**Measuring The Effect of Difficulty Labels on Problem Solving**

Qian (Cathy) Deng, Patti Degner, and Heather Heck

Some people like a challenge; others prefer to be told exactly what to do. Usually, the difficulty of a problem is something we can only truly determine on our own. However, we are often given some indication of a problem's difficulty beforehand: by the problem's reputation or source, by the opinion of others who worked on it previously, or (in the case of exams like the GMAT) because we expect the problems to get harder as we proceed.

The question that our experiment will answer is: How does the expectation of the difficulty of a problem change people's ability to solve it? Does the indication of difficulty breed discouragement if a subject is told the problem is hard? Does it cause carelessness if the subject is told the problem is easy? Conversely, does being told about difficulty inspire confidence if told the problem is easy, or extra determination if told it is hard?

The answer to this question would be useful for educational institutions, and by those in management. The large test prep industry can use this information to help individuals who wish to improve their performance on standardized tests. Students may be interested in learning if knowing the difficulty of a problem could change their ability to do well on standardized tests. The findings from this study could be applied in education more broadly: teachers could improve their students' accuracy rate on tests and homework by providing the difficulty. Similarly, many professions require problem solving abilities; companies may want to know if there is a benefit to revealing the true difficulty of a problem. If so, they may want to inform their employees of the difficulties beforehand to improve their employees' ability to solve the problems. If employees spend more time on difficult problems and increase accuracy by spending less time on easier tasks, the result could be a more efficient workplace.

Prior research notes that adaptive tests, tests that get harder as you go along, tend to improve students' learning outcomes (Heitmann). On the other hand, additional research suggests that the adaptive tests create anxiety when it comes to difficult questions (Ponsoda). Our experiment can help us understand whether the performance improvements from the difficulty-adaptive tests are due entirely to the adaptiveness of the tests, or if the psychological experience of having some idea of question difficulty can play a helpful role as well. In the case of adaptive tests, the problems are often not given a difficulty label directly, but students are aware that the test will become harder

as it goes. We plan to label the questions directly because we want to know if it is knowledge of the problem difficulty that has an effect on performance.

The experiment was delivered via online survey. Half of the survey respondents were randomly assigned to control and half were randomly assigned to treatment. Control consisted of seeing 15 critical thinking problems at once without difficulty listed. Treatment subjects were shown the same 15 questions at once with the difficulty rating included. The difficulty levels were as follows: Easy, Medium, Hard, and Very Hard. The questions are shown below. The choices offered are provided in parentheses, with the correct answer written in red.

(1) EASY - How many continents are there in the world? (5,6,7 or 8)
(2) EASY - What is regulation height for a basketball hoop? (10ft, 11ft, 12ft, 13ft)
(3) EASY - What is the sum of the angles of a triangle? (120 degrees, 160 degrees, 180 degrees, 200 degrees)
(4) MEDIUM - Who wrote the book Frankenstein? (Mary Shelley, Maurice Sendak, Edgar Allan Poe, Charles Dickens)
(5) MEDIUM - In Harry Potter, what does the Imperius Curse do? (tortures, kills, immobilizes, controls)
(6) MEDIUM - How many sides are there in a hexagon? (6, 8, 10, 12)
(7) MEDIUM - Select the correct way to write the asterisked part of the below sentence: Hospitals are increasing the hours of doctors, ***significantly affecting the frequency of surgical errors, which already are a cost to hospitals of*** millions of dollars in malpractice lawsuits.
(A) significantly affecting the frequency of surgical errors, which already are a cost to hospitals of
(B) significantly affecting the frequency of surgical errors, which already cost hospitals
(C) significantly affecting the frequency of surgical errors, already with hospital costs of
(D) significant in affecting the frequency of surgical errors, and already costs hospitals
(8) HARD - Who invented the rabies vaccination? (Louis Pasteur, Louis Cooper, Jonas Salk, John Robbins)
(9) HARD - One day, a person went to a horse racing area. Instead of counting the number of humans and horses, he counted 74 heads and 196 legs. How many humans and horses were there? (37 humans and 98 horses, 24 horses and 50 humans, 31 horses and 74 humans, 24 humans and 50 horses)
(10) HARD - What is the only US state that only touches one other state? (Florida, Michigan, Rhode Island, Maine)
(11) HARD -Which of these cars did James Bond not drive in any of the James Bond films? (Bentley, Toyota, Acura, Mercury)
(12) HARD - The average temperature of the last six days is 44 degrees. The median temperature is 36. If Tuesday was the warmest of the six days, what was the lowest possible temperature Tuesday could have had? (84, 64, 48, 60)

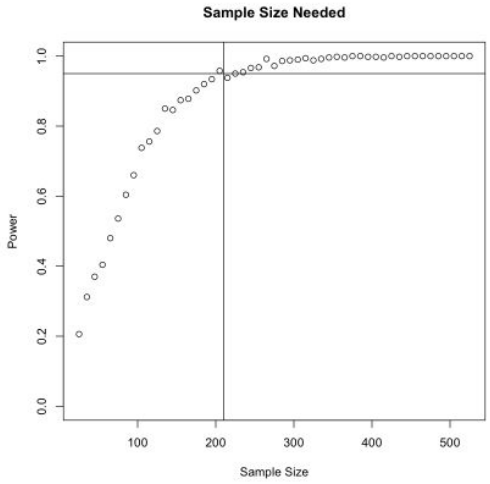
(13) VERY HARD - How many ways can the letters of the word PUZZLE be scrambled so that the first and the last letters are both vowels? (12, 24, 144, 720)

(14) VERY HARD - Which gas is formed when a hydrogen bomb is detonated? (Hydrogen, Helium, Methane, Uranium Dioxide)

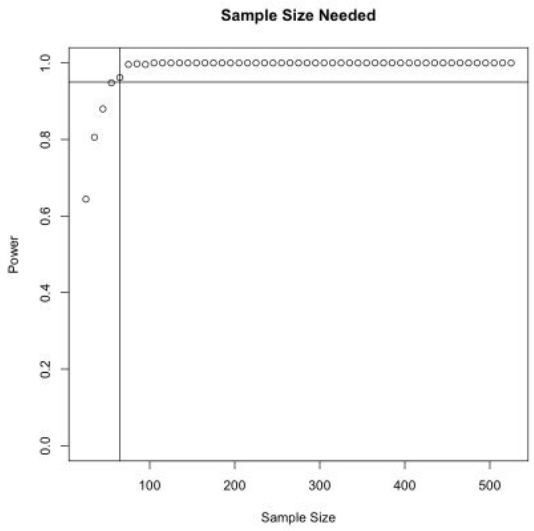
(15) VERY HARD - Rephrase this sentence so it has the same meaning: "If you don't keep calm, I shall shoot you," he said to her in a calm voice. (He warned her to shoot if she didn't keep quiet calmly, He warned her calmly that he would shoot her if she didn't keep quiet, He said calmly that I shall shoot you if you don't be quiet, Calmly he warned her that be quiet or else he will have to shoot her)

To deter cheating, we included two honesty questions. At the beginning of the survey, subjects were asked to certify that they would not search the internet for answers. At the end of the survey, they were asked if they searched the internet for answers. In addition to the honesty questions, the subjects were asked a series of demographics questions regarding: gender, years of schooling, household income, location, employment status, and age. We also asked how difficult they found the survey and how stressed they feel compared to the previous 6 months. These types of questions will allow us to use blocking during our analysis if needed.

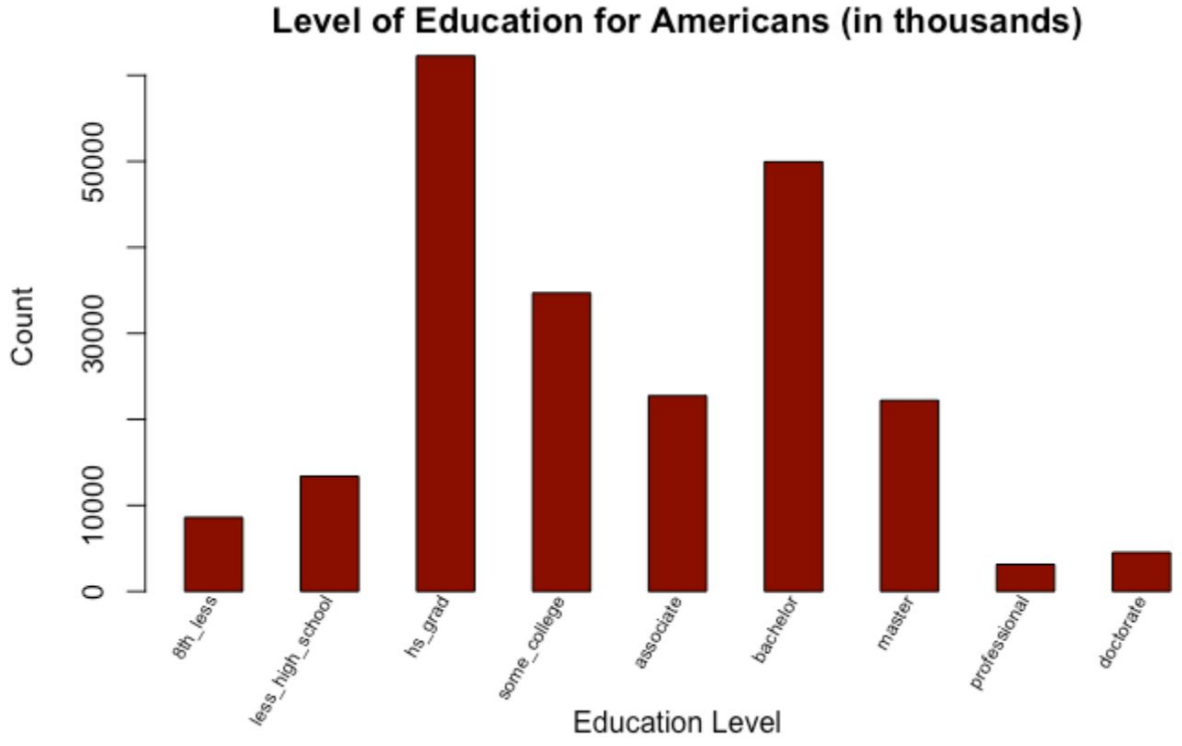
We precalculated how many observations we would need to achieve a significant result. We made the assumption that the average respondent would answer 10/15 correct without difficulty labels. We estimated a standard deviation of 2 questions or 2/15. We estimated the impact (tau) of the difficulty ratings as 1 question or 1/15. In this scenario we would need approximately 215 respondents.

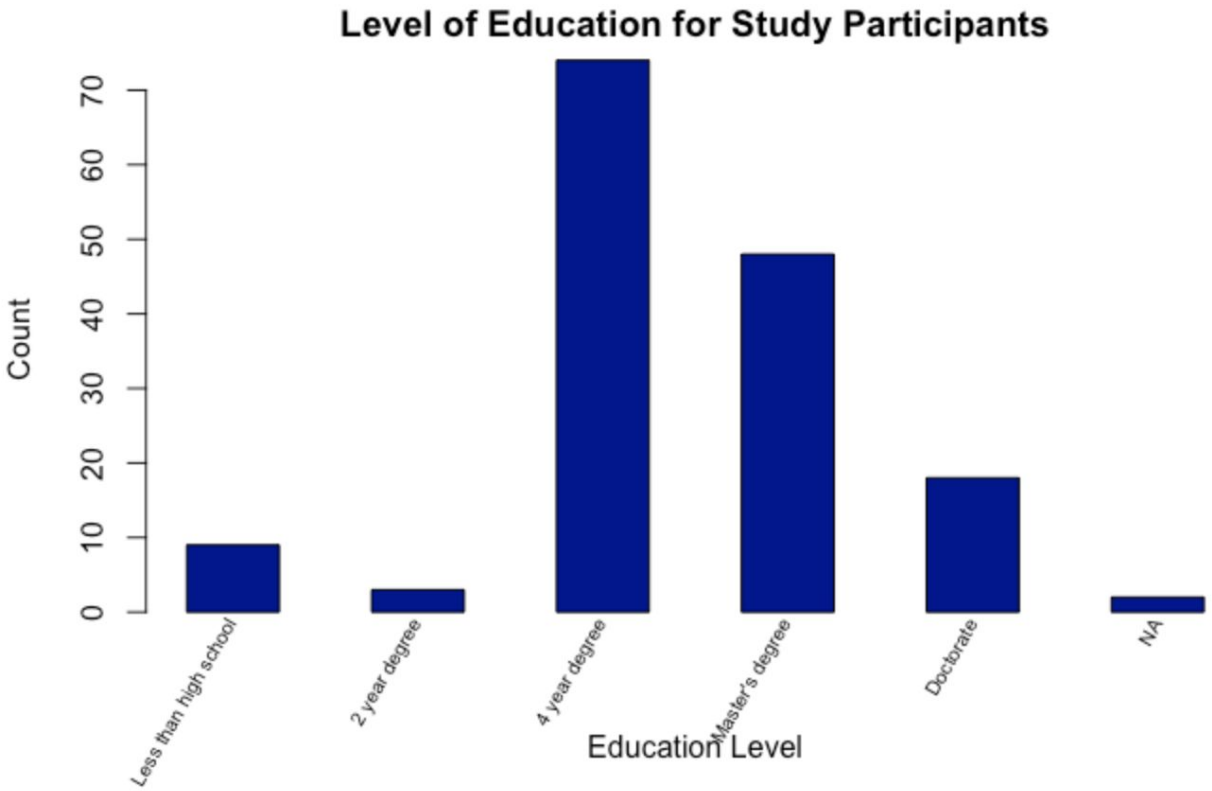


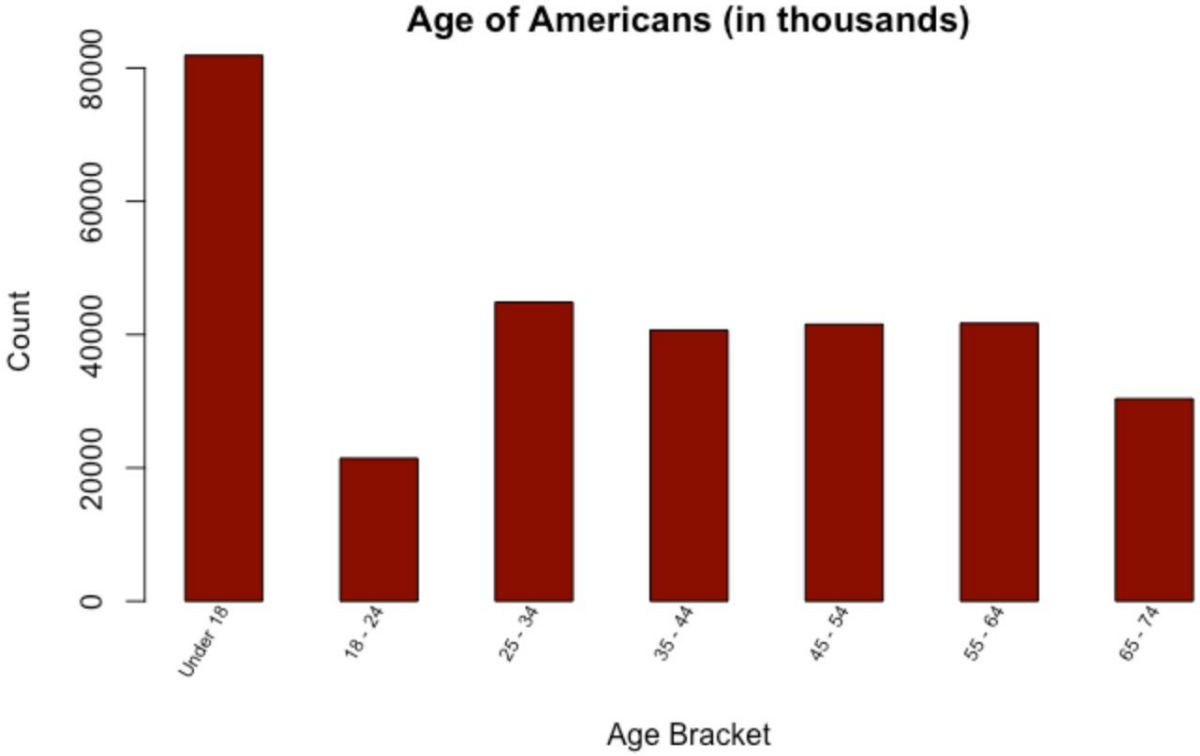

If we estimated the impact (tau) of the difficulty ratings as 2 questions or 2/15, we would need approximately 65 respondents.

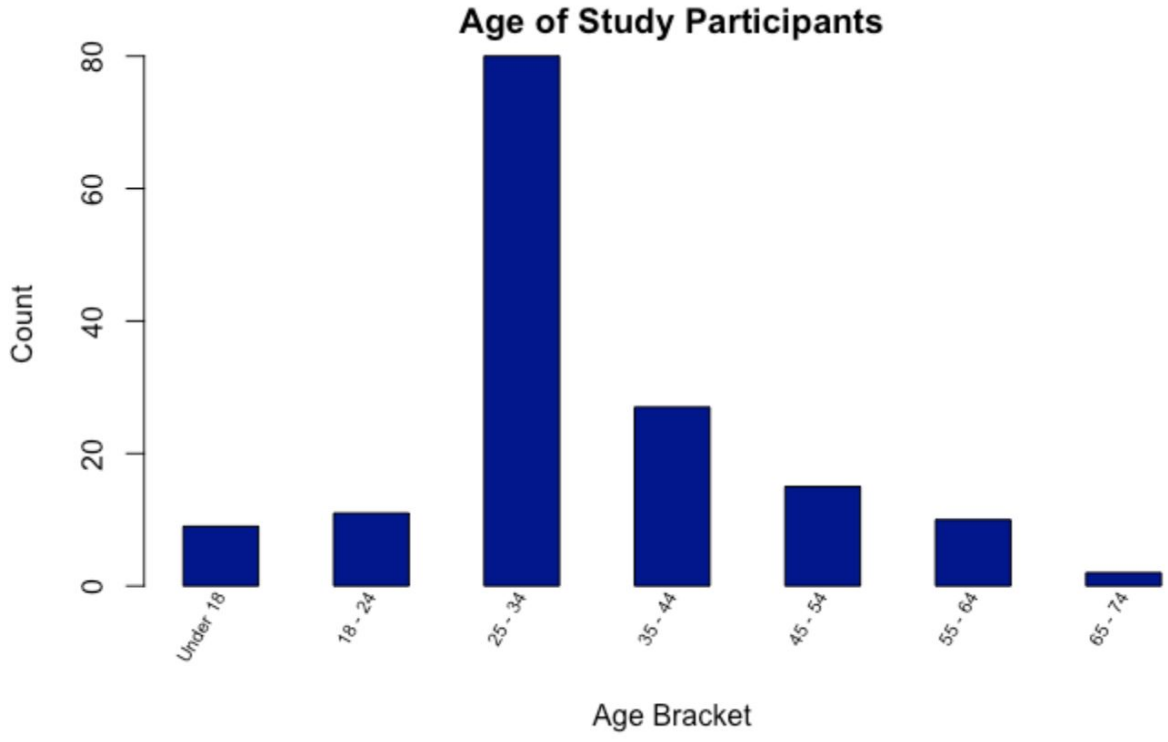



We found subjects by sending survey links to classmates from MIDS, friends, family and coworkers. This pool is not representative of the overall population. Our subjects are on average more educated, younger, and regionally clustered than the general American population. The graphs below show our population distribution compared to the USA distributions. Therefore, the scope of our conclusions should be limited to only those who are similar to our test subjects.

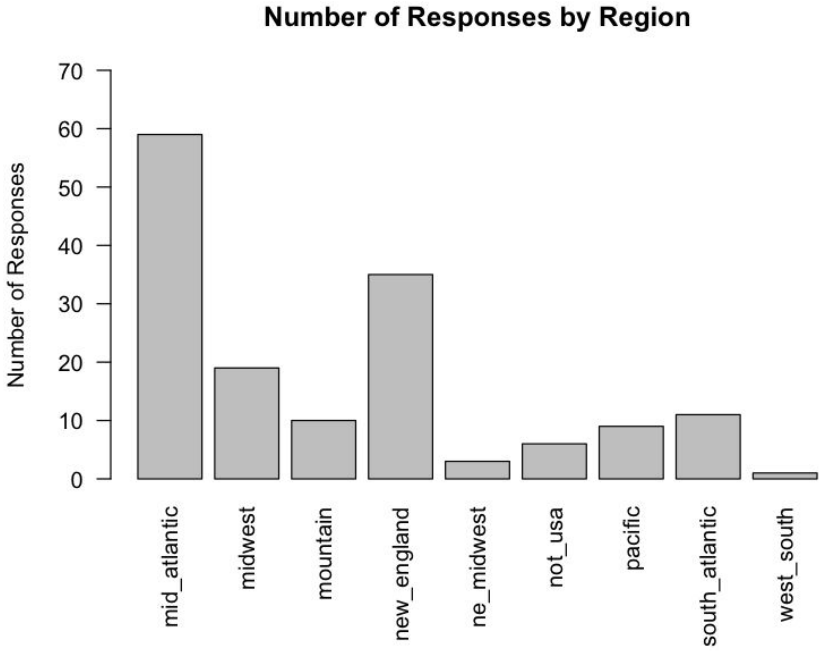

Level of Education for Americans (in thousands)
Count
0
10000
30000
50000
8th_less
less_high_school
hs_grad
some_college
associate
bachelor
master
professional
doctorate
Education Level


Level of Education for Study Participants
Count
0
10
20
30
40
50
60
70
Less than high school
2 year degree
4 year degree
Master's degree
Doctorate
NA
Education Level


Age of Americans (in thousands)
Count
0
20000
40000
60000
80000
Under 18
18 - 24
25 - 34
35 - 44
45 - 54
55 - 64
65 - 74
Age Bracket


Age of Study Participants
Count
0
20
40
60
80
Under 18
18 - 24
25 - 34
35 - 44
45 - 54
55 - 64
65 - 74
Age Bracket

In exploring our data, we saw that we had a large number of responses from the regions the three of us were from. The distribution is below. For modeling purposes, we only used Mid Atlantic and New England.




Nonetheless, our results may be of practical interest.

In order to prevent bias in the delivery of the survey, we came up with a consistent backstory to use when sending out the survey, suggesting to recipients that we were measuring the effects of stress, not difficulty labels. The text of the email template is provided below:

> *Family, Friends and Colleagues:*
>
> *As most of you know, I am currently pursuing my Masters degree. As part of our coursework this term, we are studying experiments and causality. As part of our final project, we have been asked to analyze survey data. We are looking to measure whether stress impacts your ability to solve problems. I would be so grateful if you could take 5 minutes out of your day to take the following survey: https://berkeley.qualtrics.com/jfe/form/SV_0dhrssOwS3bMQw5*
>
> *Please note, all replies to every question (including demographics questions) will be anonymous; we have no ability to link any replies to specific individuals. Please feel free to skip any questions you do not feel comfortable answering.*

*Hope everyone is doing well and staying safe and sane!*

*Thank you in advance,*
*Heather, Cathy, and Patti*

Our pilot study helped us fine tune our survey before sending out to a broader audience. In our pilot study, we received feedback that subjects wanted to know what their "score" was. We had purposefully not included a survey key to avoid any spillover effect. Because we were sending this to family and coworkers that are in close proximity to each other, we did not want people sharing answers. In our final survey, we included an answer key at the end. This seems to successfully appease our test subjects, and we hoped our honesty questions would deter cheating. Ultimately, we did not see a prevalence of extremely high scores that would suggest cheating; participants who forwarded the survey to others often wanted to see if they could score higher than their friends, and were thus unlikely to share answers.

In brainstorming how to deter cheating, we originally planned to use a timer and limit the amount of time each person had to complete the survey. Unfortunately, this timer was only administered to control during the pilot, and we were not able to overcome this technical difficulty. We removed the timer altogether in our final survey, meaning we were not able to use completion time as a control variable in our final model.

The primary outcome variable is the percentage of correctly answered questions. Each question was represented with a 1 if a question was answered accurately and a 0 otherwise. Thus, the possible outcomes are 0/15, 1/15, 2/15, … 15/15. Secondary outcome variables were how each subject did on each subset of questions grouped by their difficulty rating. As an example, there were 3 questions marked as "easy". The possible outcomes for the subset of easy questions are 0/3, 1/3, 2/3, or 3/3.

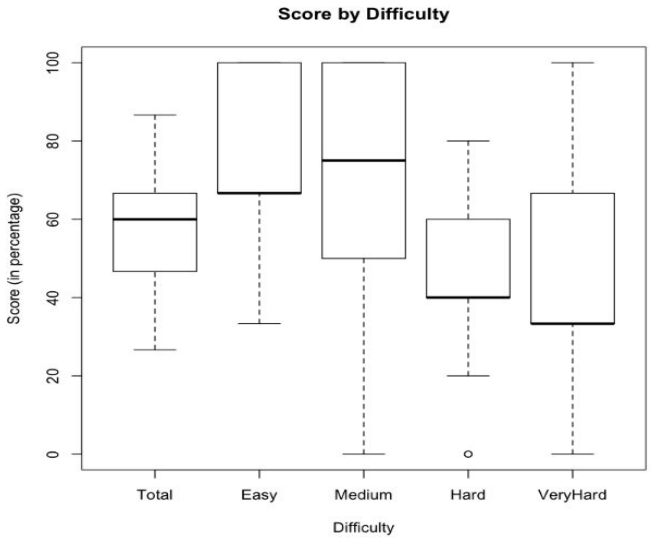
Our initial hypothesis is that providing the difficulty labels will increase the score overall. We think that it will increase confidence on easier problems and allow the survey takers to think more carefully about the harder problems. It will take the guesswork out of prioritizing which questions to spend more time on. The null hypothesis would be that the difficulty labels had no effect on the participants ability to answer the questions accurately. We also wanted to pay special attention to age and gender. It may be that younger people are more impressionable than older people when it comes to believing what they are told, or, conversely, more likely to distrust what they are told. Women may be influenced by a culture that has historically

discouraged them from tackling intellectually challenging problems. All of these factors could result in different effect sizes, and have implications as to how to implement learnings from the experiment going forward, so we also want to pay special attention to heterogeneous treatment effects in these cases.

When we began exploring our data, we first removed all responses that were recorded before the removal of the faulty time limit; these were from the pilot study. We also asked two honesty questions, one at the beginning and one at the end. If someone replied at the end that they cheated, we removed the response. Luckily, this was only three people. Lastly, we removed anyone who did not make it to the end of the survey (i.e. they did not answer any demographics questions). The data is cleaned in the "data_cleaning.ipynb" file, resulting in 154 valid responses.

To assess the effect of both our direct treatment and the covariates we are interested in, we performed a regression analysis, along with hypothesis tests using robust standard errors (especially if error distribution is not normal) to check the statistical significance of coefficients in the regression.

We explored the scores by difficulty. The median score on the very hard problems was the lowest, followed by the hard problems. We saw that the median score on the medium problems was higher than the median score on the easy problems, but the distribution of the medium scores was much wider. We were comfortable that the question difficulties were well calibrated.




We started our modeling with very simple regressions. We looked at the difference in each score type by treatment and control. We used robust standard

errors. The below figure shows one model for each score: starting from total score, and then breaking it down into: easy score, medium score, hard score, and very hard score. “Grouptreatment” refers to whether the respondents were in the treatment group. Each “Pct” variable refers to the percent score on the full test or a particular difficulty category.

Naive Models

| | Dependent variable: | | | | |
|---|---|---|---|---|---|
| | Total_Pct (1) | Easy_Pct (2) | Medium_Pct (3) | Hard_Pct (4) | Very_Hard_Pct (5) |
| grouptreatment | 4.631** | 4.759 | 2.717 | 7.775** | 1.815 |
| | (2.352) | (4.193) | (3.866) | (3.110) | (5.480) |
| Constant | 56.340*** | 76.471*** | 70.098*** | 39.216*** | 46.405*** |
| | (1.917) | (3.687) | (3.164) | (2.527) | (4.430) |
| Observations | 154 | 154 | 154 | 154 | 154 |
| R2 | 0.025 | 0.010 | 0.003 | 0.039 | 0.001 |
| Adjusted R2 | 0.019 | 0.004 | -0.003 | 0.033 | -0.006 |
| Residual Std. Error (df = 152) | 13.695 | 22.283 | 22.421 | 18.161 | 32.174 |
| F Statistic (df = 1; 152) | 3.900* | 1.556 | 0.501 | 6.251** | 0.109 |
| Note: | | | | | *p<0.1; **p<0.05; ***p<0.01 |

From the naive models, we observe that the treatment effect is significant at the $p<0.05$ level for both the total score and for the score on hard questions. The coefficients for the treatment are positive for all other labels, but none are statistically significant. One interpretation might be that the overall score improves because the labels help people prioritize questions in relation to each other. The effects may be diminished at the “easy” and “medium” range because while these labels can reduce overthinking, they might also invite carelessness. Likewise, labels of “very hard” could encourage better focus, but might also create extra anxiety. The “hard” label plausibly creates the biggest effect in helping people reach their potential, if this label can inspire extra focus and motivation to rise to the challenge, without inordinate anxiety; for a “very hard” problem beyond the participant’s abilities, motivation alone may be insufficient.

From here, we chose to dive deeper into the total score by introducing covariates. (Although the specific effect on hard questions is notable, the larger quantity of questions (15) for the total score invites less variance than would be present in any more saturated models for the 4 hard questions.) As outlined earlier, we also sought to understand whether there would be heterogeneous treatment effects for women or for younger participants. Hence, we built an additional model containing these covariates and associated interactions. Finally, we tested the robustness of our treatment effect

with a nearly saturated model using all covariates with more than 30 observations, within the data collected about region, level of stress, income, and education.

All Models

| | Dependent variable: | | |
|---|---|---|---|
| | Total_Pct | | |
| | (1) | (2) | (3) |
| grouptreatment | 4.631** | 10.576*** | 10.798*** |
| | (2.352) | (3.551) | (3.866) |
| Female | | -4.173 | -4.643 |
| | | (4.182) | (4.579) |
| age_25_34 | | 3.284 | 2.776 |
| | | (4.165) | (4.581) |
| mid_atlantic | | | -0.976 |
| | | | (2.955) |
| new_england | | | -4.658 |
| | | | (3.232) |
| employeed_FT | | | -4.344 |
| | | | (2.772) |
| income_200 | | | -1.014 |
| | | | (3.069) |
| stress_little_more | | | -0.607 |
| | | | (3.172) |
| stress_alot_more | | | -0.901 |
| | | | (3.440) |
| master | | | -0.436 |
| | | | (2.431) |
| grouptreatment:Female | | -5.150 | -5.614 |
| | | (4.916) | (5.309) |
| grouptreatment:age_25_34 | | -5.424 | -4.530 |
| | | (4.897) | (5.194) |
| Constant | 56.340*** | 56.565*** | 62.591*** |
| | (1.917) | (2.880) | (4.585) |
| Observations | 154 | 154 | 154 |
| R2 | 0.025 | 0.113 | 0.146 |
| Adjusted R2 | 0.019 | 0.083 | 0.074 |
| Residual Std. Error | 13.695 (df = 152) | 13.241 (df = 148) | 13.304 (df = 141) |
| F Statistic | 3.900* (df = 1; 152) | 3.758*** (df = 5; 148) | 2.017** (df = 12; 141) |
| Note: | | | *p<0.1; **p<0.05; ***p<0.01 |

The naive model indicates a statistically significant effect (at 95% confidence) from the treatment, where those in the treatment group scored 4.6pts higher (or just under 1 additional question out of 15). When we introduced covariates for age and gender, as well as interaction terms between these covariates and our treatment, we did not observe any statistically significant effects from these additions in the second model. Instead, only the treatment was statistically significant, now at the 99% confidence level.

Similarly, the fully saturated model did not reveal any additional statistically significant coefficients. Some of the variables, including region and gender, have coefficients that may be practically significant if valid, but the sample size is likely too small to provide enough power for us to rigorously test these effects in this particular experiment.

Overall, adding more covariates to the model more than doubled the effect size attributed to the treatment from 4.6pts to about 10.5pts out of 100 (about 1.5 questions out of 15), and decreased the level of uncertainty in the treatment effect.

To confirm these conclusions, we compared our naive model against the two other models using F-tests. In a comparison with the naive model, the test showed that Model 2 (introducing age, gender, and interactions with the treatment) does explain more variance.

**F-test Results, Naive Model vs. Model 2**

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 152 | 28508.58 | NA | NA | NA | NA |
| 2 | 148 | 25946.35 | 4 | 2562.228 | 3.653787 | 0.007197143 |

On the other hand, the addition of the remaining covariates to create Model 3 did not produce a significantly different result from Model 2, as evidenced in the F-test results below.

**F-test Results, Model 2 vs. Model 3**

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | 148 | 25946.35 | NA | NA | NA | NA |
| 2 | 141 | 24956.71 | 7 | 989.6423 | 0.7987522 | 0.5895948 |

In summary, our experiment showed a statistically and practically significant effect of showing difficulty labels to test takers. Additionally, it showed a significant effect of informing test takers of the difficulty of “hard” questions in particular.

These conclusions can help explain behaviors in certain real world situations. For instance, it is possible that some high scorers on the GMAT attain their success by evaluating the difficulty of a question prior to attempting to solve it. This expectation changes their approach to the question (e.g. the amount of “second-guessing”) and ultimately improves accuracy.

In addition, we can apply these conclusions to improve performance in various contexts. A manager in a business can motivate employees to perform a task at a higher standard by indicating that it is a hard problem. Indeed, the significance of the treatment effect on the total score in the survey suggests that comprehensive conversations about difficulty can be useful across the full range of difficulty, again because it could help individuals assess the relative priority of tasks, taking into account capacity demands and level of complexity. Based on the results, we might also hypothesize that knowing “hard” questions are hard can be motivating in itself. However, this effect seems to diminish at the highest level of difficulty, “very hard” - practically speaking, the motivation can push people to strive beyond a baseline level, but not exceed the true limits of their ability and time.

Future enhancements to our experiment could include some of the following options. We would be interested in making questions open-ended. Multiple choice questions allow survey respondents to use the process of elimination. They may eliminate answer choices that seem ‘too easy’ if a question was labeled hard. We could eliminate some of that risk by switchen to open-ended questions. If given more time, we could also better calibrate the difficulty of our questions. If we ran a robust pilot, we could separate questions that were only answered correctly by those who did well on the survey, and use those questions as the hard questions. Alternatively, we could eliminate questions that almost everyone answered correctly. If we had more resources and time, we would have gathered a larger sample of respondents and a more diverse sample of respondents. We could add a covariate that measured survey takers' background. This covariate could focus on their area of work, their major in college, or some set of questions that could isolate what they were a subject matter expert in, to account for variations caused by people’s greater ability to answer questions that they have expertise in.

It would be interesting to break this experiment into many sub-experiments. We could have one experiment that was only math questions, one experiment that was only

grammar questions, etc., to see if the effect is more significant in some fields than others. We could also go the other route and provide a survey with questions that tested more diverse subject matters: the more diverse the questions, the more likely it is that respondents would have to reach beyond their existing domain expertise.

One final possible enhancement would be a longer survey with more questions, which would allow us to measure the impact of difficulty labels on a more granular scale. With the 15 questions we used, each additional question answered correctly added 6.7% to the total percentage score. If we had 30 questions, it would cut that increase in half.

**Bibliography:**


Heitmann, Svenja, et al. "Testing Is More Desirable When It Is Adaptive and Still Desirable When Compared to Note-Taking." *Frontiers in Psychology*, Frontiers Media S.A., 18 Dec. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6305602/.

Ponsoda, Vicente, et al. "The Effects of Test Difficulty Manipulation in Computerized Adaptive Testing and Self-Adapted Testing." *Applied Measurement in Education*, vol. 12, no. 2, 1999, pp. 167–184., doi:10.1207/s15324818ame1202_4.

US Census Bureau. "Age and Sex Composition in the United States: 2018." *The United States Census Bureau*, 11 July 2019, www.census.gov/content/census/en/data/tables/2018/demo/age-and-sex/2018-age-sex-composition.html.

US Census Bureau. "Educational Attainment in the United States: 2019." *The United States Census Bureau*, 30 Mar. 2020, www.census.gov/data/tables/2019/demo/educational-attainment/cps-detailed-tables.html.