`<folder>`
`<script>`
`<data file>`

Replication
`<replication_multistate>`

1. **Raw replication data**
   1.1. `<domain_info.Rdata>` connects the domain number to its name and description of the game type.
   1.2. `<user_info.Rdata>` contains grade, date of birth, and gender data per user id.
   1.3. `<logs_replication_quitting.Rdata>` has user playing logs from all domains in the Math Garden and Language Sea learning environments, within the period 2023-05-29 and 2023-07-30.

2. **Data preparation** `<replication_data_pred.Rmd>`
   2.1. Inputs the raw data files and cleans them according to our selection criteria:
      2.1.1. Removed domain 5 as it has more than 10 items per game
      2.1.2. 88 sessions had more than 10 observations, which affects the session counter and have to be removed. Because the multistate model is dependent on accurate chronological flow of gameplay per user, we have to remove all users who are affected by this (n = 68).
      2.1.3. Non-deliberate gameplay is detected and deleted (63,362 observations of out 6,029,135 total items played)
      2.1.4. Delete all sessions which lasted shorter than one whole game (83,207 observations removed)
      2.1.5. Delete all sessions with only incorrect responses (2567 observations removed)
      2.1.6. Users in grades 1 and 2, and in special education (grade = 19), are deleted.
   2.2. Computes the relevant variables for analysis
   2.3. Saves the cleaned data as `<replication_clean.Rdata>` to the `<data_clean>` folder.

3. **Main analysis** `<replication.Rmd>`
   3.1. The first section of the markdown file, “Descriptives”, computes sample characteristics and distributions of key variables in the data. It only describes the data and is not strictly necessary for running the replication models.
   3.2. The external script, `<modelfit_msm.R>`, fits the MSSM to the clean data. It needs to be run (on `<replication_clean.Rdata>`) before continuing on to the second section of the markdown file.
   3.3. The second section, “Multistate Survival Model”, contains the main analysis.
   NOTE: `<modelfit_msm.R>` is an external model fitting script for the MSSM. It needs to be run on the clean replication data before continuing into this section of the analysis.
      3.3.1. The resulting fitted models are saved into the subfolder `<models>`
   3.4. The following figures are saved into `<replication_figures>`:
      3.4.1. Descriptive plots
      3.4.2. Hazard ratio plots, with and without representing HRs on a log scale.

4. **Extras**
      4.1.1. `<exploration>` contains analysis scripts used for exploring the data apart from the main analyses.
         4.1.1.1. `<exploration_switchovertime.R>` plots the probability of state transitions over time.
         4.1.1.2. `<model_assessment_msm.R>` plots the model fit (difference between observed and expected probabilities of being in each state) of the constrained and covariate models. The final figure in found in the supplementary materials folder.
         4.1.1.3. `<replication_check_data.R>` checks the difference between computed variables in the original study compared to this one, as the data filtering had to be done slightly different, due to a mistake in the original code.

Individual Differences in Quitting from Addition
`<individual_differences/ind_diff_addition>`

1. **Raw addition data**
   1.1. `<all_logs_addition_2022.Rdata>` (2022-2023) and `<all_logs_addition.Rdata>` (2020-2022) are the raw log files for the addition domain. `<player_ratings_addition1.Rdata>` are the latest player ratings for the players contained the 2020-2022 log data.

2. **Data preparation** `<addition_data_prep.Rmd>`
   2.1. All files are found in the folder `<data_prep>`.
   2.2. `load_addition_data.R` is sourced at the beginning of the markdown file, thus it needs to exist in the same directory. This script inputs the raw addition data and splits it into training and testing datasets. The desired dataset to be prepared ("training" or "testing") needs to be specified in the main markdown file. Crucially, this script also creates the player ratings data of the 2022 log data `(<player_ratings_addition2.Rdata>)`, and saves it into the `<data_clean>` folder.
   2.3. `<load_player_ratings.R>` is a script that inputs the player ratings data files and combines them into one large, raw data file of the player ratings. The resulting dataset, `<player_ratings_addition.Rdata>` is loaded into the main markdown file to compute the user rating variable.
   2.4. Data is cleaned:
      2.4.1. Users with faulty sessions (see 2.1.2) are removed (n = 55).
      2.4.2. Non-deliberate gameplay is marked.
      2.4.3. Users in grades higher than grade 8 are removed.
      2.4.4. Key variables are computed.
      2.4.5. Distribution of key variables is visualized. Plots are saved to `<figures/data_prep>`

3. **Main analysis** `<ind_diff_addition.Rmd>`
   3.1. The main markdown script requires the following files, which also exist in the `<main_analysis>` folder:
      3.1.1. `<fit_mm.R>` and `<fit_mm_test.R>` to run the simple Markov models on the training and testing data.
      3.1.2. `<clean_add_forLMER.R>` to apply exclusion criteria for the mixed-effect logistic regression model.
      3.1.3. `<fit_lmer.R>` and `<fit_lmer_test.R>` to fit the mixed-effect logistic regression model to the training and testing data.
   3.2. Analysis is performed in three sections:
      3.2.1. 2-state Markov model: for model comparison, looking at transition rates, hazard ratios, and main and interaction effects from the 2-state Markov model.
         3.2.1.1. Plots saved to either the `<train>` or `<test>` subfolder of the `<figures/addition>` folder.
      3.2.2. Longitudinal effects: plot probability of quitting over time, across difficulty setting and experience level
         3.2.2.1. Plot saved to either the `<train>` or `<test>` subfolder of the `<figures/addition>` folder.
      3.2.3. Mixed effects logistic regression: compares model fits and looks at fixed and random effects in the glmer model.
         3.2.3.1. Plots saved to either the `<train>` or `<test>` subfolder of the `<figures/addition>` folder.

Individual Differences in Quitting from Subtraction
`<individual_differences/ind_diff_subtraction>`

1.1. The file structure for the analysis on the subtraction data looks the same as that for the addition data, without the separation of train and test folders.

Stability of Individual Differences `<ind_diff_stability.Rmd>`
`<individual_differences/ind_diff_stability>`

1.1. The main .Rmd file contains all analysis pertaining to the stability of individual differences in the addition and subtraction data.
1.2. This file:
   1.2.1. Compares the fixed effect estimates between the datasets.
   1.2.2. Compares the distribution of random effects between the datasets.
   1.2.3. Computes the correlation of random effects between the datasets.
   1.2.4. Plots baseline quit rates (random intercepts) and effects of sequential errors on quitting (random effects) across 300 randomly sampled users.

General notes

- All figures, in .pdf form, are saved to the `<figures>` folder, within a corresponding subfolder.
- `<dependencies.R>` in a file that is sourced in most other files mentioned. It loads relevant libraries, plotting themes, and other specialized functions needed for the analyses in this paper.
- At the beginning of each script in the addition analysis, "training" or "testing" can be specified as the dataset to analyze. This will load in the correct dataset and save outputs in the correct subfolder.