Investigating the Effect of Competition on the Ability to Solve Arithmetic Problems

Andrea Soto, Sarah Neff, Juanjo Carin, and Gopala Tumuluri

MIDS – W241 (FE) – Spring semester 2015

Contents

Abstract

The primary objective of this study was to determine if competition has an effect on a person's ability to solve arithmetic problems. To evaluate this question we had participants play a math game where they had a limited amount of time to answer as many simple arithmetic problems as they could. In treatment condition, subjects were told that they were competing against a 10-year-old child while in the control condition subjects did not compete against anyone. Moreover, the treatment condition was divided into two groups. In one group, subjects played against an outstanding 10 year-old who outperformed their score and in the other group, subjects played against an unexceptional 10 year-old who fell behind their score. Our research did not find any evidence that competition had an impact on performance. However, this is not a decisive conclusion and would recommend more experiments to be done in more realistic settings. Despite our efforts, we suspect our experiment did not have high stakes to truly mimic a competitive environment. In the end, our experiment resembled a 'lab experiment' more than a 'field experiment'.

As a secondary objective, we also investigated whether competition had heterogeneous effects among men and women. Again, we did not find any treatment effects.

1 Introduction

From an early age we learn to compete in sports and to have the best grades. We then compete to get into college, to secure a job and to get higher pay, and we even promote competition to ensure fair markets. We also value winning and use it to measure success. Our competitive approach to life is stems from our primitive biological instinct to secure food and survive, which has in turned molded our modern economic and social structures. Many people consider competition an important mechanism to find motivation and to motivate others, arguing that when we measure ourselves against others we feel motivated to do better. According to Google, to compete is to "strive to gain or win something by defeating or establishing superiority over others who are trying to do the same."

With competition permeating most of our lives, one question that arises is whether competition actually improves performance. This question could be investigated for a broad set of activities, including both physical or mental tasks. However, for the course project we narrow the scope of our research to investigate how competition affects performance on a single task: solving arithmetic problems. Our field experiment consists of a game where subjects have to correctly answer as many addition, subtraction, multiplication and division problems as they can. The outcome measure of interest is the total number of correct answers that the participant gets.

This task is particularly relevant to educational settings where competition is at the core of academic achievement and distinction. Standardized tests are the norm in most countries and results are used to rank both students and schools. This had lead many educational institutions to focus their teaching efforts almost entirely on attaining high scores on standardized tests. Culturally, academic achievement is also desired. In China, pressure to score high on the college entrance exam called gaokao is so intense, that students use intravenous drips to be able to study more[1]. Getting some insights into the effectiveness of such competitive pressure on performance would be of great value to asses educational frameworks. A positive effect would favor the current status quo while a negative effect could be used to promote cooperative styles of education.

The resulting net effect of competition on solving arithmetic problems is ambiguous. Competition can motivate individuals, make them more attentive, and thus, lead them to perform better. However, competition can also cause stress and be a distraction if the person focuses too much on winning and not on the task at hand. The effect might also depend on how individuals think they are doing. Some people might feel motivated only if they are doing worse than others, but not feel motivated if they are ahead of the pack. Some might want to outrank others no matter what, and yet others might not be influenced at all by the comparison.

[1] "Inside a Chinese Test-Prep Factory - NYTimes.com." 2014. 15 Apr. 2015. http://www.nytimes.com/2015/01/04/magazine/inside-a-chinese-test-prep-factory.html

Since losing and winning are a key characteristic of competition, we investigated the effects of competition when subjects think they are winning from when they think they are losing.

We hypothesis that the number of correct answers will be higher when the person thinks they are losing because they will try to catch up. On the other hand, we do not have a clear hypothesis for the winning treatment. When winning by a high margin the environment becomes less competitive, so people can slack off because they are already winning or they might still feel internal motivation to perform at their best. Therefore, the average effect when winning is ambiguous.

Finally, we investigate heterogeneous effects among men and women. Gneezy, Niederle and Rustichini (2003)[2] studied gender differences and found that men and women have different propensities to perform under competition. A follow up study by Gupta, Poulsen, and Villeval (2005)[3] found that when given the option to choose between being rewarded based on competitive standings or self-performance, men prefer to compete while women don't. We hypothesis that competition will have a positive effect on men's performance but a small or no effect on women.

2 Experiment Overview

2.1 Experimental Design

The objective of this study is to determine if competition impacts a person's ability to solve numerical problems. We examined this research question by creating an online, competitive game environment where participants can compare their progress to the progress of what would appear to be a lesser opponent, a 10-year-old child. The experiment ran for a 10-day period from the 1st to the 10th of April of 2015.

Subjects in the experiment played a math speed game where the objective was to correctly answer as many arithmetic problems as they could in 1 minute. The problems included addition, subtraction, multiplication, and division questions. The game was distributed online and was designed to be played on a computer or laptop. We therefore instructed participants to avoid playing on a mobile device.

The game randomly assigned participants to one of three conditions: control, losing treatment or winning treatment. In treatment condition, subjects were told that they were competing against a 10-year-old child. This opponent was chosen because it appears to be a lesser opponent who can be easily beaten. The intent was to appeal to participant's egos in an effort to engage them and make them feel pressure to do better.

If assigned to the losing treatment condition, subjects competed against an outstanding 10-year-old who outperformed their score. Conversely, if assigned to the winning treatment condition, subjects competed against an unexceptional 10 year-old who fell behind their score.

In order to create a virtual competitive environment, it was important to show treated subjects how they compared to the 10-year-old while playing. This would increase the sense of competition and allow the treatment to affect subject's performance. For subjects in treatment condition, the game displayed the score of the 10-year-old above the score of the participant. At the start of the game, both scores were set to 0. When the game began, the score of the participant would increase whenever they got a correct answer while the score of the 10-year-old would increase at a predetermined rate based on the assigned treatment. The performance of the 10-year-old was fabricated to make subjects believe that they were either winning or losing.

The average answer rate of the 10-year-old in the losing and winning conditions was 0.7 and 0.3 answers per second, equivalent to a score of 42 and 18 at the end of the game. After every score increase, the answer rate was randomly varied by +/-0.1 answers per seconds to make the score less periodic and more realistic. Subjects in the control group would only see their own score.

[2] Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. "Performance in competitive environments: Gender differences." Quarterly Journal of Economics - Cambridge Massachusetts - 118.3 (2003): 1049-1074.

[3] Datta Gupta, Nabanita, Anders Poulsen, and Marie Claire Villeval. "Male and female competitive behavior-experimental evidence." (2005).

From the outset we decided that we would conduct a pre-test to control for individual differences. There are two major advantages of doing a pre-test. First, we anticipated high variability in scores due to individual differences which include math ability, familiarity with technology, keyboard speed, and understanding of the game among others. By conducting a pre-test, we would be able to account for these differences, decrease the variability of our estimands, and increasing power. Second, the pre-test would allow us to estimate a difference-in-difference treatment effect instead of a difference in means. For the difference-in-difference estimate, we need to assume that each individual's trend over time will be the same for all groups. Because the game is played back to back, we think that it is unlikely that any outside factor will influence our participants in a way that will make their score differ over time. We also need to assume that the effects of "training" in round 1 is the same for all participants. In other words, we assume that some people do not benefit more from practicing than others. This could be a strong assumption given that the effects of training might be different between participants. People's confidence will be impacted differently depending on the their expectations coming into the game and their score. For example, a person with low expectations might feel encourage after a "good" round 1, while a person with high expectations might feel the opposite if they felt the problems were challenging. This would mean that practicing in round 1 had different effects on these individuals.

In order to establish the baseline score for each participant, we had participants play two rounds of the game. We wanted to make the control and treatment experience as similar as possible to avoid biased results. If treated subjects were told from the start that they would be competing against against another person on round 2, this could potentially affect their performance on round 1. To avoid this, we decided to tell treated subjects about the 10-year-old only after they had finished round 1 and were ready to start round 2. This ensured that everyone played without an opponent on round 1 and that they would have the same experience from the initial screen to the end of the first round.

Another concern was the risk of spillovers. Our recruiting method relied in part on a convenient samples of friends, students, and colleagues. This made the risk of spillovers more prominent than if we were only using random sampling methods. Our guess was that spillovers would be more frequent when people played the game together, one after the other. To mitigate some potential spillovers we clustered treatment assignment by game session. Sessions[4], like cookies, are used by websites to store information across multiple pages. However, unlike cookies, sessions do not store data on the user's computer. Our clustering by session does not completely eliminate the potential for spillovers because it is fairly common for people in a same household to have their own computer and to therefore have different game sessions. Additionally, spillovers could also occur in offices where people are nearby but on different computers. The risk of spillovers could be minimized by using random sampling instead of a convenience sample. For the current experiment, we assumed that no spillovers occurred beyond our session cluster assignment.

[4] For more information on sessions, see http://php.net/manual/en/session.examples.basic.php.

2.2 Game Interface

In this section we describe the stages of the experiment and the game interface. Figure 1 shows an overview of the process that participants had to follow.

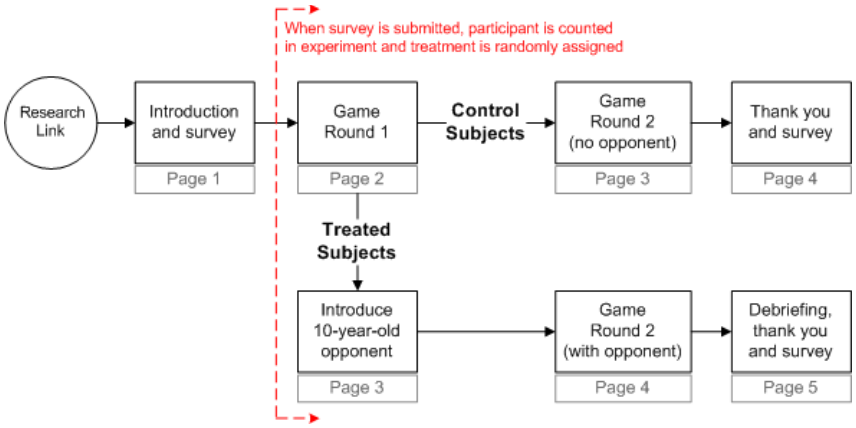

Figure 1: Flow Diagram of the Experiment

The experiment was distributed online by sharing a link to the game interface. When a person followed the link they would arrive to a welcome page with an invitation to participate in the experiment. The invitation text included the goals of the research, a brief description of what participants would be asked to do, the objective of the game and the total duration it would take to complete (no more than 5 minutes). People were instructed to complete a short, four-question survey at the bottom of the page to begin the game. The survey captured information about gender, age, education level, and comfort in completing simple arithmetic problems. Figure 2 shows a screenshot of this page.

Participants were told that they would play two rounds of the game but no one was told about the 10-year-old opponent. The opponent was introduced to subjects assigned to treatment before commencing round 2 of the game. Besides adding the score of the 10-year-old to the game interface for treatment condition, all other aspects of the game were the same. Each round had a one-minute time limit during which math problems appeared one at a time. Participants could not advance to the next problem until they had correctly answered the question asked and their score, equal to the number of correct answers, was displayed below the math problem. A sample game interface for round 1 and for the three conditions in round 2 are shown in Figure 3, Figure 4, Figure 5, and Figure 6.

At the end of round 2 participants were automatically directed to a final thank you page. Treated subjects were debriefed and notified that the score of the 10-year-old child was fabricated by the research team and that they did not played against a real person.

In addition to the thank you text, participants were asked to answer three follow-up questions regarding their experience. The follow-up questions were:

1. Did you enjoy playing the game?
2. Did you feel competitive pressure to win?
3. Did you feel you wanted to have a high score?

These question were intended to help us assess the level of engagement and competitive pressure that participants experienced while playing the game. The results will not be used in the experimental analysis but are a way to get some insight as to whether or not the competitive environment created was able to replicate a real competitive environment and if participants, especially participants in treatment, experienced a competitive urge to do well.

We also included a fourth question asking if the participant needed a Mechanical Turk confirmation code. This question would allow us to verify that the experiment was completed by participants recruited through Mechanical Turk. Figure 7 shows a screenshot of the final page.

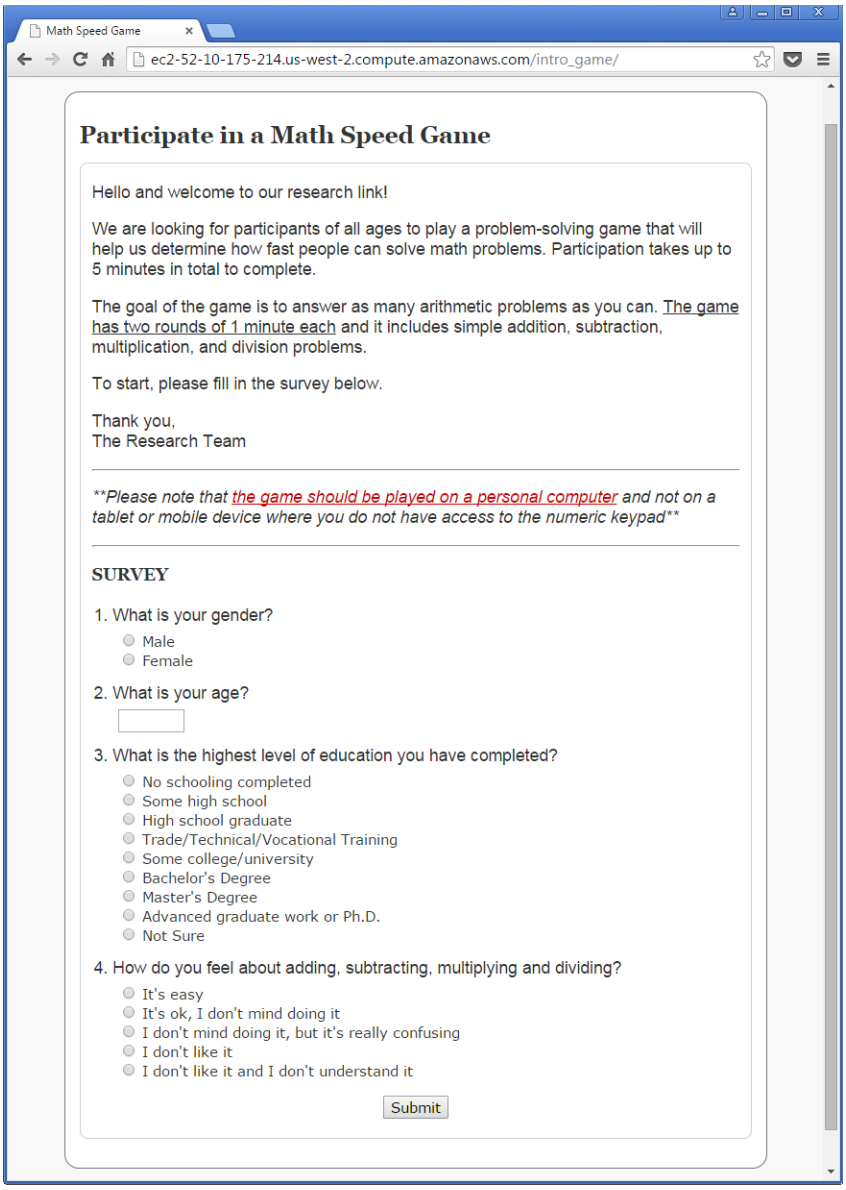

Figure 2: Welcome Page and Survey

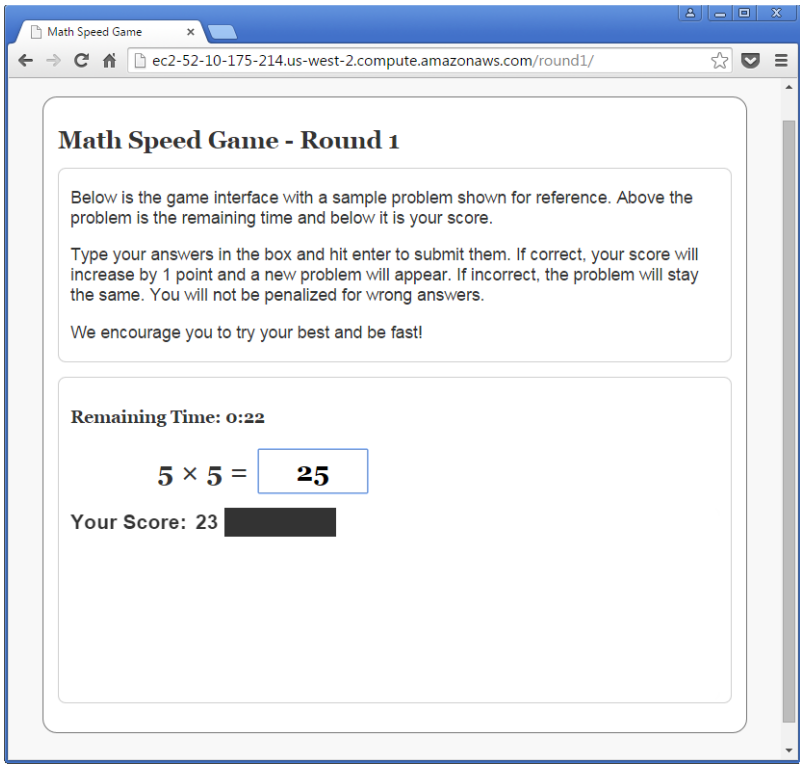

Figure 3: Sample Round 1 Interface

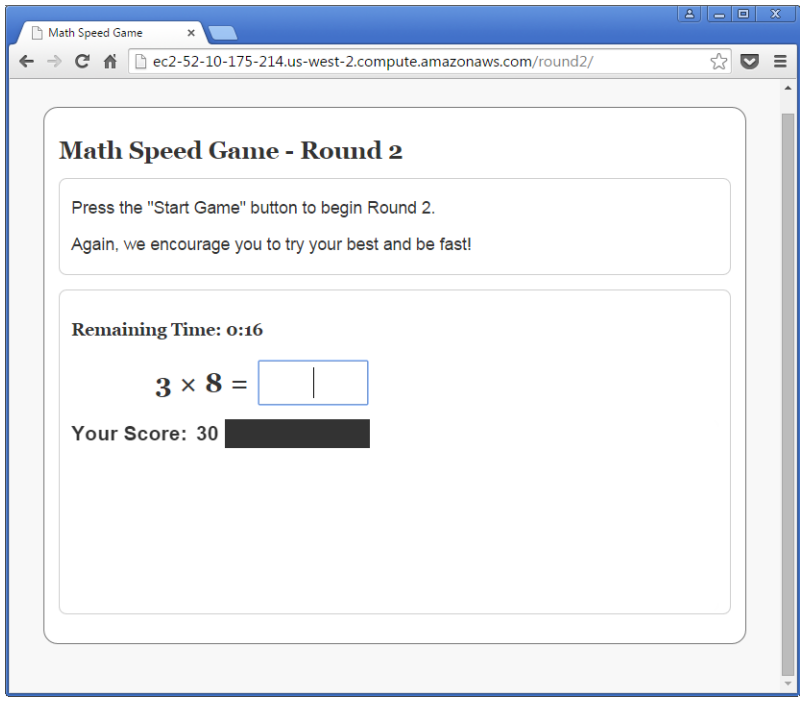

Figure 4: Control Condition Sample Round 2 Interface

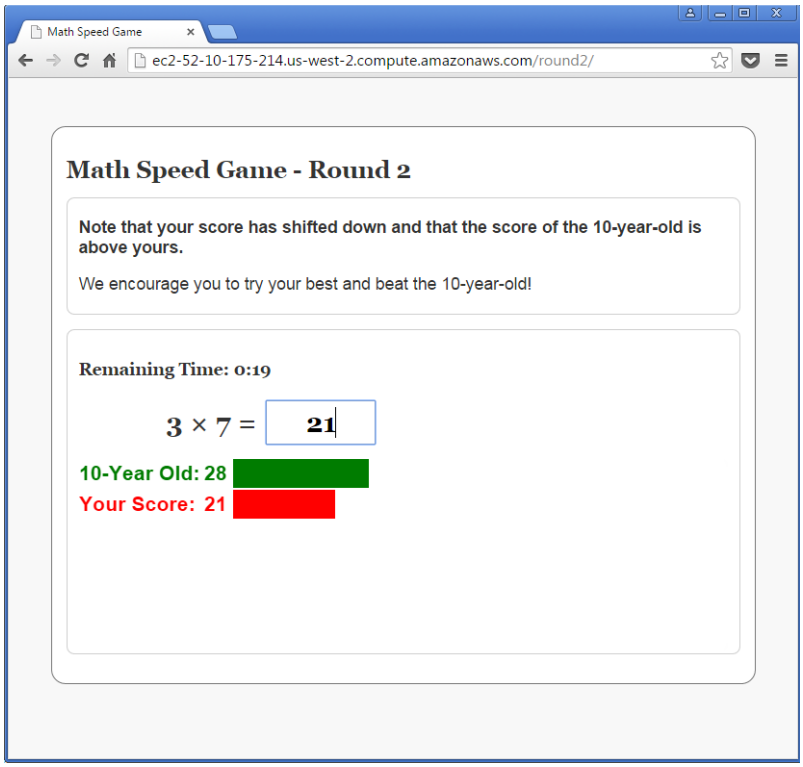

Figure 5: Losing Treatment Condition Sample Round 2 Interface

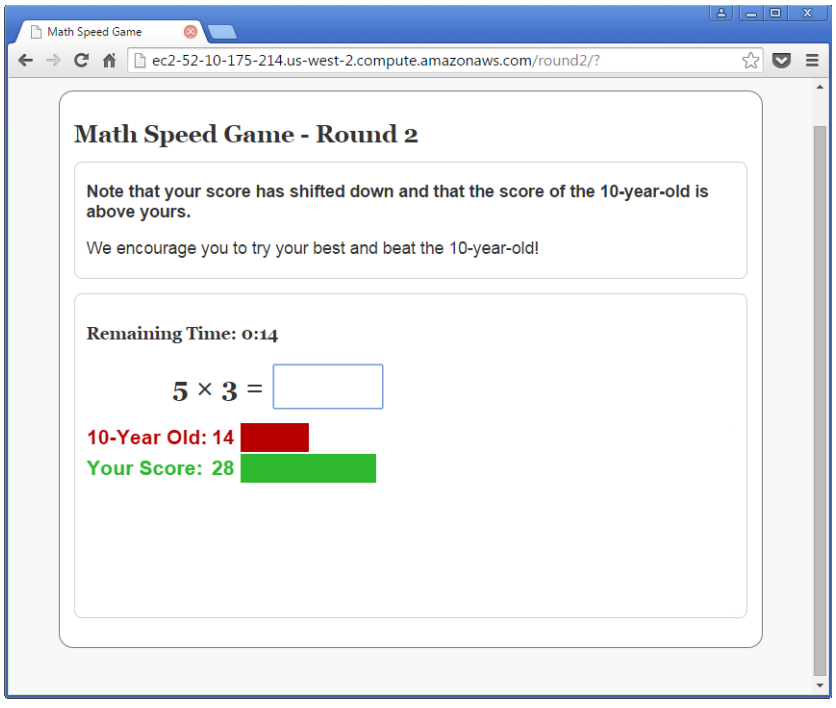

Figure 6: Winning Treatment Condition Sample Round 2 Interface

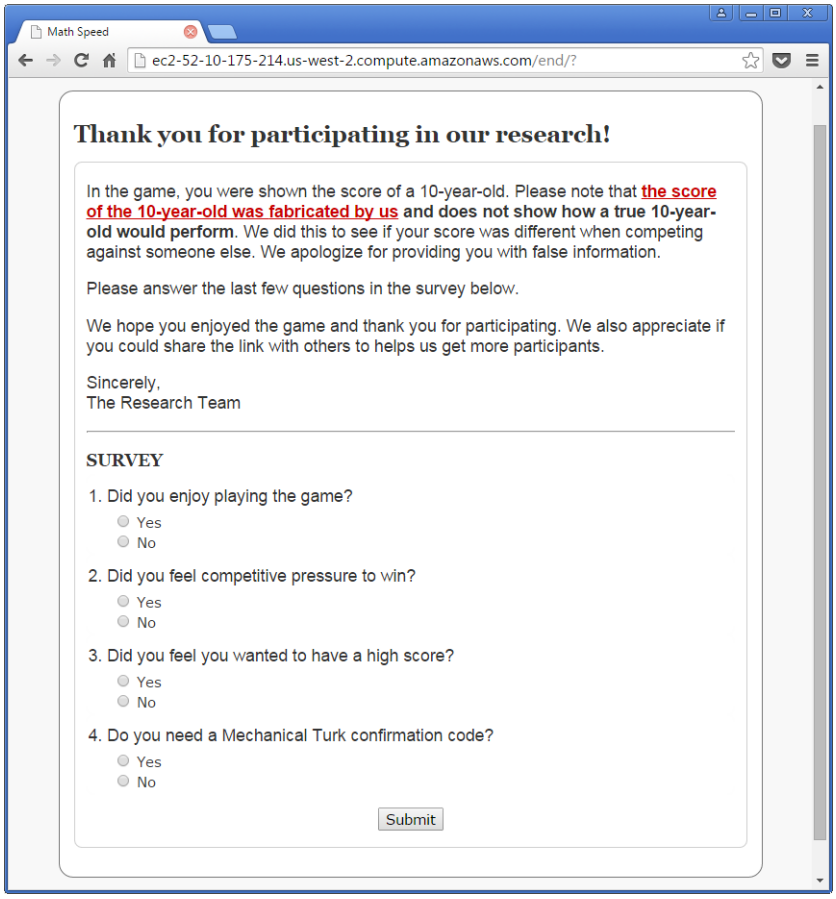

Figure 7: Debriefing, Thank You and Follow-Up Survey

2.3 Design Considerations

- *Technology*: We looked at several alternatives for the game, including Qualtrics, Python, Construct 2, Khan Academy, and HTML/Javascript. Because our treatment required showing the score of a competitor while playing the game, we needed a technology that was highly customizable. We discarded Khan Academy and Qualtrics for this reason. Looking into the other alternatives, we decided to use HTML and Javascript to develop the game because we found many online resources and found a simple tutorial for a basic game.

- *Process and Transitions*: We were concerned about how many people we were going to be able to recruit. A long experiment involving different interfaces, with surveys in one link and the game interface in another, could increase attrition, missing values, and mistakes. So we focused on making the experiment convenient for participants, with access to the survey and game through a single link with smooth transitions between instructions, survey questions, and the game itself.

- *Game Duration*: Initially the game had a time limit of 2 minutes per round, totaling 4 minutes of play time. However, after playing the game ourselves, we found 2 minutes to be a very long time; while playing, we felt like we wanted to quit, specially in the second round. We therefore cut the time of each round to last only 1 minute. This was good duration that kept us engaged in the game.

- *Types and Difficulty of Problems*: We had to decide on the type and difficulty of the problems. The initial proposal suggested simple arithmetic problems, but we discussed the option of having written problems, problems combining multiple operations (like $((9+3)/6x8)/4$), and problems with two digit

numbers (like $54 + 31$). On one side, we wanted the problems to be challenging for the average person, but on the other side, we didn't want most people to stall on a very difficult question. We foresaw two scenarios where potential outcomes could be identical for most participants. In one extreme, questions would be so easy that most people would solve all of the problems. In the other extreme, questions would be so hard that most people could not get past the first few questions. To quickly test if we were in an appropriate range and for simplicity we decided to use simple arithmetic problems consisting of addition, subtraction, multiplication and division.

- *Quantity of Problems*: The initial proposal and game design included several math problems popping up on the screen at different times. The game could hold up to 24 problems at any given time, and participants could choose which problem to answer. However, we discovered a flaw in this approach. Since participants could pick the questions to answer, the sequence of problems was not necessarily the same for everyone, and hence, the scores would not be comparable. To reduce variability in the scores and ensure all participants had to answer the same questions, we decided to show only one question at a time.

2.4 Randomization

Treatment was assigned after submitting the responses to the welcome page survey, at which point we assume that the person had agreed to participate in the study. The assignment was done using the mt_rand function of PHP[5]) which generates pseudorandom numbers using the Merseene Twister[6] algorithm. The treatment assignment persisted as long as the user did not close the browser and was lost once the browser was closed.

2.5 Pilot Study

Before conducting the actual experiment we ran a weeklong pilot study relying on Amazon Mechanical Turk (AMT) as our sole source to recruit participants. The two main goals of the pilot study were:

1) **Validate Experiment**

After locally testing the math game implementation, we embarked on ensuring the whole end-to-end experiment could be conducted at some scale, and that we would be able to gather the required data. It was also a chance to ensure random assignment to the three conditions worked as designed, and that we getting participants in each of the three groups. Additionally, the pilot was used to refine the math problems (difficulty and duration to solve) incorporated into the game based on participant's performance.

2) **Validate AMT Recruitment**

Being new to AMT, we also used the pilot as an opportunity to familiarize ourselves with the system, and to develop feasible approaches to publishing, recruiting and approving HITs. We created a template to recruit workers and experimented with it by posting a 3-question survey (unrelated to the experiment and created through Survey Monkey) and offering a small fee. For the pilot study, we repeated this procedure but incorporated our research link into the "Survey Link" template in the AMT system, and proceeded to posting the job to recruit workers to participate in our pilot by playing the math game. We experimented with a range of price incentives and auto-approval time delays. In the end, there were a total of 35 unique pilot participants out of which only two failed to complete the game.

[5] PHP is a server-side scripting language used mostly in website development.
[6] See PHP manual: http://php.net/manual/en/function.mt-rand.php.

2.5.1 Conclusions from Pilot Study

Recruiting enough participants was a big concern from the start. This concern was intensified when we were seeing little to no participation on the posted HIT which offered $0.03 payment. To rule out game difficulty/application reliability as the drivers of low participation, we raised the incentive level to $0.10 and then to $0.50 using batches. We found that any meaningful number of recruits could only be obtained at $0.50 per HIT, as only at this level we immediately saw a pickup in participation. Not only were the workers participating, they were using the game properly and generating good data to support our data analysis. The final incentive price was steeper than we had originally planned, and certainly meant that the team could not recruit hundreds of participants for the final experiment through this system. But, it gave us a very clear and quick path to a large enough sample size.

A second key concern we had in recruiting AMT workers was their seriousness in completing the game as designed to provide meaningful data. Our pilot study alleviated the first concern as most workers spent the necessary time to complete the game (3 to 5 minutes), as can be seen in the following distribution of completion times.

Figure 8: Pilot Worker Completion Times (Anticipated - 180 to 300 seconds)

Another concern was our ability to ensure unique worker participation, even across multiple batches. While there is no direct enforcement mechanism in AMT to prevent repeated participation, we found that all of the 35 workers across the three batches posted were unique, as identified by the AMT worker ID field. We also discovered that publishing a single large batch ensures unique worker participation (no repeated gamers).

During the initial pilot, we did not have a security measure to validate that each worker had truly completed the game. However, the survey link template in AMT has a provision to require workers to submit a code to prove that they completed the survey. During the last batch of pilot recruitment, we implemented a security code in the Web application and required AMT workers to submit it after completing the game in order to get their HIT approved for payment. This measure worked and all subsequent workers complied.

At the conclusion of the pilot study, we were confident that the application was working, random assignment was uniform across groups, that the data was being reported and had no errors, and that AMT was a viable recruitment platform for the final study. Data collected through the pilot study was analyzed for general completeness and usefulness. However, given the small sample size, it did not provide a meaningful basis to draw conclusions about the validity of the final experiment. We proceeded to the final experiment.

2.6 Experiment Recruitment

Our goal was to have at least 100 subjects in the experiment, with nearly 30 participants in each condition. In order to meet this goal, we decided to use a two-pronged recruitment efforts 1) using AMT and 2) by sharing the link through email and social media sites. Both methods can be called convenience samples, and

while we are aware of the limitations of convenience sampling in experiments in support of causal inference, this was the best approach we mustered given the time limitations.

1) **AMT Recruitment and Results**

AMT by default gives you access to workers rated as "masters." These are workers who acquired special status and are considered the most reliable and of higher quality. However, these workers also tend to be pickier about the work they do and the price they expect to be paid. Access to workers not designated as masters requires changing the default settings. In order to prevent any worker to take your HIT, AMT allows you to set thresholds on the number of tasks a worker completed and their task approval rate. By properly selecting these thresholds, you get access to a vastly larger pool of fairly good quality workers, and therefore a greater chance to achieve the target sample size quickly.

For the final experiment, we published two batches of 20 and 100 HITs respectively. The first batch required "masters" level qualification and the second one required high level of reputation (at least 1000 HITs completed with 95% approval rate). The second type of worker recruitment was used to speed up the process as the "masters" level resulted in a slow trickle of participants, at least for our comfort. Both batches were completed within two days.

In the end, we successfully recruited 120 participants through AMT, and were able to ensure unique participants across both batches (identified by AMT Worker ID). Below is the completion time distribution of the two AMT groups based on the data reported on that site. Out of the 120 AMT participants, 18 of them did not complete the game as specified (spending way less than 100 seconds or 1 minute and 40 seconds). There is a mean completion time difference of about 20 seconds between the masters group and the non-masters group of participants.

Table 1: AMT participants

	Master	Non-Master
Total Recruited	20	100
Number of Attritions	2	16
Mean Completion Time (sec)	200.5	230

The difference in attrition shown above is not statistically significant ($p = 0.73$), and neither is the difference in completion time ($p = 0.16$).

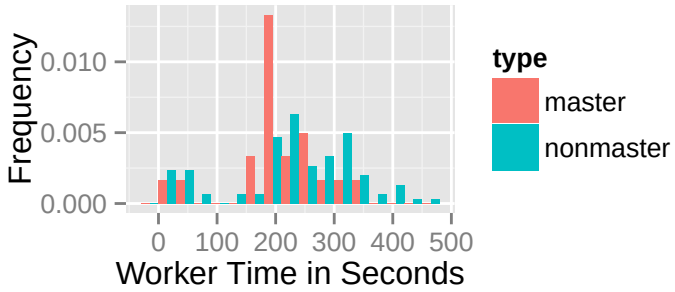

Figure 9: Work Time Distribution of Mechanical Turk Workers

2) **Social Media Recruitment and Results**

This was our second participant recruitment channel which provided **nearly 2/3 of the total participants** in our experiment. This is also the most convenient of the two samples. Recruitment through social media consisted of sharing the research link via Facebook, emailing friends and family, and reaching out to iSchool Berkeley students by email and the Slack application.

To ensure a consistent recruitment effort and avoid further bias in the sample, all of us used an identical invitation letter, solicited participation only once, and posted on the same social channels to each other's friends and family (see Table 2). These first level acquaintances were encouraged to forward the solicitation to their own friends and family. All invitees were strictly instructed not to discuss or disclose their experience with the game on any social channels. We monitored our networks for violations of confidentiality.

Table 2: Text statement used to release link of research experiment study

English	***Spanish***
Hi Everyone, As part of my UC Berkeley Data Science program I am running an experiment. Please follow this link and help me out with my research project. *ec2-52-10-175-214.us-west-2.compute.amazonaws.com/intro/*	Hola, Estoy haciendo una investigación como parte de mi maestria en Data Science en UC Berkeley y necesito participantes que me ayuden. Para participar, haz click en el siguiente link: *ec2-52-10-175-214.us-west-2.compute.amazonaws.com/intro/*
It will take you less than 5 minutes. Thanks, Researcher's Name	En total, el experimento toma menos de 5 minutos en completar. Gracias, Researcher's Name

Overall, this effort was successful in getting nearly **250 additional participants (well above the 120 participants recruited through AMT)**. For the most part, participants followed the confidentiality guidelines. Only in one instance (MIDS Slack channel) was there a minor breach through a comment posted about one participant's experience with the game. Unfortunately, our application is not designed to distinguish users or channels, and we could not exclude data related to this incident. In a more formal experiment, spillovers and non-compliance effects should be handled with more care.

Unlike AMT, these organic efforts did not allow us to compile completion time or attrition metrics as was done above for AMT workers. We did not have time to instrument the Web application to capture such granular data, including the channel through which the participant was recruited. In a real-world experiment, such capabilities can be highly valuable, and can allow for analyzing differences between different groups.

3 Data Cleaning and Descriptive Statistics

Over the 10-day period when the experiment link was open a total of 363 observations were collected[7]. We anticipated that our data would contain repeated responses because the game did not block or stopped participants from attempting to play more than once. However, to play multiple times, participants had to re-submit their responses to the first survey. This allowed us to identify repeated responses by matching the gender, age, and IP address of participants.

Screening the raw data for repeated responses resulted in 39 duplicated records, corresponding to 17 people.

We defined complete records as those who had no missing scores for the two rounds. We used this definition to clean the set of repeated responses by discarding non-complete records and keeping only the first complete record of each participant. So for example, if a person had two unfinished games with incomplete data and then a third game with complete data, we would only keep the third record. Similarly, if a person completed the whole experiment (all complete records) more than once, we would only keep the first record. Additionally, if a person with duplicate responses never finished the experiment we kept one of the records so that the observation would be accounted for when looking at attrition.

After applying the described filters to the set of repeated responses we discarded a total of 22. We also found 2 records from participants below 18, that we removed because they were underage, leaving a total of 339 observations in the dataset.

From the non-duplicate dataset with 339 observations, 32 participants had missing values in one or both scores of the game, leaving 307 with observable outcomes. Finally, 4 of those 307 responses completed the experiment using a mobile device which we identified as having a screen size of 375×667 or smaller. We removed these observations because the game was not designed or tested for mobile devices (we were unsure if the application was displayed or worked correctly) and performance on a mobile device may be very different from a computer.

We estimated the treatment effect for two datasets, one without missing values, and the other without missing and abnormal values. The second dataset is a trimmed version of the first where abnormal scores were considered as missing. This way we reduce our sample size (without attrition) from 303 to 260 participants, after discarding those who replied: We defined abnormal scores as those suspiciously low or very different between rounds. The criteria for obtaining the latter dataset was based on percentiles: observations with scores in either round below the 5th percentile, and observations whose score difference were below the 5th percentile or above the 95% percentile (i.e., a huge difference, either positive or negative, between round 1 and round 2). The trimmed dataset had a reduced sample size of 260 participants, since 43 observations met one or more of the following criteria:

- A score of 9 or less in the 1st round,
- A score of 10 or less in the 2nd round, or
- A difference in scores from the 1st to the 2nd round of -6 to -20 or 12 to 23

Table 3 shows a summary of these two datasets that were analyzed.

[7] These data (with the IP addresses conveniently anonymized), as well as the R code we used to analyze them, can be found at http://goo.gl/tJRZRF.

Table 3: Descriptive statistics of the sample under study

	w/o missing values	w/o abnormal values
Assigned to winning condition	31.0%	28.8%
Assigned to losing condition	37.6%	38.1%
Score in the 1st test (mean)	26.3	28.1
Score in the 2nd test (mean)	28.9	30.6
Male	61.7%	61.2%
Age (mean)	36.9	36.2
Education level (mean)	6.2	6.3
N (sample size)	303	260

3.1 Randomization and Covariate Balance Checks

Before proceeding to the analysis phase, we looked for red flags in our randomization procedure by checking that the proportion of individuals assigned to each group was similar[8] and that the covariates gender, age, and education level were balanced. Table 4, for instance, shows the number of participants assigned to each group, as well as the corresponding proportion.

Table 4: Subject Distribution by Group

	No. Participants	Proportion
Control	107	31.38%
Loser	126	36.95%
Winner	108	31.67%
TOTAL	341	100.00%

The proportion of participants in each group is not statistically different from each other ($\chi^2 = 1.9292$, $df = 2$, $p = 0.3811$), which gives us confidence that our randomization was correctly implemented.

Table 5 shows the results of the regression analysis checking covariate balance across experimental conditions. As expected, the experimental condition participants are assigned to does not help predict the proportion of gender, age, education level or math confidence.

Table 5: Effect of the treatment on covariates

	Gender	Age	Education level	Math Confidence
Losing treatment	-0.001	0.802	0.071	-0.126
	(0.064)	(1.615)	(0.150)	(0.098)
Winning Treatment	0.012	1.956	0.192	-0.003
	(0.066)	(1.809)	(0.150)	(0.108)
Baseline (Control)	0.377***	36.358***	6.113***	1.670***
	(0.047)	(1.124)	(0.103)	(0.073)
R^2	0.000	0.004	0.005	0.006
F	0.024	0.605	0.790	1.031
p	0.977	0.547	0.455	0.358
N	339	339	339	339

[8] To check the randomization process, we also consider the 2 participants below 18 (that's why the total number that appears in Table 4 is 341, because they used numbers from our Random Number Generator. There is no need to consider the duplicated records, because they re-used random numbers (participants that entered the website more than once were assigned their prior experimental condition).

While our sample does not have a 50-50 split of males and females, the male-female ratio is similar between the three conditions. More specifically, treatment assignment does not help predict the proportion of women. A χ^2 test also yields the same results ($\chi^2 = 0.047$, $df = 2$, $p = 0.977$).

4 Analysis and Results

4.1 Pre-Analysis

In the pre-analysis we looked at the mean scores by group and their corresponding standard errors to get a preview of the treatment effects. Whether we compare the second score or the difference with the first score, participants in the winners' group (those who are compared to a kid performing worse than them) seem to have a higher score than participants in the other two groups. However, looking at the standard errors the difference does not seem to be statistically significant.

Table 6: Mean score by group

	Control	Loser	Winner	TOTAL
Mean score on 1st test	27.26 (0.90)	25.92 (0.96)	25.63 (1.07)	26.25 (0.57)
Mean score on 2nd test	29.68 (0.90)	28.32 (0.99)	28.91 (1.07)	28.93 (0.57)
Mean score difference	2.42 (0.48)	2.40 (0.43)	3.29 (0.63)	2.68 (0.29)

We estimated clustered standard errors because we clustered treatment by game session. However, since only those participants sharing the same session—and hence the same personal computer—were grouped in a single cluster, there were very few clusters with more than one participant: only 5 out of the 298 clusters of participants that finished the experiment—256 if we consider abnormal values also as attrition). Therefore, the resulting clustered standard errors were pretty similar to robust standard errors[9].

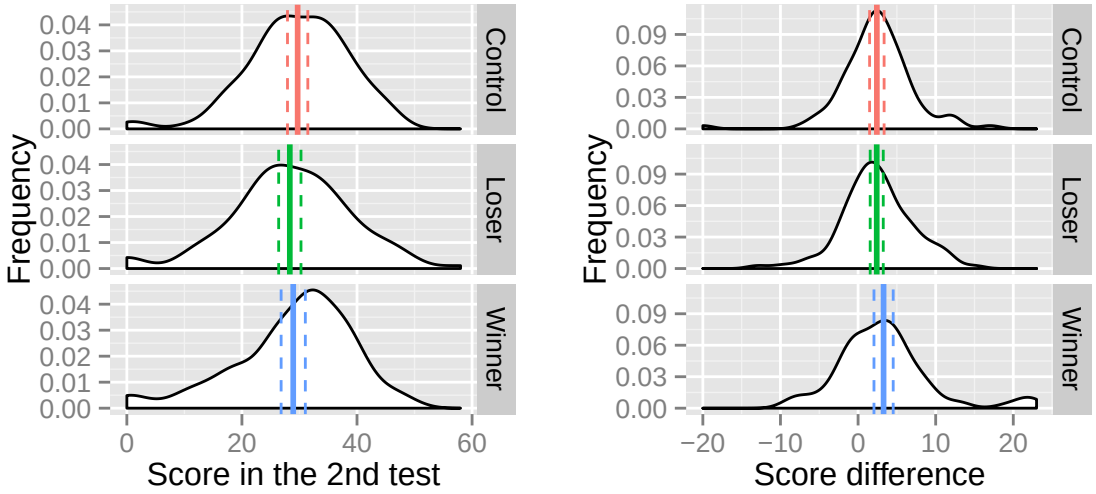

Figure 10: Density plot of 2nd test score and score difference between tests by group

[9] And these are also pretty similar to standard errors under the assumption of homoskedasticity, since that condition is met—variance across Groups is very similar.

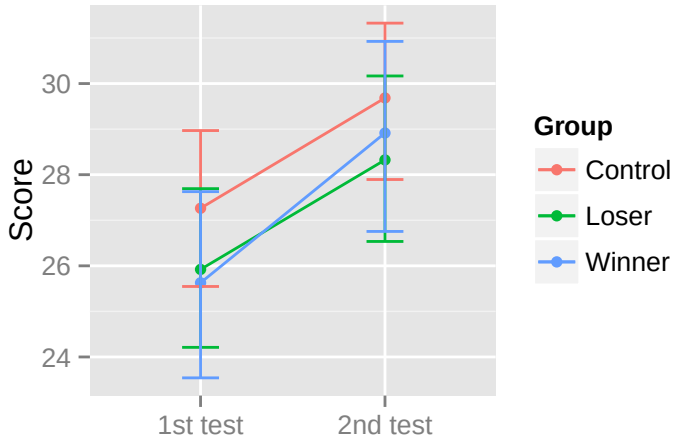

Figure 11: Error bar graph of the mean score by group

Figure 10 shows the distribution of scores and difference in score between rounds, and Figure 11 shows the mean score per round with the 95% confidence interval.

The results are very similar for the trimmed dataset where abnormal values are discarded. There does not seem to be a statistically significant difference between treatment conditions.

Table 7: Mean score by group when abnormal scores are trimmed out

	Control	Loser	Winner	TOTAL
Mean score on 1st test	28.22 (0.87)	27.69 (0.85)	28.35 (0.95)	28.05 (0.51)
Mean score on 2nd test	30.38 (0.84)	30.36 (0.87)	31.01 (0.89)	30.56 (0.50)
Mean score difference	2.16 (0.35)	2.68 (0.34)	2.67 (0.42)	2.50 (0.21)

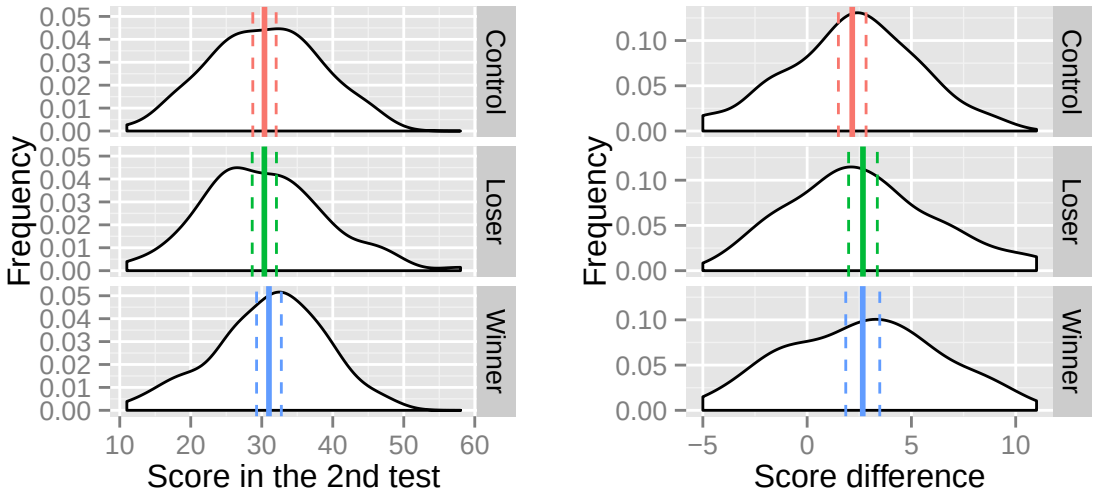

Figure 12: Density plot of 2nd test score and score difference between tests, by group, when abnormal scores are trimmed out

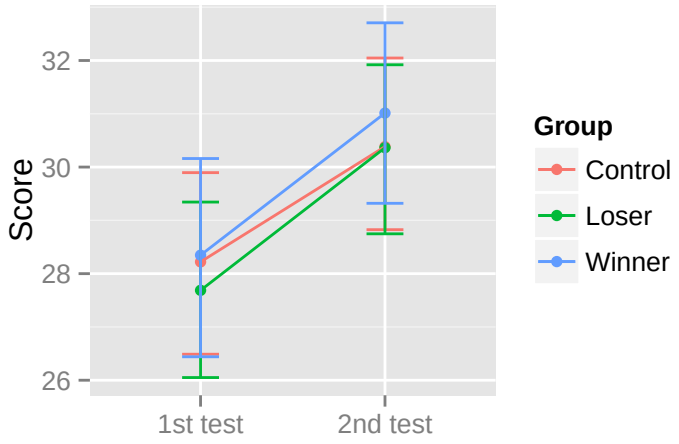

Figure 13: Error bar graph of the mean score by group when abnormal scores are trimmed out

4.2 Main Analysis

To perform our analysis, we decided to use regression to compare each of the two different treatments to the baseline/control group. Our hypothesis was that the score mean of these two groups was different from the score mean of the control group—i.e., we used planned contrasts—, without assuming whether those possible differences were positive or negative—so we used two-tailed tests.

Table 8 shows the estimated effect of treatment based on the following model specification:

$$y_{2i} = \alpha + \beta_L L_i + \beta_W W_i + \varepsilon_i$$

where i indexed each participant, y_{2i} is the score of round 2, L_i and W_i are dummy variables that denote

whether participant i was assigned to losing or winning condition respectively, and ε_i is the error or residual term .

Table 8: Effect of the treatment on the score

	w/o missing values	w/o abnormal values
Losing treatment	-1.360	-0.020
	(1.339)	(1.208)
Winning Treatment	-0.769	0.630
	(1.401)	(1.219)
Baseline (Control)	29.684***	30.384***
	(0.902)	(0.839)
R^2	0.003	0.001
F	0.480	0.167
p	0.619	0.847
N	303	260

The regression results show no significant treatment effect between conditions[10]. Although there are no significant effects, the initial model suggest that the effect of both treatment conditions had a negative impact on performance.

In order to reduce the variance in our initial results due to individual differences like math and keyboards ability, we run a difference-in-difference estimate using the score of the first round. Table 9 shows the difference-in-differences estimates, according to the following model:

$$y_{2i} - y_{1i} = \alpha + \beta_L L_i + \beta_W W_i + \varepsilon_i$$

where y_{1i} is the score of round 1.

Table 9: Effect of the treatment on the score difference

	w/o missing values	w/o abnormal values
Losing treatment	-0.018	0.514
	(0.644)	(0.490)
Winning Treatment	0.866	0.504
	(0.794)	(0.542)
Baseline (Control)	2.421***	2.163***
	(0.483)	(0.348)
R^2	0.006	0.005
F	0.944	0.641
p	0.390	0.527
N	303	260

None of the estimates are statistically significant. The effects are close to zero, suggesting competition did not have any effect on performance on solving arithmetic problems.

[10] The notation for significance is the usual one: *** for p-values smaller than 0.001, ** for p-values smaller than 0.01, * for p-values smaller than 0.05, and . for p-values smaller than 0.1.

4.3 Heterogeneous Treatment Effects

Since the 1st test score does not predict perfectly the 2nd test score[11] (even in the absence of treatment), we used the 1st test score as a covariate. We also added gender and age as possible covariates[12]. Hence, the model specification is:

$$y_{2i} = \alpha + \beta_L L_i + \beta_W W_i + \beta_C C_i + \beta_{LC}(L_i C_i) + \beta_{WC}(W_i C_i) + \varepsilon_i$$

where C_i is the value of the covariate (be it the score in the first round, the gender, or the age) for participant i.

Table 10: Heterogeneous effects

	1st score	Gender	Age
Baseline	6.083***	30.603***	28.621***
	(1.556)	(1.204)	(2.908)
Losing Treatment	-1.838	-2.034	6.523
	(2.012)	(1.778)	(4.024)
Winning Treatment	1.612	-0.077	10.899**
	(2.590)	(1.704)	(4.092)
(1 point more in) 1st score	0.866***		
	(0.050)		
Losing:1st score	0.063		
	(0.066)		
Winning:1st score	-0.038		
	(0.083)		
Female		-2.360	
		(1.787)	
Losing:Female		1.696	
		(2.680)	
Winning:Female		-1.734	
		(2.914)	
Age (per year)			0.030
			(0.070)
Losing:Age			-0.213*
			(0.095)
Winning:Age			-0.311**
			(0.100)
R^2	0.753	0.020	0.068
F	181.367	1.229	4.351
p	0.000	0.294	0.014
N	303	303	303

As expected, the score in the first round helps predict the score in the second round ($p = 1.60e-47$), but the effect of any of the two treatments is not statistically significant, neither is the interaction with the score in the first round (i.e., participants seem to be equally unaffected by the treatment, regardless of how they performed in the first round).

The treatments seem to have no effect, regardless of the gender, but when age is used as a covariate, not only

[11] In other words, its corresponding regression coefficient, β_{1i}, is close but not equal to 1, as shown later.

[12] We decided not to use education level and math confidence as covariates due to the small proportion of some levels—the combination of two or more covariates partitions our original sample into even smaller subgroups, therefore increasing uncertainty and yielding not statistically significant effects.

being assigned to the winning condition has a significant (positive) effect (10.90 (4.09), $p = 0.008$; but do note that the baseline for age is 0), but also the interaction with Age (0.31 (0.10) points less per additional year, $p = 0.002$). The interaction of Losing condition and Age is also significant (0.21 (0.09) points less per additional year, $p = 0.025$). These results are depicted in Figure 14.

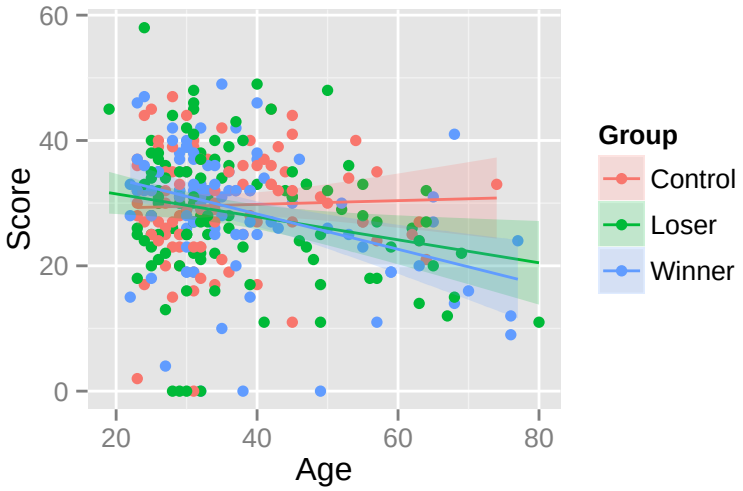

Figure 14: Scatterplot of age and score in the 2nd test split by group

Anyhow, we should be cautious about these results—since the treatments were randomly assigned, but none of the three covariates were, so interpretation remains ambiguous. We might say that the older people are, the more responsive they are to any of the two treatments, but we cannot claim causality.

The results when we discard abnormal values are pretty similar, as shown in Table 11.

Table 11: Heterogeneous effects when abnormal scores are trimmed out

	1st score	Gender	Age
Baseline	5.178***	31.765***	30.890***
	(1.080)	(1.006)	(2.995)
Losing Treatment	-0.773	-1.087	7.578.
	(1.604)	(1.521)	(3.905)
Winning Treatment	2.067	0.453	11.914**
	(1.837)	(1.398)	(3.939)
(1 point more in) 1st score	0.893***		
	(0.039)		
Losing:1st score	0.044		
	(0.055)		
Winning:1st score	-0.055		
	(0.062)		
Female		-3.393*	
		(1.721)	
Losing:Female		2.554	
		(2.472)	
Winning:Female		0.279	
		(2.589)	
Age (per year)			-0.014
			(0.080)
Losing:Age			-0.204*
			(0.101)
Winning:Age			-0.312**
			(0.103)
R^2	0.836	0.027	0.109
F	258.729	1.388	6.216
p	0.000	0.251	0.002
N	260	260	260

4.4 Attrition

Considering only missing values, the attrition rate is 10.6%.

Attrition rates are similar in each group: 10.4% in the control group, 8.8% in the Losers' group, and 13.0% in the Winners' groups—these differences are no where near statistically significant ($F = 0.531$, $p = 0.589$). Attrition also seems uncorrelated with pre-treatment variables—some levels of some covariates predict missingness within a certain group (e.g., Education Level 7 in the Control group, Education Level 9 in the Winners' group), but just because their proportion is relatively small; overall, attrition rates are not different from one group to the other.

Had these conditions not been met, attrition might still be independent of potential outcomes (and hence results could be biased), but we would be less confident. Our assumption is that attrition may be related to potential outcomes (subjects who perform poorly in this kind of tests, or who are not confident in their Math skills, may be likelier to attrit—or even more probably, to not participate, since participation was voluntary—, and thus most of the missing values would correspond to low scores), but not to the experimental group they are assigned—therefore, we assume participants did not leave the experiment (or did not try their best) because they were compared—or not—to another individual.

Table 12: Regression estimates predicting missingness as a function of covariates, by experimental group

	Control	Loser	Winner
Baseline	0.052 (0.136)	0.089 (0.093)	0.094 (0.126)
Age	0.003 (0.004)	-0.001 (0.002)	0.001 (0.003)
Female	-0.055 (0.064)	0.023 (0.060)	-0.033 (0.065)
Education level 7	-0.129* (0.051)	-0.019 (0.061)	0.099 (0.080)
Education level 8	0.091 (0.163)	-0.042 (0.078)	0.031 (0.111)
Education level 9			0.713** (0.269)
Education level 5	-0.044 (0.095)	0.066 (0.130)	0.016 (0.077)
Education level 4		-0.070 (0.079)	-0.124 (0.079)
Education level 3	0.081 (0.240)	0.068 (0.159)	0.005 (0.125)
Education level 2			0.389 (0.332)
Math confidence 2	0.007 (0.062)	0.052 (0.060)	-0.118. (0.065)
Math confidence 3	-0.018 (0.049)	-0.058 (0.137)	-0.168* (0.066)
Math confidence 4	-0.132 (0.092)	0.332 (0.234)	0.104 (0.224)
Math confidence 5	-0.212 (0.171)		-0.237* (0.118)

If we assume that missingness is not independent of potential outcomes, but that it is given the set of covariates we have collected, we might apply an *inverse probability weighting* scheme to recalculate our estimates[13]. The results we obtain are very similar, as shown in Table 13. We could also place bounds around the ATE—which requires fewer assumptions, hence reducing the risk of bias—, but that would make our estimates even more imprecise, which they already are[14]).

[13] By replacing missing outcomes by the corresponding subgroups's average outcome, and giving more weight to those subgroups with a higher attrition rate.

[14] We did it anyway. Let's suppose the maximum possible score is 60 (in our sample, it is 58). The minimum score is 0, so given the attrition rates and mean scores in each group, we have:
$y_{Loser\ lower\ bound} = [28.32 \cdot 0.912 + 0 \cdot 0.088] - [29.68 \cdot 0.896 + 60 \cdot 0.104] = -7.001$ (instead of -4.56)
$y_{Loser\ upper\ bound} = [28.32 \cdot 0.912 + 60 \cdot 0.088] - [29.68 \cdot 0.896 + 0 \cdot 0.104] = 4.51$ (instead of 1.84)
$y_{Winner\ lower\ bound} = [28.91 \cdot 0.870 + 0 \cdot 0.130] - [29.68 \cdot 0.896 + 60 \cdot 0.104] = -7.664$ (instead of -4.12)
$y_{Winner\ upper\ bound} = [28.91 \cdot 0.870 + 60 \cdot 0.130] - [29.68 \cdot 0.896 + 0 \cdot 0.104] = 6.34$ (instead of 2.58)

Table 13: Heterogeneous effects after reweighting the estimates

	1st score	Gender	Age
Baseline	6.352***	30.599***	28.744***
	(1.558)	(1.174)	(2.827)
Losing Treatment	-1.947	-2.120	6.319
	(2.010)	(1.752)	(3.973)
Winning Treatment	1.741	0.063	10.825**
	(2.687)	(1.679)	(4.025)
(1 point more in) 1st score	0.858***		
	(0.050)		
Losing:1st score	0.067		
	(0.066)		
Winning:1st score	-0.043		
	(0.086)		
Female		-2.366	
		(1.753)	
Losing:Female		1.849	
		(2.659)	
Winning:Female		-1.716	
		(2.875)	
Age (per year)			0.026
			(0.068)
Losing:Age			-0.209*
			(0.093)
Winning:Age			-0.305**
			(0.098)
R^2	0.752	0.021	0.068
F	179.959	1.258	4.300
p	0.000	0.286	0.014
N	339	339	339
Attrition Rate	10.6%	10.6%	10.6%

If we also discard abnormal values (following the criteria mentioned in Section 3), the attrition rate is 23.3%. Now attrition rates are not so similar: 18.9% in the control group, 20.8% in the Losers' group, and 30.6% in the Winners' group—attrition in the latter group is near significantly different than that of the Control group ($p = 0.048$, but it should be almost 3 times smaller because of the multiple comparisons). Anyway, that should be further explored: the abnormal values we should be concerned about are those that involve a huge negative difference between the score in the second test and the score in the first test. An analysis of those records reveals that most of those participants (almost half of them) were assigned to the Winners' group, and that difference in the proportions is, again, near statistically significant ($p = 0.054$)—maybe that treatment did have an effect in attrition (a possible hypothesis, merely speculative, is that being compared to someone who is performing worse may have a negative conditional effect on a certain subset of subjects, that lose interest). On the other hand, the results mentioned above are just a result of the arbitrary values that we chose to trim out; had we chosen other ones, attrition rates in each group could have been rather different (but it's worth noticing this issue).

Table 14: Regression estimates predicting missingness as a function of covariates, by experimental group, when abnormal scores are trimmed out

	Control	Loser	Winner
Baseline	0.046 (0.169)	0.274. (0.140)	0.004 (0.166)
Age	0.006 (0.005)	0.001 (0.003)	0.011** (0.004)
Female	-0.100 (0.074)	0.002 (0.081)	0.029 (0.101)
Education level 7	0.002 (0.100)	-0.158. (0.091)	-0.031 (0.107)
Education level 8	0.013 (0.169)	-0.197* (0.094)	-0.169 (0.142)
Education level 9			-0.205 (0.370)
Education level 5	0.001 (0.103)	-0.016 (0.160)	-0.264. (0.155)
Education level 4		-0.060 (0.243)	-0.252* (0.102)
Education level 3	0.217 (0.261)	0.156 (0.193)	0.546** (0.184)
Education level 2			-0.214 (0.338)
Math confidence 2	-0.057 (0.080)	-0.112 (0.079)	-0.163 (0.099)
Math confidence 3	-0.199* (0.099)	0.137 (0.337)	-0.435*** (0.085)
Math confidence 4	-0.165. (0.099)	0.166 (0.220)	0.148 (0.288)
Math confidence 5	-0.204 (0.183)		-0.644*** (0.176)

Table 15: Heterogeneous effects after reweighting the estimates, when abnormal scores are trimmed out

	1st score	Gender	Age
Baseline	5.541***	31.868***	31.355***
	(1.045)	(0.984)	(3.062)
Losing Treatment	-0.933	-1.045	7.260.
	(1.551)	(1.564)	(3.970)
Winning Treatment	2.279	0.198	11.660**
	(1.721)	(1.454)	(3.835)
(1 point more in) 1st score	0.884***		
	(0.038)		
Losing:1st score	0.051		
	(0.055)		
Winning:1st score	-0.064		
	(0.059)		
Female		-3.517*	
		(1.710)	
Losing:Female		2.581	
		(2.490)	
Winning:Female		0.827	
		(2.553)	
Age (per year)			-0.023
			(0.083)
Losing:Age			-0.196.
			(0.103)
Winning:Age			-0.299**
			(0.100)
R^2	0.838	0.025	0.119
F	262.898	1.286	6.867
p	0.000	0.278	0.001
N	339	339	339
Attrition Rate	23.3%	23.3%	23.3%

4.5 Statistical Power

Not considering either the scores in the first test or any covariate (i.e., considering the results shown in Table 8), the effect size we have found, as measured by Cohen's d is as low as -0.139 for the Losers' group, and -0.080 for the Winners' group, so the statistical power is 0.10. For such small effects to be detected with a statistical power of 0.8, the sample size (without attrition) would have to be at least $2,975$.

If we discard abnormal values—using the same method—, the effect size is -0.002 for the Losers' group, and 0.081 for the Winners' group, yielding a statiscal power of 0.05. In this case, the sample size (without attrition) would have to be at least $7,330$ to achieve a statistical power of 0.8.

5 Conclusions and Discussion

Our study did not find any causal effect of significance. However, it is important to discuss this issue in the context of the sampling techniques and the operationalization of our experiment.

The goal of the experiment was to see if competition had an impact on people's ability to solve arithmetic problems. Realistic competition is an important aspect when trying to measure its effects on performance and we suspect that real competitive pressure might not have been present in our experiment. Because there is no risk or reward associated with the outcome of the game for individual participants, it is quite possible that they did not feel compelled to win or cared if they lost. We relied on subjects' curiosity, self-motivation and/or ego to play the game and feel competitive pressure. We did not have any incentives or rewards that would also motivate subjects to perform better or feel that they have something to lose or gain. There weren't any mechanisms to identify if people were taking the experiment seriously, if they understood the instructions, if they felt engaged and more importantly, if they felt any competitive pressure. Participants recruited through AMT were especially susceptible to being disengaged because their primary motivation is completing the task quickly to get paid and move on to the next HIT.

Despite our efforts, we suspect our experiment did not have high stakes to truly mimic a competitive environment. In the end, our experiment resembled a 'lab experiment' more than a 'field experiment'. The simulated setting limits our ability to identify an effect if such an effect were to exists. Intuitively, we believe that competition does have an impact on people's behaviour and performance, but the net effect is ambiguous as it will vary depending on the context and individual.

The effects of competition on performance is still a relevant issue in many fields, especially in business and education. Businesses pay employees generous bonuses based on performance and schools evaluate students based on relative performance. The latter is notably true of standardized tests which are widespread and are used to rank both students and schools.

An alternative experiment in a real educational setting for example, could have classrooms assigned to a competitive teaching style vs a non-competitive teaching style, and then see how both groups perform on the same test at the end of the year. However, such an experiment would raise significant ethical concerns. The experiment could be seen as unfair if one style is perceived to be significantly worse than the other, and parents would not want to have their kids enrolled in a class were they think that teaching standards are lower. There could also be side consequences on the students who are pressured to compete, including stress and anxiety. Putting ethical considerations aside, defining and implementing "competitive" and "noncompetitive" styles is not a trivial task. In any experiment of competition, it is essential to think through what make us feel competitive. Does competition arise as a desire for status, a want to outperform others and say we are the best? Or does competition arise as a desire to win a prize? While this questions are elusive, looking at this nuances would help design a better experiment to study the causal impact of competition in educational outcomes.

Some other critiques to our experiment include limited generalizability, reliance on self-reported characteristics, and the use of a single task to quantify performance.

In terms of generalizability of our experiment, although treatment assignment was done randomly, we relied exclusively on a convenience sample – AMT and Social Media – to recruit participants. Given our backgrounds

and associations, it is quite likely that the people we recruited through our social channels have better math skills and/or generally higher education than the general population. For these reasons, we don't believe the experiment as conducted is generalizable to the broader population.

Participant attributes like gender and education level are self-reported through an in-game survey. They can't be validated. So, even while we perform between-groups analysis (gender for example), and believe there is a balanced sample, we can't be 100% sure about it. Ideally, we would need a more reliable way to gather data about the participants.

Finally, we measured performance on a single task related to simple arithmetic problems. This is a very narrow operationalization of performance. This might have also inadvertently put people who do not enjoy arithmetic problems at a disadvantage. Future research should take into account a variety of subjects and create competition where a variety of skills are put to the test.

To conclude, while our research did not find any evidence that competition had an impact on performance, this is not a definitive answer and we would recommend more experiments to be done, preferably in more realistic settings.