Modeling Student Response Times as a Function of Question Difficulty, Student Ability, and Student Resiliency

Sophie Sommer (NYU)

Research goals

Investigate the degree to which variance in response times (for online quiz questions) can be explained by question difficulty and student ability. Then, see if other student-level attributes (specifically, “resilience/perstistence”) can improve this model. Previous research (and probably your intuition) would suggest:

a. Harder questions take longer, on average
b. For easier questions, there is a positive association between ability and response time; For harder questions, there (may be) a negative association between ability and response time

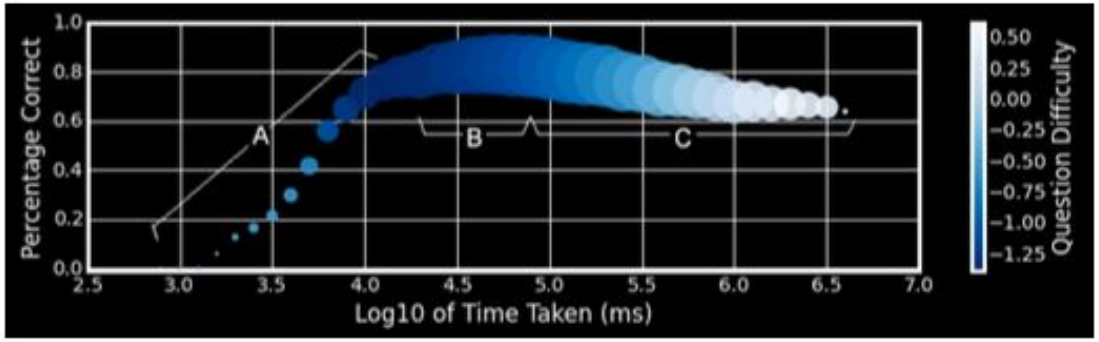


Figure 2: The relationship between the time taken (log scale) to respond to an item, and correctness. Color is used to denote item difficulty.

Note: this figure is from a similar research paper by Yijun Ma, et. al., who investigated this in the context of determining cut-off points in response time (response times below some cutoff might indicate disengagement, while response times above some cut-off might indicate that a student is struggling)

Ma Y, Baker R, Agnihotri L, Plaza P, Mojarad S (2016) Effect of student ability and question difficulty on duration. In: Proceedings of the 9th international conference on educational data mining, pp 135–142

Potential uses for this research

- Teachers, online course developers, assessment developers, etc. might find this relevant because a) it provides a framework for predicting how long students will spend on an assignment or quiz question and b) it is helpful to understand how much variability in response times can be explained by ability, question difficulty, and other measured traits (and how much cannot be explained)

- With the increasing popularity of online learning systems, there is interest in using student behavior to identify when intervention might be necessary (i.e., if a student is struggling, an online system could interject with a hint; if a student is rushing through questions, an online system could lock them out of submitting a response for some amount of time). Response times could also be used to identify cheating if expected response times are well understood.

Data source

- A massive open online course (MOOC) called Money in Business 1. I have a second run of the course that I am still in the process of analyzing (but preliminarily, seems to give similar results)
- Students respond to quiz questions that are multiple choice. Some are single-answer questions, some are multiple-answer questions. Response times seem generally similar across questions, so I pooled them
- Students have unlimited time to complete each question and can make as many attempts as they would like
- Dataset includes 2387 students responding to 30 questions (36861 rows)
- Omitted: more than 4 attempts (rare), attempt times greater than 2400 seconds (40 mins) (also rare, but some extreme outliers were possibly related to the mechanism by which I had to recover attempt times)

Raw data

Enrollments file

learner_id	enrolled_at	unenrolled_at	role	fully_participa	purchased_s	gender	country	age_range	highest_education_level	employment_status	employment_area
f7c3d94d-73	2015-05-07 11:33:18 UTC		learner			Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
bb284b11-d:	2015-05-08 09:08:05 UTC		organisation_admin			Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
20e6ec35-0f	2015-05-21 12:26:06 UTC		admin			Unknown	Unknown	Unknown	Unknown	Unknown	Unknown

Question response file

learner_id	quiz_question	week_number	step_number	question_number	response	submitted_at	correct
eb4a74e3-da	2.11.1	2	11	1	2	2015-07-03 13:(	FALSE
6e9bbed1-75	1.12.1	1	12	1	6	2015-07-03 16:(	FALSE
6e9bbed1-75	1.12.2	1	12	2	3	2015-07-03 16:(	FALSE
6e9bbed1-75	1.12.2	1	12	2	2	2015-07-03 16:(	TRUE
6e9bbed1-75	1.12.3	1	12	3	2	2015-07-03 16:(	FALSE

Step activity file

learner_id	step	week_number	step_number	first_visited_at	last_completed_at
4a5f9fcf-fef3	1.1	1	1	2015-05-05 12:46:55 UTC	
ef274229-03	1.1	1	1	2015-05-19 10:06:22 UTC	2015-07-08 15:17:44 UTC
6e9bbed1-75	1.1	1	1	2015-06-08 09:19:50 UTC	2015-10-09 09:55:19 UTC

Key variables (in the final, cleaned dataset)

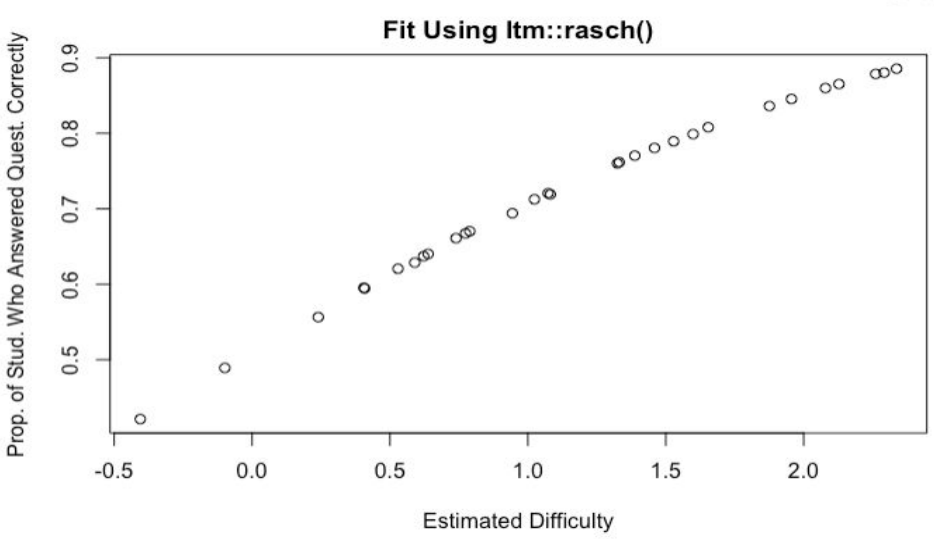
- **Att1_Time**: Time spent on first attempt of the question (mean ~81 seconds/ 4.4 in log units)
- **RaschAbility**: Student ability as estimated by a Rasch model (fit using a Bayesian approach -- see next slide). Ability is estimated using only correct/incorrect on first attempt
- **RaschDiff**: Difficulty of the question as estimated by a Rasch model (again, fit using a Bayesian approach; again, based on correct/incorrect on first attempt)
- **PropUsedOpp**: Proportion of opportunities to retry a question that the student used
  - If a student made 4 attempts and never got the question right, then they had 4 chances to retry the question after an incorrect attempt and they used ¾ (because they gave up on the last one)
  - If a student made 4 attempts and got the question right on the 4th attempt, then they had 3 chances to retry the question after a wrong answer and used all 3
  - If a student got all questions right on the first try, they had no attempts to retry a question after getting it wrong, and therefore they are coded as NAs.

Why a Bayesian approach?

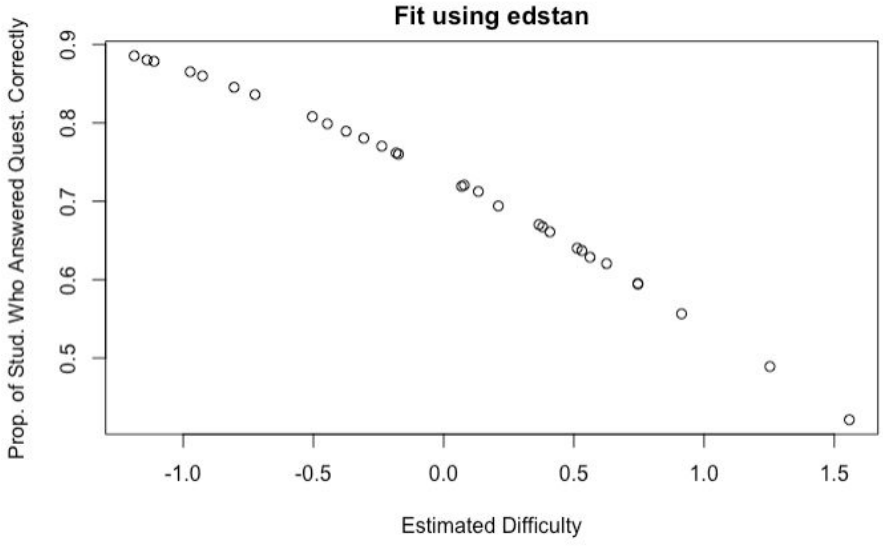
TLDR: The model fit using marginal maximum likelihood (in ltm) gave me issues. (mirt as well)

Note: If you think you know why this is happening, please let me know!

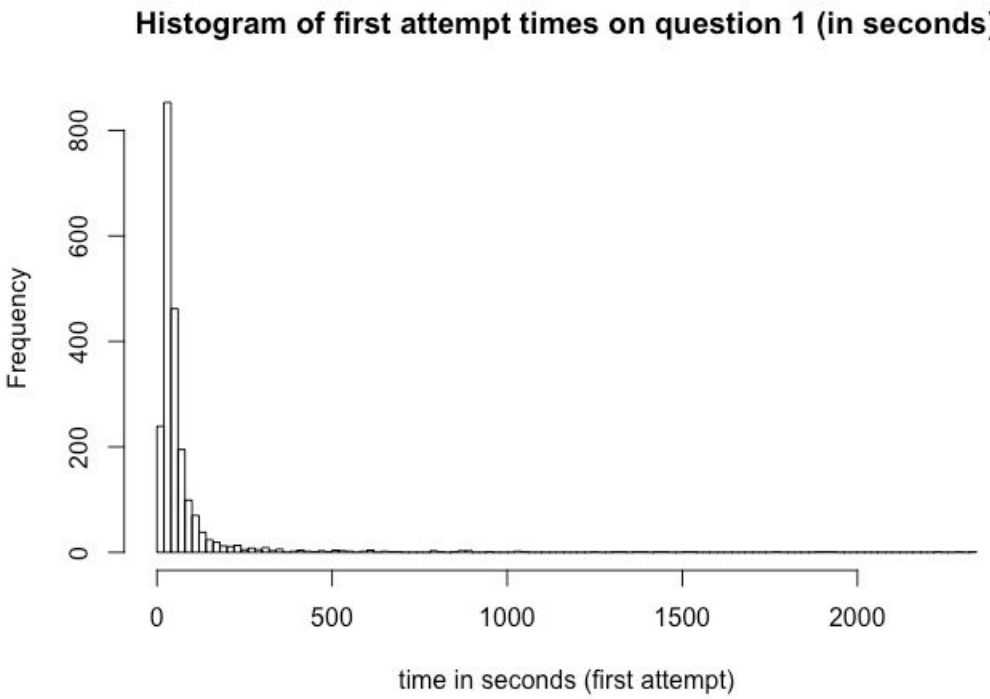
When I fit the model using ltm::gpcm() with constraint= “rasch”, estimates for all beta and theta parameters were negative. When I fit using ltm::rasch() with no constraint (i.e., a 1 Parameter Logistic Model), the betas (i.e., difficulty estimates) were opposite from my

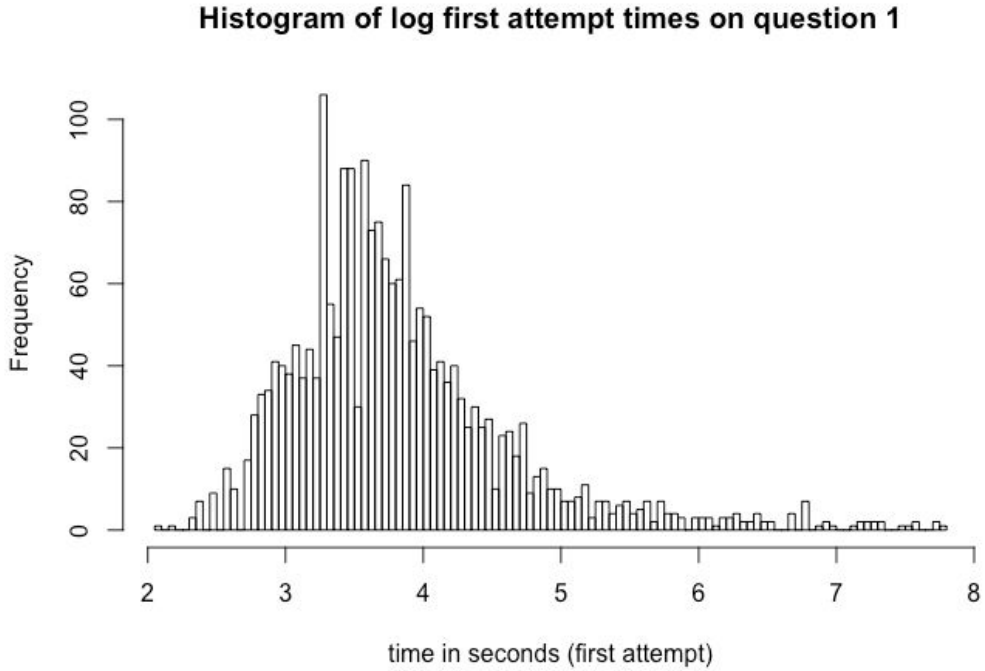


ed this

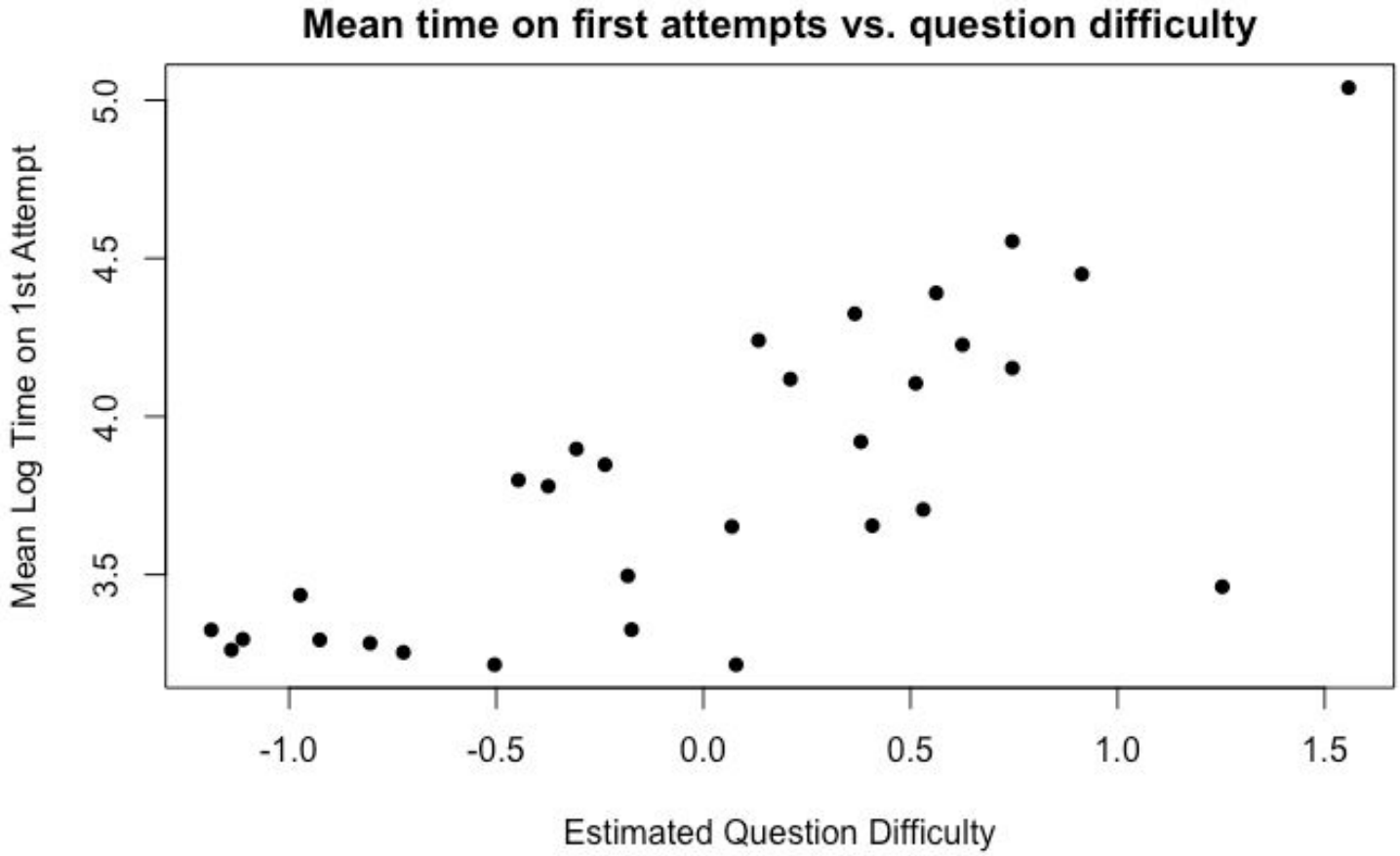


Choice to model log (base e) response times

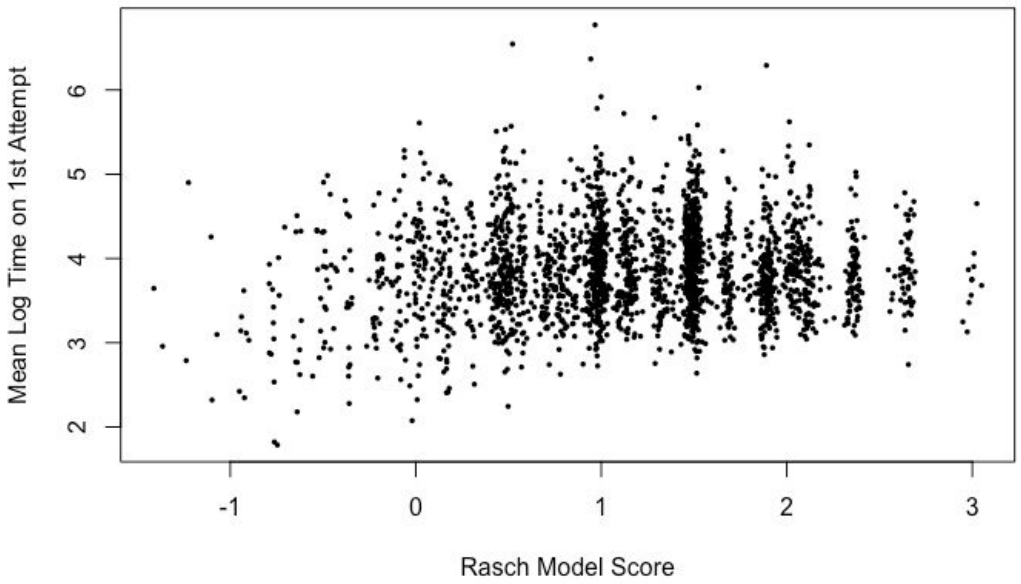


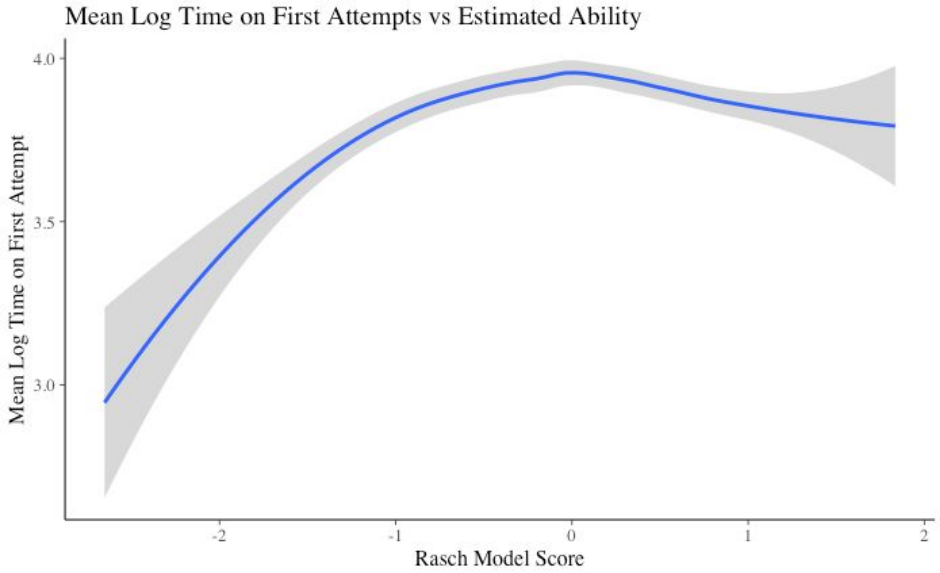


Relationship between first attempt response times and question difficulty



Relationship between student ability and first attempt response time





Baseline model (first attempt time)

```
## Call:
## lm(formula = log(Att1_Time) ~ (RaschAbility + RaschAbilitySq) *
##     RaschDiff, data = finaldata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8959 -0.5717 -0.0806  0.4873  4.4185
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.650343   0.008272  441.30   <2e-16 ***
## RaschAbility               0.298709   0.012157   24.57   <2e-16 ***
## RaschAbilitySq            -0.096250   0.005118  -18.81   <2e-16 ***
## RaschDiff                  0.353912   0.011178   31.66   <2e-16 ***
## RaschAbility:RaschDiff     0.268715   0.016535   16.25   <2e-16 ***
## RaschAbilitySq:RaschDiff -0.074993   0.006963  -10.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8748 on 36855 degrees of freedom
## Multiple R-squared:  0.1852, Adjusted R-squared:  0.1851
## F-statistic:  1675 on 5 and 36855 DF,  p-value: < 2.2e-16
```

Re-run without missing data for PropUsedOpp:

```
## Call:
## lm(formula = log(Att1_Time) ~ (RaschAbility + RaschAbilitySq) *
##     RaschDiff, data = finaldata2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8956 -0.5748 -0.0801  0.4888  4.4274
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.650104   0.008313 439.100   <2e-16 ***
## RaschAbility               0.300549   0.012506  24.032   <2e-16 ***
## RaschAbilitySq            -0.096928   0.005516 -17.572   <2e-16 ***
## RaschDiff                  0.354219   0.011234  31.532   <2e-16 ***
## RaschAbility:RaschDiff     0.263256   0.017015  15.472   <2e-16 ***
## RaschAbilitySq:RaschDiff -0.072547   0.007520  -9.647   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.878 on 35587 degrees of freedom
## Multiple R-squared:  0.1821, Adjusted R-squared:  0.1819
## F-statistic:  1584 on 5 and 35587 DF,  p-value: < 2.2e-16
```

Model with resiliency (first attempt time)

```
## Call:
## lm(formula = log(Att1_Time) ~ (RaschAbility + RaschAbilitySq) *
##     RaschDiff + PropUsedOpp, data = finaldata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9028 -0.5758 -0.0809  0.4891  4.4163
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.502324   0.031581 110.901  < 2e-16 ***
## RaschAbility              0.291742   0.012633  23.093  < 2e-16 ***
## RaschAbilitySq           -0.095186   0.005526 -17.225  < 2e-16 ***
## RaschDiff                 0.353976   0.011230  31.520  < 2e-16 ***
## PropUsedOpp               0.162703   0.033544   4.850 1.24e-06 ***
## RaschAbility:RaschDiff    0.263400   0.017010  15.485  < 2e-16 ***
## RaschAbilitySq:RaschDiff -0.072536   0.007518  -9.649  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8777 on 35586 degrees of freedom
##   (1268 observations deleted due to missingness)
## Multiple R-squared:  0.1826, Adjusted R-squared:  0.1825
## F-statistic:  1325 on 6 and 35586 DF,  p-value: < 2.2e-16
```

Note: this measure of resiliency is a significant, positive predictor of log response time

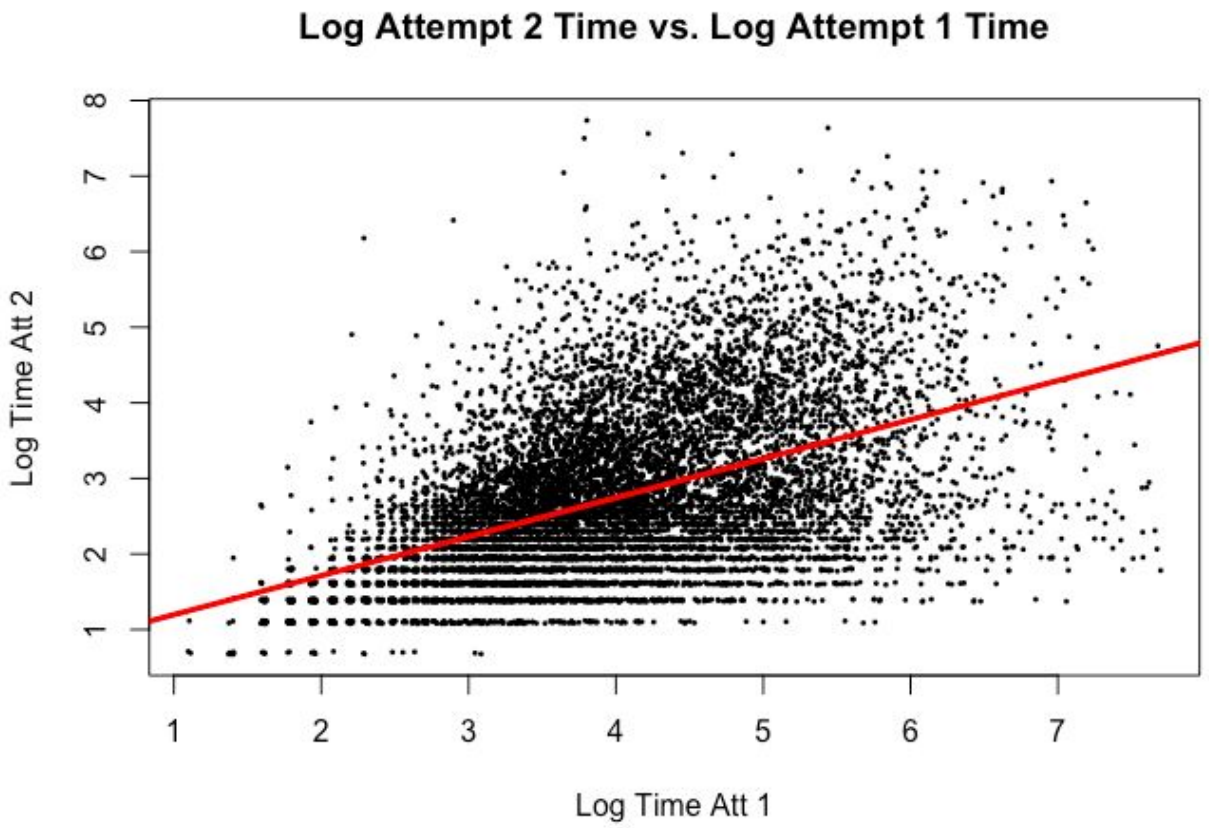
```
## Analysis of Variance Table
##
## Model 1: log(Att1_Time) ~ (RaschAbility + RaschAbilitySq) * RaschDiff
## Model 2: log(Att1_Time) ~ (RaschAbility + RaschAbilitySq) * RaschDiff +
##     PropUsedOpp
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  35587 27432
## 2  35586 27414  1    18.124 23.526 1.237e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Other findings, possibly of interest for future work

The following table shows mean log response times for 1st through 4th attempts. As expected, students tend to spend less time on every subsequent attempt:

1st Att	2nd Att	3rd Att	4th Att
4.4	3.58	2.96	2.43

Students who spend comparatively more time on a first attempt also spend comparatively more time on second attempts



Future interests

- Is this replicable with the second Money in Business course run?
- Is this replicable in other MOOC courses in different disciplines?
- Are there other ways to measure resilience/other student level characteristics that would also improve the model?
- Is this replicable when students are under time pressure? Maybe resilient/persistent students tend to spend more time on questions when they are given unlimited time but can respond faster if necessary.

Questions?

Please also feel free to reach out with any additional questions, thoughts, or ideas!

(email: sjs908@nyu.edu)