# Analysis on Addition Data

2024-05-07

This document is intended for the analysis of individual differences in the effect of sequential errors on quitting, in the addition game within Prowise Learn. It is accompanied by other scripts needed to properly analyse the data:

- <load_addition_data.R> to load in the raw addition data.
- <addition_data_prep.Rmd> to clean the raw data and compute the necessary variables.
- <dependencies.R> to load the relevant libraries, plotting themes, and plot colors.
- <fit_mm.R> and <fit_mm_test.R> to run the simple Markov models on the training and testing data.
- <clean_add_forLMER.R> to apply exclusion criteria for the mixed-effect logistic regression model.
- <fit_lmer.R> and to fit the mixed-effect logistic regression model to the training and testing data.

When compiling, plots are saved, in LaTeX-friendly format, to the folder . This document also compiles into a pdf document: ind_diff_addition.pdf.

First, I load in the data that was prepped in the script <addition_data_prep.Rmd>. Here, specify "training" or "testing" to load in the desired dataset.

I look only at gameplay that we consider to be deliberate. Exclusion criteria:

(1) sessions where the game was started but exited immediately; (2) sessions with long sequences of incorrect responses; or (3) sessions with three fast incorrect responses in a row (as detected by the system, resulting in an automatic ending of a game).

We also excluded users in grades 1 and 2, as most games are not suited for children in this age, thus they provide very little data.

## Sample characteristics

Total sample size and sample size per grade:

```
## Total Sample Size:
```

```
## [1] 106170
```

```
##    grade     N
##    <num> <int>
## 1:     3 29554
```

```
## 2:     4 21387
## 3:     5 16291
## 4:     6 14363
## 5:     7 13472
## 6:     8 11103
```

How many sessions, on average, were quit prematurely?

```
## [1] 0.3171328
```

# 2-State Markov model

NOTE: Markov models are fitted in an external script, <fit_mm.R>. These need to be run separately before continuing the analysis in this document.

Here, I load in and analyse model fit statistics for all models. Then, I analyse the results of the best-fitting model. This was the model including all possible covariates and their interactions with sequential errors. Analysis of all fitted Markov models, on training and testing data, is found in <mm_supp.Rmd>

Variables are defined in the following manner:

- error_seq = Sequential errors. How many errors have been committed in a row? Categorical variable with 5 levels: 0, 1, 2, 3, >3.
- difficulty = What is the predicted probability that the next item will be correct, for the individual player? Categorical variable with 3 levels: 0 = easy (90% correct), 1 = medium (75% correct), 2 = difficult (60% correct).
- weekend_evening = Playing during school hours. Did the user play during school hours, or in the weekend/evening? Binary variable: 0 = school-time, 1 = weekend or evening hours.
- grade = What school grade is the user enrolled in? Categorical variable with 3 levels: 3-4, 5-6, 7-8.
- RT = Response time. Was the response slow (0) or fast (1)? Computed based on median response time (faster than median = fast, and vice versa).

## Model comparison

Loading in all Markov models and comparing their fit statistics. It is not advisable to run this chunk all at once, as the markov data files take up a lot of space. Run one at a time, and keep an eye on available R storage, to avoid crashing.

```
##         model     AIC minus2loglik
## 1    baseline 2904426      2904424
## 2   covariate 2371495      2371473
## 3 interaction 2364586      2364516
```

```
## Likelihood ratio test
##
## Model 1: state ~ item_count
## Model 2: state ~ item_count
```

```
## Model 3: state ~ item_count
##   #Df   LogLik Df    Chisq Pr(>Chisq)
## 1   1 -1452212
## 2  11 -1185736 10 532950.9  < 2.2e-16 ***
## 3  35 -1182258 24   6956.2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Transition rates

The Q-matrix represents the transition intensities between different states in a Markov model, indicating the instantaneous rate at which transitions occur from one state to another.
The P-matrix represents the transition probabilities over 10 items in a session. This is derived from the Q-matrix, showing the likelihood of moving from one state to another within this interval.

```
## Baseline transition Intensities and Probabilities for the 2-State Markov Model
```

```
## (When seq. errors = 0, difficulty = medium, grade = 7-8, RT = slow, and play hours = schooltime)
```

```
## Q-Matrix
```

```
##             State 1    State 2
## State 1 -0.01821554 0.01821554
## State 2  0.00000000 0.00000000
```

```
##
## P-Matrix
```

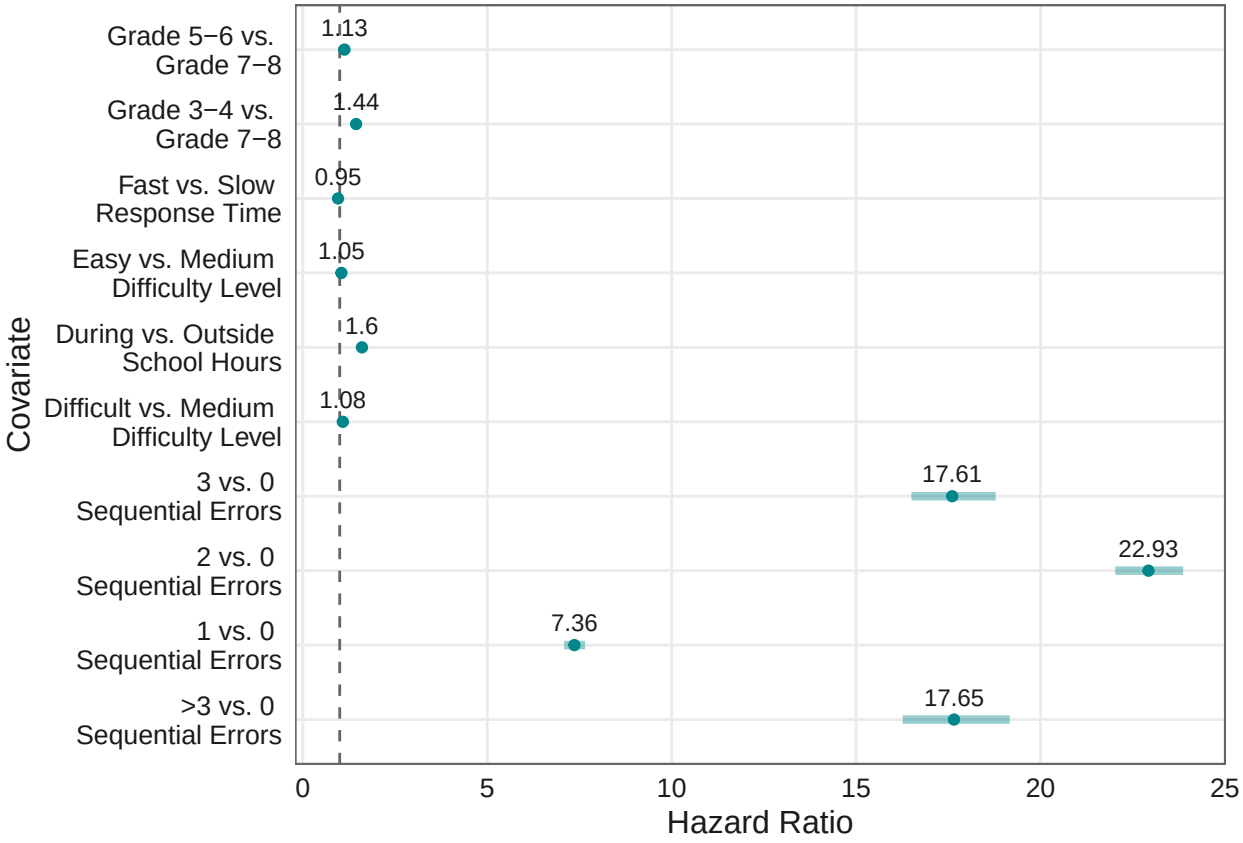
```
##           State 1   State 2
## State 1 0.8334718 0.1665282
## State 2 0.0000000 1.0000000
```

In the addition domain, there is an instantaneous probability of quitting of 1.7% when all covariates are controlled for.

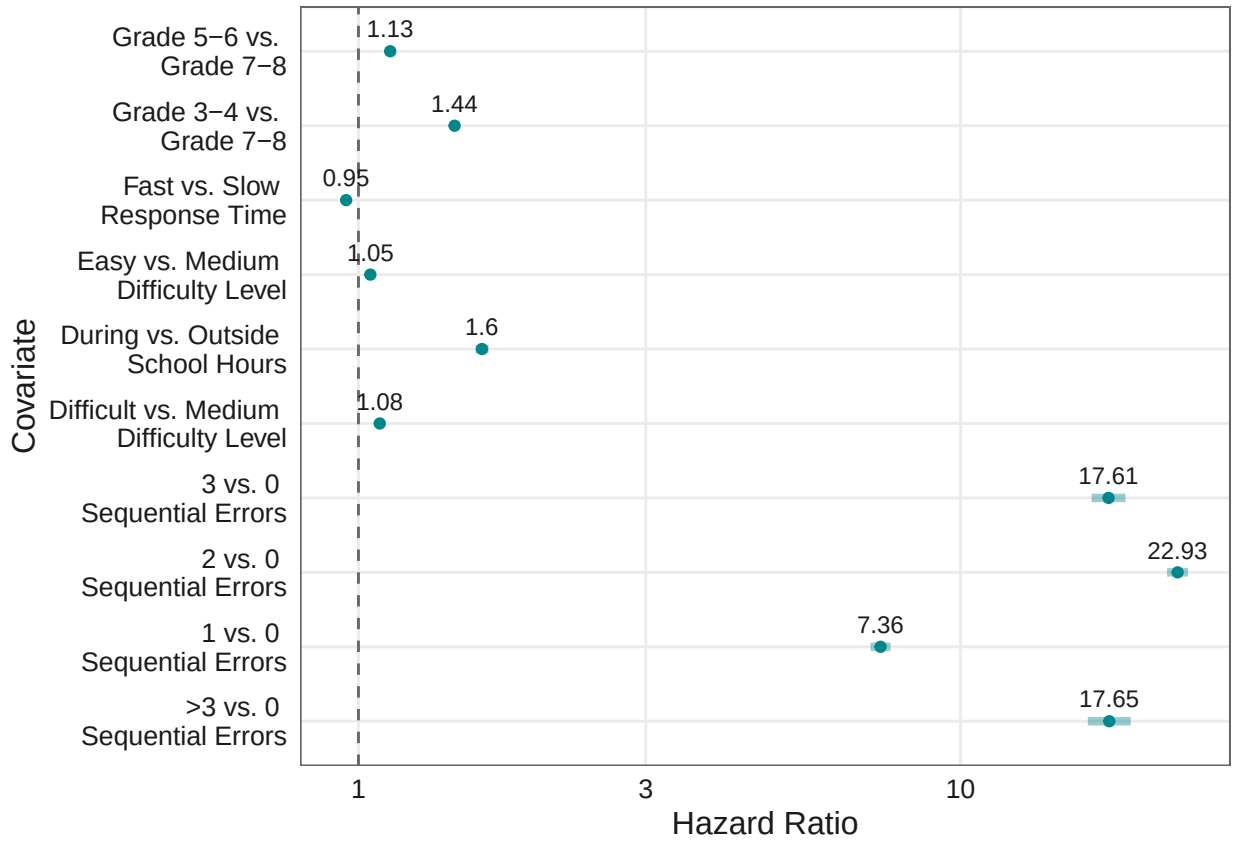
## Hazard ratios

Hazard ratios are derived from the proportional hazards model, and represent the relative risk of transitioning from a persisting into a quitting state, between two groups, each differing by one unit of a covariate, while holding all other variables constant. In this model, we derive main effects hazards ratios (the relative increase in risk of a state transition given one value of a covariate) and interaction effect hazard ratios (the relative risk of a state transition given one value of a covariate, across each level of sequential errors.)

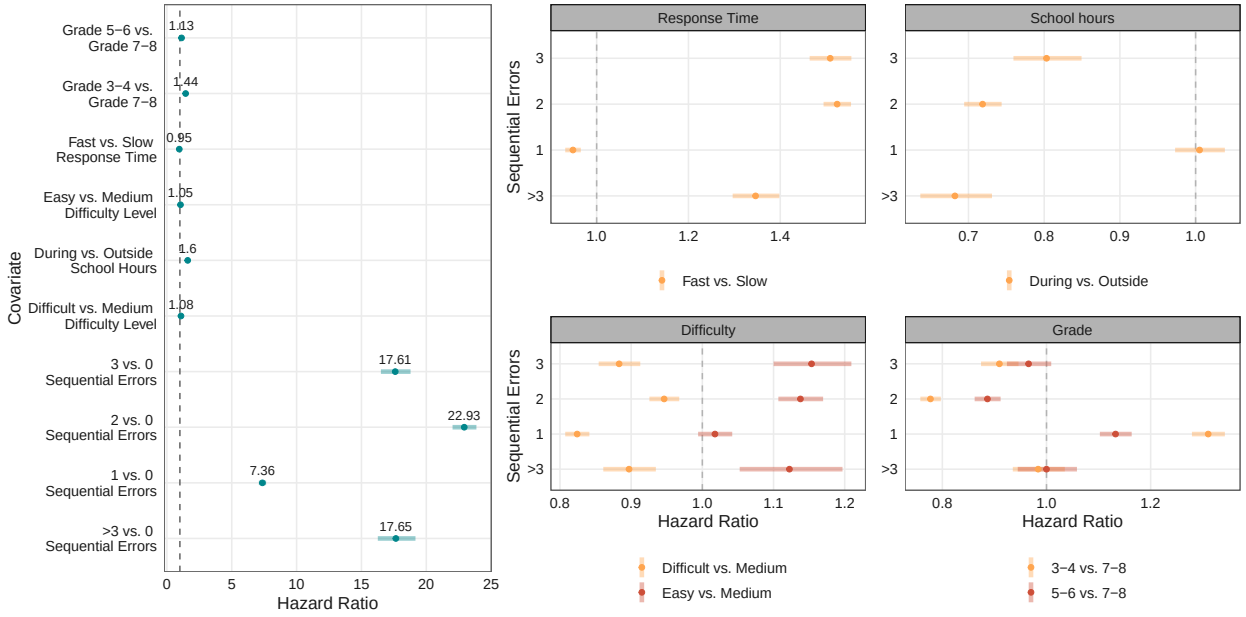
**Main effects**

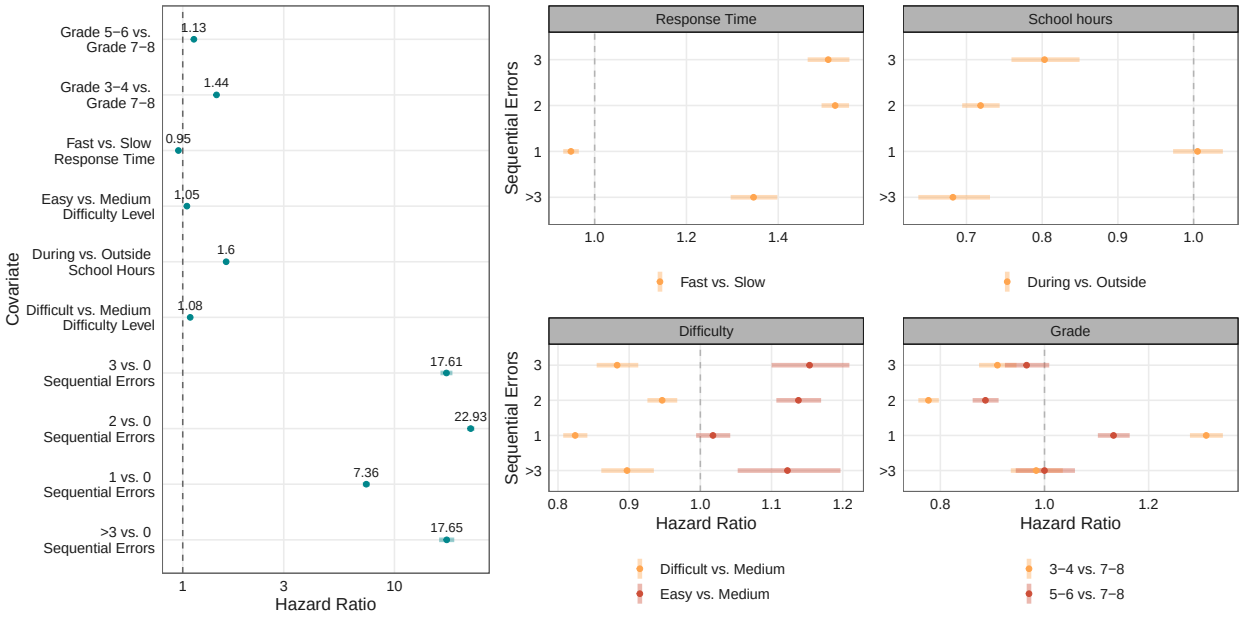



And displayed in a log scale:




## Interaction effects

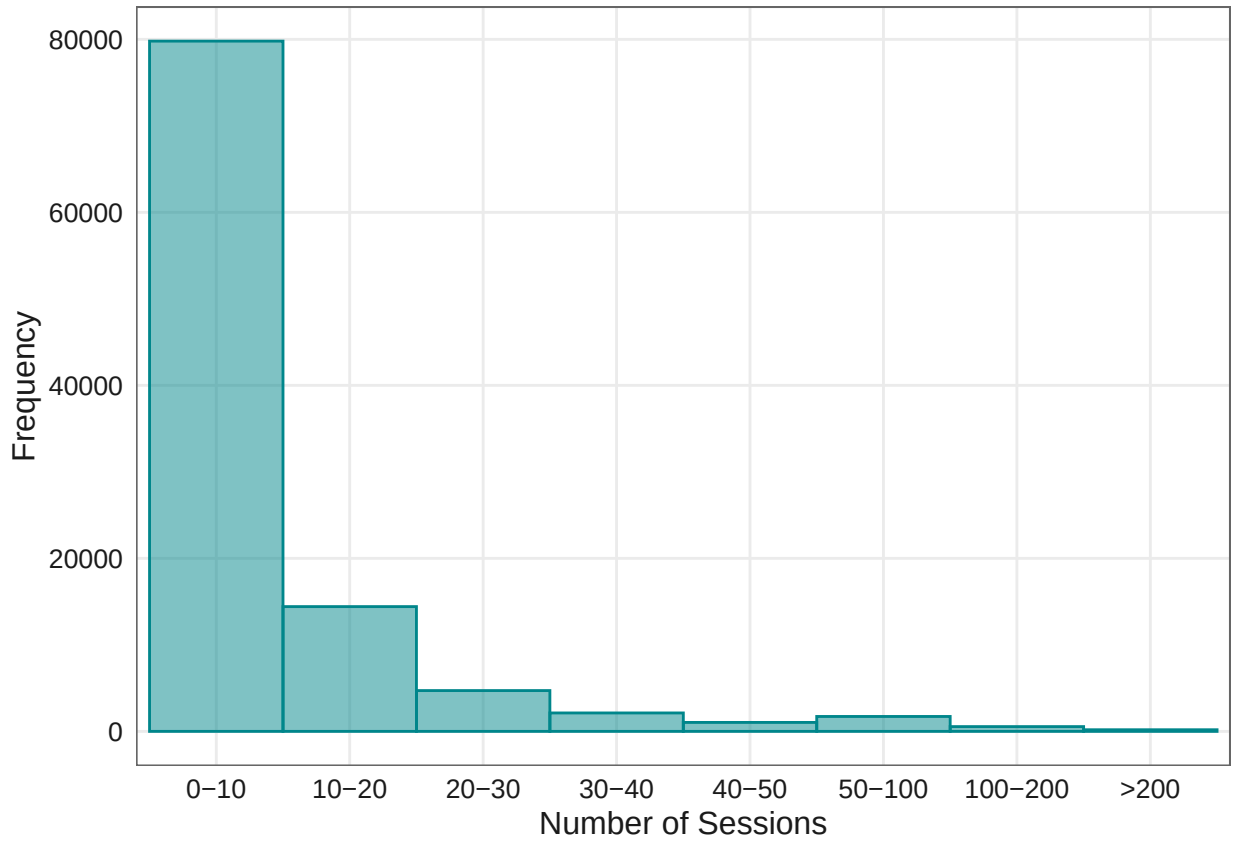





# Longitudinal effects: Error-induced quitting over time

## Histogram completed sessions

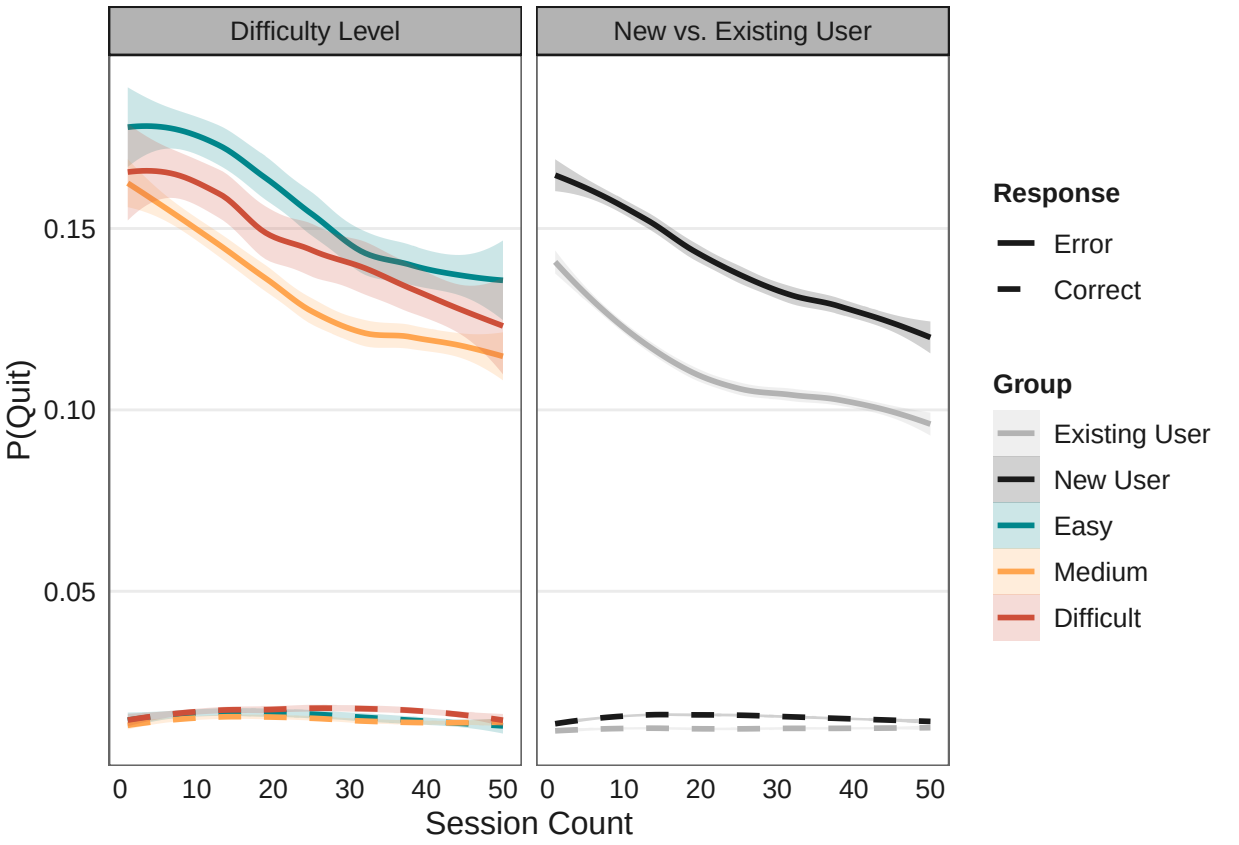
This distribution shows the total amount of full sessions completed within the addition domain per user.




### Session plots

Here, I examine how the quitting rate changes over different session counts. Each person in our data is assigned a session count, indicating which session number they're currently on since the start of our data collection. By aggregating the quitting events across these session counts and then averaging them out, I'm able to see how the quitting behavior evolves over time, giving us insights into the trends of quitting behavior as people progress through sessions. I separate quitting probabilities across difficulty setting, and whether a user was new throughout the scope of our data or not (representing a more experienced user).

I look at these trends across the first 50 sessions, as there is not many users who complete more than a total of 50 sessions.




# GLMER Model

## Fit LMER Models

I apply more stringent data selection criteria to this subset of data. Datasets for LMER model fitting are created in a separate script: <clean_add_forLMER.R>. In there, training or testing datasets are cleaned according to the exclusion criteria:

- at least 10 total quits per user
- at least 50 total sessions per user

and saved to the directory (folder: data_clean) as <addition_train_forLMER.Rdata> or <addition_test_forLMER.Rdata>.

GLMER models are also fit in a separate script: <fit_lmer.R> or <fit_lmer_test.R>. These should be run separately before analysis in this document is continued.

Here, I load all of the models, and compare their fit to the data. Then, I plot the results of the best fitting model, the random intercept and slope model with covariates. Analysis of all fitted glmer models (training and testing) is found in <glmer_supp.Rmd>.

- Random intercept and slope model with covariates: quit ~ 1 + error_seq + rating + grade + (1 + error_seq | user_id)

Note that, in contrast to the previous Markov models, the sequential error variable here is continuous. That is because the categorical sequential error variable leads to convergence issues in the glmer modelling.

## Sample characteristics

```
## Total Sample Size:
```

```
## [1] 3998
```

```
## Average quit rate, glmer data (addition): 0.3454802
```

## Model comparison

```
##                model     AIC     BIC    logLik
## 1           ran_int 1047301 1047340 -523647.3
## 2     ran_int_slope 1032731 1032796 -516360.3
## 3 ran_int_slope_cov 1013950 1014041 -506968.0
```

## Fixed effects

Model coefficients for the fixed effects measures:

### Odds Ratio

When modelling logistic regression, we can derive an odds ratio for each coefficient. In the context of this analysis, the odds ratio expresses the relative odds of quitting given a one unit increase of each covariate, compared to the odds of quitting when the covariate is at its reference level.

```
## Odds Ratios:
##  Sequential errors:  2.242011
##  Rating:  0.9072007
##  Grade:  0.9477148
```

## Random effects

Variance estimates for the glmer model on the addition data.

### ICC

Here, I calculate intra-class correlations (ICC), which provides insight into the proportion of total variability attributable to the grouping structure within the subtraction model. The ICC for the intercept indicates the proportion of variance in each individuals' baseline quitting rate (meaning quitting in the absence of error) that can be attributed to between-subject variability, while the ICC for the slope represents the proportion of variance in each individuals' effect of sequential errors on quitting, which can be attributed to between-subject variability.

### Distribution of random effects

I extract the random effects and plot their distribution.