regression simply compares behavior above and below the reference point within the score band considered, controlling for a trend in effort.

Our approach assumes that there is a common model for the reference point in the population, which is the approach that much of the empirical literature on reference-dependence has adopted. In Section 4.1 we consider the possibility of individually-based reference points from students' personal academic records, which also follows this general approach. [5]

Our estimation framework is

$$effortmeasure_i = \alpha + \beta_1 \cdot gainloss_i + \beta_2 \cdot loss_i + \varepsilon_i$$

[5]A particular reference point candidate, whether universal or individual-specific, needs to be utilized 'commonly enough' in the population in order to yield significant coefficients in the estimation. This is the pre-condition for much of the empirical reference-dependence literature which proposes a common cross-individual model of the reference point, including Crawford and Meng (2011), Eil and Lien (2014), Lien and Zheng (2015), as well as the aforementioned papers Pope and Schweitzer (2011), Berger and Pope (2011), and Pope and Simonsohn (2011). An alternative potential approach which allows for reference point model heterogeneity would be a mixture model as in ElGamal and Grether (1999), which tends to require richer data on individual choice behavior compared to what we have available here.

where *effortmeasure* refers to some measurement of student *i*'s effort between the midterm and the end of the semester, *gainloss* is the difference between student *i*'s score and the candidate reference point in curved numerical scores, and *loss* is a dummy variable which equals 1 if the student is in the loss frame based on the candidate reference point. Here, β_1 captures a linear trend in the effort measure as a function of midterm score, while β_2 captures the additional effort expended when a student is in the loss frame. If students are further motivated by being slightly behind compared to being slightly ahead, this will appear as a significantly positive β_2 estimate, and we further expect the coefficient on the linear term β_1 to be weakly negative.

The estimation for each candidate reference point is conditional on a student falling within a particular narrow band around the reference point in the curved score of the midterm exam. That is, each regression only includes such individuals in the estimation. We note that the average score is announced to the class in curved format, and students' midterm result is also reported in curved format. In fact, students do not have access to their raw score unless they specifically ask the instructor or teaching assistants. Therefore, if students do actually form any common numerical reference point, we expect it to be in terms of curved scores.

We consider six possible "universal" reference point candidates: the fail threshold (60), the class average as reported to students, 70, 80, 90 and 95. We consider 95 due to its numerical salience, but more importantly, as an 'arbitrarily high' reference point candidate. We cannot implement the same specification with a reference point of 100, as in that circumstance there would be no students in the gain frame for comparison. We also consider three bands of proximity around the candidate reference point, also in curved score points: 10 points, 5 points, and 3 points. Among these proximity bands, there is a tradeoff between sample size and testing the effect of being in the "marginal" loss frame. We implement all three proximity bands to test for qualitative robustness as suggested by Imbens and Lemieux (2008). In Section 4.1, as a robustness check, we consider three possible "individual specific" reference point candidates: a student's cumulative grade point average at the point of enrolling in Intermediate Microeconomics, a student's grade in the Principles of Economics sequence of the previous year, and a student's worst academic grade obtained in a course thus far.

For measures of effort, we consider three measures which involve students' improvement in raw score from the midterm to the final exam, meant to be a (noisy) measure of their improvement in knowledge gained via effort spent studying. These measures are: 1. *Improvement (raw)*: each student's raw score difference compared to the average raw score on that exam, expressed in raw exam points; 2. *Improvement ratio*: each student's improvement ratio of raw exam points compared to the class average, expressed as a ratio; 3. *Improvement indicator*: an indicator variable for whether a student's improvement ratio is greater than 1. As a fourth measure of effort we consider whether or not students participated in the optional *extra credit essay* due at the end of the semester.

4. Empirical Results

Our results can be seen in Tables 3 through 5. While the nominal *gainloss* variable controls for trends in the effort measure across midterm score realizations, we are most interested in the coefficient on the loss indicator. If a reference point candidate is valid in the loss aversion framework, we should expect to see the loss coefficient being significantly positive. That is, higher effort realization occurs when a student is just below the reference level.

We now discuss our results, starting with Table 3, which shows the estimation results for the band width of 10 curved score points in either direction, from the reference point candidate. Several observations can be made from Table 3: 1. Most of the coefficients are statistically indistinguishable from zero; 2. When disregarding the statistical significance of the coefficients, the coefficient on the *gainloss* variable tends to generally be negative, suggesting an overall negative relationship between initial performance and subsequent effort. 3. The coefficient on the loss indicator tends to be negative and substantially larger than the coefficient on the gainloss variable, suggesting that effort is substantially lower on the loss side, contrary to the prediction if these benchmark scores were serving as reference points.

The exception to these general patterns is the candidate reference point of 95, in which the coefficient on loss is positive, but very imprecisely estimated. In the range of 85 and above, the likelihood of participating in the essay is significantly increasing in initial score. The results for the reference point of 95 with a band range of 10 should be interpreted cautiously because the band is asymmetric in this case, as the maximum possible score is cut off at 100.

Table 4 shows the results for the same specifications with a band width of 5 on each side. Overall, we have reason to believe that a band width of 5 is more reasonable than the band width of 10, for the assumption that being categorized into the gains versus the loss frame is due to randomness and not due to ability differences. The estimates to notice in this table are for the reference point candidates of 90 and 95. The same general patterns are exhibited here as in Table 3, except that the coefficients are statistically significant (10% level or better). The only group to exhibit the pattern of loss aversion are those students near the 95 point mark. Students scoring slightly below 95 were significantly more likely to improve by all measures, including the improvement indicator measure, which corrects to some degree for the potentially higher difficulty of increasing one's score past 95. However, the results for the essay participation do not follow this pattern.

Table 5, which shows the results for a band width of 3 on each side, shows largely the same story. Here, students scoring slightly below the midterm average are significantly less improved than those scoring slightly above (this pattern in the other band widths were at lower significance levels, around the 15% to 20% level). The results for 90 and 95 reference points maintain the same pattern as in Table 4, but with reduced precision. With a band of 3, 70 seems to stand out as another significant benchmark, but again with the *opposite* effect on effort as predicted by loss aversion and goal setting.

Tables 3 through 5, which display results under different band ranges serve as an important robustness check on one another. Although the traditional local RD approach generally favors very small band ranges around the discontinuity, in our setting, we would like to account for the possible psychological approaches of the students as much as possible. For example, students may view different score ranges such as 3 points, 5 points, or 10 points as being 'close' to a benchmark, and choose their effort in the course accordingly.

To summarize, we tend to observe reduced effort in response to initial performance below several salient benchmarks, rather than increased effort. These lower effort levels are sometimes statistically significant. The exception is when we consider a very high benchmark such as 95. Students in the loss frame in this case, seem to exert more effort towards improvement, suggestive of 'aiming' for a top score. Students centered around lower reference point candidates by contrast, seem to not be aiming higher, an indication of lower motivation. Since we focus on small ranges at a time around salient scores, the result is not well-explained by student heterogeneity in ability levels.

Table 3: Dependent Variables: Effort Measures, band range of 10 points

(OLS: improvement and improvement ratio; Probit: improvement indicator and essay participation)

band = 10	**improvement (raw)**		**improvement ratio**		**improvement indic**		**essay participation**		**Obs**
	coeff	*p-val*	coeff	*p-val*	coeff	*p-val*	coeff	*p-val*	
midterm gain loss (60, fail)	-0.8817*	*0.054*	-0.0143*	*0.081*	-0.0191	*0.604*	0.0377	*0.350*	130
	(0.4534)		(0.0081)		(0.0369)		(0.0404)		
midterm loss indic (60, fail)	-8.2940	*0.112*	-0.1402	*0.133*	-0.1471	*0.727*	0.6069	*0.189*	
	(5.1807)		(0.0927)		(0.4213)		(0.4622)		
midterm gain loss (average)	-0.0442	*0.848*	0.0007	*0.820*	-0.0107	*0.630*	-0.0221	*0.329*	384
	(0.2307)		(0.0033)		(0.0223)		(0.0226)		
midterm loss indic (average)	-3.8126	*0.154*	-0.0548	*0.152*	-0.3262	*0.207*	-0.1825	*0.485*	
	(2.6717)		(0.0381)		(0.2586)		(0.2614)		
midterm gain loss (70)	0.3716	*0.276*	0.0067	*0.208*	0.0737**	*0.011*	0.0361	*0.219*	248
	(0.3407)		(0.0053)		(0.0289)		(0.0294)		
midterm loss indic (70)	-5.8677	*0.119*	-0.0941	*0.108*	-0.7118**	*0.025*	-0.0489	*0.879*	
	(3.7472)		(0.0584)		(0.3173)		(0.3221)		
midterm gain loss (80)	0.2168	*0.389*	0.0052	*0.137*	0.0261	*0.280*	0.0155	*0.530*	384
	(0.2516)		(0.0035)		(0.0241)		(0.0246)		
midterm loss indic (80)	-0.7977	*0.771*	-0.0181	*0.635*	-0.2524	*0.336*	-0.3103	*0.246*	
	(2.7363)		(0.0381)		(0.2626)		(0.2675)		
midterm gain loss (90)	-0.5286**	*0.028*	-0.0056*	*0.063*	-0.0608**	*0.033*	0.0172	*0.554*	344
	(0.2394)		(0.0030)		(0.0285)		(0.0290)		
midterm loss indic (90)	-3.9976*	*0.070*	-0.0463*	*0.097*	-0.4559*	*0.082*	-0.5003*	*0.059*	
	(2.2026)		(0.0278)		(0.2620)		(0.2651)		
midterm gain loss (95)	-0.1458	*0.559*	-0.0014	*0.645*	-0.0407	*0.186*	0.0958***	*0.002*	241
	(0.2490)		(0.0031)		(0.0308)		(0.0312)		
midterm loss indic (95)	1.0789	*0.689*	0.0155	*0.645*	0.2916	*0.380*	-0.2232	*0.508*	
	(2.6955)		(0.0331)		(0.3322)		(0.3375)		

*Standard errors in parentheses, p-values in italics; *90% significance, **95% significance, ***99% significance*

Table 4: Dependent Variables: Effort Measures, band range of 5 points

(OLS: improvement and improvement ratio; Probit: improvement indicator and essay participation)

band = 5	improvement (raw)		improvement ratio		improvement indic		essay participation		Obs
	coeff	*p-val*	Coeff	*p-val*	coeff	*p-val*	coeff	*p-val*	
midterm gain loss (60, fail)	0.9086	*0.568*	0.0171	*0.556*	0.0320	*0.793*	0.0994	*0.452*	62
	(1.5832)		(0.0289)		(0.1222)		(0.1320)		
midterm loss indic (60, fail)	-0.7116	*0.930*	-0.0124	*0.933*	0.1638	*0.793*	0.7858	*0.255*	
	(8.0470)		(0.1469)		(0.6237)		(0.6910)		
midterm gain loss (average)	-0.2753	*0.683*	-0.0022	*0.820*	-0.0499	*0.433*	-0.0756	*0.242*	187
	(0.6733)		(0.0096)		(0.0636)		(0.0646)		
midterm loss indic (average)	-4.9048	*0.205*	-0.0677	*0.218*	-0.4948	*0.175*	-0.4505	*0.221*	
	(3.8523)		(0.0547)		(0.3650)		(0.3682)		
midterm gain loss (70)	1.0116	*0.346*	0.0165	*0.326*	0.1315	*0.148*	0.0276	*0.764*	122
	(1.0701)		(0.0167)		(0.0910)		(0.0920)		
midterm loss indic (70)	-7.7625	*0.159*	-0.1232	*0.153*	-0.8990*	*0.056*	-0.0096	*0.984*	
	(5.4825)		(0.0857)		(0.4709)		(0.4710)		
midterm gain loss (80)	0.1334	*0.867*	0.0037	*0.734*	0.0432	*0.579*	0.0218	*0.783*	175
	(0.7955)		(0.0109)		(0.0779)		(0.0789)		
midterm loss indic (80)	0.1356	*0.973*	-0.0019	*0.972*	-0.3243	*0.415*	-0.3330	*0.409*	
	(4.0617)		(0.0554)		(0.3981)		(0.4030)		
midterm gain loss (90)	-1.0476*	*0.085*	-0.0123*	*0.102*	-0.2128***	*0.004*	-0.0051	*0.945*	202
	(0.6058)		(0.0075)		(0.0734)		(0.0730)		
midterm loss indic (90)	-6.1107*	*0.051*	-0.0736*	*0.057*	-1.1140***	*0.003*	-0.5811	*0.122*	
	(3.1091)		(0.0385)		(0.3787)		(0.3756)		
midterm gain loss (95)	-1.5812**	*0.011*	-0.0180**	*0.015*	-0.2833***	*0.003*	0.0700	*0.443*	117
	(0.6130)		(0.0073)		(0.0952)		(0.0913)		
midterm loss indic (95)	5.6885*	*0.073*	0.0684*	*0.070*	1.1391**	*0.017*	-0.1803	*0.700*	
	(3.1384)		(0.0375)		(0.4778)		(0.4675)		

*Standard errors in parentheses, p-values in italics; *90% significance, **95% significance, ***99% significance*

Table 5: Dependent Variables: Effort Measures, band range of 3 points

(OLS: improvement and improvement ratio; Probit: improvement indicator and essay participation)

band = 3	**improvement (raw)**		**improvement ratio**		**improvement indic**		**essay participation**		**Obs**
	coeff	*p-val*	coeff	*p-val*	coeff	*p-val*	coeff	*p-val*	
midterm gain loss (60, fail)	2.7912	*0.368*	0.0526	*0.361*	0.0992	*0.732*	0.3116	*0.348*	38
	(3.0591)		(0.0568)		(0.2895)		(0.3321)		
midterm loss indic (60, fail)	9.8991	*0.269*	0.1797	*0.279*	0.7388	*0.393*	1.0550	*0.287*	
	(8.8086)		(0.1635)		(0.8641)		(0.9905)		
midterm gain loss (average)	-1.7378	*0.259*	-0.0228	*0.289*	-0.2787*	*0.066*	-0.1146	*0.444*	112
	(1.5304)		(0.0214)		(0.1515)		(0.1497)		
midterm loss indic (average)	-8.4877*	*0.093*	-0.1181*	*0.094*	-1.0477**	*0.035*	-0.5018	*0.305*	
	(5.0070)		(0.0700)		(0.4978)		(0.4892)		
midterm gain loss (70)	1.8258	*0.526*	0.0281	*0.538*	0.4209*	*0.081*	0.0046	*0.985*	63
	(2.8654)		(0.0453)		(0.2412)		(0.2377)		
midterm loss indic (70)	-10.4920	*0.175*	-0.1637	*0.181*	-1.6302**	*0.014*	-0.0925	*0.885*	
	(7.6448)		(0.1210)		(0.6617)		(0.6370)		
midterm gain loss (80)	-0.6712	*0.688*	-0.0085	*0.706*	-0.1356	*0.445*	-0.1254	*0.480*	100
	(1.6681)		*0.022*		(0.1776)		(0.1777)		
midterm loss indic (80)	1.9157	*0.707*	0.0253	*0.711*	0.0146	*0.978*	0.0474	*0.930*	
	(5.0822)		(0.0682)		(0.5370)		(0.5413)		
midterm gain loss (90)	-0.0695	*0.963*	0.0005	*0.978*	-0.1086	*0.537*	-0.0741	*0.670*	103
	(1.4848)		(0.0181)		(0.1757)		(0.1739)		
midterm loss indic (90)	-5.1126	*0.226*	-0.0614	*0.234*	-0.9929**	*0.047*	-0.7263	*0.144*	
	(4.1973)		(0.0513)		(0.5000)		(0.4970)		
midterm gain loss (95)	-0.9169	*0.512*	-0.0097	*0.555*	-0.0601	*0.784*	-0.0096	*0.965*	66
	(1.3897)		(0.0163)		(0.2194)		(0.2183)		
midterm loss indic (95)	3.7410	*0.362*	0.0442	*0.360*	0.6138	*0.340*	-0.0950	*0.881*	
	(4.0759)		(0.0479)		(0.6429)		(0.6357)		

*Standard errors in parentheses, p-values in italics; *90% significance, **95% significance, ***99% significance*

4.1 Individual reference points

One possibility is that instead of taking commonly known benchmarks, such as the class average or round number scores as motivators, students may instead refer to some personal benchmark, such as scores they have previously obtained in other classes. To explore this possibility, we conduct a similar analysis as in the previous section, with several candidate reference points from students' individual academic history. We consider this as a robustness check on our main findings because we do not expect students in our setting to be more sensitive to individual-specific benchmarks compared to publicly announced or universal benchmarks, as considered in the previous section.

Our data access on individual academic records is limited to students whose primary major is in the business school. This subset of students is indeed the standard student population of the course, which accounts for the majority of students who enroll in the course each year. These are also the students for whom we should be most likely to find an effect, assuming that students tend to care more about courses in their own major. For simplicity, we focus on the 163 students who were in their second year of study in the intermediate microeconomics class cohort of 2012.[6]

We consider three possible individually based candidate reference points, based on courses completed prior to the semester of taking intermediate microeconomics: 1. The student's average final course score (numerical, 0 to 100) obtained across all courses completed; 2. The student's final course score in Principles of Economics. 3. The lowest academic course score of the student prior to the semester of taking Intermediate Microeconomics. All three measures are using the final course scores, since we do not have access to students' midterm performances (if applicable) in those courses.

In taking the student's average score across completed courses, we exclude the required physical education courses, since students' scores in this course are often outliers in their general grade average. The average score variable should thus reflect the students' performance in prior academic courses. However, in case students may not particularly care about some of their required first year courses (and thus may exclude them from their reference point formation), we also consider the average of his or her scores in the two prior semesters of introductory economics. Finally, we consider the lowest out of the student's final grades as a possible reference point. This is a potential lower-end benchmark for a student's personal standard of "doing poorly", in case students are not aware of or using their personal grade average as a benchmark.

The effort measures and regression specifications are identical to those in the previous section for common reference point candidates. The measures of gain and loss however, are calculated based on individual students' academic records, as discussed above, so that each student has his or her own personal candidate reference point in each regression.

Table 6 shows the results for personal reference points for band width of 5 raw score points on the midterm exam. While very few coefficients are statistically significant, the visible exception is for the essay participation variable. Students have the opposite participation behavior compared to what is expected if either personal grade point average, or lowest grade obtained is a reference point in the loss aversion framework. Students falling slightly below these two benchmarks are significantly *less* likely to participate in the essay compared to students slightly above the benchmark. We do not find significant effects of the introductory economics course grade on essay participation.

[6] We have little reason to believe that the results on individual reference points differ across class years, so we arbitrarily choose 2012 as our sample for the individual reference point analysis.

Table 6: Individual Reference Points: Dependent Variables: Effort Measures, band range of 5 points

(OLS: improvement and improvement ratio; Probit: improvement indicator and essay participation)

band = 5	**improvement (raw)**		**improvement ratio**		**improvement indic**		**essay participation**		**Obs**
	coeff	*p-val*	Coeff	*p-val*	coeff	*p-val*	coeff	*p-val*	
Average grade gain loss	-1.8560*	*0.064*	-0.0250*	*0.085*	-0.0866	*0.401*	0.2984***	*0.006*	65
	(0.9842)		(0 .0143)		(0.1030)		(0.1082)		
Average grade loss indic	4.8096	*0.422*	0.0616	*0.478*	-0.1131	*0.855*	-1.1680*	*0.068*	
	(5.9554)		(0.0864)		(0.6194)		(0.6410)		
Introductory course gain loss	-0.6389	*0.444*	-0.0055	*0.654*	-0.1802	*0.112*	-0.0456	*0.680*	71
	(0.8297)		(0.0122)		(0.1135)		(0.1105)		
Introductory course loss indic	2.8082	*0.567*	0.0224	*0.757*	0.8244	*0.216*	0.3153	*0.628*	
	(4.8809)		(0.0721)		(0 .6663)		(0.6503)		
Lowest grade gain loss	-.1711	*0.886*	0.0036	*0.846*	-0.1893	*0.196*	0.3694**	*0.015*	46
	(1.1825)		(0 .0186)		(0.1463)		(0.1518)		
Lowest grade loss indic	-2.36242	*0.665*	-0.0575	*0.502*	0.7785	*0.239*	-1.7080**	*0.017*	
	(5.410902)		(0.0849)		(0.6616)		(0.7164)		

*Standard errors in parentheses, p-values in italics; *90% significance, **95% significance, ***99% significance*

These individual results are consistent with the findings in the main results, confirming that the students are not especially motivated by the intuitive benchmarks. If anything, we find some evidence suggesting that students performing slightly below a benchmark tend to 'give up', exerting *less* effort when performing below a standard.

5. External validity and interpretations

An important natural follow-up question is whether the results we find in this particular field setting extend readily to other settings. Our interpretation is that our results are likely to be representative of highly ranked universities that *also* use the 0 to 100 grade reporting scheme. More work analyzing the effort provision of students from differently ranked institutions with different grade reporting schemes is needed.

We interpret our results in a cautious manner, as being representative of highly-ranked universities, due to the potential reference point selection by students as a function of university admissions difficulty.[7] That is, one explanation for our finding of a high reference point is that the students in our field setting are among the top students in the country. It is possible that among 'high-achievers', there could be a mentality that the only way that scores matter is if a top score is obtained, and incentives for improvement if one is not near the top, may be negligible. However, it is possible that in a low ranked university, a passing score may serve as a salient reference point. At a middle-ranked university, it is possible that the average score could serve as the predominant reference point.[8] However, given that students in our field context have already acclimated to university level grading standards for at least one year, it was surprising to us that students did not noticeably respond to the announced average. Analysis using the individually-based candidate reference point did not yield support for individual benchmarks as significant goals, except in the essay participation domain, where students performing below their personal standards were significantly less likely to make this extra effort in improving their grade.

The other major factor which we believe is likely to have an influence on students' effort provision is the grade reporting structure at the university we study. Since grades for each course taken by students are ultimately reported as a numerical score 0 to 100, round number scores such as 70, 80, or 90 may not hold the same weight in students' minds as the case when grades are reported as A, B, C, etc. Although the evidence from Pope and Simonsohn (2011) suggests that round numbers serve as natural psychological goals in several field contexts, some kind of external valuation of the round numbers may be needed to enforce their use as reference points. Our analysis seems to show that in our field context, the marginal loss in valuation from receiving an 89 instead of a 90 is the same magnitude as the marginal gain in valuation from receiving a 91 instead of a 90. One possibility is that students believe that anyone looking at their transcript would value the grading points approximately linearly over most of the grading scale. We suspect that students at the university we study might be less likely to have such a perception if grades on the transcript were reported in A, B, C, D, F format. Grant and Green (2013) and Oettinger (2002) provide some evidence on this, and together their studies suggest that the sensitivity of the student population to such grading thresholds may be sensitive to the culture of the school studied.

---

[7] Indeed, educators we spoke to at another top Chinese university were not surprised by the patterns we found, and reported similar behavior among their students.

[8] However, we note that the results of Grant and Green (2013) checking students' effort around score thresholds in the A, B, C, D, F grading system at two universities in Texas, does not seem to support this possibility.

Our analysis has assumed that students will respond to their midterm score relative to a reference midterm score, when this is potentially a simplification of students' thinking on how hard they need to work. A plausible forward-looking reasoning, is that a student may ask, "Given my score on the midterm, what score do I need to obtain on the final (and other course assignments), in order to attain my academic goal?"

A similar analysis could be implemented assuming forward-looking students, and indeed many students may use this type of reasoning. However, we note that the qualitative prediction with respect to our regressions, is essentially similar to the current version. That is, suppose that a student's goal for his total course grade is 90, and his midterm score is 85. Since his midterm score was below his numerical goal for the total course grade, he still needs to exert more effort in order to reach it than if he had scored above 90 on the midterm. Overall, our finding that students in the marginal loss frame seemed to work *less* towards higher performance, is counter to this basic prediction of goal-setting, even when students consider their future expected scores as benchmarks.

6. Conclusion

In this paper we have investigated the possibility of reference-dependent effort provision among university students in a core class in their major. We find that for all reference point candidates except an arbitrarily high one, students' efforts either show no significant response to the candidate reference point, or show an *opposite* pattern compared what would be expected under loss aversion. By opposite pattern, we mean that students show less effort in the loss frame than they do in the gain frame, suggestive of a discouragement from being 'behind'. We conclude by discussing some policy implications of our findings, as well as potential insights for research directions with regard to goal setting in the academic domain.

First, we discuss the implication of our results under the Chinese context, in light of a few different policy concerns. The first policy concern pertains to the state of the educational system in China which is well-known for being heavily score-oriented and competitive. The students in our analysis are generally the top academic performers in their high schools, and are top scorers in their provincial College Entrance Examination. However based on our findings, once these students arrive at university, they seem not particularly motivated for self-improvement unless they are very well-positioned to obtain a top score.

A related concern has to do with the general motivation for effort in Chinese society. The results in this paper imply that the extrinsic and intrinsic reward structure in modern China may emphasize too heavily on top placements and near perfect performances, perhaps without sufficient emphasis on localized goals or personal improvements. In such an environment, our results indicate that even students at a very top university are not particularly motivated to improve their performance, unless they *already* have a very high score. Under these circumstances, we would expect the dispersion in performances to get larger over time, at each level of education, the majority of students may give up, and a small number of top performers excel to the next level. Our study is just a snapshot taken at one of the highest education levels, however, majorities of students at lower education levels may have already been discouraged by such a system.

In terms of research on loss aversion in the educational setting, our study points to several important avenues for additional research. First, analysis on similar effort patterns from other schools of different ranking levels need to be considered, to test whether students at these schools do in fact have reference points at non-arbitrarily-high score marks. Second, similar analysis can be conducted at institutions which have different grading schemes to test the sensitivity of students' effort to the implied incentives. It may be plausible that students are motivated by round number scores when they are actually given different categorical labels on the transcript. Finally, this raises an optimal design issue: What is the optimal grading scheme which can maximize students' effort in practice? [9] While this is an important issue which will be of general interest to educators and policy-makers, we leave the further exploration of this topic to future work.

[9] Dubey and Geanakoplos (2010) provide a game theoretic analysis of this issue, showing that coarse grading is advantageous for effort when utilities depend on score rankings.

References:

Azmat, G. and Iriberri, N., "The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment using High School Students", *Journal of Public Economics*, Vol. 94 (2010), p. 435 – 452.

Berger, J. and Pope, D., "Can Losing Lead to Winning?" *Management Science*, Vol. 57 (2011), No. 5, p. 817 – 827.

Betts, J.R. and Grogger, J., "The Impact of Grading Standards on Student Achievement, Educational Attainment, and Entry-Level Earnings", *Education of Economics Review*, Vol. 22 (2003), No. 4, p. 343 – 352.

Camerer, C., L. Babcock, G. Loewenstein, and R. Thaler, "Labor supply of New York City cabdrivers: One day at a time," *The Quarterly Journal of Economics*, Vol. 112 (1997), No. 2, 407-441.

Costrell, R., "A Simple Model of Educational Standards", *American Economic Review*, Vol. 84 (1994), No. 4, p. 956 – 971.

Crawford, V.P. and J. Meng, "New York City Cabdrivers' Labor Supply Revisited: Reference-Dependence Preferences with Rational-Expectations Targets for Hours and Income," *American Economic Review*, Vol. 101 (2011), No. 5, p. 1912 – 1932.

De Fraja, G., Oliveira, T., and Zanchi, L., "Must Try Harder: Evaluating the Role of Effort in Educational Attainment", *The Review of Economics and Statistics*, Vol 92 (2010), No. 3, p. 577 – 597.

Dubey, P. and Geanakoplos, J., "Grading exams: 100, 99, 98, … or A, B, C?", *Games and Economic Behavior*, Vol 69 (2010), p. 72 – 94.

Eil, D. and Lien, J.W., "Staying Ahead and Getting Even: Risk Attitudes of Experienced Poker Players", *Games and Economic Behavior*, Vol. 87 (2014), p. 50 – 69.

El-Gamal, M.A. and Grether, D.M., "Are people Bayesian? Uncovering behavioral strategies", *Journal of the American Statistical Association*, Vol. 90 (1995), Issue 432, p. 1137 – 1145.

Farber, H.S., "Is Tomorrow Another Day? The Labor Supply of New York City Cab Drivers." *Journal of Political Economy*, Vol. 113(2005) No.1, p. 46-82.

Farber, H.S., "Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers." *American Economic Review*, Vol. 98 (2008), No. 2, p. 1069-1082.

Fehr, E. and L. Goette, "Do workers work more if wages are high? Evidence from a randomized field experiment," *The American Economic Review*, Vol. 97 (2007). No. 1, p. 298-317.

Galdon, P. P., and Gonzalbez, J.L.N., "Testing prospect theory in students' performance", *Revista de Educacion*, January – April Issue, (2013), p. 645 – 665.

Gill, D. and Prowse, V., "A Structural Analysis of Disappointment Aversion in a Real Effort Competition", *American Economic Review*, Vol. 102, No. 1, p. 469 – 503.

Grant, D. and Green W.B., "Grades as Incentives", *Empirical Economics*, Vol. 44 (2013), Issue 3, p. 1563 – 1592.

Hadsell, L. "Achievement goals, locus of control, and academic success in economics", *American Economic Review, Papers and Proceedings*, Vol. 100 (2010), p. 272 – 276.

Imbens, G. and Lemieux, T., "Regression discontinuity designs: A guide to practice", *Journal of Econometrics*, Vol. 142 (2008), p. 615 – 635.

Lien, J.W. and Zheng, J., "Deciding When To Quit: Reference-Dependence Over Slot Machine Outcomes", *American Economic Review, Papers and Proceedings*, Vol. 105 (2015), No. 5, p. 366 – 370.

Liu, H. and Wu, Q., "Consequences of college entrance exams in China and the reform challenges", *KEDI Journal of Educational Policy*, Vol. 3 (2006), No. 1, p. 7 – 21.

Oettinger, G.S., "The Effect of Nonlinear Incentives on Performance: Evidence from 'Econ 101'", *The Review of Economics and Statistics*, Vol. 84 (2002), No. 3, p. 509 – 517.

Pope, D. and Simonsohn, U., "Round Numbers as Goals: Evidence from Baseball, SAT Takers and the Lab", *Psychological Science*, Vol. 22 (2011), No. 1, p. 71 – 79.

Tran, A. and Zeckhauser, R., "Rank as an inherent incentive: Evidence from a field experiment", *Journal of Public Economics*, Vol. 96 (2012), p. 645 – 650.

Wang, L.C., "The Effect of High-stakes Testing on Suicidal Ideation of Teenagers with Reference-Dependent Preferences", *Working paper*, 2013.