1-2011

Round Numbers as Goals Evidence From Baseball, SAT Takers, and the Lab

Devin Pope
University of Pennsylvania

Uri Simonsohn
University of Pennsylvania

Recommended Citation

Pope, D., & Simonsohn, U. (2011). Round Numbers as Goals Evidence From Baseball, SAT Takers, and the Lab. *Psychological Science, 22* (1), 71-79. http://dx.doi.org/10.1177/0956797610391098

Round Numbers as Goals Evidence From Baseball, SAT Takers, and the Lab

Abstract
Where do people's reference points come from? We conjectured that round numbers in performance scales act as reference points and that individuals exert effort to perform just above rather than just below such numbers. In Study 1, we found that professional baseball players modify their behavior as the season is about to end, seeking to finish with a batting average just above rather than below .300. In Study 2, we found that high school students are more likely to retake the SAT after obtaining a score just below rather than above a round number. In Study 3, we conducted an experiment employing hypothetical scenarios and found that participants reported a greater desire to exert more effort when their performance was just short of rather than just above a round number.

Keywords
judgment, decision making, reference points, goals

Disciplines
Other Psychology | School Psychology

ROUND NUMBERS AS GOALS:
EVIDENCE FROM BASEBALL, SAT TAKERS, AND THE LAB

Devin Pope[*]
University of Pennsylvania
The Wharton School
dpope@wharton.upenn.edu

Uri Simonsohn
University of Pennsylvania
The Wharton School
uws@wharton.upenn.edu

Word count: 3,826

Abstract.
How do individuals evaluate their performance? We conjecture that round numbers in performance scales act as reference points and predict that performing just short of them motivates people to improve in the future. We find substantial and significant evidence of this in how batting averages influence professional baseball players, SAT scores impact retaking decisions by high school students, and laboratory participants react to hypothetical scenarios.

[*] Both authors contribute equally, order of authorship was determined with a coin-toss (Devin accurately predicted heads).

Where do reference points come from? Previous research has shown that goals (Heath, Larrick, & Wu, 1999; Larrick, Heath, & Wu, 2009), expectations (Feather, 1969; Mellers, Schwartz, Ho, & Ritov, 1997) and counterfactuals (Kahneman & Miller, 1986; Medvec, Gilovich, & Madey, 1995; Medvec & Savitsky, 1997), among others, can all act as reference points when evaluating outcomes.

We hypothesize that because round numbers are “cognitive reference points” in numerical scales (Rosch, 1975), they also act as reference points in the process of subjectively judging outcomes. We conjecture that as a consequence, performing just short of a round number motivates future improvement more than performing just above it does.

We test this notion studying professional baseball players and SAT takers. As predicted, we find that those performing just short of a round number are more driven to improve their performance. The effects we document are significant both statistically and in practical terms. For example, baseball players with a batting average of .299 before the last play of the season are almost *twice* as likely to hit the ball on that last play relative to players with a .300 average. Similarly, high school students are approximately 20% more likely to retake the SAT if their total score ends in 90 (e.g. 1190) than the most proximate 00 (e.g. 1200). We obtain similar findings in a scenario experiment that rules out alternative explanations for the field studies.

The studies of (Medvec & Savitsky, 1997) and (Heath et al., 1999) are the closest to the current research. The former shows that participants report greater satisfaction imagining barely meeting a category of performance than comfortably exceeding it. The latter argues that explicitly set goals act as reference points which, among other consequences, lead to greater expressed motivation for improvement when people are just short of meeting a goal.

We contribute to these findings by demonstrating that a round number – a "goal" that has not been explicitly set and which is not attached to a direct consequence – is a powerful motivator of behavior both inside and outside the laboratory.

STUDY 1: BASEBALL PLAYERS

Method

Study 1 examines the behavior of professional baseball players. In addition to sports' general advantage of having readily available performance data that players themselves are likely to pay attention to, baseball has the additional advantage of possessing a particularly salient measure of performance that varies with great granularity: *batting average*.[†]

We examine how players respond to their season batting average being just below relative to just above a round number. For ease of exposition, we refer to performance that is exactly equal to a round number as being above it.

We test two predictions that arise from the hypothesis that round numbers act as motivating goals: (1) fewer than expected players end the season just below a round number and more than expected above it, and (2) players that are very close to a round number as the season is ending adjust their behavior to ensure ending it above it.

[†] Batting average is the number of times a player has successfully hit the ball divided by the number of at bats. The Baseball Almanac refers to it as "easily the most common statistic in baseball and the most understood" (http://www.baseball-almanac.com/stats.shtml).

To test these predictions we obtained baseball data from Retrosheet.com. These data contain play-by-play results for all players in Major League Baseball from 1975 to 2008.[‡] To ensure having a sufficiently granular batting average we restrict our sample to players who had at least 200 at bats during the season (reducing overall sample from 11,430 player-seasons to 8,817). Batting averages for professional baseball players are almost never below .200 or above .400, so we focus on batting averages around .300, the most round number in that range.

Results

Most of our statistical analyses employ the full range of batting averages in the data, however, for ease of exposition, we graphically present results in the .280 to .320 range (N=3,083 season*player observations).

Figure 1 depicts the relative-frequency of batting averages at the end of the season and with five plays left in the season. We include the latter as a control that seeks to account for a mechanical blip in the frequency of .300 arising from rounding to three decimals, and for the possibility that the distribution of batting ability is discontinuous at .300.

*** Figure 1***

Consistent with the notion that players use round numbers as goals, we see that season averages are markedly *less* likely to be just below .300 than just above it. For example, the proportion of players ending the season with a .298 or .299 (P=0.97%) is lower than with a .300 or .301, P=2.30%, Z=7.35, p<.001. Furthermore, the marked increase between .299 and .300 ($P_{.299}$=0.38%, $P_{.300}$=1.40%, Z=3.54, p=.001) is the

[‡] Season-level results were computed by aggregating the individual plays. Our calculations were validated against season-level data from Sean Lahman's Baseball Archive.

only increase between consecutive observations in the figure with *p*<.10. No such pattern is observed for the control distribution of season average with five plays left in the season. These results suggest that players find a way in their last few at bats during a season to ensure that they finish above .300. How exactly they do this is examined by looking at our second prediction in Figure 2.

Panel A shows that a higher percentage of players hit the ball on their last at bat of the season when their average was .298 or .299, P_{hit}=35.2%, rather than .300 or .301, $P_{hit=}$22.4%, *Z*=2.36, *p*=.018. Furthermore, a higher percentage of players with a .299 did so (P_{hit}=43%) than the average player in the figure did P_{hit}= 22.8%, *Z*=3.62, *p*<.001. P_{hit}=43% is also significantly higher than the expected rate of .299 (*Z*=2.17, *p*=.029). These differences suggest that players that are just below .300 adjust their behavior on the last play of the season in order to increase their chance of getting a hit.

*** Figure 2 **

Like hitting, getting a "walk" helps the team get closer to scoring, but unlike a hit, "walking" does not count as an at bat and thus cannot increase a player's batting average. Thus players may be reluctant to walk when they are just under .300 at the end of the season. Panel B shows that players are in fact less likely to walk when their batting average is .298 or .299, P_{walk}=2.5%, than when it is .300 or .301 P_{walk}=8.6%, Z=2.14, *p*=.032. Furthermore, not a single player (out of 61) 'walked*'* when their season average was .299.

One final way for a player to try to finish the season just above .300 is to have a substitute ("pinch hitter") hit instead of the player if the player is just above the round number. Panel C shows that players with a .298 or .299 are less likely to be substituted (P_{pinch}=4.1%) than those with an average of .300 or .301, P_{pinch}=19.7%, Z=3.85. p<.001. Furthermore, the rate of substitution for players with an average of .300 (P_{pinch}=34.3%), is by far the highest of all players in the figure, who have an overall P_{pinch}=7.0%, Z=8.29, p<.001.

Discussion

Overall, the performance of baseball players proves consistent with the notion that a round number, a batting average of .300, can act as a goal and influence behavior. The baseball analyses have two notable limitations: (1) there is only one relevant round number, and (2) players' actions to improve performance take place on the last plays of the seasons and hence have relatively minor consequences. Our next study seeks to address both limitations.

STUDY 2: SAT RE-TAKERS

Method

The SAT is a standardized test for college admissions in the United States. Until 2006 (and for the entirety of our sample period), SATs were scored between 400 and 1600 in intervals of 10. Students are allowed to retake the test and a high proportion of them do (about 50%, according to (Vigdor & Clotfelter, 2003)). We conjecture that if round numbers act as performance goals for test takers, then those scoring just below a round number would be more likely to retake the test than those scoring just above a round number would.

For Study 2 we use data from the College Board's Test Takers Database. Our dataset is a random 25% sample of all SAT test takers graduating between 1994 and 2001. It also includes, for those same years, a 100% sample of all SAT test takers from California and Texas, and all takers self-reported as African American or Hispanic (N=4,323,906 overall).

The dataset only includes the score and date of the students' last test ever taken and it does not include information on whether the student has taken the test before. Hence, we do not directly observe which students retook the SAT. Instead, as we describe in detail below, our analyses identify retaking *rates* by looking at the SAT-score distributions for juniors and seniors. Juniors or seniors account for 99.5% of the data, so we focus on these two groups only.

Results and Discussion

Test Retaking

Figure 3 plots the distribution of SAT scores separately for juniors and seniors. The distribution of juniors is centered to the right of seniors, indicating that, on average, juniors (who did not retake the SAT as seniors) did better on the SAT than seniors did. We are mostly interested in gaps in the frequency of scores around round numbers for juniors (as compared to seniors).

*** Figure 3 ***

The majority of seniors that take the test do not have a chance to receive their scores and then retake the SAT before they send out their college applications and as would be expected the distribution of their scores is smooth. The majority of juniors that take the SAT, in contrast, can see their scores and then have the option to retake the test before sending out college applications. Thus, if students are more likely to retake the exam if they score just below a round number, we would expect to see discrete jumps in the frequency of juniors with scores below and above round numbers, which is what Figure 3 shows.

The most visually striking gap occurs between 990 and 1000. There are 18,134 juniors obtaining a 990 and 20,057 obtaining a 1000, a difference of 1,923 students or 10.6%. Among seniors, these numbers are 58,714 and 58,716 respectively, that is, just *two* more seniors obtained a 1000 than a 990. Using seniors as a control, in other words, there are roughly 11% too few 990s in the distribution of SAT scores by juniors. There is a problem with intuitive comparison, however, as the distribution of seniors peaks before that of juniors; we address this in our statistical analyses that follow by employing a more conservative baseline for 'expected' score frequencies.

While these gaps in the distribution of juniors' scores are apparent to the naked eye for most round numbers above 700 in the figure, they appear to be smaller for scores further from 1000. These visual comparisons are misleading, however, because the baseline upon which they occur is also getting smaller as one moves away from 1000.

To account for this, Figure 4 plots the ratio of relative frequencies between consecutive scores (the slope of the density function). For example, in line with the numbers from the example of SAT=1000 vs. SAT=990 discussed above, Figure 4 shows that the ratio of relative frequencies for juniors at 1000 is 1.11 while the same ratio for seniors is 1.00. The figure shows sizeable gaps at every round number between 900 and 1500, with those between 1000 and 1400 being particularly large. From 1390 to 1400, for example, the ratio of relative frequencies for juniors is 1.08 while that of seniors is .89.

*** Figure 4***

As mentioned above, the previous comparisons suffer from the problem that the distribution of seniors peaks before that of juniors. A simple way around this problem is to exploit the fact that for scores above 1000 one expects the frequency of juniors to drop as the score is increased (e.g. that there should be fewer juniors with a score of 1250 than with a score of 1240).

We can hence conduct a conservative test of the notion that scoring just short of a round number increases the odds of retaking by testing the null that at a round number the relative frequency of students getting that score is the same as getting the immediately lower score. We do this with a simple difference of proportions tests for these two contiguous frequencies and assess whether we reject the null that these two proportions are equal.

For example, there are 1,656 *more* juniors with a score of 1100 than with a score of 1090 ($P_{1100=}$1.74%, P_{1090} =1.60%, Z=8.61, p<.001), which we interpret as favoring the notion that falling short of 1100 provided greater motivation to retake the SAT than getting 1100 did. Analogous conservative calculations reject the null for 1200, 1300 and 1400, all p's<.001.

In sum, we find a systematic pattern consistent with students being more likely to retake an exam if their score is just short of a round number. We interpret this as evidence of SAT takers using round numbers as implicit goals for performance.

A related though alternative account involves test takers, either correctly or incorrectly, believing that surpassing a round number disproportionately increases their odds of being admitted to college or of receiving financial aid. This is not implausible considering that, (i) universities and scholarships often do impose minimum SAT thresholds for consideration (although not as high as 1300 or 1400) and (ii) the admission process has a considerable component of subjective judgment that may respond to round numbers as well.

While this account is very closely related to the notion that SAT takers use round numbers as goals (it proposes SAT *evaluators* do), we attempt to tease apart the two in the next subsection and follow with Study 3 that rules out such concerns by design.

Score Sending

The dataset that we use contains information about the colleges to which students send their scores. We use this information to indirectly study whether they believe that a score just short of a round number is disproportionately worse than a score just above it. Intuitively, if they did, the set of schools to which they send their scores would differ.

To measure the quality of schools to which scores are sent, we first compute the average SAT score sent to each school by all applicants in the sample. We then compute the mean of these school averages for the set of schools where each applicant sent scores. In short, we use the average SAT score of *other* students who sent their scores to such schools as a proxy for the quality of schools to which a given test taker chooses to apply.[§]

[§] This measure correlates highly with that provided by the *US News and World Report* rank of top-50 schools for 1998 (the mid-year of our sample), r=.71, p<.0001.

Figure 5 shows the average quantity and quality of schools senior SAT scores were sent to as a function of the score obtained.[**] Both lines are upward sloping, indicating that students with higher SAT scores apply to more and better schools. More importantly, the figure shows that the quality and quantity of schools scores are sent to increase smoothly as scores pass round numbers. We interpret this as suggesting that round-number scores influence retaking decisions because of their motivating effect towards improvement, rather than because of their direct impact on the beliefs of the outcomes students can obtain with such scores.

Admission decisions

Figure 5 provides indirect evidence suggesting test takers do not *believe* their odds of admission change discontinuously by passing a round number. Data on admission decisions are required to assess if these beliefs are correct. While there is no centralized dataset of admission decisions across universities, one of the authors of this paper has used admissions data from two institutions in previous (and completely unrelated) work and we use those data here to partially address this question.

The first dataset is from a highly competitive private university for which just over 1100 undergraduate admission decisions are available. We estimate regressions with the admission decision as the dependent variable and with dummy variables for SAT scores as key predictors, controlling for other academic characteristics of the applicants (intuitively, we obtain conditional average admission probabilities for different SAT scores). We test the null that the probability of admission changes by the same amount for a given increase in SAT from just below and from just above a round number (e.g.,

[**] We must analyze score sending behavior by seniors because the sample of juniors has selection bias; we only observe school choices by students who felt their score was high enough not to merit re-taking the test.

that probability of admission increases by the same when going from 1390 to 1400, as when going from 1400 to 1410). This is a conservative test because we expect higher scores to have smaller marginal effects on admission decisions.[††] We test this null for scores around 1200, 1300, 1400, and 1500 and fail to reject the null for all four scores (N's=55, 97, 144 and 87 respectively, p-values=.96, .99, .20 and .92 respectively).

The second dataset is from a business school and hence consists of GMAT rather than SAT scores. We tested the equivalent null for scores around 600 and 700 (the maximum in the GMAT is 800) and also failed to reject the null (N's=3,432 and 1,404 respectively, p-values=.09, .93 respectively).

In sum, the admissions data suggest that probability of admission does not discretely change as scores pass round numbers, and the score-sending data suggests that students behave accordingly. Our next and final study is a scenario experiment that further eliminates concerns about third parties driving the motivating effects of round numbers.

*** Figure 7***

STUDY 3: SCENARIOS

Method and Procedure

This study was part of a sequence of unrelated experiments conducted at a behavioral laboratory in exchange for a flat payment fee. It was computer-based and presented each participant with three scenarios, always in the same order (N=172).

[††] Due to the small sample size we collapse contiguous scores, e.g. one dummy for SAT=1380 or SAT=1390.

In each scenario participants were presented with a situation that included feedback on partial performance and they were asked how motivated they thought they would be to improve upon it. The sole manipulation in the study was the level of partial performance. Each scenario had a 3 (distance-from-round-number: far below, just-below, just-above) x 2 (round-number: low, high) design

To ensure a *between*-subject design, each participant was assigned to the same condition for all three scenarios (e.g., a participant would be given partial performance measures that were far below the low round number for all three scenarios).

The scenarios were:

Scenario 1: Imagine that in an attempt to get back in shape, you decide to start running laps at a local track. After running for about 30 minutes and having done [*18/19/20/28/29/3*0] laps, you start feeling quite tired and are thinking that you might have had enough. How likely do you think it is that you would run one more lap?

Scenario 2: Imagine you are participating in a basketball tournament and that your current free-throw average is [*48.2/49.2/50.2/58.2/59.2/60.2*]%. Before your next game you calculate that if you made two free throws (and missed none) your season average would go up by just over 1%, to <*add 1.1% to previous number*>. During the game, you are fouled and walk to the line to shoot two free throws as your team is ahead by 5 points. How motivated do you think you would be to make those free throws?

Scenario 3: Suppose that you got a temp job to make some extra money. The job is tedious as it consists of copying and pasting from Acrobat (.pdf) files into an Excel spreadsheet, and then manually fixing cells that did not copy properly. You get paid roughly $3.50 per table copied. You are considering whether to do one more table or whether to head home directly. You check out your computer screen and see that today you have so far made $[*84.16/88.16/92.16/94.16/98.16/102.16*]. How likely do you think it is that you would do another table before going home?

Respondents answered these questions employing 1-9 Likert Scales.

Results and Discussion

We present results for the average 1-9 score given by each participant to all three scenarios as the dependent variable (*M*=6.34, *SD*=1.45, *MIN*=2.3, *MAX*=9). The results are shown in Figure 6.

*** Figure 6 ***

A one-way ANOVA revealed a main effect for distance-from-goal (F=15.31, p<.0001), and no main effect for high vs low round number (F =.2, p=.653) nor for an interaction (F=1.02, p=.362).[‡‡] The main effect for distance-from-goal arises because people reported greater motivation for improvement when their performance is just below a round number (M=7.14), compared to far below (M=6.06, t=4.61, p<.001) or just above (M=5.84, t=4.94, p<.001).

The fact that the round-number level did not influence motivation indicates that participants were not making inferences about overall effort or exhaustion based on that level, further suggesting they respond to distance from a round number because of its motivating effect rather than because of other inferences.

CONCLUSIONS

The present paper shows that round numbers in performance scales act as goals which more strongly motivate performance improvements among those performing just below vs. just above them. These findings contribute to the literature on reference points in at least two ways. First, they propose a new source of reference points that are naturally occurring and present in many different situations, and second, they provide often difficult to obtain evidence that reference points matter for real-life decisions.

We believe that round numbers acting as goals lead to at least two interesting questions for future research. The first is how these implicit goals, round numbers, interact with explicitly set ones. It is common for explicit goals to be set at round numbers; if round numbers are goals even without incentives, could it be that explicit

[‡‡] If the scenarios are analyzed separately the qualitative pattern of results is observed in all three; they are statistically significant for scenarios 1 and 2 (p's<.001) but not for scenario 3 (p=.4).

goals set at other numbers would lead to greater overall motivation? For instance, if people would naturally want to lose 30 pounds, would it be more effective to set an explicit goal at losing 33 rather than 30 pounds? Or is it the case that explicit goals are particularly effective if they are combined with implicit ones?

Another question for future research is the precise mechanism by which round numbers are motivating. One possibility is that people performing below a round number subjectively assess the odds of better performance in the future as greater (a judgment-mediated effect). Another possibility is that performing short of a round number is simply disproportionately aversive (a utility-mediated effect).

References

Feather, N. T. (1969). Attribution of Responsibility and Valence of Success and Failure in Relation to Initial Confidence and Task Performance. *Journal of Personality and Social Psychology, 13*(2), 129-&.

Heath, C., Larrick, R. P., & Wu, G. (1999). Goals as Reference Points. *Cognitive Psychology, 38*(1), 79-109.

Kahneman, D., & Miller, D. T. (1986). Norm Theory - Comparing Reality to Its Alternatives. *Psychological Review, 93*(2), 136-153.

Larrick, R. P., Heath, C., & Wu, G. (2009). Goal-Induced Risk Taking in Negotiation and Decision Making. *Social Cognition, 27*(3), 342-364.

Medvec, V. H., Gilovich, T., & Madey, S. F. (1995). When Less Is More - Counterfactual Thinking and Satisfaction among Olympic Medalists. *Journal of Personality and Social Psychology, 69*(4), 603-610.

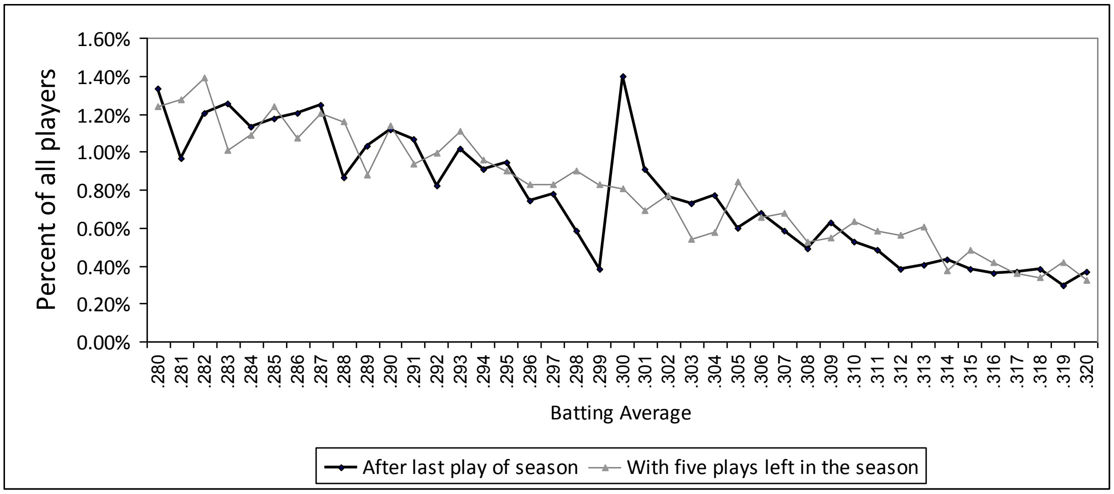
Medvec, V. H., & Savitsky, K. (1997). When Doing Better Means Feeling Worse: The Effects of Categorical Cutoff Points on Counterfactual Thinking and Satisfaction. *Journal of Personality and Social Psychology, 72*(6), 1284-1296.

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision Affect Theory: Emotional Reactions to the Outcomes of Risky Options. *Psychological Science, 8*(6), 423-429.

Rosch, E. (1975). Cognitive Reference Points. *Cognitive Psychology, 7*(4), 532-547.

Vigdor, J. L., & Clotfelter, C. T. (2003). Retaking the Sat. *Journal of Human Resources, 38*(1), 1-33.

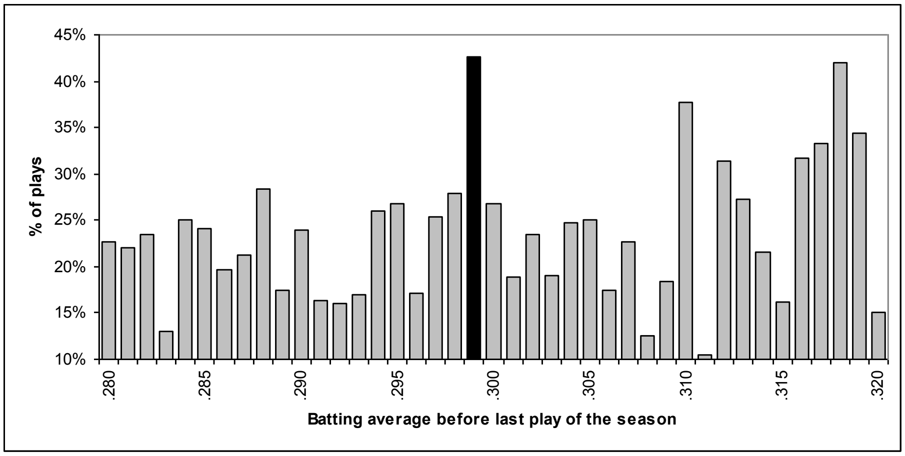
Figure 1. Distribution of batting averages in Major League, 1975-2008



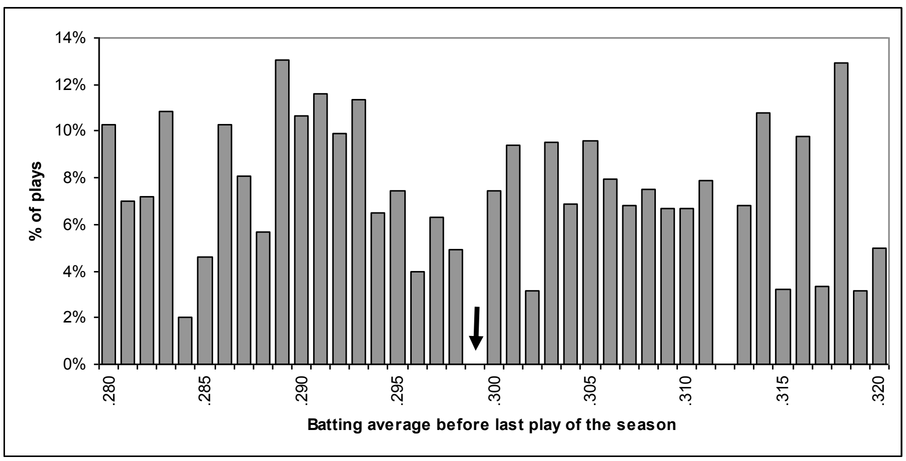
Note: Figure includes only player-seasons with at least 200 at bats.

Figure 2. Outcome of final plate appearance of season

Panel A – Hits



Panel B – Walks (cannot increase batting average)



Panel C – Substitutions (pinch hitter brought in)

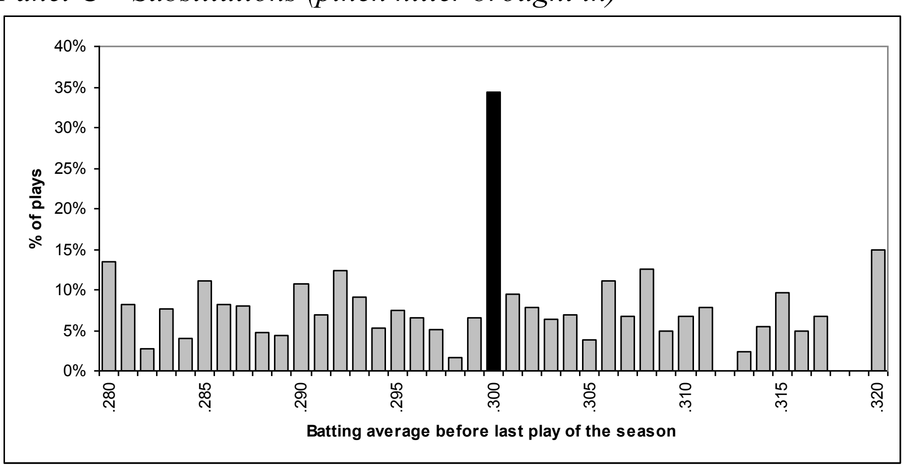
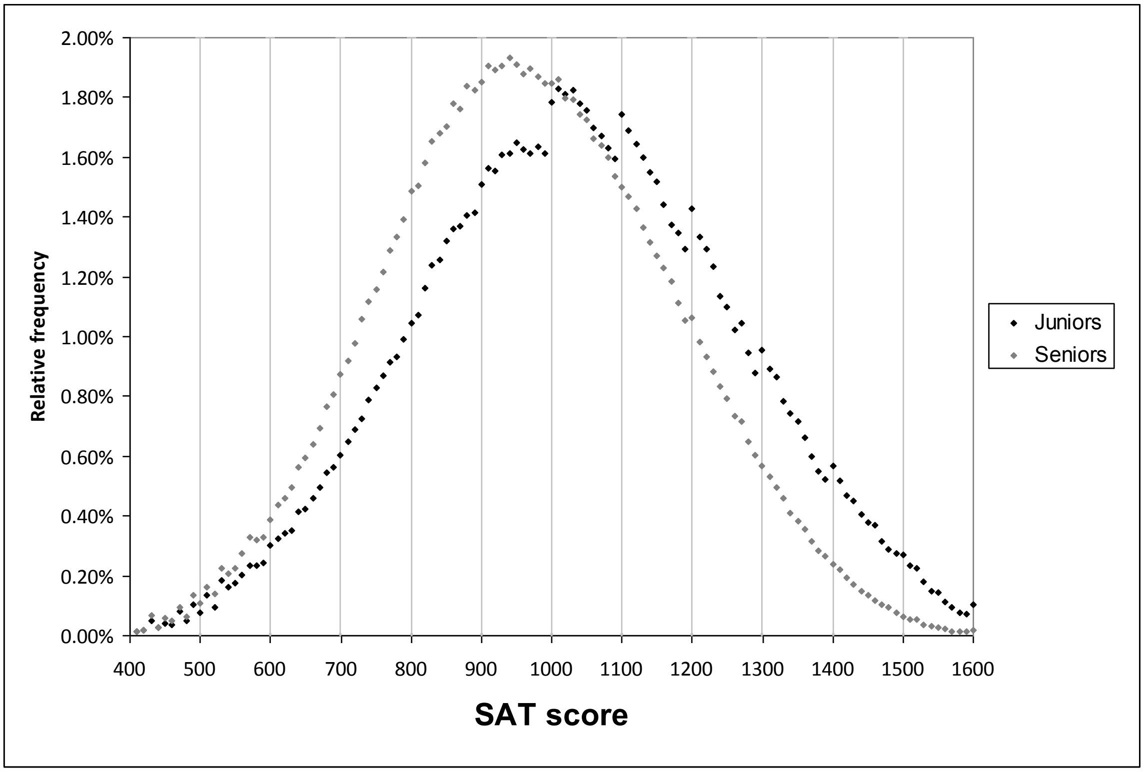


Figure 3. Distribution of SAT scores by high school juniors and seniors



Notes: Only final SAT scores are included in dataset. Prior tests taken by same students are not. Retaking is inferred from gaps in the distribution for juniors.

Figure 4. Ratio of frequencies in SAT scores for contiguous scores, for juniors and seniors *(slope of Figure 3)*

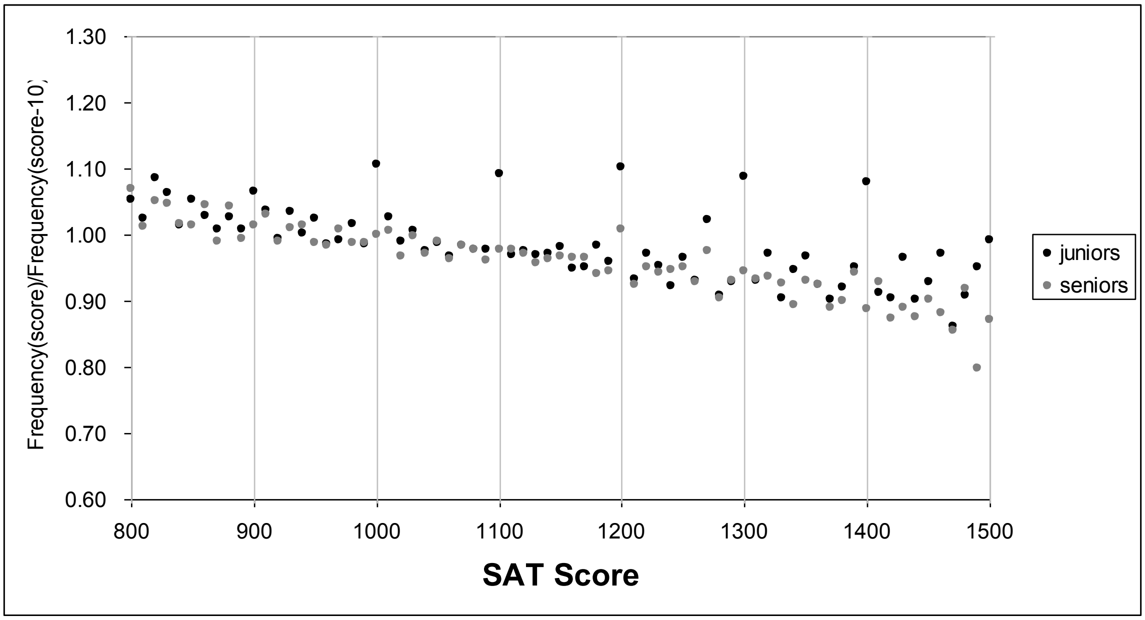
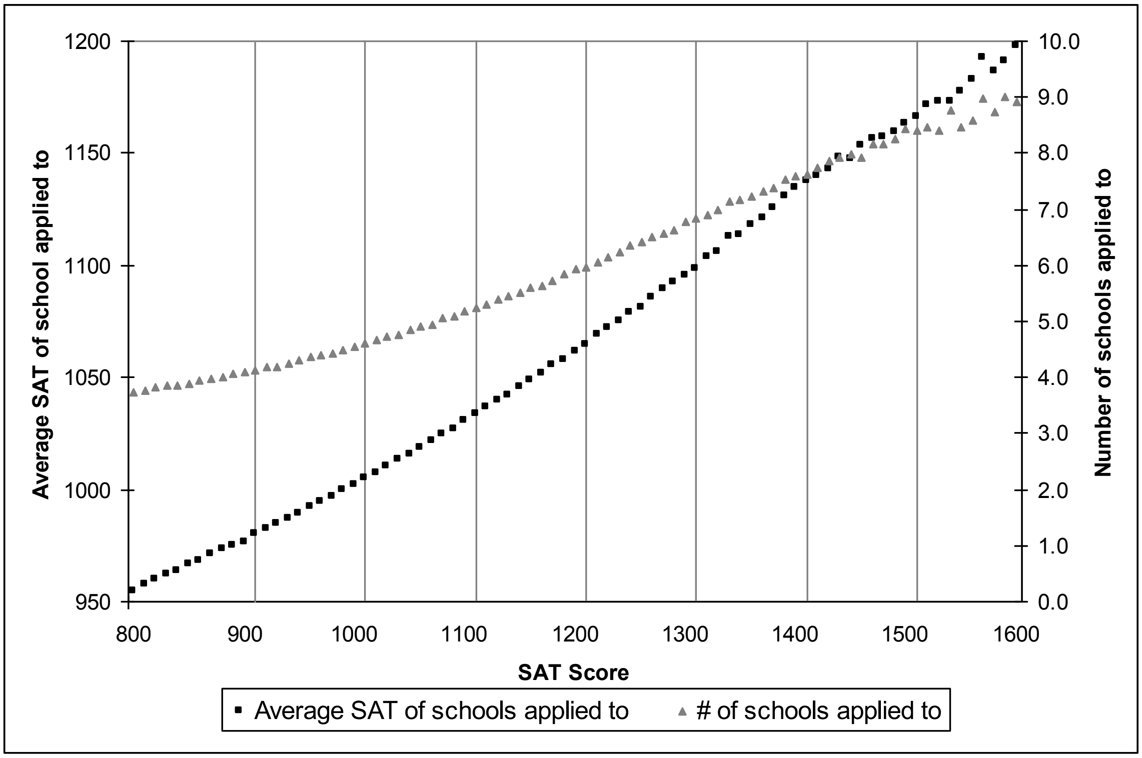
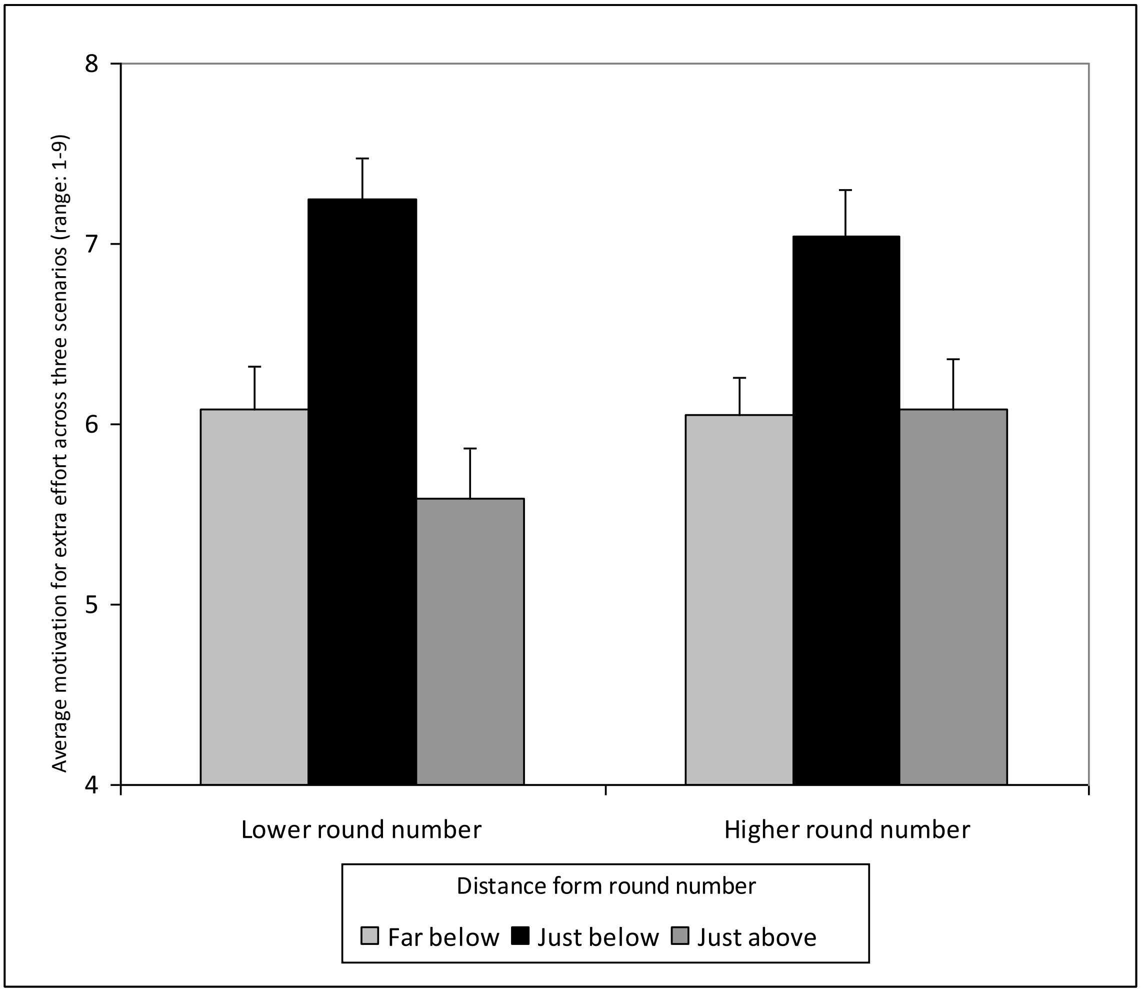


Figure 5. Set of schools SAT takers send their scores to as a function of their score



Notes: Left vertical axis (black line) indicates the mean SAT score of set of schools to which scores are sent to ("quality"). The right vertical axis (grey line) indicates the total number of schools applicants sent scores to ("quantity").

Figure 6. Study 3 – Reported motivation for more effort as a function of performance



Note: Results averaged across three scenarios, each with a 3 (distance from round number) x 2(high/low round number) design. All scenarios for given subject assigned to same condition.
Vertical bars indicate one standard error for the mean.