UCLouvain

epl

École polytechnique de Louvain

Modelling Motor Learning in the Elderly using a Serious Virtual Reality Game

Author: **Lore Lernoux**
Supervisors: **Estelle Massart, Gauthier Everard**
Readers: **Thierry Lejeune, Marco Saerens**
Academic year 2025–2026
Master [120] in Mathematical Engineering

Abstract

This work aims to study the evolution of subjects' performance in an experiment in immersive virtual reality relying on a serious game, inspired by the *Whac-a-Mole* arcade game. This was performed by modelling two outcomes: the action time required to hit the target and the probability of a successful hit. To account for repeated observations within subjects, mixed-effects models with a random intercept per subject were used. Separate models were fitted for the first and second halves of the experiment in order to compare early and late performance.

The action time results showed that subjects tended to become faster as the experiment progressed, when task difficulty and subject baseline differences were taken into account. The late model achieved lower prediction errors than the early model in the repeated 80–20 test procedure, suggesting that action times became more predictable in the second half of the experiment.

For the binary hit outcome, logistic mixed-effects models were used to predict the probability of hitting the target. The models achieved moderate predictive performance: they ranked successful trials above unsuccessful trials better than chance, but trial-level success remained difficult to predict precisely. The early model performed slightly better than the late model at the trial level, whereas the block-level bias remained close to zero in both cases. Per-subject analyses showed that some subjects were well predicted, while others had abrupt drops in hit rate or irregular performance patterns that were not fully explained by the available predictors.

Finally, two model exploitation and generalisation procedures were performed: leave-one-subject-out analyses and difficulty-level simulations. The action time models generalised better than the binary hit models to unseen subjects. The simulations suggested that the late models generally predicted shorter action times for some subjects, and higher hit probabilities for all of them under comparable task configurations, but the magnitude of these changes varied between subjects.

Acknowledgments

I sincerely thank both of my supervisors, Prof. Estelle Massart and Dr. Gauthier Everard, for their advice, their time, and their encouragement throughout this thesis.

I would also like to thank Prof. Thierry Lejeune and Prof. Marco Saerens for agreeing to read and evaluate my work.

I warmly thank my close friends and family, with a special mention to my dad, for their unwavering support during periods of doubt and for their invaluable encouragement throughout this journey. I am also grateful to my fellow students who were working on their theses at the same time, especially Simon, for sharing this experience with me, for the mutual support, and for making the challenges feel lighter.

Last but not least, I would like to thank Pierre, my boyfriend, for being by my side throughout this journey. Thank you for listening to me whenever I needed to talk, for supporting me, and for motivating me when I doubted myself. Your patience, kindness, and encouragement helped me more than you know. I am deeply grateful for your presence, your love, and the strength you gave me along the way.

Acronyms

AT Action Time

AUC Area Under the Curve

BBT-VR Box and Block Test

FE Fixed Effects

iVR immersive virtual reality

LMEM Linear Mixed-Effects Model

LMEMs Linear Mixed-Effects Models

LOSO Leave-One-Subject-Out

MAE Mean Absolute Error

RE Random Effects

RMSE Root Mean Squared Error

VR Virtual reality

Use of AI

I used the generative AI chatbot *ChatGPT* for the realization of the code for some figures based on my ideas, to point out possible mistakes or to help me rephrase some part of the following text. Every AI contribution was carefully reread.

General aid

I used the *geeksforgeeks* website on multiple occasions to comprehend concepts or models.

Contents

1 Introduction

Motor learning is a complex process through which individuals improve their motor performance with repeated practice and experience [BPD18]. It is fundamental for humans as it is involved in a wide range of activities, from everyday movements to the acquisition of complex motor skills, and it plays an important role in rehabilitation. Therefore, understanding how motor performance evolves over repeated practice is important to design effective training protocols and characterise learning mechanisms.

Motor skill acquisition can be defined as the gradual improvement in rapid action selection and execution [Kra+19]. There are two major motor learning categories: sequence learning and sensorimotor adaptation [Sei10]. The former refers to combining different movements into a smooth, purpose-directed action. The latter corresponds to behavioural changes as a response to modified sensory inputs. Modelling learning can be challenging as it is a complex process, influenced by multiple factors such as variability, feedback, and task difficulty. The human body also has its own fluctuations, such as fatigue or injury, leading to similar motor commands having possibly different outcomes [DSÖ17].

Modelling motor learning has been studied for a long time, but is still relevant nowadays as the proportion of older adults (more than 65 years old) is only increasing [Wor23]. Global life expectancy has increased significantly since 1950, going from 46 years to 70 in 2023, with considerable variability between countries of different incomes. This increase implies a rapid growth of older adults across the world, with an estimation of more than 2.1 billion people older than 60 years by 2050 [Gia+25].

But ageing is also associated with a progressive decline in motor and cognitive abilities, as well as a decrease in bone density, which increases the risk of falls and reduced functional capacity. This directly impacts autonomy and quality of life, limiting their activities and social interactions [HKP25]. The rise in the number of older adults is also positively correlated with an increase in the total number of chronic and neurodegenerative diseases, or a coexistence of several pathologies

in the same patient [Gia+25]. This motivates the research to find a baseline in motor learning patterns by investigating performance in healthy older adults, and could eventually lead to designing appropriate experimental protocols for impaired ones. Indeed, those studies can establish reference models to further improve rehabilitation strategies.

To do so, immersive virtual reality (iVR) has become a promising tool. Let us first define Virtual reality (VR); VR allows users to interact with another version of reality, which can be explored in real time [Sve04]. The term VR is quite broad and can refer to different types of systems, from simple screen-based environments to more immersive systems using a head-mounted display [Cip+18]. These environments have been increasingly studied in recent years, as more and more papers related to VR are published every year, as shown in Figure 1.

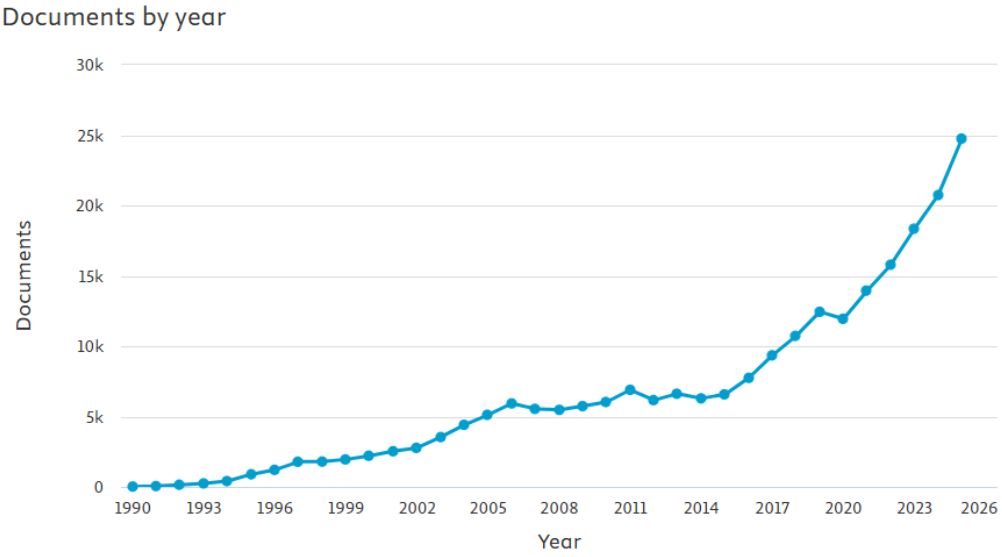

Figure 1: Evolution of the number of publications per year related to VR. Graph found on Scopus using "Virtual Reality" as keywords [Sco26].

Indeed, recent advances in this technology, and in particular in iVR, have opened new perspectives for studying and enhancing motor learning. iVR differs from the standard VR technology by using displays that fully immerse users in a 3D virtual world. These systems are equipped with sensors that track the user's movements to give them the possibility to look around and interact with the environment. Several applications in iVR provide controlled, interactive, and reproducible environments in which task parameters can be adapted to the different users [Lui+25].

iVR environments may also increase the subjects' motivation, as their rehabilitation is closer to a game-like session. The setup may also provide real-time feedback, which can guide performance and allow for adapting the task or its difficulty to the user's performance, and is thus particularly interesting in the context of rehabilitation and personalised training. This motivation boost may result in longer practice sessions, and thus induce quicker learning [LS14].

Modelling motor learning is still a challenging task because of all the variability that comes with trial-to-trial practice in a self-adaptive environment.

To study the effects of iVR on motor learning in older adults, Everard et al. [Eve+25] developed REAsmash-iVR, a self-adaptive serious game inspired by the *Whac-a-Mole* arcade game. They then conducted an experiment on 20 healthy older adults, to assess the transfer of performance from the practice on REAsmash-iVR for seven days to other iVR tasks. During the experiment, older adults had to hit a target as fast as possible while avoiding distractors, and the difficulty of the game was adapted according to their performance throughout the experiment.

Their work aligns with previous research on the use of VR for motor rehabilitation in older adults. Indeed, iVR environments are very flexible and can be used for neurological disorders rehabilitation, such as chronic pain or stroke [Cer+24; Mek+21]. In their systematic review, Høeg et al. [Høe+21] showed that VR has been increasingly explored as a tool for motor rehabilitation, while also indicating that the term VR is often used in an ambiguous way. Indeed, many studies classified as iVR interventions are actually based on non-immersive systems, while only a small number of them truly use iVR systems. The work of Everard et al. [Eve+25] is thus aligned with their line of research, as REAsmash-iVR is clearly based on an immersive system.

The main objective of this work is to model motor learning by investigating how the performance of the subjects evolved as the experiment progressed. This extends the work from Everard et al. by studying subjects' performance during the experiment itself, rather than focusing only on retention and transfer after the seven days of practice. More specifically, we seek to characterise the evolution of performance across the experiment while accounting for task difficulty. We used two metrics to represent performance:

- the Action Time (AT), which is the time taken by the subjects to complete a successful trial,
- a binary success/failure metric, in which passed trials were coded as successful, while omissions and distractor hits were coded as failures.

To achieve this goal, we first performed exploratory analyses of the raw data, including visualisations of AT across blocks, computations of the proportions of the three possible trial outcomes, and spatial heatmaps of target and distractor locations. We then used Linear Mixed-Effects Models (LMEMs) to predict log-transformed AT on successful trials, and logistic mixed-effects models to predict binary success on all trials. These models were chosen because they account for subject-specific baseline differences. Separate models were fitted for the first and second halves of the experiment, an early and a late one, to assess changes over time. To see whether the models could predict both performance metrics on an unseen subject, we then performed Leave-One-Subject-Out (LOSO) analyses. Finally, simulations of the game's difficulty levels were used to interpret the fitted models, while keeping plausible game configurations.

By combining all these analyses, this work contributes to a better understanding of motor performance in iVR-based training. More broadly, it aims to support the development of adaptive virtual environments that could be used to personalise motor learning and rehabilitation protocols for older adults.

2 Background and Preliminary Work

2.1 Experiment Overview

This work is based on the experiment presented in *A Self-Adaptive Serious Game to Improve Motor Learning Among Older Adults in Immersive Virtual Reality: Short-Term Longitudinal Pre-Post Study on Retention and Transfer*, by Everard et al. [Eve+25]. The study investigated motor learning in older adults using REAsmash-iVR, a self-adaptive serious game developed in iVR. The goal of the experiment was to evaluate whether repeated practice during seven consecutive days in a virtual environment could support motor learning, retention, and transfer. Demographic information about the subjects was recorded, such as their age, sex, weight, and height.

This experiment was conducted on 20 older adults who were all between 69 and 89 years old, and who had no visual impairments or difficulty understanding simple tasks. They were also able to hold and use the manual controller of the iVR setup, as people with orthopaedic or neurological disorders that could affect upper-limb movement or controller use were excluded from the study. The experiment was based on REAsmash-iVR, which is an iVR adaptation of the *Whac-a-Mole* arcade game: participants were placed in a virtual environment where they had to find and hit a target mole while ignoring other types of moles acting as distractors. The experiment took place for seven consecutive days, during which subjects were asked to play for at least 15 minutes per day. A visualisation of the game can be found in Figure 2. Although the task is simple to understand, it combines several abilities that are important for motor learning, such as visual search, attention, coordination, movement speed, and movement accuracy.

The protocol followed a short-term longitudinal pre-post design. Participants were evaluated at three time points: immediately before the experiment, immediately after the experiment, and seven days later. Two iVR tools were used to assess motor learning transfer: KinematicsVR and the iVR version of the Box and Block Test (BBT-VR).

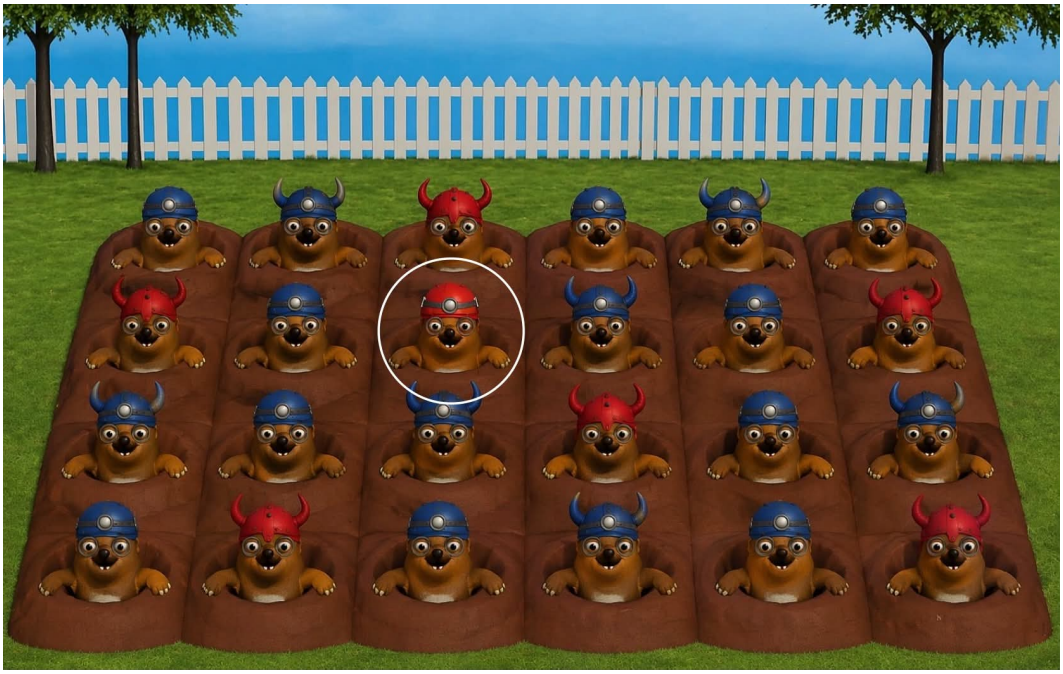

Figure 2: Visualisation of the game[1].

All iVR applications in this experiment used Unity `2019.3.15` software on the Oculus Quest 2 (Meta) as shown in Figure 3. Subjects were equipped with a headset, which allowed them to see the virtual environment, in addition to one or two handheld controllers, in order to interact with it.

Figure 3: Meta Quest 2 headset and hand controllers [Ste21].

[1]OpenAI. (2026). ChatGPT. `https://chat.openai.com/chat`, used on 20/02/2026.

2.2 Serious Game Overview

REAsmash-iVR is a self-adaptive serious game. A serious game is a game designed for a purpose other than entertainment only; in this case, the purpose is to train motor abilities in older adults [Fen+18]. The game is self-adaptive because its difficulty changes automatically according to the participant's performance. The virtual environment contains a six-column and four-row ground, from which moles could come out, as shown in Figure 2. At each trial, the subject must identify the target mole and hit it with the virtual hammer. The target is the mole wearing a red miner's helmet, which is circled in Figure 2. Other moles appear as distractors and must be ignored, and differ from the target by their colour and/or shape:

- a mole with a blue helmet;
- a mole with a blue-horned helmet;
- a mole with a red-horned helmet.

This design makes the task both motor and cognitive. The participant must first find the correct mole among similar distractors, then move the controller toward it.

The game is organised by blocks. A block is a sequence of trials played at a constant difficulty level. After each block, the game analyses the participant's performance and adapts the next block accordingly. This structure is represented in Figure 4.

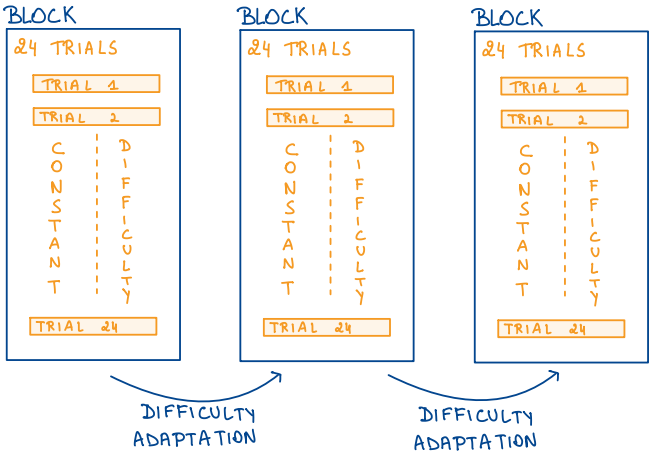

Figure 4: Representation of the blocks and trials structure.

The main performance indicator used by the adaptation algorithm is the success rate, defined as the proportion of target moles correctly hit. The objective of the regulator is to keep the task challenging but feasible, with an average success rate close to 75%. If the success rate is higher than 75%, the game considers the task too easy and increases the difficulty. If the success rate is between 50% and 75%, the game considers the task difficult but still manageable, so there is only a small decrease in difficulty. If the success rate is below 50%, the game considers the task too difficult and decreases the difficulty considerably.

However, the adaptation is not only based on the success rate. The algorithm also considers more detailed information about the participant's errors, including:

- omissions, meaning trials where the subject did not hit either the target or a distractor;
- the location of omitted targets;
- false positives, meaning distractor moles that were hit;
- the location of the distractors that were incorrectly hit.

To adapt the difficulty, several variables of the game could change from one block to the other, as listed below.

- The size of the virtual ground. Easier difficulty levels had fewer rows in each column, while harder levels had the full 4×6 ground.
- The number of distractors of each type could vary from none to a maximum of 10 distractors of a type.
- The target had a given lifetime, which refers to the time it stayed above the ground during a trial. This lifetime could vary from 8 to 2.5 seconds, where longer lifetimes correspond to easier game configurations.
- Three types of cues could be given to the subject to help them locate the target. Those cues were, from the most explicit to the least explicit:
  - an arrow pointing to the target;
  - an animation: the target would move while the distractors stayed still;
  - a spatialised sound cue through the headset, indicating the direction of the target location.

Figure 5 shows the REAsmash-iVR environment and the experimental setup. Subjects see the moles through the headset and use virtual hammers controlled by the handheld controllers to hit them. During each trial, the system recorded several variables, including the result of the trial and the AT.

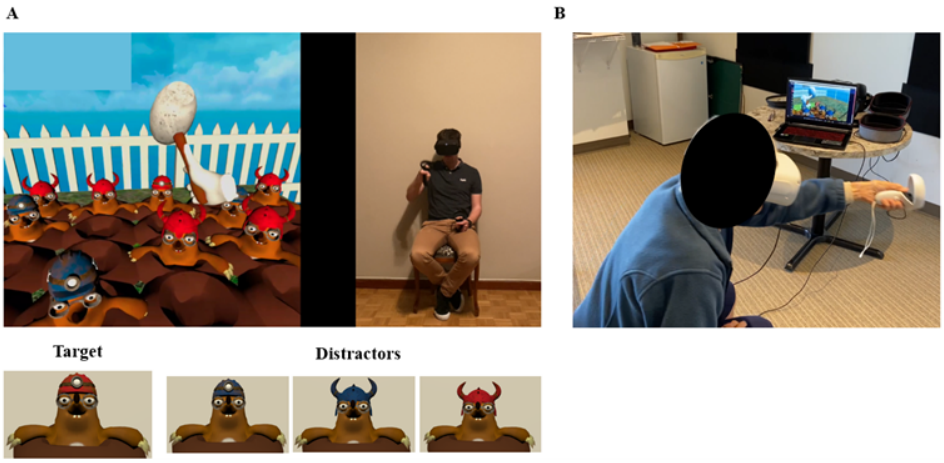

Figure 5: Game visualisation from [Eve+25]. (A) REAsmash-iVR environment as seen by the subjects on the left, with the corresponding movement to be done on the right. Just under it is a representation of the target and the three different distractors. (B) Subject taking part in the experiment.

2.3 Technical Background

2.3.1 Linear Mixed-Effects Models

Most statistical models, such as classical multiple linear regression, rely on the independence of observations, as they assume residuals to be independently and identically distributed. The goal of multiple linear regression is to capture how several features influence a dependent variable, the target, by fitting a hyperplane through data points [Mic25]. The classical multiple linear regression equation is shown in Equation 1, where β_0 is the intercept, the value of the dependent variable whenever all features are equal to zero. The terms $\mathbf{X}_k$ correspond to the p different predictors which all have an associated coefficient β_k corresponding to their slope, and ϵ is the random error term [UG13].

$$\mathbf{Y} = \beta_0 + \sum_{k=1}^{p} \beta_k \mathbf{X}_k + \epsilon \tag{1}$$

In our case, the experiment involves many observations per subject, which induces intra-subject correlation and thus violates the assumption of independence of ob-

servations. Indeed, we can expect trials from the same subjects to be more similar than those of someone else.

This problem of non-independent data is tackled by LMEMs as they account for these sources of dependencies [Bat+15]. They are used in various types of analysis, as they can take into account not only the different subjects participating in an experiment, but they can also compare clinics in the case of a multi-location experiment, for example, or students from the same classroom answering multiple questions.

The different variables of the analysis can thus be split into two categories: Fixed Effects (FE) and Random Effects (RE).

The first category includes variables for which all possible values they can take may have been observed throughout the data. In other words, they regroup the population-level trends, as in a classical regression model. Measured predictors such as the `number of distractors`, or demographic data such as `age`, are fixed variables as we assume the study regroups all the levels of interest for those variables.

The variables from the *random* category regroup those for which not all levels are present in the data and several observations are collected from those same levels, creating non-independence in the data. In our case, the `subject` variable can refer to any subject from the population target by the experiment, but only a sample of them is included in the data. If different subjects were chosen in a similar study, there would be no way to anticipate their a priori contributions to the data variation, as there is no way to know their prior knowledge and capabilities. Furthermore, the levels of the *random* category are not interpreted as true categories in themselves. In the case of the `subject` variable, being labelled *subject 1* or *subject 15* only serves to identify the participant [BC18; OM07].

We used the *MixedLM* function from the `statsmodels` library in *Python* to estimate the parameters of the mixed-effects models [stab]. In this framework, models are specified using a formula-based interface, similar to classical regression. The FE are written directly in the regression formula, while the RE are specified separately through grouping variables, such as `subject`. In practice, `statsmodels` internally builds two design matrices: a FE design matrix X, containing the predictors associated with the population-level coefficients β, and a RE design matrix Z, containing the predictors whose effects are allowed to vary between groups.

LMEMs regression equation can therefore be written as seen in Equation 2, where

we now take into account the random effects of subject j.

$$\mathbf{Y}_j = \beta_0 + \underbrace{\sum_{k=1}^{p} \beta_k \mathbf{X}_{jk}}_{\text{FE}} + \underbrace{Z_j u_j}_{\text{RE}} + \boldsymbol{\epsilon}_j \tag{2}$$

Contrary to a classical regression model with one coefficient estimated independently for each subject, LMEMs assume that the subject-specific effects are random draws from a common distribution:

$$\boldsymbol{u}_j \sim \mathcal{N}(\mathbf{0}, \Sigma_u), \qquad \boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0}, \sigma^2 I). \tag{3}$$

This means that the model does not only estimate the average effects $\boldsymbol{\beta}$, but also the amount of variability between subjects, represented by the covariance matrix Σ_u, and the remaining unexplained within-subject variability, represented by σ^2.

The parameters are estimated by maximising the likelihood of the observed data. More precisely, `statsmodels` can use either Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML). ML maximises the probability of the observed responses with respect to all model parameters, while REML corrects for the uncertainty introduced by estimating the FE and is often preferred when the main interest lies in the variance components. During fitting, the algorithm searches for the values of $\boldsymbol{\beta}$, Σ_u, and σ^2 that make the observed data most probable under the mixed-effects model.

For example, when only a random intercept is specified for each subject, the model assumes:

$$\mathbf{Y}_j = \beta_0 + \sum_{k=1}^{p} \beta_k \mathbf{X}_{jk} + u_{0j} + \epsilon_j, \qquad u_{0j} \sim \mathcal{N}(0, \sigma_u^2). \tag{4}$$

β_0 is the average intercept across the population, and u_{0j} is the deviation of subject j from that average intercept. Thus, the subject-specific intercept is:

$$\beta_0 + u_{0j}. \tag{5}$$

A positive value of u_{0j} indicates that subject j tends to have a higher baseline response than the average across subjects, whereas a negative value indicates a lower baseline response.

Importantly, the random intercepts u_{0j} are not estimated in the same way as fixed coefficients for each subject. Instead, once the FE coefficients and variance components have been estimated, `statsmodels` computes the most likely subject

deviation, for each subject, given the data and the variance. Those estimates are partially shrunk toward zero, meaning toward the population average. This shrinkage is one of the advantages of LMEMs, as it makes subject-specific estimates more stable. Subjects with fewer or noisier observations tend to have RE estimates closer to zero, whereas subjects with more informative repeated observations can deviate more strongly from the population mean.

More generally, random slopes can also be included. In that case, not only the baseline level but also the effect of a predictor is allowed to vary across subjects. For example, if the effect of block number is allowed to differ between subjects, the model becomes:

$$Y_{ij} = \underbrace{\beta_0 + \sum_{k=1}^{p} \beta_k X_{ijk}}_{\text{FE}} + \underbrace{u_{0j} + u_{1j}\texttt{block}_{ij}}_{\text{RE}} + \epsilon_{ij}. \tag{6}$$

with:

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u_0} & \sigma_{u_0 u_1} \\ \sigma_{u_0 u_1} & \sigma^2_{u_1} \end{pmatrix} \right). \tag{7}$$

In this formulation, u_{0j} represents the random intercept of subject j, while u_{1j} represents the random slope of subject j. The covariance term $\sigma_{u_0 u_1}$ indicates whether subjects with higher baseline values also tend to have stronger or weaker slopes.

In the formula interface, a random intercept per subject is obtained by defining the grouping variable as `subject`. When no additional random-effect formula is specified, `statsmodels` uses a random intercept for each group by default. For example, a model with a fixed effect of `block` and a random intercept per subject can be written as:

```
Y ~ block,    groups = subject.
```
(8)

A model with both a random intercept and a random slope of `block` per subject can be written as:

```
Y ~ block,    groups = subject,    re_formula = "1 + block".
```
(9)

In the expression `1 + block`, the `1` represents the random intercept, while the `block` term represents the random slope.

The final output of the model contains several types of information. The FE estimates describe the average relationship between the predictors and the dependent variable across all subjects. The variance components describe how much subjects differ from one another in terms of their baselines and, when included, their slopes. Finally, the subject-specific RE indicate how each subject deviates from the population trend after taking into account both the FE and the estimated variability between the subjects.

2.3.2 Binomial Logistic Mixed-Effects Models

While LMEMs are appropriate when the dependent variable is continuous and approximately Gaussian, they are not whenever the prediction task involves a binary outcome, such as success or failure. In that case, the response variable can only take two values, usually 0 and 1. A classical linear model is therefore not appropriate, as it could predict values outside of the interval $[0, 1]$ [Bol+09]. For this reason, we used the `BinomialBayesMixedGLM` class from `statsmodels`, which extends the mixed-effects models to binary outcomes [staa].

This model belongs to the family of generalised linear mixed models. Instead of modelling the response variable directly as a linear combination of predictors, it models a transformation of the probability of success. Let H_{ij} denote the binary outcome for observation i from subject j, where:

$$H_{ij} = \begin{cases} 1, & \text{if the trial is successful,} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The model assumes:

$$H_{ij} \sim \text{Bernoulli}(p_{ij}), \tag{11}$$

where p_{ij} is the probability of success for observation i of subject j.

Since probabilities are restricted to the interval $[0, 1]$, the model uses a link function to connect the probability of success to a linear predictor. In the binomial case, this is typically the logit link:

$$\text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = X_{ij}\boldsymbol{\beta} + Z_{ij}\boldsymbol{u}_j. \tag{12}$$

Equivalently, the predicted probability of success is obtained by applying the inverse-logit function:

$$p_{ij} = \frac{1}{1 + \exp\left[-\left(X_{ij}\boldsymbol{\beta} + Z_{ij}\boldsymbol{u}_j\right)\right]}. \tag{13}$$

The logit function is defined only for probabilities strictly between 0 and 1. In practice, this is not a problem for the model, because the inverse-logit transformation maps any finite linear predictor to $(0, 1)$.

In this formulation, $X_{ij}\boldsymbol{\beta}$ represents the FE part of the model, while the term $Z_{ij}\boldsymbol{u}_j$ represents the RE part of the model.

A random-intercept version of the model can be written as:

$$\text{logit}(p_{ij}) = \beta_0 + \underbrace{\sum_{k=1}^{p} \beta_k X_{ijk}}_{\text{FE}} + \underbrace{u_{0j}}_{\text{RE}}, \tag{14}$$

where β_0 is the average population-level intercept and u_{0j} is the deviation of subject j from this average intercept. Contrary to the LMEMs, no additive Gaussian residual term is included in the linear predictor. Indeed, the observation-level variability is described by the Bernoulli distribution of the response variable itself.

In `statsmodels`, a random intercept for each subject can be specified as:

```
vc_formulas = {"subject":  "0 + C(subject)"}.
```
(15)

The term `0 + C(subject)` creates one random intercept for each subject. Adding a random slope per subject for the block variable would then be:

```
vc_formulas = {"subject":  "0 + C(subject)",
               "slope":  "0 + C(subject):block"}.
```
(16)

Contrary to the `MixedLM` model, `BinomialBayesMixedGLM` uses a Bayesian estimation framework. This means that the model does not only rely on the likelihood of the observed data, but also places prior distributions on the parameters. The posterior distribution therefore combines the information coming from the observed data with the assumptions encoded in the priors.

In practice, `statsmodels` can estimate the model using either `fit_map()`, which finds the parameter values that are most plausible after combining the observed data and the prior distributions, or `fit_vb()`, which approximates the posterior distribution with a simpler one. Indeed, the exact posterior is generally too hard to compute directly.

As mentioned above, the output of the model can be interpreted either on the log-odds scale or on the probability scale. The probability scale is usually more intuitive, since it directly represents the estimated probability that a trial results in success.

For example, if the model predicts $\hat{p}_{ij} = 0.80$, this means that, given the values of the different predictors and the random effect of the current subject, the model estimates an 80% probability of success for that observation. This means that the model predicts probabilities, not binary 0 and 1. To obtain final class labels, we used a threshold of 0.5 to classify the result.

This approach is particularly useful in our experimental setting because it combines two important aspects of the data. First, it models the binary nature of the dependent variable correctly through a binomial distribution and a logit link. Second, it accounts for the repeated-measures structure of the experiment by allowing observations from the same subject to share a subject-specific random effect.

3 Exploratory Data Analysis

Before introducing the modelling part of this work, we present some visualisations of raw data in order to get more familiar with the game and its features.

First, to summarise the amount of available data for the following analyses, Table 1 reports the number of blocks per subject at our disposal. There is a great heterogeneity between subjects in number of blocks (`min = 31, max = 262, mean = 96.4, std = 65.15`), leading to a first note that modelling learning on subjects with fewer trials may be less accurate, as the amount of practice is crucial in motor learning [LS14].

Subject	Number of blocks	Subject	Number of blocks
1	262	11	72
2	230	12	65
3	121	13	64
4	141	14	75
5	40	15	74
6	62	16	58
7	69	17	83
8	54	18	40
9	226	19	73
10	88	20	31

Table 1: Number of blocks per subject.

We then explore the raw AT per block on passed trials only, for all subjects reunited. It can be seen in Figure 6. Note that, for a large block number, the observations only come from a small sample of subjects, as detailed in Table 1, which is one explanation for the smaller distributions for those blocks. Also, as we are only interested in successful trials, the number of observations per block can vary, which can also explain the differences in distributions. Overall, most passed trials were completed within a few seconds, with an average AT of approximately 2 seconds. Nevertheless, several extreme values can be observed, with some successful trials

reaching almost 8 seconds. These long AT indicate that, even among passed trials, some responses were much slower than the typical behaviour. Such observations may correspond to hesitation, distraction, or individual variability. They are also important from a modelling perspective, because extreme AT can increase prediction errors and may be harder for the model to capture accurately.

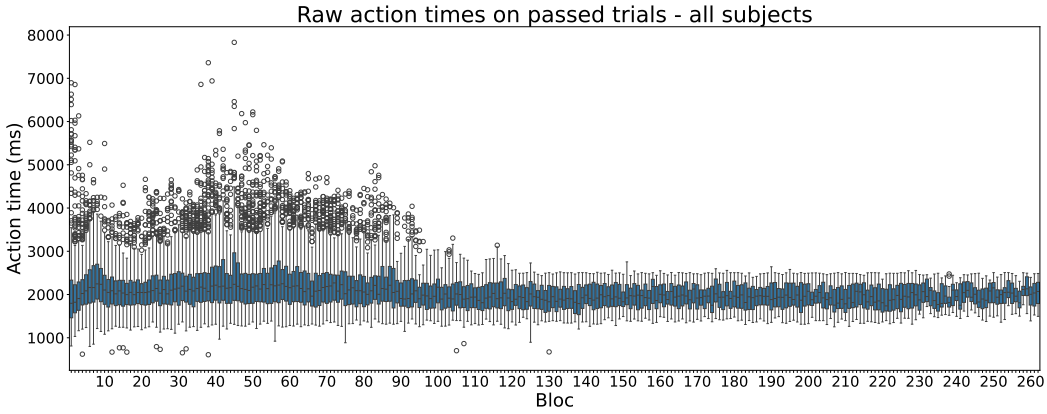

Figure 6: Evolution of the time needed to hit the target for all subjects.

The same visualisations were made for all subjects separately, to see if the AT tends to decrease as the experiment progressed. We chose three different subjects to analyse here, whereas the rest of the plots can be found in Appendix A.
Figure 7 shows the raw AT for subjects 3, 6, and 13, as they showed different behaviours. The same scale was used for all subjects, to make the comparison easier, even if some subjects have drastically smaller AT throughout their sessions compared to others.

- Subject 3's AT starts relatively low, then gradually increases with the block number to remain approximately stable in the middle part, but with high variability within blocks, as shown by the size of the boxplots. Then, it decreases again in later blocks.
- Subject 6's AT shows a big early peak, before decreasing and staying globally stable until the end.
- Subject 13's AT starts low, then has a first increase and decrease pattern, then increases slowly again until the later blocks where it shows a big final increase. Furthermore, the variability also seems bigger near the end, due either to a bigger number of observations per block, or to a bigger variability in AT in general.

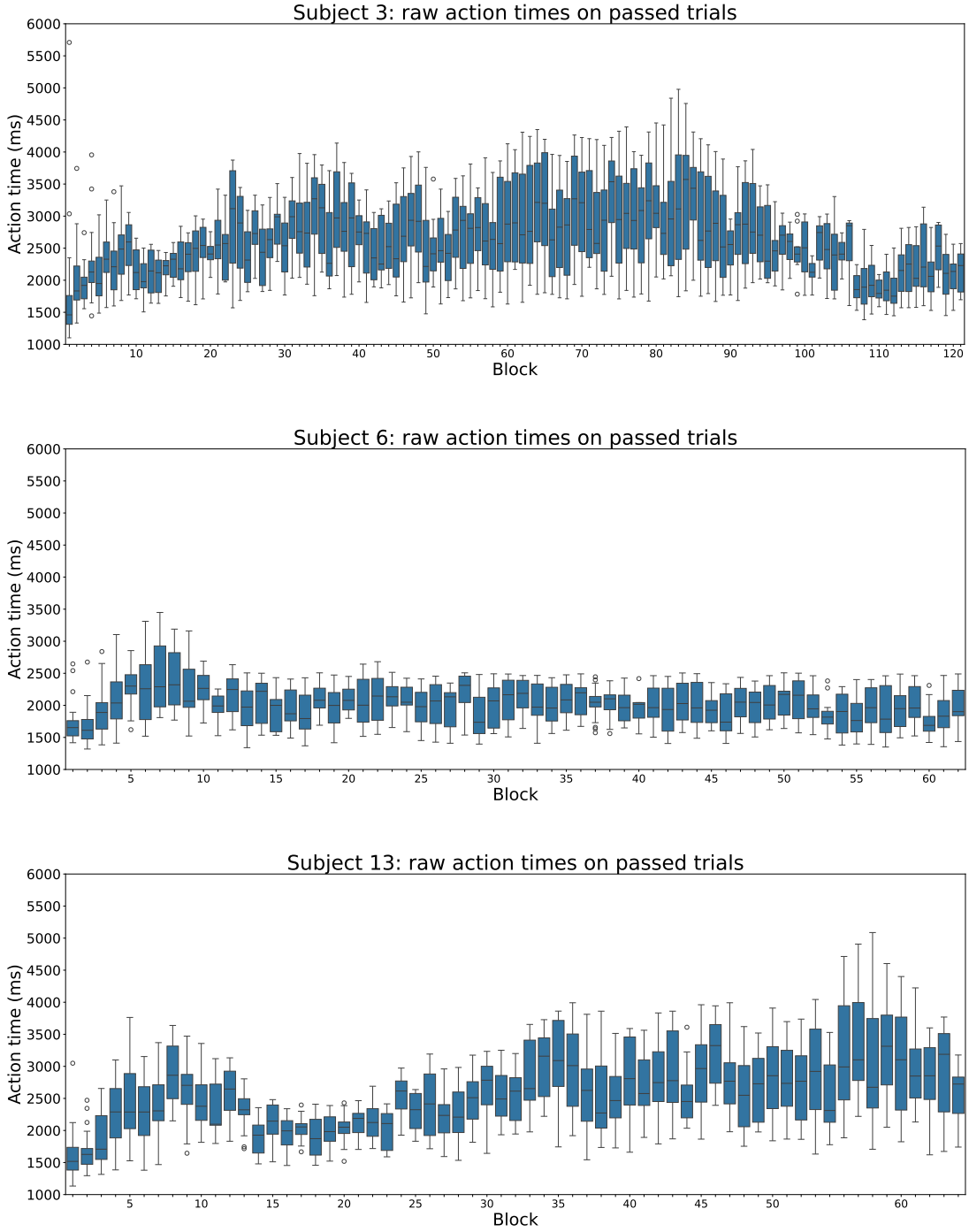

Figure 7: AT on passed trials for subjects 3, 6, and 13.

Overall, subject 6 is much quicker than the two other subjects presented here, and seems to be more consistent during blocks, as indicated by the shorter whiskers around the boxes. More consistency in the context of those boxplots can mean two things: either he did not have a lot of passed trials per block throughout the experiment, or he hit the target with relatively the same AT for each of them.

Table 2 reports the mean raw AT per subject, from the biggest positive difference in AT to the smallest in the two parts of the experiment. The early half is based on the first 50% of each subject's block sequence, while the late one is based on the remaining 50%, and this will be detailed in the Modelling and Validation section. This table describes the observed AT directly, before controlling for variables such as distractors, target position, target lifetime, and block progression.

Subject	Early mean AT (ms)	Late mean AT (ms)	Late − early (ms)
12	2272.21	3394.87	1122.66
13	2247.83	2875.95	628.12
10	2120.95	2662.24	541.29
14	2633.41	2966.18	332.78
16	2679.72	2912.53	232.81
3	2537.03	2690.65	153.62
11	1960.31	2057.89	97.58
1	1939.24	1966.90	27.66
9	1937.56	1956.58	19.01
2	1912.44	1912.88	0.44
4	1958.54	1954.62	-3.92
17	1946.89	1935.02	-11.88
5	2110.19	2078.55	-31.65
15	2019.51	1960.29	-59.21
6	2072.88	1973.10	-99.78
19	2173.27	2068.98	-104.29
18	2249.78	2096.39	-153.38
8	2087.32	1912.68	-174.63
7	2118.09	1903.40	-214.69
20	2580.71	1943.12	-637.60

Table 2: Raw mean AT in the early and late halves of the experiment for each subject. Red cells indicate higher late AT, while green cells indicate lower late AT.

Overall, the raw mean AT was higher in the late half than in the early half for

subjects 12, 13, 10, 14, 16, 3, 11, 1, 9, and 2. However, this raw comparison should not be interpreted directly as evidence that these subjects became slower or did not improve. The late part of the experiment may also contain harder trials. Therefore, raw early–late differences combine at least two effects: changes in subject behaviour and changes in task difficulty. This motivates the use of the LMEMs, which adjusts for several predictors related to the difficulty of the task before estimating the association between block progression and AT.

The same comparison was performed for the raw hit rates. These values are reported in Table 3. Here, the hit rate is computed over all trials, meaning that both omissions and distractor hits are counted as unsuccessful trials.

Subject	Early hit rate	Late hit rate	Late − early
19	0.706	0.529	-0.177
18	0.715	0.548	-0.167
20	0.756	0.604	-0.151
11	0.672	0.536	-0.137
15	0.653	0.529	-0.124
8	0.716	0.634	-0.082
1	0.735	0.659	-0.076
12	0.788	0.725	-0.063
4	0.639	0.586	-0.052
6	0.667	0.633	-0.034
3	0.657	0.625	-0.032
17	0.650	0.623	-0.027
13	0.745	0.733	-0.012
2	0.659	0.669	0.010
9	0.633	0.646	0.012
7	0.659	0.675	0.016
10	0.724	0.746	0.022
16	0.687	0.724	0.037
5	0.704	0.758	0.054
14	0.704	0.766	0.062

Table 3: Raw hit rates in the early and late halves of the experiment for each subject, sorted from the largest decrease to the largest increase in hit rate. Red cells indicate a decrease in hit rate, while green cells indicate an increase.

Several subjects showed increased raw hit rates in the late half, while others showed

decreases. The largest decreases were observed for subjects 19, 18, 20, 11, and 15, whereas the largest increases were observed for subjects 14, 5, and 16. Overall, more subjects showed a lower raw hit rate in the late half than a higher one, which may be due to the adaptive nature of the game.

To have a better idea of the subjects' overall performance, not only hit rates, we counted the number of trials ending in each possible outcome: passed trials, omissions, and distractor hits. By making the analysis per-subject, it allows us to characterise individual performance. These are reported in Table 4.

Subject	Total trials	Passed	Omission	Distractor
1	6288	4383 (69.7%)	1671 (26.6%)	234 (3.7%)
2	5496	3650 (66.4%)	1774 (32.3%)	72 (1.3%)
3	2904	1861 (64.1%)	1020 (35.1%)	23 (0.8%)
4	3384	2072 (61.2%)	1103 (32.6%)	209 (6.2%)
5	960	702 (73.1%)	236 (24.6%)	22 (2.3%)
6	1488	967 (65.0%)	494 (33.2%)	27 (1.8%)
7	1656	1105 (66.7%)	418 (25.2%)	133 (8.0%)
8	1296	875 (67.5%)	338 (26.1%)	83 (6.4%)
9	5424	3469 (64.0%)	1938 (35.7%)	17 (0.3%)
10	2112	1553 (73.5%)	489 (23.2%)	70 (3.3%)
11	1728	1044 (60.4%)	643 (37.2%)	41 (2.4%)
12	1560	1179 (75.6%)	363 (23.3%)	18 (1.2%)
13	1536	1135 (73.9%)	376 (24.5%)	25 (1.6%)
14	1612	1189 (73.8%)	363 (22.5%)	60 (3.7%)
15	1776	1050 (59.1%)	599 (33.7%)	127 (7.2%)
16	1392	982 (70.5%)	343 (24.6%)	67 (4.8%)
17	1992	1268 (63.7%)	500 (25.1%)	224 (11.2%)
18	960	606 (63.1%)	336 (35.0%)	18 (1.9%)
19	1752	1080 (61.6%)	654 (37.3%)	18 (1.0%)
20	744	504 (67.7%)	212 (28.5%)	28 (3.8%)
Total	46060	30674 (66.6%)	13870 (30.1%)	1516 (3.3%)

Table 4: Number and proportion of trials ending in each possible outcome for each subject. Darker shades correspond to higher proportions within each outcome category.

Globally, most of the trials resulted in hitting the target (66.6%). Among un-

successful trials, the majority were caused by omissions (30.1%) compared to hitting a distractor (3.3%). This suggests that the difficulty came mainly from failing to find or reach the target on time, rather than mistaking it with a distractor.

At the subject-level proportions, we can see a lot of variability. Some subjects, such as subjects 5, 10, 12, 13, 14, and 16, have more than 70% of their trials resulting in correctly hitting the target. In contrast, the proportions of omissions for some subjects are relatively high. For example, subjects 3, 9, 11, 18, and 19 have more than 35% of their personal trials resulting in omissions. In the global analysis, we observed that the proportion of distractors hit was small, while this subject-level analysis shows that it is not the case for subjects 4, 7, 8, 15, and 17, who have a higher distractor-hit rate than the global average.

We also wondered if the target's location on the ground had an impact on the outcomes of the trials. Before any other analysis, we first made sure that the distribution of target appearances was uniform, to be sure that every position was more or less equally present, and to be able to use normalised outcome rates instead of raw counts. This was indeed the case: the number of target appearances ranged from 1885 times to 1927 times across the grid. The visual heatmap can be found in Figure A.1 of Appendix A.

Then, we created heatmaps representing the ground layout, where each cell shows the proportion of trials with a given outcome, proportionally to the target's appearance at that location. This means that, for a particular location on the ground, the sum of the proportions of the three outcomes is one. We did it first globally (for all subjects together), and for the three possible outcomes of trials separately. The visualisations are shown in Figure 8 and suggest that the target's location has a big impact on the result of the trial.

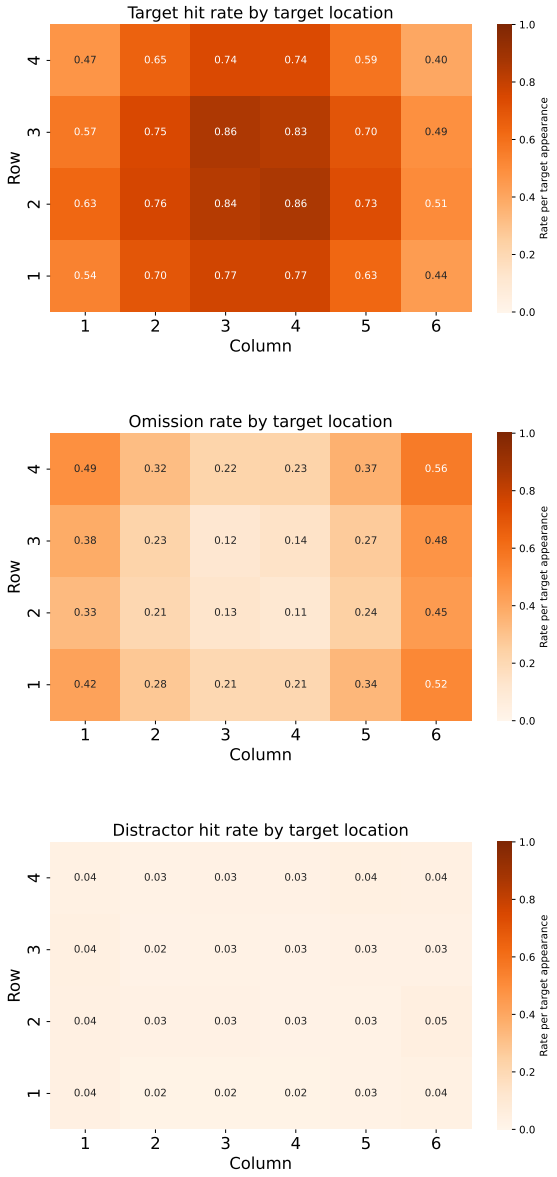

Figure 8: Outcome rates by target location for all subjects.

When proportions are computed relative to the number of times the target appeared at each location, the target seems to be hit more often near the centre of the ground. This could mean that subjects perform better when the target appears in the centre of their visual field. Conversely, omissions happen more often in the first and sixth columns of the ground, on the boundary of the visual field. Finally, there is no clear pattern for hitting distractors when taking into account the target's location in the trials, maybe due to the small number of trials ending in hitting a distractor

compared to the two other outcomes.

As for the raw AT visualisation, we chose three relevant subjects, based on the previous proportions analysis, to show their different subject-level heatmaps. We chose subjects 12, 19, and 17, as they were the subjects with the highest proportions of respectively passed trials, omissions, and distractor-hit trials. The other subjects' visualisations can be found in Appendix B.

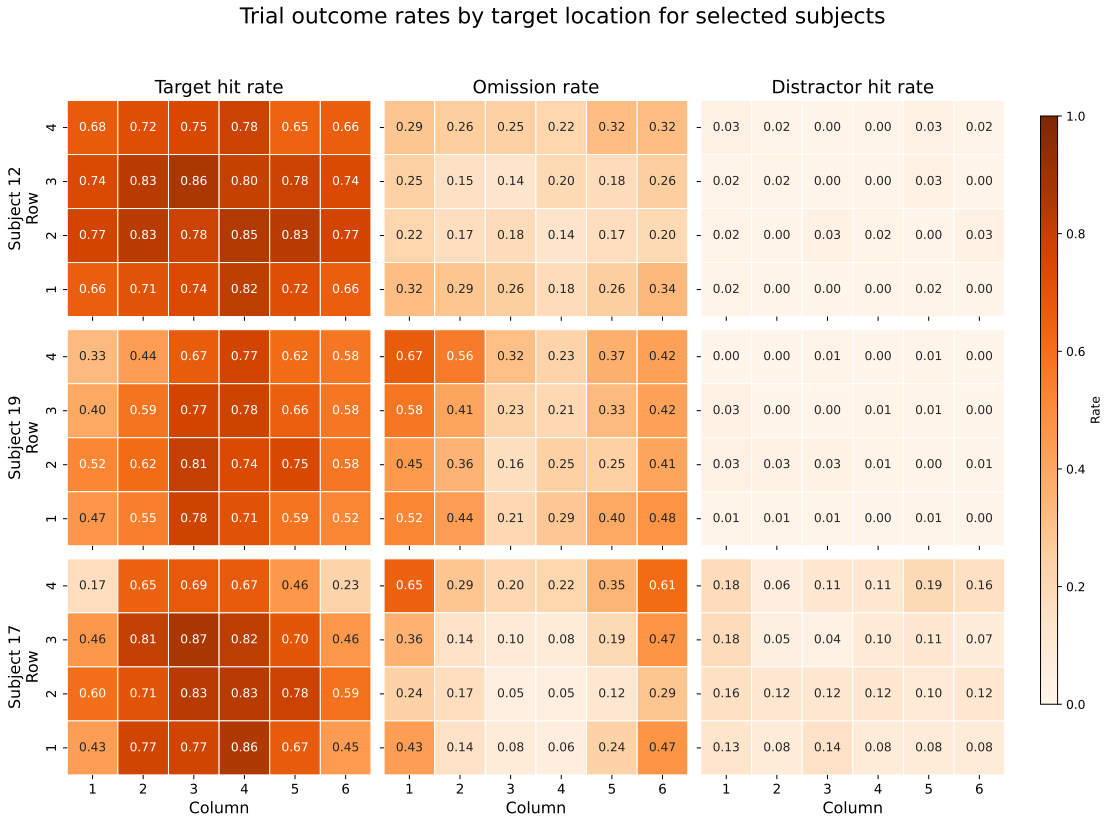

Figure 9: Trial outcome rates by target location for subjects 12, 19, and 17.

Figure 9 shows interesting results, as it shows the subject's specific behavioural profile: successful target hits for subject 12, omissions for subject 19, and distractor hits for subject 17.

Starting with subject 12, who has approximately 75% of trials resulting in a correct target hit, we can see that the proportions fit the global trend. Indeed, the target hit rate is higher whenever the target appears near the centre of the ground field. Furthermore, on the left side, subject 12 shows higher success rates in the fourth row, so further away than targets appearing closer to him, and higher success rates

in the first row on the right side.

Subject 19 has a high omission rate (37%), and is also following the global trend, with proportions higher when the target appears on the outer-most columns, even more on the left side. More than 50% of the time, whenever targets appear at the top left or just below/next to it, subject 19 tends to omit them.

Finally, subject 17 has the highest distractor hit rate of 11.2%, and tends to hit distractors when the target appears further away from him, particularly on the fourth row. Indeed, the top corners only show 0.17 and 0.23 of target hit rates, mostly explained by omissions but with a non-negligible distractor hit rate (0.18 and 0.16).

We next investigated where the distractors were located when they were hit, and if the results depended on the type of distractor that was hit. We thus did some heatmaps where distractors were separated by type. However, because the dataset did not contain them, we could not normalise the counts with the number of times a distractor appeared at that particular location. Indeed, the location was only recorded when a distractor was hit, not for all trials. This is why the results on Figure 10 are interesting but not truly reliable, as distractors could be appearing on particular locations significantly more often than others, so higher counts may reflect either a greater tendency to hit distractors at those locations, or that more distractors appeared there compared to other locations. Nevertheless, we observe that all three types of distractors were mistaken for the target, even if the red ones were the most hit. Also, all three types of distractors have their highest count of hits in the same particular location (first row, fourth column).

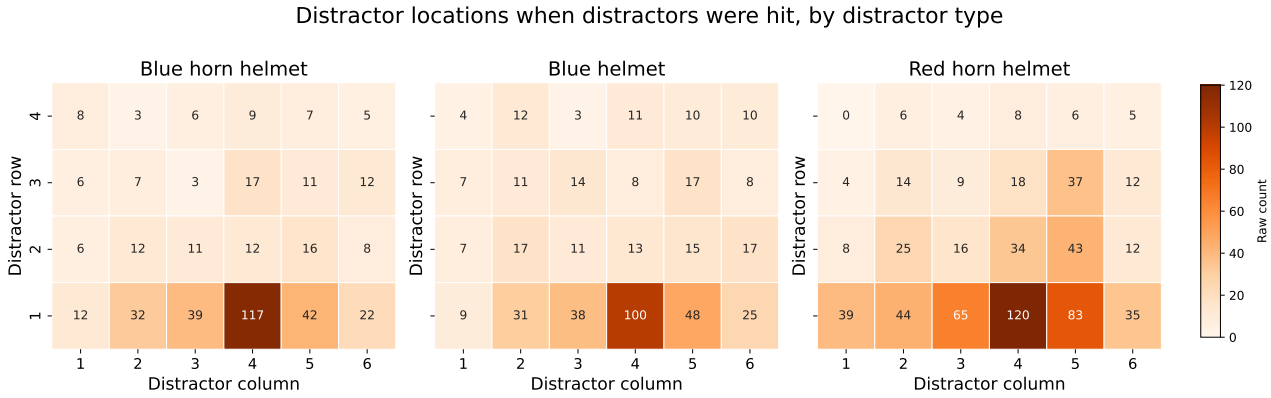

Figure 10: Number of distractor hits at each distractor location, for all subjects and by type.

4 Modelling and Validation

4.1 Data Preprocessing

For all 20 subjects of the experiment described in the Experiment Overview subsection, the data were first preprocessed in order to have continuous block numbering for each subject, as this was not the case before (the block number was reset multiple times during the experiment). For it to be more flexible, we also standardised files and column names across subjects.

Several features were gathered during the experiment and those used in this work are described in Table 5.

Feature	Type	Description
subject	integer	Subject identifier (from 1 to 20)
block	integer	Block index (24 trials per block)
trial	integer	Trial index within block
result	categorical	Passed, Omission or Distractor
time	continuous	Action time (milliseconds)
target lifetime	continuous	Target lifetime (seconds)
target position	integer	Row & column of target
distractor type	categorical	Type of distractor hit (if any)
distractor position	integer	Row & column of the distractor hit (if any)
number distractors	integer	Number of distractors present (by category)
cue	categorical	Cue type (no cue, sound, animation, or arrow)

Table 5: Features gathered in the experiment.

The variables were then checked for missing values, but there were none in the variables we used. Categorical variables were also recoded into a new format, which was better for the modelling part. In particular, the cue variable was transformed using one-hot encoding, so that every cue was separate, resulting in binary indicators for the sound cue, animation cue, and arrow cue, with the absence of cue used as the reference condition. The trial outcome was also changed to match each analysis: for the AT model, only passed trials were retained, whereas for the binary

hit model, successful target hits were coded as 1 and omissions or distractor hits were coded as 0. The number of blue distractors and the number of blue horn distractors were always identical in the dataset, so including both variables would create perfect collinearity and make the design matrix singular. We decided to keep only the blue horn distractor variable in the models.

For the binary hit model, a previous-hit variable was also created to track if the preceding trial of the same subject was successful or not. For the first trial of each subject, for which no previous trial was available, the previous-hit value was set to 0.

4.2 Action Time

4.2.1 Model Specification

The objective of the AT model was to characterise the time required by a subject to complete a successful trial, by taking into account the variables related to the difficulty of the game as well as the progression of the experiment. We used LMEMs to achieve this objective. Let AT_{ij} denote the action time, in milliseconds, for observation i of subject j. Since the AT distribution was right-skewed, as shown in Figure 11, we log-transformed it before modelling to better approximate the assumption of normality of residuals of the model:

$$Y_{ij} = \log(AT_{ij}). \tag{17}$$

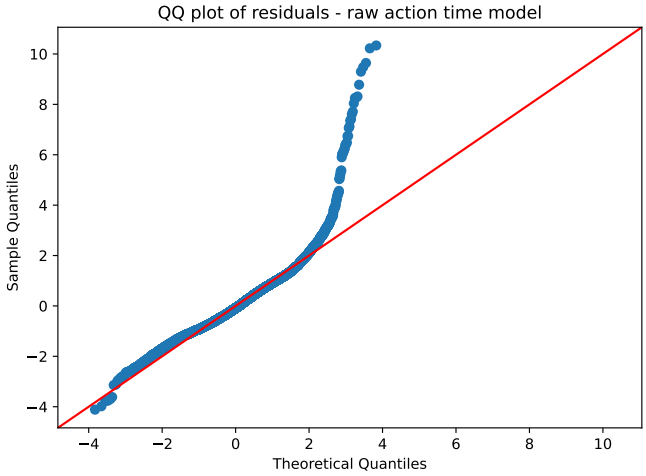

Figure 11: Q–Q plot of residuals from the LMEMs fitted on raw AT.

The residual diagnostics, shown in Figure 12, supported the use of the log transformation. The Q–Q plots of the residuals showed that most points followed the theoretical normal line, especially in the central part of the distributions.

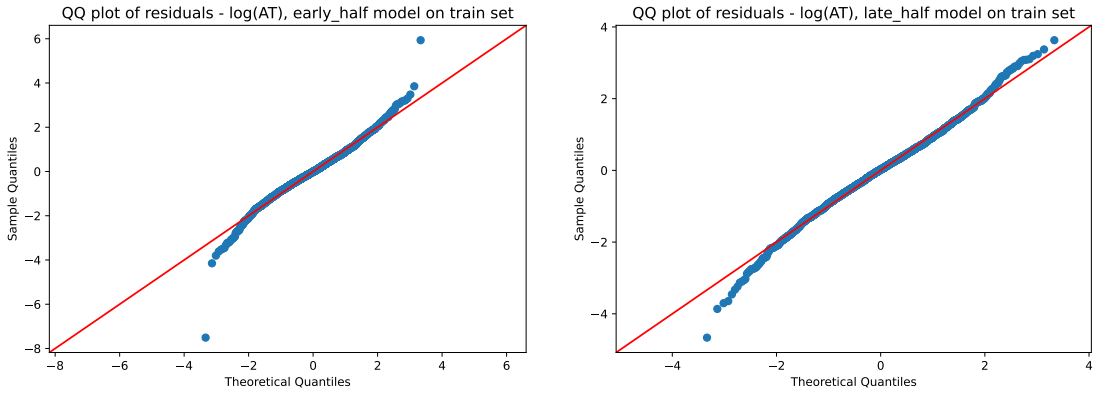

Figure 12: Q–Q plots of residuals for the early and late models in the training set fitted on log-transformed AT.

To investigate learning effects, two separate models were trained. The first model was fitted on the first half of each subject's block sequence, corresponding to the early part of the game, and is referred to as the *early model*. The second model was fitted on the second half of each subject's block sequence, corresponding to the later part of the game, and is referred to as the *late model*. This split was performed separately for each subject, since the total number of blocks was not identical across subjects.

Before fitting the final models, we performed a feature-selection step by fitting the LMEMs including all available predictors. To reduce the dependence of this selection on one particular train–test split, the candidate models were fitted 50 times with different random seeds, as described in the Test Procedure part. Feature relevance was assessed by considering the magnitude of the estimated coefficients and the frequency with which their p-values indicated statistical significance. Based on the results shown in Table 6, we discarded the *Trial* predictor, since it was not significant in either model and was associated with a small coefficient.

	Early model			Late model		
Variable	Mean coef.	SD	Range	Mean coef.	SD	Range
Intercept	7.5931***	0.0083	$^{+0.0150}_{-0.0141}$	7.5705***	0.0059	$^{+0.0111}_{-0.0115}$
Block	−0.0121**	0.0015	$^{+0.0025}_{-0.0028}$	−0.0477***	0.0026	$^{+0.0044}_{-0.0048}$
Arrow cue	−0.1864***	0.0153	$^{+0.0265}_{-0.0307}$	−0.0136	0.0059	$^{+0.0101}_{-0.0111}$
Sound cue	0.0250***	0.0047	$^{+0.0085}_{-0.0084}$	0.0044	0.0034	$^{+0.0058}_{-0.0063}$
Animation cue	−0.0277*	0.0082	$^{+0.0138}_{-0.0138}$	−0.0110	0.0051	$^{+0.0074}_{-0.0106}$
Red horn distractor	0.0809***	0.0033	$^{+0.0054}_{-0.0060}$	0.0946***	0.0044	$^{+0.0070}_{-0.0080}$
Blue horn distractor	0.0038	0.0031	$^{+0.0055}_{-0.0058}$	0.0352***	0.0028	$^{+0.0050}_{-0.0060}$
Lifetime	0.0683***	0.0042	$^{+0.0079}_{-0.0066}$	0.0620***	0.0033	$^{+0.0063}_{-0.0067}$
Target row	0.0259***	0.0024	$^{+0.0039}_{-0.0041}$	0.0270***	0.0018	$^{+0.0033}_{-0.0037}$
Target column	0.0063	0.0016	$^{+0.0035}_{-0.0028}$	0.0127***	0.0013	$^{+0.0023}_{-0.0028}$
Trial	−0.0005	0.0001	$^{+0.0002}_{-0.0003}$	−0.0001	0.0001	$^{+0.0002}_{-0.0003}$
Subject Variance	0.0049	0.0004	$^{+0.0011}_{-0.0009}$	0.0043	0.0004	$^{+0.0007}_{-0.0003}$

Table 6: Average coefficient estimates of the AT LMEMs across 50 random train–test splits. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

The target row and target column were treated as continuous predictors, so they are represented as ordered spatial coordinates rather than as purely different categories. It therefore assumes that the effect of target location, on rows and columns, can be explained by a linear relationship. This is a simplification, but it has the advantage of keeping the model easier to interpret, since only two coefficients are required to adjust for target position.

As a sensitivity analysis, we also fitted the same LMEMs by treating target row and target column as categorical predictors. This alternative specification allowed each row and column to have its own effect relative to the reference position `(1,1)`. The categorical model showed that the spatial effect was not perfectly linear, in particular because the last row and some columns had distinct effects. However, the main conclusions of the model were not changed: the block effect remained negative, red horn distractors and target lifetime remained positively associated with AT, and the comparison between the early and late models remained similar. Moreover, the gain in predictive performance was limited compared with the increase in model complexity.

We also considered a more complex RE structure by allowing the effect of the block variable to vary between subjects, by including a random slope on this variable.

This specification was tested because subjects may not all progress at the same rate during the experiment. However, the random-slope models did not improve predictive performance in the validation analyses and led to a more complex model structure, where the results can be found in Appendix E. The model appeared to fit subject fluctuations more closely without improving generalisation, suggesting possible overfitting on their data.

For these reasons, and in order to keep the model simpler and more interpretable, we retained the continuous specification of target row and target column in the final model, as well as only the per-subject random intercepts and not the random slopes.

The same set of predictors was then kept for both the early and late models, in order to make the two models directly comparable. This is why predictors that were significant in only one of the two models were retained. The row `Subject Variance` corresponds to the estimated random-intercept variance and is not a FE coefficient. The final variables of the FE part described the difficulty and progression of the game:

- the number of red horn distractors,
- the number of blue horn distractors,
- the visual arrow cue,
- the sound cue,
- the animation cue,
- the lifetime of the target,
- the target row,
- the target column,
- the block number.

The block variable was included as a metric to keep track of the progression in the experiment.

The model was implemented in *Python* using the `MixedLM` class from the `statsmodels` library. The parameters were estimated by maximum likelihood, where the final model formula was:

$$\begin{aligned}\texttt{log\_AT} \sim \texttt{block} + \texttt{Arrow\_cue} + \texttt{Sound\_cue} + \texttt{Animation\_cue} + \texttt{\#\_dist\_red\_horn} \\ + \texttt{\#\_dist\_blue\_horn} + \texttt{lifetime} + \texttt{target\_row} + \texttt{target\_col},\end{aligned} \tag{18}$$

with `subject` used as the grouping variable for the random intercept.

Before fitting the model, the data were aggregated within each block, which reduces trial-level noise and limits the influence of outliers. Each block contained 24 trials; however, as we only retained successful trials for this model, the number of observations per block could vary and be lower than 24. This aggregation thus takes into account the actual number of observations, and we defined the chunks with a balanced length of at most 6 observations. When the number of observations was not a multiple of 6, the observations were distributed as evenly as possible across chunks, in order to avoid very small final chunks. For each subject, block, and chunk, we computed the mean of the AT and of the numerical predictors. This produced a maximum of 4 observations per block instead of 24, while preserving the temporal structure of the game.

4.2.2 Test Procedure

The standard 80–20 procedure was used to evaluate whether the model could predict unseen observations from subjects already represented in the training set. The split was performed at the subject-block pair level. This ensured that all observations belonging to a given block of a given subject were assigned entirely either to the training set or to the test set. This avoids leakage between training and test sets due to repeated observations within the same block.

For each early or late dataset, 80% of the subject-block pairs were randomly assigned to the training set and the remaining 20% to the test set. This random split was repeated 50 times using different random seeds.

All continuous predictors were standardised before fitting the model by using the `StandardScaler` class from the *Scikit-Learn* library. To prevent data leakage, the scaler was fitted only on the training set and then applied to the test set. This ensured that no information from the test set was used during preprocessing.

Since the model was trained on the logarithm of the AT, predictions were first obtained on the log scale. Let $\hat{Y}_{ij}$ denote the predicted log-AT. To evaluate the model in the original unit (milliseconds), predictions were transformed back to the real AT scale. Because the exponential of a normally distributed variable follows a log-normal distribution, a bias correction was applied:

$$\widehat{AT}_{ij} = \exp\left(\hat{Y}_{ij} + \frac{1}{2}\hat{\sigma}^2\right), \tag{19}$$

where $\hat{\sigma}^2$ is the estimated residual variance of the Linear Mixed-Effects Model (LMEM). This correction estimates the expected AT on the original scale rather

than the median AT, which corresponds to the exponential without correction [Maj22].

Model performance was assessed using both the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). On the original AT scale, the RMSE was computed as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(AT_i - \widehat{AT}_i\right)^2}. \tag{20}$$

The RMSE measures the average magnitude of the prediction error, while giving more weight to large errors because the residuals are squared. It is therefore particularly sensitive to poorly predicted observations and extreme AT.

To complement this measure, the MAE was also computed on the original AT scale as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|AT_i - \widehat{AT}_i\right|. \tag{21}$$

The MAE measures the average absolute prediction error (in milliseconds). Compared with the RMSE, it is less sensitive to extreme errors and therefore provides a more direct interpretation of the typical prediction error.

Both metrics were computed globally over all chunks of each set for each split. The reported values correspond to the mean of the RMSE and MAE across the 50 repeated splits. In addition, per-subject test RMSE and MAE values were computed for each split and then averaged across splits. This subject-level analysis was used to assess whether the global error was representative of all subjects, especially since the number of available blocks differed substantially between subjects.

4.2.3 Action Time Results

The AT model was evaluated using the repeated 80–20 test procedure described above. The results averaged across the 50 random splits are reported in Table 7.

Model	Train RMSE (ms)	Test RMSE (ms)	Test RMSE (log)	Test MAE (ms)
Early model	246.86	246.29	0.109	176.97
Late model	229.15	228.81	0.092	163.15

Table 7: Mean repeated 80–20 test performance of the AT LMEMs across 50 random splits.

Both models achieved similar train and test errors, suggesting that there was no strong evidence of overfitting in this train and test sets separation choice. The model trained on the second half of the experiment obtained a lower test error than the model trained on the first half, with an RMSE of 228.81 ms compared with 246.29 ms. As expected, the RMSE also decreased on the log scale, from 0.109 in the early model to 0.092 in the late model. The test MAE also decreased from 176.97 ms to 163.15 ms. This suggests that AT may be more predictable in the later part of the experiment, where one plausible reason is that subjects became more consistent as the experiment progressed, or that the difficulty was more stable. It can also be due to the difference in variability of AT between the two halves.

	Early model			Late model		
Variable	Mean coef.	SD	Range	Mean coef.	SD	Range
Intercept	7.5874^{***}	0.0083	$^{+0.0167}_{-0.0147}$	7.5699^{***}	0.0060	$^{+0.0107}_{-0.0114}$
Block	-0.0121^{**}	0.0015	$^{+0.0025}_{-0.0028}$	-0.0477^{***}	0.0026	$^{+0.0044}_{-0.0048}$
Arrow cue	-0.1867^{***}	0.0154	$^{+0.0266}_{-0.0307}$	-0.0136	0.0059	$^{+0.0101}_{-0.0112}$
Sound cue	0.0250^{***}	0.0047	$^{+0.0084}_{-0.0084}$	0.0044	0.0034	$^{+0.0058}_{-0.0063}$
Animation cue	-0.0277^{*}	0.0082	$^{+0.0139}_{-0.0139}$	-0.0110	0.0051	$^{+0.0074}_{-0.0106}$
Red horn distractor	0.0810^{***}	0.0033	$^{+0.0054}_{-0.0060}$	0.0946^{***}	0.0044	$^{+0.0070}_{-0.0080}$
Blue horn distractor	0.0039	0.0031	$^{+0.0055}_{-0.0059}$	0.0352^{***}	0.0028	$^{+0.0050}_{-0.0060}$
Lifetime	0.0684^{***}	0.0042	$^{+0.0079}_{-0.0066}$	0.0620^{***}	0.0033	$^{+0.0063}_{-0.0067}$
Target row	0.0260^{***}	0.0024	$^{+0.0039}_{-0.0041}$	0.0271^{***}	0.0019	$^{+0.0034}_{-0.0037}$
Target col	0.0062	0.0016	$^{+0.0035}_{-0.0028}$	0.0127^{***}	0.0013	$^{+0.0023}_{-0.0028}$
Subject Variance	0.0049	0.0004	$^{+0.0011}_{-0.0009}$	0.0043	0.0004	$^{+0.0007}_{-0.0003}$

Table 8: Average coefficient estimates of the AT LMEMs across the 50 random 80–20 splits. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

The estimated coefficients for both models are reported in Table 8. Since the dependent variable is log-transformed AT, negative coefficients indicate shorter predicted AT, while positive coefficients indicate longer predicted AT. In both halves of the experiment, the coefficient of the block variable was negative and statistically significant, indicating that AT tended to decrease as the experiment progressed and after accounting for other variables and subject-specific baseline differences. This effect was stronger in the late model than in the early model, suggesting a stronger acceleration or learning effect later in the experiment.

This should be distinguished from the raw action times reported in Table 2. Some subjects had higher observed AT in the late half than in the early half, but this does not necessarily mean that they failed to improve. The negative block coefficient means that, for trials with comparable observed difficulty variables, and after accounting for subject-specific baseline differences, AT tended to decrease as block number increased.

The number of red distractors with horns had a positive and highly significant effect in both models, indicating that trials containing more red horn distractors were associated with longer AT. The number of blue horn distractors was positive but not significant in the early model, then became significant in the late model. The target lifetime also had a positive and significant effect in both models. This does not necessarily mean that a longer lifetime made the task harder. It may reflect the fact that targets staying longer above ground allow slower successful responses to occur, whereas targets with lower lifetimes must be hit quickly to be counted as passed trials.

The arrow cue showed a strong negative association with AT in the first half, but this effect disappeared in the second half. This may suggest that the arrow cue was particularly useful during the early part of the experiment, when subjects were still adapting to the game. The sound cue showed a small positive effect in the early model, but was no longer significant in the late model. This positive coefficient may be interpreted in the same way as the target lifetime, where cues were given in conditions where longer responses were possible. The animation cue showed a small negative effect in the early model, but this effect also disappeared in the late model. This difference in the cue coefficients may be due to fewer cues in the later part than in the early part of the experiment. Finally, target row and target column were both positively associated with AT, suggesting that the spatial position of the target also influenced the time needed to hit it.

Since continuous predictors were standardised before model fitting, the coefficients correspond to the effect of a one-standard-deviation increase in the corresponding predictor.

The observed versus mean predicted AT across the 50 repeated splits are shown in Figure 13. Because the train–test split was repeated randomly 50 times, most observations appeared in the training set in some splits and in the test set in others. Therefore, when showing all predictions across all splits, the train and test scatter plots are expected to look very similar.

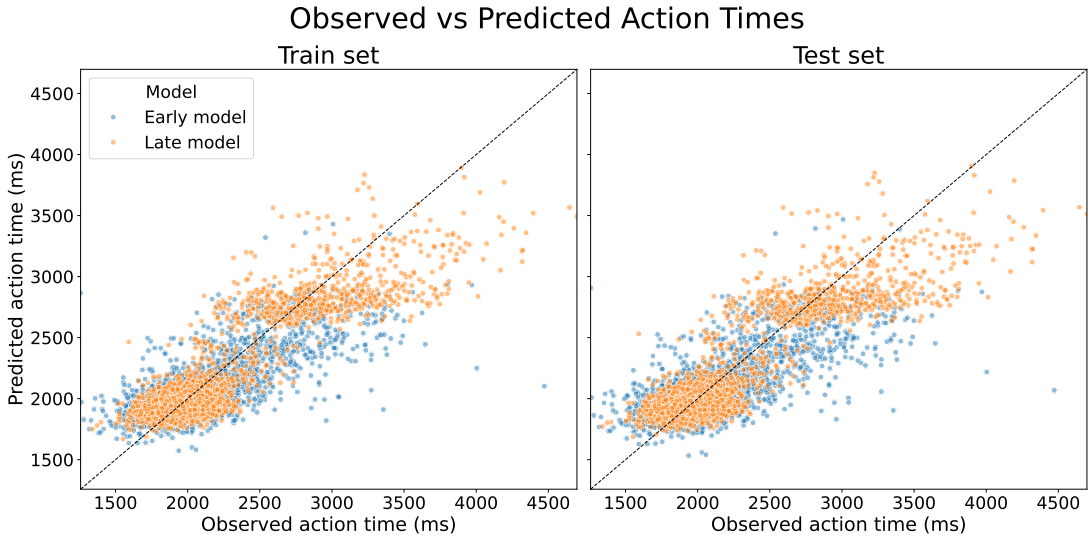

Figure 13: Observed versus predicted AT for the training and test sets across the 50 repeated train–test splits. The dashed line corresponds to perfect prediction.

In both the training and test sets, the predictions are more or less aggregated around the diagonal, suggesting that the predicted AT were generally consistent with the observed AT. The similarity between the training and test plots is also consistent with the RMSE values and indicates that the models did not strongly overfit the training data. Some deviations from the diagonal are still occurring, especially for extreme action times, which shows that unusual responses were harder to predict.

In the late model, some subjects showed higher raw AT values than in the early model, as reported in Table 2. A notable pattern in the late model is the presence of two main clouds of predicted values, with relatively few predictions around approximately 2500 ms. This suggests that the fitted late model separates observations into two regimes: a faster group, with predicted AT around 1800–2200 ms, and a slower or more difficult group, with predicted AT mostly above 2600 ms. This separation may reflect subject heterogeneity, differences in task difficulty, or a combination of both. It could also be due to the great difference in number of blocks between subjects, where blocks from the late model from one subject correspond to blocks from the early model for another (for example, subject 20 has only 31 blocks, so the late model contains blocks 16 to 31, while for subject 1, the late model contains blocks 131 to 262).

As discussed previously, there is a high variability between subjects. We wanted to investigate whether the global test error was representative of all subjects. We computed the test RMSE and MAE separately for each of them, as reported in

Table 9. The prediction error varied a lot across subjects, indicating that the model did not perform equally well for all individuals. Some of the largest RMSE values were observed for subjects with very few test observation chunks, making these subject-level errors less stable and more sensitive to individual poorly predicted chunks. This is the case of subject 20, who only has 11 test chunks in the early model, and subject 18, who only has about 10 test chunks in the late model. For this reason, the table reports the mean number of test observation chunks rather than the mean number of test blocks, because the number of passed trials, and therefore the number of aggregated observations, could differ a lot between blocks.

	Early model			Late model		
Subject	Mean RMSE (ms)	Mean MAE (ms)	Mean test chunks	Mean RMSE (ms)	Mean MAE (ms)	Mean test chunks
1	163.8 ± 21.9	125.0 ± 12.5	91.6	150.2 ± 9.1	120.6 ± 7.4	83.6
2	168.2 ± 14.5	134.2 ± 12.5	66.4	133.7 ± 11.9	107.0 ± 9.8	68.7
3	263.2 ± 34.4	203.3 ± 25.0	36.0	296.2 ± 45.9	233.0 ± 35.7	35.9
4	171.8 ± 20.1	138.2 ± 16.2	41.9	154.1 ± 19.3	123.7 ± 15.6	40.3
5	133.2 ± 49.6	105.4 ± 40.1	12.7	136.7 ± 41.2	109.3 ± 33.9	16.1
6	208.1 ± 48.0	165.7 ± 42.3	18.1	147.4 ± 20.0	120.9 ± 18.5	18.2
7	226.2 ± 82.0	170.3 ± 52.0	18.3	137.3 ± 30.4	110.4 ± 20.2	21.9
8	189.5 ± 46.9	145.2 ± 33.8	17.1	171.2 ± 40.0	138.3 ± 33.2	15.9
9	175.6 ± 10.8	143.1 ± 9.2	66.2	142.0 ± 11.8	112.7 ± 10.3	68.4
10	253.3 ± 51.7	192.8 ± 39.2	27.0	351.6 ± 45.8	284.8 ± 39.2	29.4
11	166.4 ± 33.5	135.5 ± 31.9	20.6	180.2 ± 21.2	151.6 ± 19.5	18.1
12	373.5 ± 95.4	308.1 ± 84.9	20.7	465.7 ± 57.6	372.8 ± 59.2	23.5
13	256.5 ± 53.5	204.5 ± 47.2	22.2	284.8 ± 49.5	231.4 ± 42.3	21.0
14	523.9 ± 161.7	386.6 ± 90.4	18.6	419.5 ± 55.7	329.9 ± 49.8	24.4
15	222.3 ± 42.6	177.1 ± 33.1	22.4	158.7 ± 22.2	127.7 ± 20.3	19.2
16	432.2 ± 55.3	370.5 ± 51.0	17.8	378.5 ± 82.5	300.7 ± 70.1	18.2
17	209.0 ± 76.7	155.3 ± 35.0	23.4	146.1 ± 25.8	118.4 ± 23.4	23.3
18	211.4 ± 72.1	161.4 ± 52.9	12.8	149.6 ± 30.4	123.4 ± 28.4	10.3
19	258.0 ± 65.8	198.6 ± 42.2	22.7	153.6 ± 22.5	125.3 ± 19.8	19.5
20	417.0 ± 113.1	346.3 ± 98.7	11.0	157.0 ± 42.4	126.8 ± 36.8	10.1

Table 9: Per-subject mean test RMSE and MAE for the AT models at the chunk level across 50 train–test splits. Darker red cells indicate larger prediction errors.

Subject 14 was among the subjects with the largest test errors, as also illustrated in Figure 14. This should first be interpreted in relation to the small number of observations available in the test set for this subject, since the repeated splits contained, on average, only around 18 test observation chunks for the early model. Consequently, a few poorly predicted ones can strongly affect the RMSE. This indicates that subject 14 may have had a performance profile that was less well captured by the population-level FE and the subject-specific random intercept.

These subject-level errors should also be interpreted in light of the raw early–late differences reported in Table 2. Since late-half AT may be affected by changes in task difficulty, larger late-half errors for some subjects do not necessarily imply

worse learning, but rather that their behaviour or the difficulty they faced was less well captured by the current FE and random-intercept structure. In general, the errors are bigger for the early model, which confirms previous results.

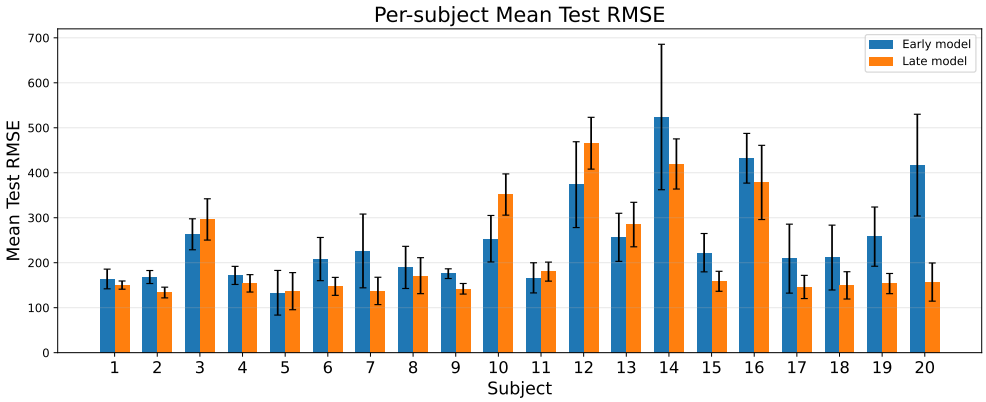

Figure 14: Test RMSE computed separately for each subject, at the chunk level, for the early and late AT models.

To illustrate the behaviour of the 80–20 test procedure, we selected three representative subjects, subjects 1, 12, and 20, and compared their observed and predicted AT on the test set, this time at the block level. Indeed, the AT were this time aggregated within each block for better visualisation. The reported RMSE in the titles is computed between the observed mean AT and the predicted mean AT for each test block, which differs from the aggregated observation level RMSE reported in the global performance table, Table 9. The corresponding plots are shown in Figure 15, whereas the plots for the other subjects can be found in Appendix C. The scale on those plots is not the same for the three subjects, as their personal AT were very different.

Subject 1 was selected because both the early and late models had relatively low test RMSE values for this subject. This subject also completed many more blocks than the others, leading to an average of 91.6 and 83.6 chunks in the test sets, compared with 20.7 and 23.5 for subject 12 and 11.0 and 10.1 test chunks for subject 20. As a result, the curves for subject 1 are based on a larger number of observations and are therefore more stable. Subject 20 was chosen because the test error was high for the early model but much lower for the late model, and subject 12 was selected because both models showed large test errors compared to other subjects. This subject completed 65 blocks, giving more test information than subject 20 but still far fewer blocks than subject 1. For subject 12, the difference between observed and predicted AT remains relatively large in both models, suggesting that this

subject's behaviour was less well explained by the predictors included in the model.

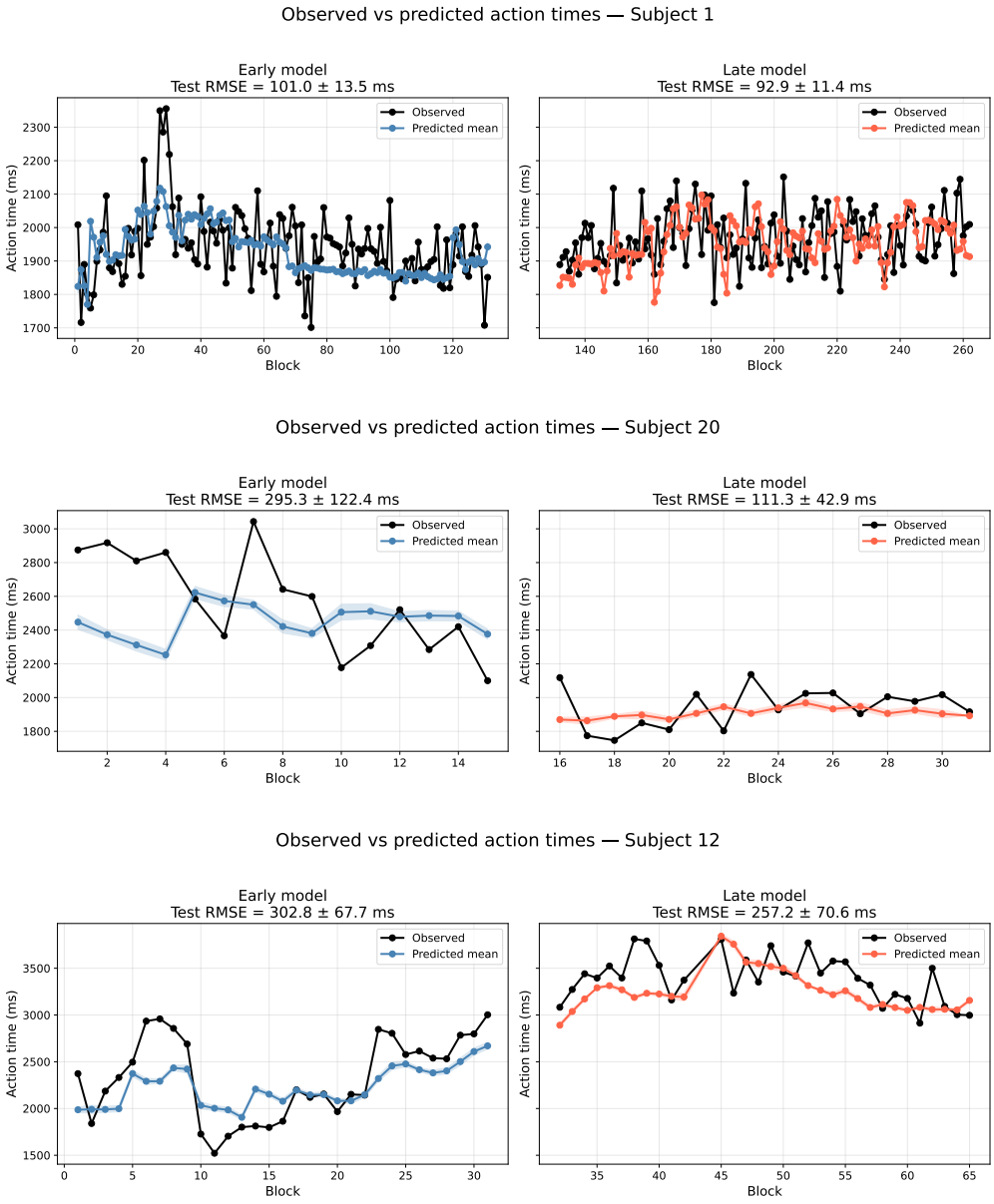

Figure 15: Observed and predicted AT on the test set for three selected subjects in the repeated 80–20 test procedure.

Overall, these examples illustrate that prediction quality varies between subjects, but also that this variability must be interpreted with respect to the number of available blocks.

Since the AT model was fitted only on passed trials, its interpretation is only based on successful hits. Therefore, the model describes how quickly subjects completed successful trials, not the overall speed in the whole experiment. This distinction is important because changes in the probability of success may also affect which and how many trials are included in the AT analysis.

4.3 Binary Hit

4.3.1 Model Specification

In addition to modelling the AT on passed trials, we also wanted to model the probability that a trial resulted in a successful target hit. For this analysis, the response variable was binary. Let H_{ij} denote the outcome of trial i for subject j, defined as:

$$H_{ij} = \begin{cases} 1, & \text{if the target was hit,} \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

Omission trials and distractor-hit trials were grouped together in the unsuccessful class. This choice was supported by the exploratory analysis, which showed that distractor hits represented only a very small proportion of the trials. Therefore, the binary model was mainly designed to distinguish successful target hits from the two other trial outcomes [Smi+04].

The objective of this model was to predict the probability of hitting the target, while accounting for both game difficulty variables and inter-subject variability. Since the response variable is binary, we used a logistic mixed-effects model.

The model was implemented in *Python* using the `BinomialBayesMixedGLM` class from the `statsmodels` library. This class fits a binomial generalised linear mixed model. In this model, similarly to the LMEMs described above, the FE describe the population-level effect of the predictors on the log-odds of hitting the target, while the random intercept captures subject-specific baseline differences. Positive FE coefficients increase the predicted probability of success, whereas negative coefficients decrease it.

As for the AT model, two separate models were fitted: an *early* and a *late* one, based on each subject's block sequence.

A feature selection procedure was also performed for the binary hit model, following the same logic as for the AT model. Predictors that were consistently informative in at least one of the two models were kept, so that the early and late models could be compared using the same set of variables. Since the binary mixed-effects

model was fitted using `BinomialBayesMixedGLM`, significance was assessed using approximate p-values computed from the posterior mean and posterior standard deviation of each FE coefficient. Indeed, the model output only gives posterior estimates, not the usual regression output with the associated p-values. Therefore, to assess the significance of a predictor, we computed a z-statistic, which compares the estimated coefficient with its uncertainty by dividing the posterior mean by the posterior standard deviation:

$$z = \frac{\widehat{\beta}}{\mathrm{SD}(\widehat{\beta})}. \tag{23}$$

A large absolute value of z indicates that the coefficient is far from zero relative to its posterior uncertainty, whereas a value close to zero indicates that the estimate is small compared with its uncertainty. We then converted it into a two-sided p-value by using the standard normal distribution [Agr13]. These p-values should therefore be interpreted as approximate indicators of coefficient stability, rather than as exact p-values as those of the AT model.

	Early model			Late model		
Variable	Mean coef.	SD	Range	Mean coef.	SD	Range
Intercept	-1.7126^{***}	0.1736	$^{+0.3252}_{-0.4640}$	-1.0684^{***}	0.2260	$^{+0.3392}_{-0.5041}$
Block	0.0421	0.0155	$^{+0.0291}_{-0.0310}$	0.1922^{***}	0.0284	$^{+0.0555}_{-0.0718}$
Blue horn distractor	0.0088	0.0312	$^{+0.1108}_{-0.0662}$	-0.0802^{**}	0.0240	$^{+0.0570}_{-0.0561}$
Red horn distractor	-0.5325^{***}	0.0229	$^{+0.0489}_{-0.0480}$	-0.5069^{***}	0.0321	$^{+0.0617}_{-0.0870}$
Animation cue	0.3355^{***}	0.0783	$^{+0.2324}_{-0.1570}$	0.0797	0.0479	$^{+0.0931}_{-0.1453}$
Arrow cue	-1.1067^{***}	0.2479	$^{+0.4375}_{-0.6600}$	0.2417^{**}	0.0586	$^{+0.1200}_{-0.1432}$
Sound cue	0.2470^{***}	0.0308	$^{+0.1031}_{-0.0644}$	0.0343	0.0310	$^{+0.0671}_{-0.0957}$
Lifetime	0.8055^{***}	0.0613	$^{+0.1669}_{-0.1141}$	0.6441^{***}	0.0791	$^{+0.1957}_{-0.1155}$
Previous hit	0.4470^{***}	0.0181	$^{+0.0372}_{-0.0418}$	0.5119^{***}	0.0223	$^{+0.0567}_{-0.0457}$
Target column	-0.0623^{***}	0.0044	$^{+0.0104}_{-0.0111}$	-0.1090^{***}	0.0036	$^{+0.0077}_{-0.0086}$
Target row	-0.0552^{***}	0.0067	$^{+0.0141}_{-0.0149}$	-0.0811^{***}	0.0050	$^{+0.0116}_{-0.0104}$
Trial	0.0184	0.0088	$^{+0.0215}_{-0.0165}$	-0.0425^{*}	0.0075	$^{+0.0137}_{-0.0147}$

Table 10: Average coefficient estimates of the binary hit mixed-effects models across the 50 random 80–20 train–test splits. Coefficients are reported on the log-odds scale. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

The results of the feature selection are shown in Table 10. In this case, all predictors were significant in at least one of the two halves.

The animation, sound, and arrow cues were treated as binary categorical predictors, and the target row and target column were kept as numerical predictors. This was

preferred to a full one-hot encoding, as in the AT models, in order to keep the binary model less complex and easier to compare with the AT model. Continuous predictors were standardised before model fitting. The subject variable was not included as a FE, but was used to define the random intercept. We also added a random slope per subject on the block variable, but as for the AT models, it made the models more complex and worsened the results. The metrics associated with the random slope version can be found in Appendix E.

The final model formula was:

$$\texttt{success} \sim \texttt{block} + \texttt{trial} + \texttt{previous\_hit} + \texttt{sound\_cue} + \texttt{arrow\_cue} + \texttt{animation\_cue} + \texttt{dist\_red\_horn} + \texttt{dist\_blue\_horn} + \texttt{lifetime} + \texttt{target\_row} + \texttt{target\_col}.$$

4.3.2 Test Procedure

For the standard test procedure, an 80–20 split was performed at the subject-block level, exactly as for the AT models, with 50 repetitions with different random splits.

Although the model was fitted at the trial level, the evaluation was also performed at the block level. Trial-level classification accuracy is not always the most informative metric in this context, because the exact outcome of a single trial may be highly variable. Instead, we were also interested in whether the model could predict the overall success rate of a block. Therefore, for each subject and block, we computed:

$$\widehat{p}_{\text{block}} = \frac{1}{n_{\text{block}}} \sum_{i \in \text{block}} \widehat{p}_{ij}, \tag{24}$$

where n_{block} is the number of observations in the block. We compared it with the observed hit rate:

$$p^{obs}_{\text{block}} = \frac{1}{n_{\text{block}}} \sum_{i \in \text{block}} H_{ij}. \tag{25}$$

The block-level prediction error was then evaluated using the RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left(p_b^{obs} - \widehat{p}_b \right)^2}, \tag{26}$$

where B is the number of test blocks, p_b^{obs} is the observed hit rate in block b, and $\widehat{p}_b$ is the predicted hit probability for that block. This RMSE is expressed in probability units. For example, an RMSE of 0.10 means that the predicted hit probability differs from the observed block-level hit rate by approximately 0.10.

The block-level MAE was also computed:

$$\text{MAE} = \frac{1}{B} \sum_{b=1}^{B} \left| p_b^{obs} - \widehat{p}_b \right| . \tag{27}$$

The MAE measures the average absolute difference between the observed and predicted block-level hit rates. Like the RMSE, it is expressed in probability units.

In addition to the block-level RMSE and MAE, three trial-level metrics were computed: the Area Under the Curve (AUC), the Brier score, and the accuracy [Ste+10]. These metrics provide complementary information. The AUC evaluates how well the model ranks successful trials above unsuccessful trials. The Brier score evaluates the quality of the predicted probabilities. Accuracy, in contrast, evaluates the final binary classification after applying a probability threshold.

The AUC score measures how well the model discriminates between successful and unsuccessful hits. It can be interpreted as the probability that the model assigns a higher predicted probability to a randomly chosen successful hit than to a randomly chosen failed trial:

$$\text{AUC} = P\left(\widehat{p}_{\text{hit}} > \widehat{p}_{\text{miss}}\right), \tag{28}$$

where $\widehat{p}_{\text{hit}}$ is the predicted probability for a randomly chosen successful hit and $\widehat{p}_{\text{miss}}$ is the predicted probability for a randomly chosen failed trial. A value of 0.5 corresponds to random discrimination, whereas a value of 1 corresponds to perfect discrimination. The AUC does not directly assess whether the predicted probabilities are well calibrated. Instead, it evaluates whether the model correctly ranks successful trials above failed trials.

The Brier score measures the mean squared difference between the observed binary outcome and the predicted probability of success. It is defined as [Gle+50]:

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^{N} \left(H_i - \widehat{p}_i\right)^2 , \tag{29}$$

where $H_i \in \{0, 1\}$ is the observed hit outcome and $\widehat{p}_i$ is the predicted probability of a successful hit. A lower Brier score indicates better probabilistic predictions. Since it is based on squared errors, it can be interpreted as measuring how close the predicted probabilities are to the observed binary outcomes.

Accuracy measures the proportion of correctly classified trials after converting predicted probabilities into binary class predictions. In this work, a trial was

classified as successful when the predicted probability was at least 0.5:

$$\widehat{H}_i = \begin{cases} 1, & \text{if } \widehat{p}_i \geq 0.5, \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

The accuracy is then defined as:

$$\text{Accuracy} = \frac{\text{number of correctly classified trials}}{\text{total number of trials}}. \tag{31}$$

Accuracy is easy to interpret, but it depends on the chosen threshold and does not indicate whether the predicted probabilities are well calibrated.

Together, the AUC, the Brier score, and the accuracy provide complementary trial-level information: discrimination, probabilistic prediction quality, and classification performance. The block-level RMSE and MAE, on the other hand, evaluate whether the model can predict the average hit rate at the block level.

4.3.3 Binary Hit Results

The trial-level prediction results are reported in Table 11.

Model	Set	AUC	Brier score	Accuracy
Early model	Train	0.688 ± 0.004	0.195 ± 0.001	0.689 ± 0.003
Early model	Test	0.684 ± 0.014	0.196 ± 0.005	0.687 ± 0.011
Late model	Train	0.639 ± 0.003	0.215 ± 0.001	0.661 ± 0.003
Late model	Test	0.630 ± 0.012	0.217 ± 0.004	0.658 ± 0.009

Table 11: Mean performance of the binary hit mixed-effects models across the 50 random 80–20 train–test splits.

The trial-level AUC values were above 0.5, indicating that the models ranked successful trials above unsuccessful trials better than chance. However, the discrimination ability remained moderate. The early model obtained a test AUC of 0.684, while the late model obtained a lower value of 0.630. This suggests that successful and unsuccessful trials were only partially separable from the available predictors, and that this separation was weaker in the second half of the experiment.

The Brier test scores were also moderate, since lower Brier scores are better, with test values of 0.196 for the early model and 0.217 for the late model. The accuracy values followed the same pattern, with a test accuracy of 0.687 for the

early model compared with 0.658 for the late model. These indicate that the models correctly classified around two-thirds of the trials, which is useful but not perfect.

For both models, the train and test values were very similar, suggesting that the models did not strongly overfit the training data. Therefore, the main limitation does not appear to be overfitting, but rather the limited ability of the available predictors to fully explain trial-level success and failure. Overall, the early model performed slightly better than the late model at the trial level.

Model	Set	Block RMSE	Block MAE	Bias
Early model	Train	0.119 ± 0.003	0.087 ± 0.001	-0.001 ± 0.000
Early model	Test	0.126 ± 0.011	0.092 ± 0.006	-0.001 ± 0.009
Late model	Train	0.128 ± 0.003	0.091 ± 0.001	-0.001 ± 0.000
Late model	Test	0.134 ± 0.013	0.094 ± 0.006	-0.003 ± 0.010

Table 12: Block-level prediction performance of the binary hit mixed-effects models across the 50 random 80–20 train–test splits.

At the block level, the results are shown in Table 12. For the early model, the test block-level RMSE was 0.126, while the late model obtained a slightly higher test block-level RMSE of 0.134. The block-level MAE showed a similar pattern, with a test value of 0.092 for the early model and 0.094 for the late model.

The bias was close to zero in both cases, suggesting that the predicted block hit rates were well calibrated on average. Since the bias was defined as observed hit rate minus predicted hit probability, the small negative values indicate a slight overestimation of the hit rate. For both models, the train and test errors were relatively close, suggesting again that the models did not strongly overfit the training data. Overall, the early model again performed slightly better than the late model, although the difference remained small at the block level.

The estimated coefficients for the final binary hit models are the same as the ones reported in Table 10, as the feature selection step did not exclude any predictor.

The block coefficient was positive in both models, but was only statistically significant in the second half of the experiment. In the early model, there was no strong evidence of a systematic change in success probability across blocks once the other predictors and subject-specific baseline differences were taken into account. In contrast, during the second half of the experiment, the block variable was positive and significant. This may indicate that learning effects became more visible in the

later part of the experiment.

The previous-hit variable had a positive and significant effect in both models. This means that, when the previous trial was successful, the probability of success on the current trial was higher, after accounting for the other predictors. The effect was slightly larger in the late model than in the early model. This may reflect more stable subject behaviour later in the game, where subjects who were performing well tended to remain successful from one trial to the next.

Target lifetime was also positively associated with success in both models. This is consistent with the interpretation that targets remaining longer above ground were easier to hit. The effect was significant in both halves, although it was stronger in the early model, which may be because longer lifetimes may have occurred more often in the early model. The number of red horn distractors had a negative and highly significant effect in both models, indicating that more red horn distractors were associated with a lower probability of hitting the target. This effect was stable across the two halves of the experiment. The number of blue horn distractors had no clear effect in the early model, but became negative and significant in the late model, meaning that blue horn distractors became more predictive of failure in the second half.

The cue effects differed between the two halves, maybe since cues were not randomly assigned but were related to the adaptive difficulty of the game. The animation cue had a positive and significant effect in the early model, but was not significant in the late model. The sound cue was also positive and significant in the early model, but not significant in the late model. The arrow cue showed the strongest contrast: it had a large negative and significant coefficient in the early model, but a positive and significant coefficient in the late model. The arrow cue was the most explicit one, but it may have appeared in early situations where subjects were still adapting to the task or where the difficulty of the game was low in response to lower performance. The negative early coefficient may therefore reflect this context rather than a harmful effect of the cue.

Target row and target column both had negative and significant effects in the two models. This indicates that some target positions were associated with a lower probability of success. The effects were stronger in the late model, especially for target column, suggesting that spatial target position became more predictive of failure in the second half of the experiment. This is consistent with the exploratory heatmaps, which showed that performance depended on where the target appeared on the grid.

Finally, the trial number was not significant in the early model, but had a small negative and significant effect in the late model. This may reflect a decrease in success probability within a local sequence of trials, for example, due to short-term fatigue, attention fluctuations, or other within-block dynamics.

Since continuous predictors were standardised before model fitting, their coefficients correspond to the effect of a one-standard-deviation increase in the corresponding predictor.

As for the AT model, we also computed the block-level test RMSE separately for each subject. For each subject, the block-level test RMSE was first computed within each train–test split and was then averaged across the 50 splits. The results are shown visually in Figure 16, while the corresponding numerical values are reported in Table 13.

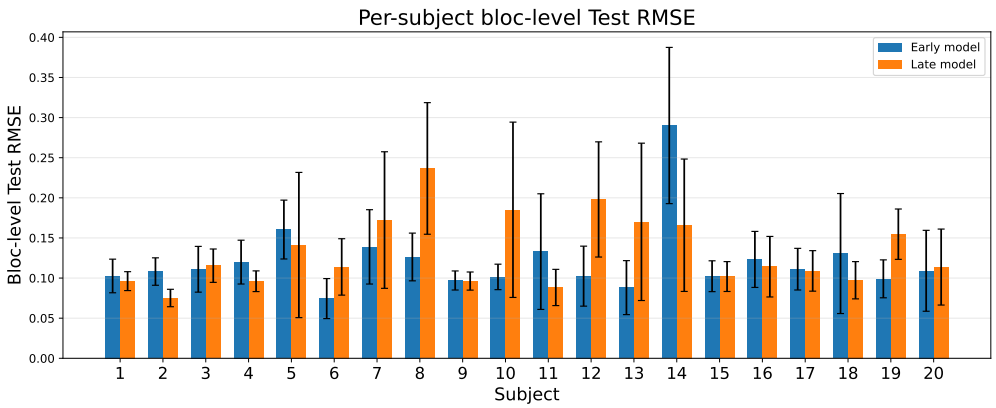

Figure 16: Per-subject block-level test RMSE for the binary hit prediction models. Error bars represent the standard deviation across the 50 random 80–20 train–test splits.

	Early model			Late model		
Subject	Block RMSE	Block MAE	Bias	Block RMSE	Block MAE	Bias
1	0.103 ± 0.021	0.079 ± 0.013	0.002 ± 0.023	0.096 ± 0.012	0.078 ± 0.011	0.004 ± 0.020
2	0.108 ± 0.017	0.088 ± 0.014	0.000 ± 0.024	0.075 ± 0.011	0.061 ± 0.009	-0.000 ± 0.017
3	0.111 ± 0.029	0.083 ± 0.022	0.007 ± 0.037	0.115 ± 0.021	0.096 ± 0.019	-0.001 ± 0.039
4	0.120 ± 0.027	0.094 ± 0.018	0.006 ± 0.037	0.096 ± 0.013	0.081 ± 0.013	-0.009 ± 0.027
5	0.160 ± 0.037	0.144 ± 0.038	-0.015 ± 0.092	0.141 ± 0.091	0.118 ± 0.060	0.028 ± 0.098
6	0.074 ± 0.025	0.062 ± 0.019	0.011 ± 0.032	0.114 ± 0.035	0.096 ± 0.028	0.001 ± 0.050
7	0.139 ± 0.046	0.104 ± 0.039	-0.003 ± 0.061	0.172 ± 0.085	0.125 ± 0.049	-0.018 ± 0.086
8	0.126 ± 0.030	0.107 ± 0.031	-0.002 ± 0.062	0.237 ± 0.082	0.184 ± 0.061	-0.046 ± 0.111
9	0.097 ± 0.012	0.078 ± 0.010	-0.000 ± 0.022	0.096 ± 0.011	0.079 ± 0.010	-0.003 ± 0.022
10	0.101 ± 0.016	0.085 ± 0.015	-0.001 ± 0.040	0.185 ± 0.109	0.126 ± 0.056	-0.012 ± 0.079
11	0.133 ± 0.072	0.096 ± 0.036	-0.010 ± 0.062	0.088 ± 0.023	0.074 ± 0.018	0.004 ± 0.043
12	0.102 ± 0.037	0.080 ± 0.025	-0.006 ± 0.048	0.198 ± 0.072	0.139 ± 0.035	0.009 ± 0.079
13	0.088 ± 0.034	0.070 ± 0.023	0.010 ± 0.043	0.170 ± 0.098	0.119 ± 0.050	-0.015 ± 0.075
14	0.290 ± 0.097	0.218 ± 0.063	-0.022 ± 0.133	0.166 ± 0.083	0.114 ± 0.045	0.011 ± 0.075
15	0.102 ± 0.019	0.087 ± 0.017	-0.001 ± 0.042	0.102 ± 0.019	0.092 ± 0.019	-0.012 ± 0.038
16	0.123 ± 0.035	0.101 ± 0.027	-0.007 ± 0.054	0.114 ± 0.038	0.089 ± 0.030	-0.008 ± 0.053
17	0.111 ± 0.026	0.091 ± 0.020	-0.006 ± 0.042	0.109 ± 0.025	0.097 ± 0.022	-0.004 ± 0.040
18	0.131 ± 0.075	0.101 ± 0.054	-0.021 ± 0.071	0.097 ± 0.023	0.088 ± 0.021	-0.007 ± 0.056
19	0.099 ± 0.024	0.083 ± 0.021	-0.000 ± 0.044	0.155 ± 0.031	0.126 ± 0.027	0.013 ± 0.063
20	0.109 ± 0.050	0.090 ± 0.044	-0.017 ± 0.075	0.114 ± 0.047	0.096 ± 0.038	-0.016 ± 0.080

Table 13: Per-subject block-level prediction performance of the binary hit mixed-effects models on the test set, averaged across the 50 random 80–20 train–test splits. Darker red cells indicate larger prediction errors for the RMSE and MAE.

Overall, the per-subject results show that block-level prediction errors were moderate for most subjects. The bias values were generally close to zero, so the model did not strongly overestimate or underestimate hit rates for most of them on average. The errors were not uniform across subjects, for example subject 14 showed the largest prediction error in the early model, with a block-level RMSE of 0.290, while subject 8 showed the largest error in the late model, with a block-level RMSE of 0.237. Other relatively large late-model errors were observed for subjects 12, 10, 7, and 13. These cases suggest that the model did not capture all subject-specific changes in hit probability equally well. This may reflect individual strategies, attention fluctuations, fatigue effects, or other sources of variability that are not fully described by the available predictors.

As for the AT model, these subject-level metrics are computed on a small number of test blocks for some subjects, so a few poorly predicted blocks can strongly affect the RMSE. This is especially relevant for subjects 20, 18, 5, 12, and 14, who have respectively 6, 8, 8, 13, and 15 test blocks. Still, the higher errors for some subjects indicate that the model captured the general hit-rate tendency better for some

subjects than for others.

To visualise the behaviour of the binary model, we selected representative subjects and compared the observed and predicted hit probabilities on the test set. The corresponding plots are shown in Figure 17, while the other subjects' visualisations can be found in Appendix D. Each curve summarises the test predictions obtained across the 50 random train–test splits. Note that the vertical scale is not the same across the three subjects, in order to better highlight the subjects' specific patterns. In each plot, the observed curve corresponds to the actual proportion of successful trials in each test block, whereas the predicted curve corresponds to the mean predicted hit probability in that block. This visualisation shows whether the model captures the subject's block-level performance evolution, rather than individual trial-by-trial outcomes.

We chose to show the curves of subjects 9, 14, and 8. Subject 9 was selected as an example of relatively good predictive performance. This subject also had many more blocks than subjects 14 and 8, which makes the comparison between observed and predicted block-level hit rates more informative and more stable. The corresponding plots show that the model captured the general evolution of the hit rate reasonably well. In contrast, subject 14 showed a more irregular pattern, with some blocks having much lower observed hit rates, including blocks with a hit rate equal to zero. These abrupt drops are difficult for the model to predict because they are not fully explained by the available predictors. Finally, subject 8 was selected because this subject had a particularly high RMSE in the late model. In the late part of the experiment, the first observed hit rates were lower than those observed in the early part. These changes were not well captured by the predictors included in the model, which tended to predict smoother and higher hit probabilities for those blocks.

Overall, the binary hit prediction models showed moderate predictive performance at the block level. The comparison between the early and late models did not show a clear improvement in predictive accuracy. Moreover, some subjects showed larger differences between observed and predicted hit rates, especially when their performance contained abrupt drops or specific patterns not explained by the available predictors. This suggests that the included predictors explain part of the probability of success, but that passed versus failed trials remain influenced by additional trial-level and subject-specific variability not fully captured by the model.

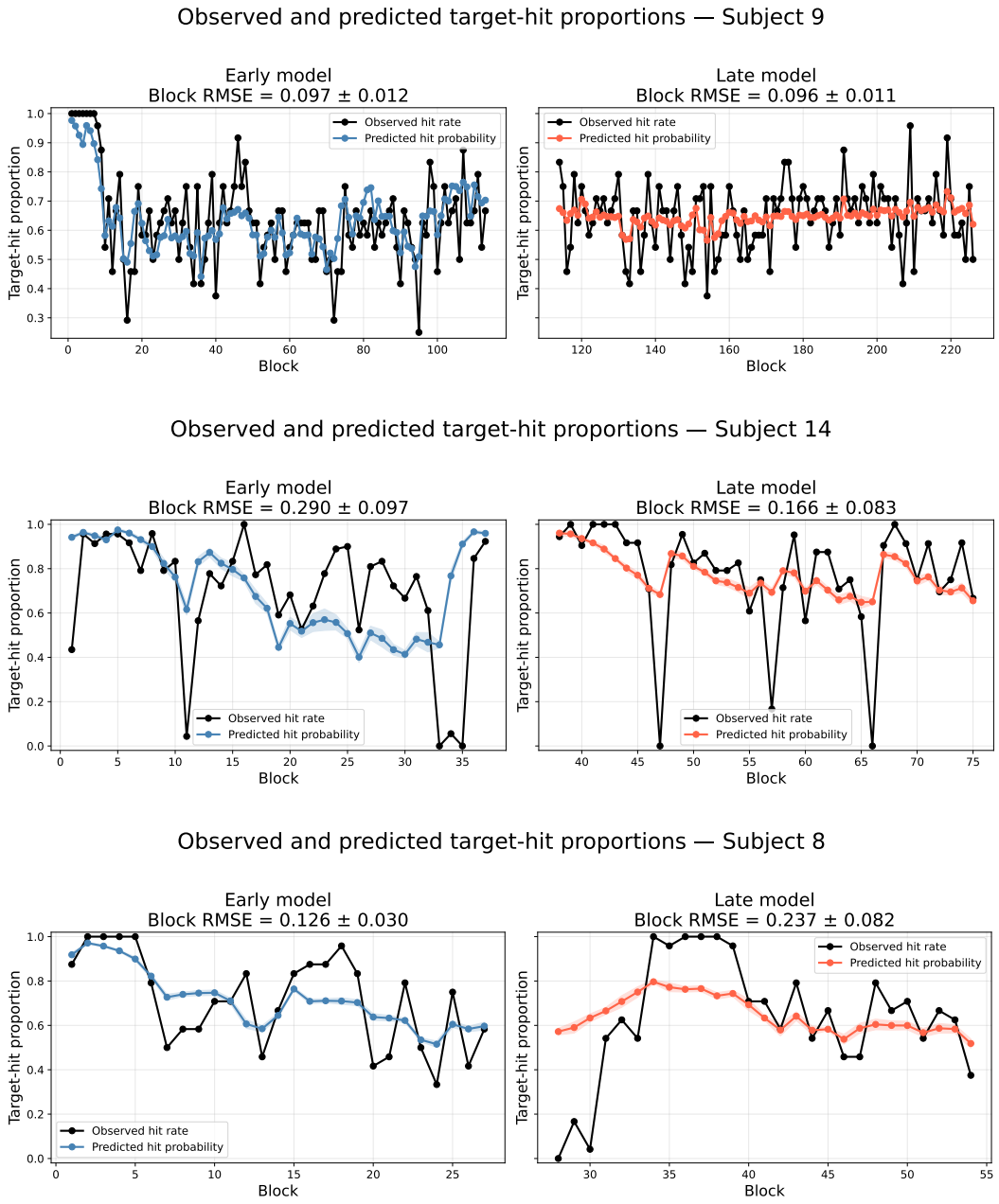

Figure 17: Observed hit rates and predicted hit probabilities on the test set for selected subjects in the repeated 80–20 test procedure.

5 Model Exploitation

5.1 Leave-One-Subject-Out: Action Time Models

In addition to the standard 80–20 test procedure described above, we also evaluated the ability of the AT model to generalise to completely unseen subjects. For this purpose, a LOSO procedure was used. In each fold, one subject was entirely removed from the training set, the model was fitted on the remaining subjects, and predictions were then computed for the held-out subject. This procedure was repeated for the 20 subjects of the dataset.

This setting is more difficult than the test procedure, because no observation from the held-out subject is available during training. Consequently, the random intercept of the held-out subject cannot be estimated directly by the LMEMs. Thus, the LOSO analysis evaluates whether the population-level effects learned from the other subjects are sufficient to predict the AT of a completely new subject, based on their observed predictors at each block.

As before, the early and the late models were considered separately. The table containing the average coefficient estimates obtained across the LOSO folds is reported in Appendix F.

Three prediction strategies were compared. First, a baseline prediction was computed as the mean AT in milliseconds of the training subjects, so a constant AT was predicted for every block of the held-out subject. Second, FE-only predictions were obtained from the LMEMs. These correspond to population-level predictions for an unseen subject j:

$$\widehat{\mathbf{Y}}_j^{FE} = \widehat{\beta}_0 + \sum_{k=1}^{p} \widehat{\beta}_k \mathbf{X}_{jk}, \tag{32}$$

where $\widehat{\mathbf{Y}}_j^{FE}$ denotes the predicted log-transformed AT of subject j. In this strategy, the held-out subject's random intercept is not estimated and is therefore set to zero. Third, we investigated whether predictions could be improved by estimating the held-out subject's random intercept from demographic variables, which are the

subject's sex, height, weight, and age. For this, the random intercepts estimated on the training subjects were used as response variables in a Ridge regression, with those variables as predictors. The L_2 penalty of the Ridge regression shrinks the regression coefficients toward zero, which reduces the risk of obtaining unstable coefficient estimates from the limited number of training subjects. The fitted Ridge model was then used to estimate an approximate random intercept for the held-out subject:

$$\widehat{\mathbf{Y}}_j^{FE+\widehat{RE}} = \widehat{\beta}_0 + \sum_{k=1}^{p} \widehat{\beta}_k \mathbf{X}_{jk} + \underbrace{\widehat{b}_{0j}}_{\text{Estimated RE}} . \tag{33}$$

The predictions were then back-transformed to milliseconds before computing the prediction errors.

The global LOSO results are shown in Table 14. The values correspond to the mean block-level performance across held-out subjects. Bias is defined as observed AT minus predicted AT, so that a positive bias indicates that the model tends to predict shorter AT, while a negative bias indicates that the model tends to predict longer AT.

	Early model			Late model		
Strategy	RMSE (ms)	MAE (ms)	Bias (ms)	RMSE (ms)	MAE (ms)	Bias (ms)
Baseline	310.01	263.97	72.72	427.61	395.48	49.24
FE-only	267.06	220.34	35.42	222.12	187.99	42.60
FE + estimated RE	275.91	230.53	31.61	225.99	192.02	44.81

Table 14: Mean LOSO block-level prediction performance for the AT models.

In both halves of the experiment, the FE-only model improved over the baseline. In the early model, the mean block-level RMSE decreased from 310.01 ms for the baseline to 267.06 ms for the FE-only model. The improvement was larger in the late model, with the RMSE decreasing from 427.61 ms to 222.12 ms. This suggests that the predictors related to the game's difficulty learned from the other subjects contained useful information for predicting the AT of an unseen subject. However, these LOSO results should not be compared directly with the repeated 80–20 test errors, because the evaluation setting is different. In the 80–20 procedure, the errors were computed at the chunk level, while LOSO results reported here are computed at the block level. Block-level aggregation can smooth part of the observation-level variability. Therefore, the main interpretation is not the comparison between the two procedures, but the fact that the FE-only model clearly improved over the baseline even for unseen subjects.

Adding an estimated random intercept based on demographic variables did not improve the predictions on average. In the early model, the mean RMSE increased slightly from 267.06 ms for the FE-only model to 275.91 ms for the FE + estimated RE model. In the late model, it also increased slightly, from 222.12 ms to 225.99 ms. Therefore, although the demographic correction improved the prediction for some individual subjects, it did not provide a robust average improvement in the LOSO setting. This is likely due to the small number of subjects available to train the regression from demographic variables to random intercepts. For this reason, the FE-only predictions were retained as the main LOSO prediction strategy, while the demographic random-intercept estimation was considered exploratory.

The per-subject RMSE values are reported in Table 15. The results again show heterogeneity across subjects. Some subjects were predicted relatively well by the FE-only model, for example, subjects 5 and 17, while others had much larger errors, especially subjects 14, 16, and 20. This confirms that even if the FE captured part of the general relationship between task variables and AT, some subject-specific AT profiles remained difficult to predict without data in the training set.

	Early model RMSE (ms)			Late model RMSE (ms)		
Subject	Baseline	FE-only	FE + estimated RE	Baseline	FE-only	FE + estimated RE
1	217.75	155.91	163.08	288.82	254.81	275.14
2	237.35	289.97	268.58	345.06	188.59	179.36
3	558.20	238.96	219.16	634.04	213.58	222.12
4	175.71	208.77	279.64	286.96	169.46	154.85
5	107.63	100.20	107.91	183.17	157.20	189.87
6	178.30	247.01	198.46	253.92	164.59	131.67
7	252.38	203.63	239.79	336.29	128.43	100.56
8	119.19	127.38	160.03	350.01	168.46	169.80
9	215.21	233.53	216.55	295.47	82.83	109.95
10	388.18	210.20	222.36	608.14	368.55	355.36
11	188.35	235.62	219.39	220.30	267.16	263.22
12	490.59	329.70	324.06	1253.32	369.54	383.53
13	334.05	227.32	211.50	707.54	232.01	217.68
14	685.63	537.62	539.85	856.75	400.68	401.10
15	201.54	142.50	176.56	279.37	189.80	219.43
16	727.93	610.96	590.47	745.72	536.14	498.42
17	208.54	132.65	204.41	294.87	81.09	135.69
18	171.08	196.67	275.28	145.68	139.36	192.61
19	191.98	219.12	212.18	171.97	224.46	211.86
20	550.55	693.50	688.95	294.73	105.61	107.62
Mean	310.01	267.06	275.91	427.61	222.12	225.99

Table 15: Per-subject LOSO block-level RMSE values for the AT models. Darker red cells indicate larger prediction errors.

The demographic random-intercept estimation improved the RMSE for 10 out of 20 subjects in the early model and for 9 out of 20 subjects in the late model. This indicates that the demographic variables did contain some information for certain subjects, but not enough to reliably reconstruct subject-specific baseline AT. In some cases, the estimated random intercept moved the prediction in the wrong direction and increased the error.

To illustrate the behaviour of the LOSO predictions, representative subjects were compared visually, as shown in Figure 18. The figures corresponding to the other subjects can be found in section 1 of Appendix F. Note that the vertical scale is different across subjects to better visualise the per-subject fluctuations. Subjects with relatively low RMSE, such as subjects 5 and 17, illustrate cases where the FE captured the observed AT reasonably well. Their observed AT profiles were relatively stable across blocks, which made their trajectories easier to predict. In contrast, subjects 14 and 16 illustrate more difficult cases, where the observed AT values deviated more strongly from the population-level prediction. Their

AT varied over a much wider range, approximately from 1400 ms to 3400 ms for subject 14 and from 1900 ms to 3500 ms for subject 16. These larger subject AT fluctuations indicate less stable performance over the experiment and are therefore harder for the model to predict.

Overall, the LOSO analysis shows that the AT model generalises partially to unseen subjects. The FE-only model clearly improved over the baseline prediction on average, especially in the late half of the experiment. The attempt to improve predictions by estimating the random intercept from demographic variables was not consistently successful. This suggests either that the number of subjects was too low for this estimation, or that the available demographic variables were not sufficient to explain the differences in baseline AT between subjects, and that additional information on their performance is needed to improve predictions for completely unseen subjects.

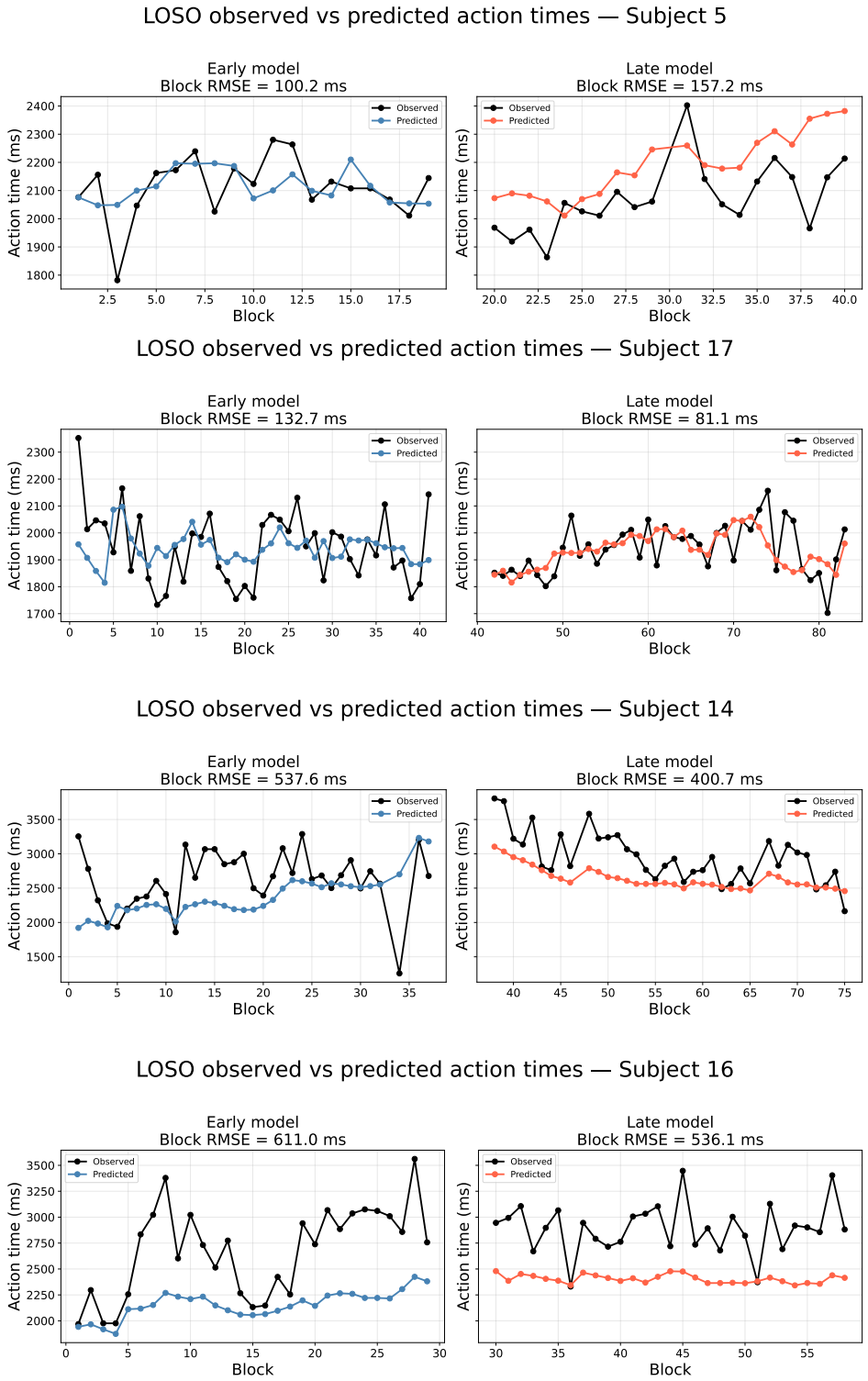

Figure 18: LOSO AT predictions for representative subjects.

5.2 Leave-One-Subject-Out: Binary Hit Models

As for the AT models, a LOSO analysis was also performed for the binary hit models in order to evaluate their ability to generalise to completely unseen subjects.

As in the AT LOSO analysis, three prediction strategies were compared. First, a baseline prediction was computed using the mean hit rate of the training subjects. This single value was then assigned as the predicted probability for all trials of the held-out subject. Second, FE-only predictions were obtained from the population-level variables of the binary mixed-effects model:

$$\widehat{\eta}_j^{FE} = \widehat{\beta}_0 + \sum_{k=1}^{p} \widehat{\beta}_k \boldsymbol{X}_{jk}. \tag{34}$$

Third, the same demographic correction as for the AT model was explored. For each LOSO fold, the random intercepts estimated on the training subjects were used to estimate the random intercept of the held-out subject, using a Ridge regression with the demographic variables as predictors. The corresponding linear predictor was:

$$\widehat{\eta}_j^{FE+\widehat{RE}} = \widehat{\beta}_0 + \sum_{k=1}^{p} \widehat{\beta}_k \mathbf{X}_{jk} + \underbrace{\widehat{b}_{0j}}_{\text{Estimated RE}}. \tag{35}$$

As before, the demographic random-intercept estimation should be interpreted as exploratory, because the number of subjects available to train the Ridge model is small.

Trial-level metrics, such as the AUC, Brier score, and accuracy, were computed. However, the block-level RMSE and MAE remained the main metrics of interest because they evaluate whether the model captures the average success rate of unseen subjects across blocks.

	Early model						Late model					
Strategy	AUC	Brier	Accuracy	Block RMSE	Block MAE	Bias	AUC	Brier	Accuracy	Block RMSE	Block MAE	Bias
Baseline	0.500	0.213	0.693	0.197	0.165	0.009	0.500	0.229	0.647	0.182	0.144	0.000
FE-only	0.690	0.202	0.678	0.160	0.127	0.009	0.616	0.224	0.640	0.168	0.130	0.001
FE+$\widehat{RE}$	0.690	0.205	0.671	0.168	0.135	0.011	0.616	0.224	0.639	0.169	0.130	0.000

Table 16: LOSO performance of the binary hit models.

The global LOSO results are reported in Table 16, while the averaged coefficient estimates are also reported in Appendix F. As expected, the baseline obtained an AUC of 0.500, since it assigns the same predicted probability to all trials of the held-out subject. The mixed-effects models improved the AUC, reaching 0.690 in

the early model and 0.616 in the late model. This indicates that the FE learned from the other subjects provided some ability to rank successful and unsuccessful trials for unseen subjects.

The Brier score also improved slightly compared with the baseline. In the early model, it decreased from 0.213 for the baseline to 0.202 for the FE-only model. In the late model, it decreased from 0.229 to 0.224. This suggests that the predicted probabilities were slightly closer to the observed binary outcomes when using the mixed-effects model. However, the accuracy did not improve compared with the baseline.

At the block level, the FE-only model improved over the baseline in both halves. In the early model, the block-level RMSE decreased from 0.197 to 0.160, and the MAE decreased from 0.165 to 0.127. In the late model, the improvement was smaller, with the RMSE decreasing from 0.182 to 0.168 and the MAE decreasing from 0.144 to 0.130. This suggests that the binary hit model captured some useful information for predicting the average hit rate of unseen subjects, especially in the early half.

Adding the demographic estimate of the random intercept did not improve the predictions. In both halves of the experiment, the FE + estimated-RE strategy produced results that were nearly identical to, or slightly worse than, the FE-only strategy.

	Early model			Late model		
Subject	Baseline	FE-only	FE + estimated RE	Baseline	FE-only	FE + estimated RE
1	0.161	0.197	0.250	0.120	0.143	0.149
2	0.166	0.135	0.131	0.091	0.119	0.121
3	0.192	0.257	0.254	0.157	0.193	0.189
4	0.181	0.140	0.155	0.148	0.098	0.099
5	0.221	0.169	0.170	0.244	0.263	0.263
6	0.219	0.133	0.127	0.150	0.137	0.139
7	0.195	0.159	0.193	0.192	0.177	0.177
8	0.206	0.144	0.140	0.269	0.219	0.220
9	0.166	0.105	0.103	0.111	0.113	0.114
10	0.166	0.101	0.105	0.238	0.203	0.203
11	0.232	0.166	0.164	0.160	0.099	0.100
12	0.230	0.101	0.101	0.258	0.219	0.220
13	0.173	0.104	0.104	0.223	0.202	0.201
14	0.271	0.310	0.309	0.274	0.193	0.193
15	0.183	0.150	0.155	0.183	0.228	0.231
16	0.189	0.236	0.233	0.165	0.138	0.140
17	0.182	0.152	0.145	0.135	0.124	0.124
18	0.252	0.160	0.194	0.150	0.135	0.136
19	0.199	0.102	0.150	0.219	0.205	0.202
20	0.160	0.175	0.172	0.141	0.153	0.150

Table 17: Per-subject block-level RMSE in the LOSO binary hit analysis. Darker red cells indicate larger prediction errors.

The per-subject results in Table 17 show that generalisation performance again varied across subjects. In the early model, the FE-only strategy gave relatively low block-level errors for subjects 10, 12, 19, 13, and 9, with RMSE values close to 0.10. In contrast, subjects 14, 3, and 16 were harder to predict, with FE-only RMSE values ranging from 0.236 to 0.310. In the late model, the lowest FE-only errors were observed for subjects 4, 11, 9, 2, and 17, while the largest errors were observed for subjects 5, 15, 12, 8, and 19.

The comparison with the baseline shows that the FE-only model improved prediction for several subjects, but not for all of them. For example, clear improvements were observed for subjects 12, 19, 10, 13, 8, and 6 in the early model, and for subjects 4, 11, 8, 10, and 12 in the late model. For other subjects, such as subjects 1, 3, 14, 16, and 20 in the early model, or subjects 2, 3, 5, 15, and 20 in the late model, the FE-only prediction was worse than the baseline.

Adding the demographic estimate of the random intercept did not provide a consistent improvement. In some cases, it slightly reduced the RMSE, but in others it increased the error. This supports the conclusion that the demographic correction

was not reliable enough to improve prediction for unseen subjects.

To visualise the behaviour of the LOSO binary hit model, we also compared the observed and predicted hit probabilities for the same representative subjects as in the standard 80–20 test analysis. The corresponding curves are shown in Figure 19, while the others are shown in section 2 of Appendix F. Only the FE-only predictions are shown, since this strategy represents the main LOSO generalisation result, and the y-scale is different across subjects to better show per-subject variations. As in the 80–20 test procedure, the model captured part of the general block-level evolution, but some blocks remained difficult to predict accurately. In particular, when the observed hit rate changed abruptly or deviated strongly from the other blocks, the model tended to produce smoother predictions and did not fully reproduce these local variations.

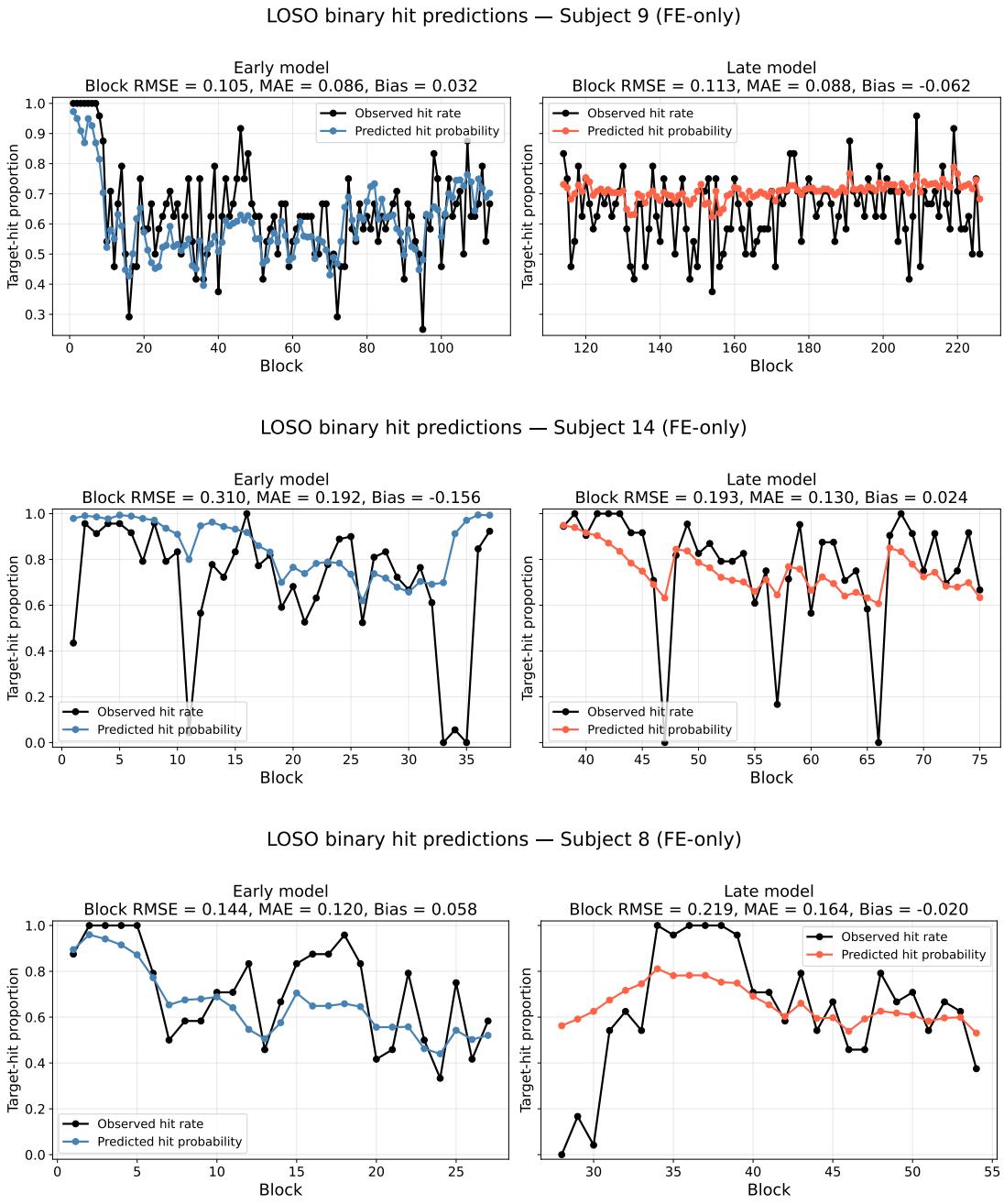

Figure 19: Observed hit rates and predicted hit probabilities for selected held-out subjects in the LOSO binary hit analysis. Only the FE-only predictions are shown.

This confirms that the main limitation of the binary hit model is not specific to the LOSO setting, and that the available predictors only partly explain the variability in trial success. As a result, it cannot always reproduce abrupt changes in observed

hit rate, especially when these changes may be driven by subject-specific behaviour, attention fluctuations, fatigue, or other factors not directly included in the model.

Overall, the LOSO binary hit analysis suggests that the FE captured some general relationships between task variables and the probability of success, as shown by the increase in AUC compared with the baseline. The block-level results also showed some improvement over the baseline, especially in the early model, where the RMSE decreased from 0.197 to 0.160 for the FE-only predictions. In the late model, the improvement was smaller, with the RMSE decreasing from 0.182 to 0.168. However, these improvements remained moderate, indicating that predicting hit probability for a completely unseen subject remains difficult. This contrasts with the AT LOSO analysis, where the FE produced a clearer improvement over the baseline. Therefore, predicting whether an unseen subject will successfully hit the target appears more difficult than predicting their AT.

5.3 Difficulty-Level Simulations

To investigate whether subjects became faster after accounting for the difficulty of the game, both the early and late AT models were evaluated on the same simulated difficulty configurations. We used a resampling approach with the `sample` function from the `pandas` library for the difficulty levels to be plausible. This means that, instead of creating artificial predictor values independently, we sampled 1000 observations with replacement from the original dataset. This allowed us to preserve the structure of the game, including the relationships between the cues, the number of distractors, the target lifetime, and the target location. The variables used to describe the simulated difficulty were the same as those used to train the models.

The same simulated configurations were then given to all 50 of the early and late splits models, resulting in 50000 predictions per subject. Using the same difficulty levels ensures that differences in predicted AT were not caused by differences in the distribution of task difficulty. The block variable was fixed to the median value of the corresponding half. Therefore, the simulation compares predictions under the same sampled task configurations, but at the typical progression level of each half. For each subject, predictions were obtained using the subject-specific random intercept estimated by the LMEMs.

The resulting predicted distributions are shown in Figure 20, while Table 18 is the global table containing the simulations results. For each subject, we computed the mean predicted AT under the early model and under the late model. The difference

between the two predictions was then computed as:

$$\Delta\widehat{AT} = \widehat{AT}_{late} - \widehat{AT}_{early}. \tag{36}$$

This means that negative values indicate that the late model predicted faster responses than the early model under comparable difficulty conditions. The last column refers to the percentage of splits for which the late model was faster.

Subject	Early $\widehat{AT}$ (ms)	Late $\widehat{AT}$ (ms)	Mean Δ (ms)	SD Δ (ms)	Median Δ (ms)	Late faster (%)
20	2753.66	2226.42	-527.24	153.38	-551.82	100.0
7	2247.02	2088.28	-158.74	119.40	-181.89	90.0
1	2007.39	1874.44	-132.95	105.69	-152.49	90.0
2	1915.43	1826.19	-89.25	101.47	-107.51	88.0
5	2197.19	2122.40	-74.79	117.79	-96.58	85.0
14	2385.50	2321.41	-64.08	131.21	-87.69	80.0
8	2225.48	2180.95	-44.53	120.24	-67.02	78.0
16	2579.75	2543.18	-36.57	142.73	-60.10	72.0
4	2019.62	1994.44	-25.17	108.69	-43.71	73.0
18	2346.48	2327.54	-18.95	127.39	-41.37	68.0
11	1978.64	1961.26	-17.38	107.07	-36.45	69.0
17	2172.40	2164.92	-7.48	117.31	-28.37	64.0
13	2263.38	2264.89	1.51	124.27	-18.90	59.0
3	2084.51	2096.36	11.85	114.46	-7.03	54.0
6	2038.89	2056.43	17.54	112.08	-2.46	51.0
19	2292.29	2344.66	52.37	127.97	30.83	37.0
9	1964.87	2020.89	56.02	108.85	38.43	32.0
12	2253.09	2329.02	75.93	130.05	54.54	29.0
15	2235.39	2344.50	109.11	126.84	87.78	17.0
10	2113.33	2275.89	162.56	122.89	143.65	4.0

Table 18: Simulated predicted AT under comparable task-difficulty configurations for the early and late models, sorted from the largest predicted improvement in the late model to the largest predicted deterioration. Green cells indicate lower predicted AT in the late model, while red cells indicate higher predicted AT.

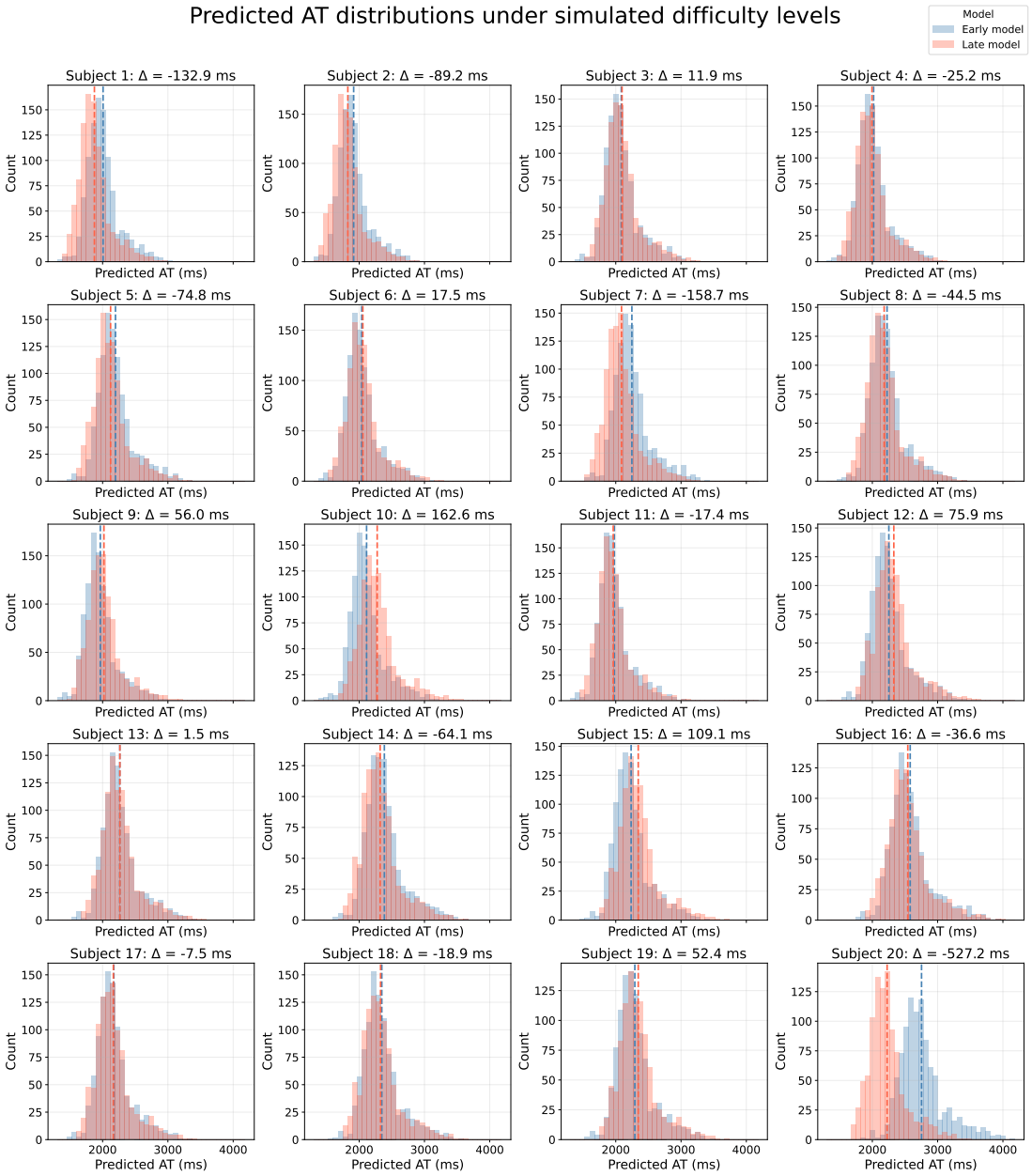

Figure 20: Predicted AT distributions under simulated difficulty levels. For each subject and simulated configuration, predictions were averaged across the 50 repeated train–test splits before plotting.

The simulated distributions show that the predicted late–early difference was not homogeneous across subjects. The largest negative differences were observed for subjects 20, 7, 1, and 2, with difference values of approximately −527 ms, −158 ms,

−133 ms, and −90 ms, respectively. These subjects showed the strongest tendency toward shorter predicted AT in the late part of the experiment. Smaller negative differences were also observed for subjects 5, 14, 8, 4, 11, and 16.

In contrast, the largest positive differences were observed for subjects 10 and 15, with difference values of approximately 163 ms and 109 ms, respectively. Subjects 12, 9, and 19 also showed positive differences, with increases of approximately 76 ms, 56 ms, and 52 ms. Finally, several subjects had differences close to zero, such as subjects 3, 6, 13, 17, and 18, suggesting that the early and late models produced very similar predicted AT for them under the simulated difficulty levels.

Overall, the simulated late–early differences ranged from approximately −527 ms to 163 ms. There was no clear global shift in one direction: 12 subjects had negative values and 8 subjects had positive values, with an average difference of approximately −35.51 ms, where this average is strongly driven by subject 20. This suggests that the early–late change predicted by the models highly depends on the subjects, and is not reflecting a uniform global effect across all subjects.

However, for all subjects except subject 20, the absolute value of those differences was smaller than the prediction error of the model. This means that they should not be interpreted as reliable evidence that a given subject truly became faster or slower. Instead, the simulation should be retained as descriptive: it shows how the early and late models behave when applied to the same distribution of task difficulty, and it highlights subjects' tendencies in the predicted trajectories, but it does not provide precise conclusions on learning.

We then applied the same difficulty-level simulation procedure to the binary hit models. The simulation was performed across the 50 repeated train–test splits, as this reduces the dependence of the simulation results on any particular random split. When including the subject random intercepts, the mean predicted hit probability increased from approximately 0.594 for the early model to approximately 0.649 for the late model.

Subject	Early $\hat{p}$	Late $\hat{p}$	Δ	SD Δ	Late > Early
1	0.605	0.717	0.111	0.087	96.7%
2	0.603	0.703	0.100	0.085	95.6%
9	0.603	0.702	0.099	0.085	95.4%
4	0.597	0.665	0.068	0.082	88.4%
3	0.595	0.656	0.060	0.081	85.3%
10	0.593	0.641	0.048	0.080	78.9%
17	0.593	0.638	0.046	0.080	77.7%
14	0.591	0.635	0.044	0.080	76.2%
15	0.592	0.635	0.042	0.080	75.5%
19	0.592	0.634	0.042	0.080	75.2%
11	0.592	0.634	0.042	0.080	75.0%
7	0.592	0.632	0.040	0.080	74.2%
12	0.591	0.630	0.039	0.080	73.2%
13	0.591	0.630	0.039	0.080	73.0%
6	0.591	0.629	0.038	0.080	72.4%
16	0.591	0.627	0.036	0.080	71.3%
8	0.590	0.625	0.035	0.080	70.1%
5	0.590	0.619	0.029	0.080	65.6%
18	0.590	0.619	0.029	0.080	65.6%
20	0.589	0.614	0.026	0.080	62.7%

Table 19: Simulated predicted hit probabilities under comparable difficulty configurations, sorted from the largest to the smallest predicted improvement in the late model. Darker green cells indicate larger predicted improvement in the late model.

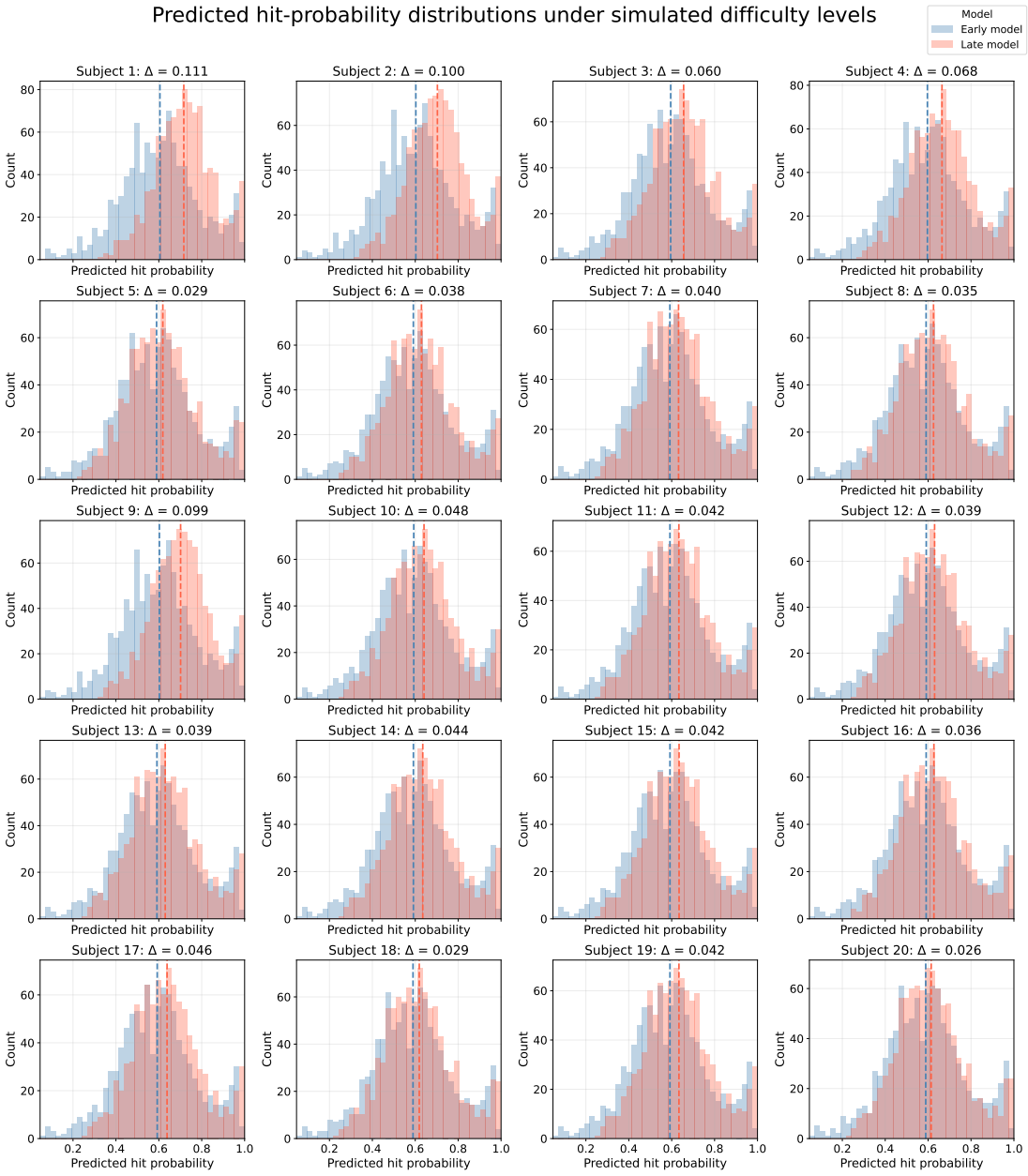

Figure 21: Predicted hit probability distributions per subject under simulated difficulty levels for both early and late models.

The predicted hit probabilities are reported in Table 19, while the subject-specific differences are shown in Figure 21. All subjects showed a positive mean difference between the late and the early model simulations, meaning that the late model predicted a higher probability of success under comparable simulated difficulty

configurations. The mean predicted probabilities were relatively close across subjects, especially in the early model. Most early predicted hit probabilities were around 0.59–0.60, so the simulation produced similar average early predictions for most subjects. The differences between subjects were more important in the late model predictions. The largest predicted improvements were observed for subjects 1, 2, and 9, with increases of approximately 0.111, 0.100, and 0.099 hit probability, respectively. Subjects 3 and 4 also showed positive differences, with increases of approximately 0.06–0.07.

For the remaining subjects, the predicted improvements were smaller but still positive. The smallest mean differences were observed for subjects 20, 5, and 18, with increases of approximately 0.026 to 0.029. The percentage of simulations in which the late prediction was higher than the early prediction was above 50% for all subjects, ranging from 62.7% for subject 20 to 96.7% for subject 1.

The binary hit simulation suggests an improvement between the two halves of the experiment. The average predicted probability of success increased under comparable difficulty conditions, and this increase was observed for all subjects. However, the standard deviation remained relatively large for many subjects.

As a final comparison, we can combine the simulation results from the AT and binary hit models. The clearest combined improvement was observed for subjects 1 and 2, for whom the late model predicted both shorter AT and clearly higher hit probability under comparable simulated difficulty conditions. Subjects 7 and 20 also showed strong decreases in predicted AT, together with positive changes in predicted hit probability. For subject 20, the predicted decrease in AT was particularly large, while the predicted increase in hit probability was smaller, suggesting that the main improvement for this subject was related to speed rather than accuracy.

Several subjects showed more moderate improvements. Subjects 4, 5, 8, 14, 16, 17, and 18 had shorter predicted AT in the late model, while their predicted hit probability also increased, although the increase was generally modest. These subjects can therefore be interpreted as showing some evidence of improvement, but with a smaller magnitude than subjects 1, 2, 7, and 20.

Other subjects showed mixed patterns. Subjects 3, 6, 11, and 13 had predicted AT differences close to zero, but still showed positive changes in predicted hit probability. This suggests relatively stable response speed and a modest improvement in predicted success probability. In contrast, subjects 9, 10, 12, 15, and 19 had longer predicted AT in the late model, while also showing higher predicted hit probability.

For these subjects, the simulations may suggest a speed–accuracy trade-off: the late model predicted slower responses, but also a higher probability of success under comparable difficulty conditions.

Overall, the combined simulation results suggest that the evolution of performance was not uniform across subjects. The binary hit model predicted higher success probabilities for all subjects, whereas the AT model showed both positive and negative changes depending on the subject. This indicates that some subjects may have improved mainly in speed, others mainly in accuracy, and others may have traded speed for accuracy.

6 Methodological Challenges and Limitations

Throughout the realisation of this work, we encountered several difficulties linked to the experiment or the game it used.

First, the number of subjects included in the experiment was small, which is clearly a limitation for this work, as the different models do not have a lot of separate sources to learn from. Also, the number of observations per subject was very heterogeneous, as some subjects had played a lot more than others, or some subjects' data were incomplete. Indeed, even if they all participated in the experiment for a total of seven days, some would play a lot more per day than others, or some of the data for some players were missing. To be more precise, two complete days were missing for subject 3, one day for subjects 5, 6, 8, 12, and 18, and four days for subject 20. In the worst case, a subject has eight times less data than another (this can be seen in Table 1), which implied that the comparison between subjects was very hard, and also we had to make sure the models we used were not taking into account this big gap of observations (i.e., that the models would not overfit on more represented subjects).

Another difficulty we encountered in the between-subject analysis was that subjects did not start with the same difficulty level. Indeed, some would start with a very low difficulty level, meaning that they had few or even no distractors in their first block of trials, with the arrow cue, which is the most helpful cue a subject could get. This is the case with most of the subjects. In parallel, three subjects started the experiment with a higher level of difficulty, as they had a very short target lifetime, a significant number of distractors, and a less obvious cue.

Then, during the experiment, the game was supposed to adapt the difficulty to the subject's performance, which is already hard to model, as each subject will encounter different difficulty levels that completely depend on their own performance. But, we observed that for some subjects, the difficulty would stay the same even if the performance was higher than 75%. Indeed, the difficulty parameters (number of each type of distractors, target's lifetime, and cue) did not change, and the appearance of the target on the ground was very similar in both blocks.

Worse, sometimes the game would become harder even if the subject's performance was very low. This could bias our analysis, as performance in this case could be linked to a difficulty adjustment issue, not a real (un)improvement of the subject.

Finally, one of the limitations of this work is the way the early and late models were defined. The split was made separately for each subject, at 50% of their own block sequence. However, because the total number of blocks differs considerably between subjects, the notions of *early* and *late* do not correspond to the same amount of practice for everyone. As a result, a block considered *early* for one participant may correspond to a more advanced stage of practice for another participant, which can make between-subject comparisons more difficult.

7 Conclusion

The objective of this work was to study subjects' performance in the serious game inspired by the *Whac-a-Mole* arcade game. This was done by modelling two outcomes: the AT required to successfully hit a target, and the probability of successfully hitting the target. Mixed-effects models were used in order to account for both predictors related to the game's difficulty and repeated observations within subjects, which are correlated. The AT models were fitted using LMEMs on log-transformed AT, while the binary hit models were fitted using logistic mixed-effects models. In both cases, separate early and late models were trained in order to investigate how the relationship between task variables and performance changed with practice.

The AT analysis showed that the models were able to capture meaningful relationships between task features and the time needed to hit the target. The block coefficient was negative in both the early and late models, indicating that, after taking the task difficulty and subject-specific baseline differences into account, AT tended to decrease as subjects progressed through the experiment. The effect was stronger in the late model, which may indicate that subjects became more consistent after gaining experience with the task, and so the presence of a learning or adaptation effect.

The repeated 80–20 validation procedure showed that the AT models generalised well to unseen blocks from subjects already present in the training data. Train and test errors were close to each other, suggesting no strong evidence of overfitting. The late model obtained lower prediction errors than the early model, with a lower test RMSE and MAE, indicating that AT were more predictable in the second half of the experiment. However, some subjects were predicted accurately, whereas others had much larger errors, especially when the number of available test observations was small or when their behaviour differed from the population-level pattern.

The binary hit analysis provided a different perspective on performance. The models showed moderate discrimination ability, with AUC values above chance, but not close to perfect discrimination. The early model performed slightly better than the late model at the trial level, with a higher test AUC, a lower Brier score,

and a higher accuracy, but the results indicate that the available predictors could not predict all trial-level success and failure. At the block level, the bias remained close to zero, so the models predicted the average hit rate reasonably well, although the RMSE and MAE showed that errors remained for some subjects and blocks.

The binary hit analysis also captured meaningful relationships between task features and the probability of success. The block coefficient was positive in both halves, but significant only in the late model, suggesting clearer learning or adaptation effect on hit probability in the second half of the experiment.

The LOSO analyses evaluated a more difficult setting, where the held-out subject was completely absent from the training data. For the AT models, the FE-only predictions improved clearly over the baseline, especially in the late model. Estimating the random intercept from demographic variables did not improve performance on average. This suggests that sex, age, height, and weight were not sufficient to reliably explain the subjects' baseline AT, at least with the limited number of subjects available.

For the binary hit models, the LOSO results were more modest. The FE improved the AUC compared with the baseline, showing that the model retained some ability to rank successful and unsuccessful trials for unseen subjects. The block-level results also improved over the baseline, especially in the early model, where the RMSE and MAE decreased more clearly. In the late model, the improvement was smaller. The accuracy did not improve over the baseline, which highlights the limitation of classification in this setting. The binary hit models' ability to predict the block-level hit rate of completely unseen subjects remained moderate. As for the AT model, estimating the held-out subject's random intercept from demographic variables did not provide a consistent improvement.

Finally, the simulation analysis allowed the early and late models to be compared under the same sampled difficulty configurations. For AT, the simulation did not show a uniform global improvement from early to late predictions. Some subjects were predicted to become faster, while others were predicted to become slower under the same task conditions, so the evolution of AT was not homogeneous among subjects. For the binary hit model, the simulation showed a clearer positive tendency: the late model predicted a higher hit probability for all subjects under comparable difficulty configurations. The early predicted probabilities were very close across subjects, so the differences in improvement mainly come from the late model.

This work shows that mixed-effects models are useful for analysing performance in this experiment because they allow the difficulty parameters of the game and subject differences to be separated. The AT models provided the clearest evidence of learning or adaptation: after accounting for task difficulty, AT tended to decrease with progression, and the late model produced more accurate predictions, so it seems to be the more sensitive measure of performance improvement, while binary success captures another aspect of performance that is more affected by local variability, subjects' behaviour, and the adaptive nature of the game.

In conclusion, the analyses suggest that subjects' performance evolved during the experiment, but that this evolution was expressed differently depending on the outcome considered, and that this evolution was different among subjects. The models indicate that subjects generally became faster under comparable task conditions, while the binary simulations suggest an increase in predicted success probability in the late model.

Bibliography

[Agr13] Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2013, pp. 11–13.

[Bat+15] Douglas Bates et al. "Fitting linear mixed-effects models using lme4". In: *Journal of statistical software* 67 (2015), pp. 1–48.

[BC18] Markus Brauer and John J Curtin. "Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items." In: *Psychological methods* 23.3 (2018), pp. 389–411.

[Bol+09] Benjamin M Bolker et al. "Generalized linear mixed models: a practical guide for ecology and evolution". In: *Trends in ecology & evolution* 24.3 (2009), pp. 127–135.

[BPD18] Eva Berlot, Nicola J Popp, and Jörn Diedrichsen. "In search of the engram, 2017". In: *Current Opinion in Behavioral Sciences* 20 (2018). Habits and Skills, pp. 56–60. ISSN: 2352-1546. DOI: https://doi.org/10.1016/j.cobeha.2017.11.003.

[Cer+24] Matteo Ceradini et al. "Immersive VR for upper-extremity rehabilitation in patients with neurological disorders: a scoping review". In: *Journal of NeuroEngineering and Rehabilitation* (2024).

[Cip+18] Pietro Cipresso et al. "The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature". In: *Frontiers in psychology* 9 (2018), p. 2086.

[DSÖ17] Ashesh K Dhawale, Maurice A Smith, and Bence P Ölveczky. "The role of variability in motor learning". In: *Annual review of neuroscience* 40 (2017), pp. 479–498.

[Eve+25] Gauthier Everard et al. "A Self-Adaptive Serious Game to Improve Motor Learning Among Older Adults in Immersive Virtual Reality: Short-Term Longitudinal Pre-Post Study on Retention and Transfer". In: *JMIR AGING* (2025).

[Fen+18] Zhenan Feng et al. “Immersive virtual reality serious games for evacuation training and research: A systematic literature review”. In: *Computers Education* 127 (2018), pp. 252–266. ISSN: 0360-1315. DOI: https://doi.org/10.1016/j.compedu.2018.09.002.

[Gia+25] Vincenza Gianfredi et al. “Aging, longevity, and healthy aging: the public health approach”. In: *Aging Clinical and Experimental Research* 37.125 (2025).

[Gle+50] W Brier Glenn et al. “Verification of forecasts expressed in terms of probability”. In: *Monthly weather review* 78.1 (1950), pp. 1–3.

[HKP25] Yonggu Han, Seonggil Kim, and Sunwook Park. “Effects of virtual reality intervention on motor function in community-dwelling older adults: A systematic review and meta-analysis”. In: *Medicine* 104.30((e43488)) (2025).

[Høe+21] Emil Rosenlund Høeg et al. “System Immersion in Virtual Reality-Based Rehabilitation of Motor Function in Older Adults: A Systematic Review and Meta-Analysis”. In: *Frontiers in Virtual Reality* 2 (2021). ISSN: 2673-4192. DOI: 10.3389/frvir.2021.647993. URL: https://www.frontiersin.org/journals/virtual-reality/articles/10.3389/frvir.2021.647993.

[Kra+19] John W Krakauer et al. “Motor learning”. In: *Comprehensive physiology* 9.2 (2019), pp. 613–663.

[LS14] Danielle E Levac and Heidi Sveistrup. “Motor learning and virtual reality”. In: *Virtual reality for physical and motor rehabilitation*. Springer, 2014.

[Lui+25] Lasse F Lui et al. “The efficacy of adaptive training in immersive virtual reality for a fine motor skill task”. In: *Virtual Reality* 29.20 (2025).

[Maj22] Maja Pavlovic. *Log-normal Distribution – A simple explanation*. 2022. URL: https://towardsdatascience.com/log-normal-distribution-a-simple-explanation-7605864fb67c/ (visited on 04/29/2026).

[Mek+21] Destaw B. Mekbib et al. “A novel fully immersive virtual reality environment for upper extremity rehabilitation in patients with stroke”. In: *Annals of the New York Academy of Sciences* 1493.1 (2021), pp. 75–89. URL: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.14554.

[Mic25] Michael Brenndoerfer. *Multiple Linear Regression: Complete Guide with Formulas, Examples Python Implementation*. 2025. URL: https://mbrenndoerfer.com/writing/multiple-linear-regression-complete-guide-math-formulas-python-scikit-learn-implementation (visited on 11/05/2026).

[OM07] Ann L Oberg and Douglas W Mahoney. "Linear mixed effects models". In: *Topics in biostatistics* (2007), pp. 213–234.

[Sco26] Scopus. *Documents by year with keywords 'Virtual Reality' - From 1990 to 2025*. 2026. URL: https://www.scopus.com/term/analyzer.uri?sort=plf-f&src=s&sid=83d9cf3fde9a91d1ed829281381459b6&sot=a&sdt=a&sl=30&s=TITLE-ABS-KEY%28virtual+reality%29&origin=resultslist&count=10&analyzeResults=Analyze+results (visited on 04/20/2026).

[Sei10] Rachael D Seidler. "Neural correlates of motor learning, transfer of learning, and learning to learn". In: *Exercise and sport sciences reviews* 38.1 (2010), pp. 3–9.

[Smi+04] Anne C Smith et al. "Dynamic analysis of learning in behavioral experiments". In: *Journal of Neuroscience* 24.2 (2004), pp. 447–461.

[staa] statsmodels Developers. *BinomialBayesMixedGLM*. https://www.statsmodels.org/stable/generated/statsmodels.genmod.bayes_mixed_glm.BinomialBayesMixedGLM.html. Accessed 2026-05-11.

[stab] statsmodels Developers. *Linear Mixed Effects Models*. https://www.statsmodels.org/stable/mixed_linear.html. Accessed 2026-05-11.

[Ste+10] Ewout W Steyerberg et al. "Assessing the performance of prediction models: a framework for traditional and novel measures". In: *Epidemiology* 21.1 (2010), pp. 128–138.

[Ste21] Steven M. *Où acheter un Oculus Quest 2?* 2021. URL: https://www.news-console.fr/acheter-oculus-quest-2/ (visited on 07/04/2026).

[Sve04] Heidi Sveistrup. "Motor rehabilitation using virtual reality". In: *Journal of neuroengineering and rehabilitation* 1.10 (2004). DOI: https://doi.org/10.1186/1743-0003-1-10.

[UG13] Gülden Kaya Uyanık and Neşe Güler. "A study on multiple linear regression analysis". In: *Procedia-Social and Behavioral Sciences* 106 (2013), pp. 234–240. DOI: https://doi.org/10.1016/j.sbspro.2013.12.027.

[Wor23] World Bank. *Life expectancy at birth, total (years)*. 2023. URL: https://data.worldbank.org/indicator/SP.DYN.LE00.IN (visited on 05/02/2026).

Appendix A

Raw Action Times – Other Subjects

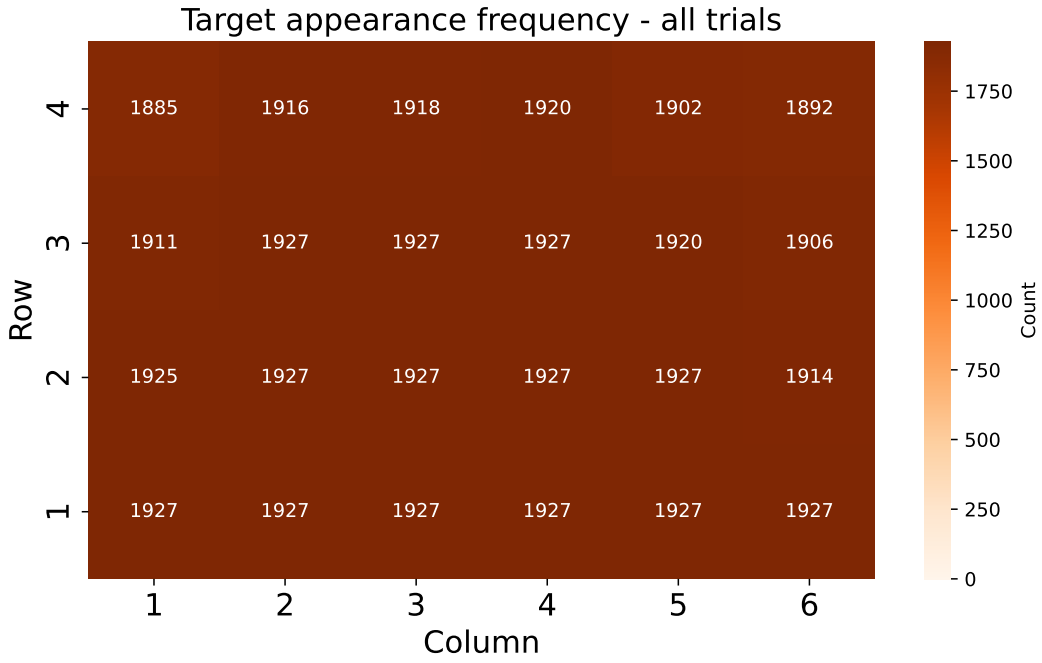

Figure A.1: Raw counts of the target's appearance locations.

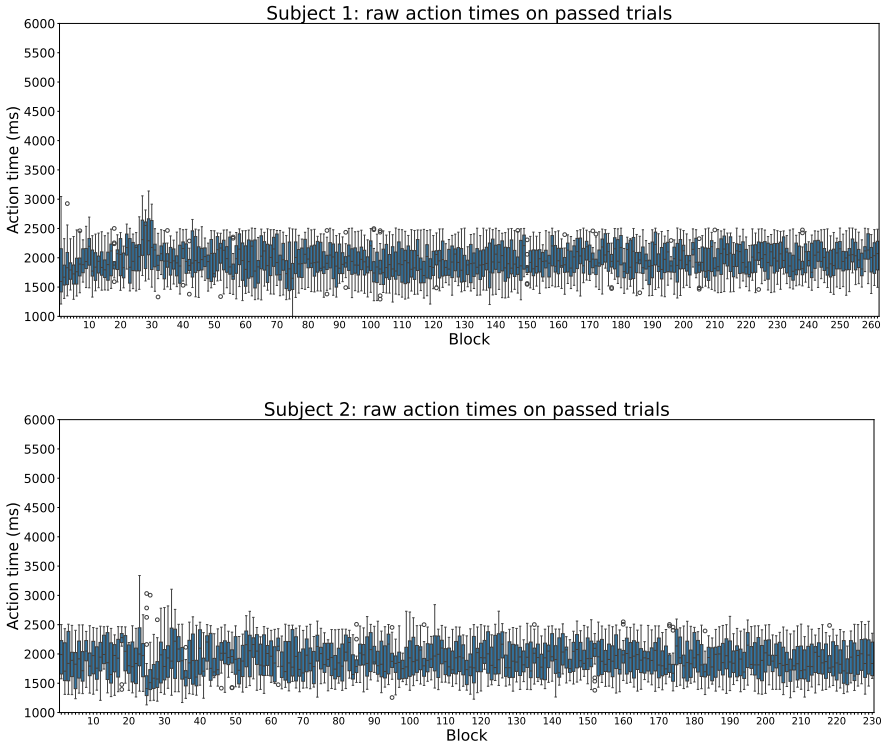

Figure A.2: AT on passed trials for subjects 1 and 2.

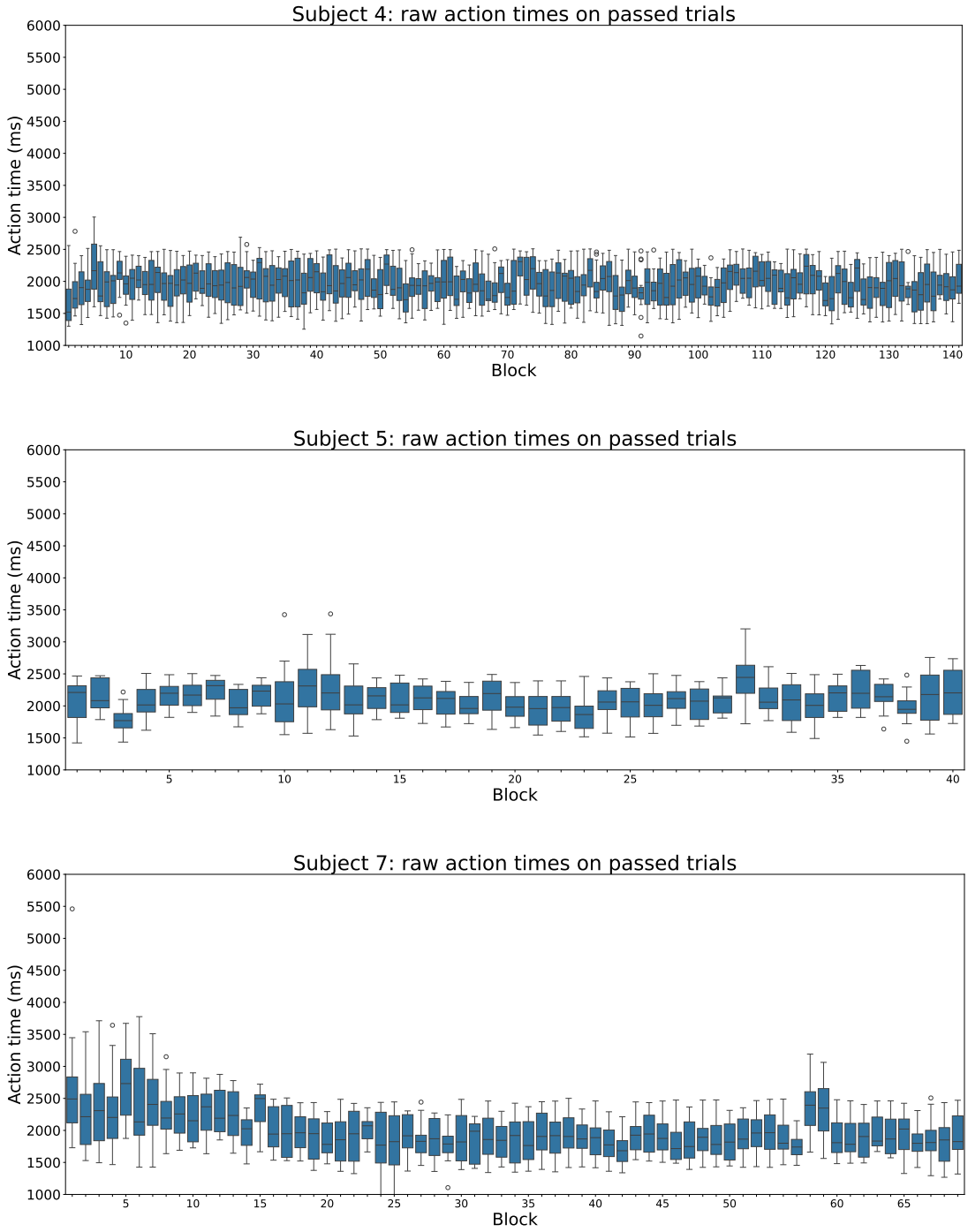

Figure A.3: AT on passed trials for subjects 4, 5, and 7.

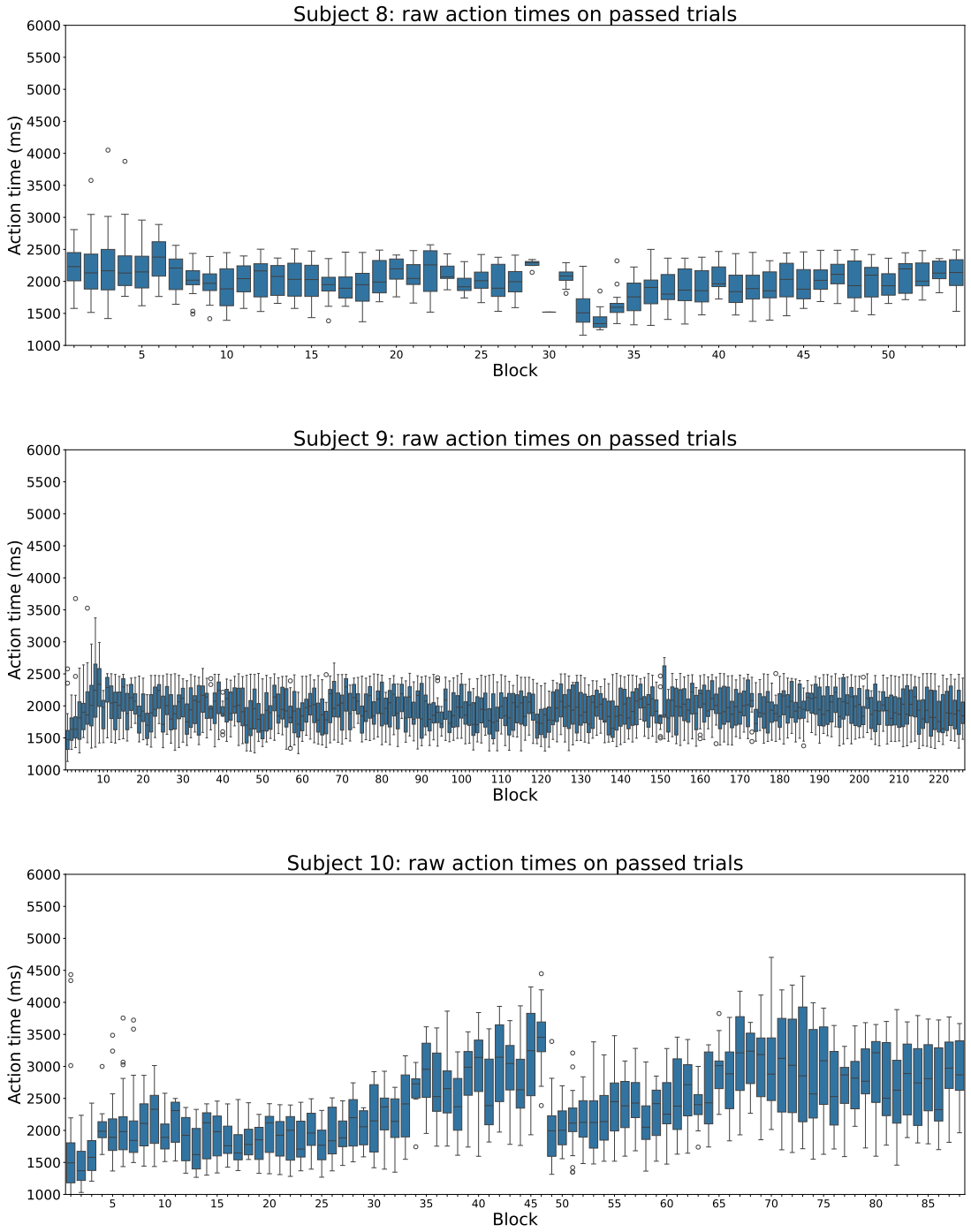

Figure A.4: AT on passed trials for subjects 8, 9, and 10.

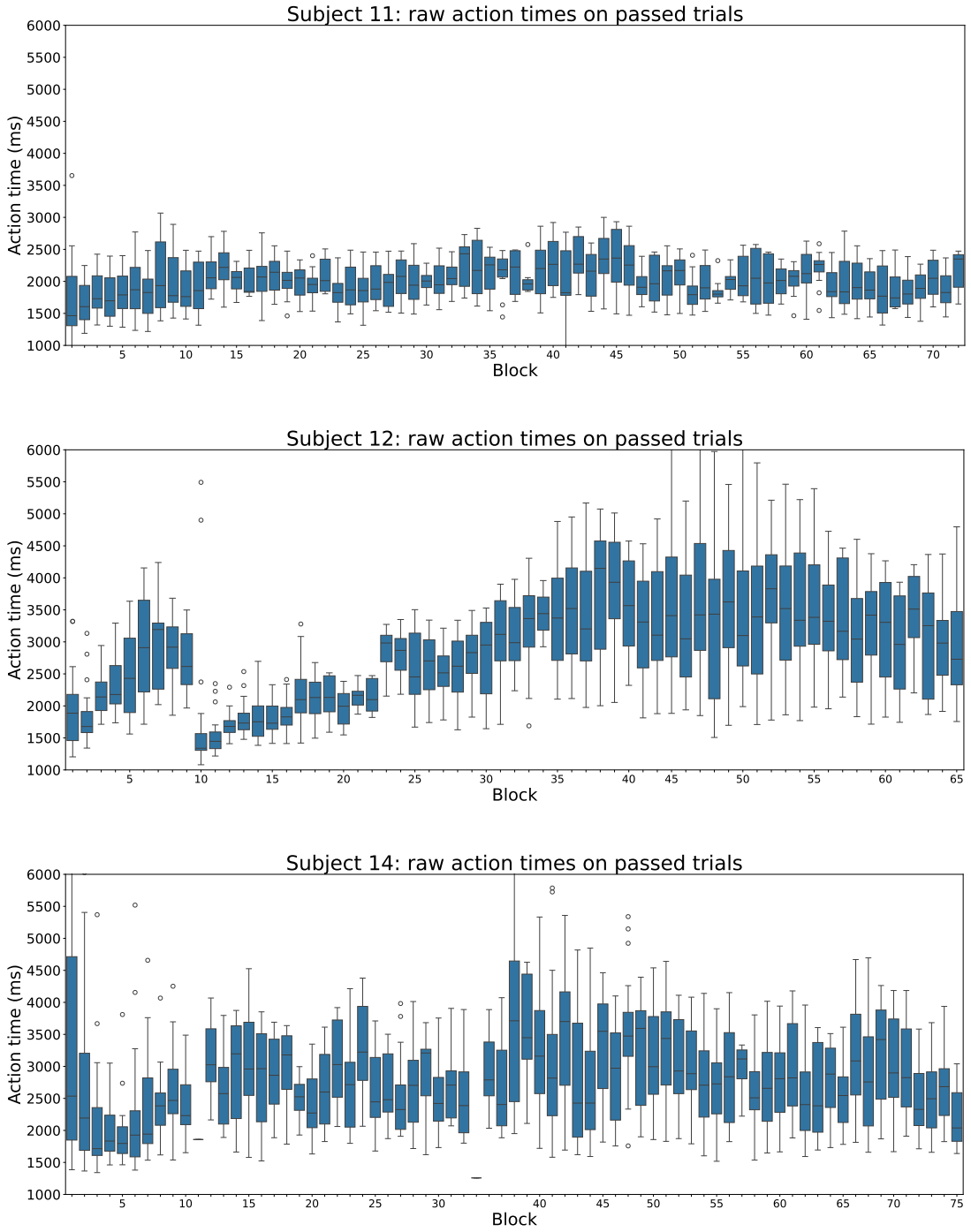

Figure A.5: AT on passed trials for subjects 11, 12, and 14.

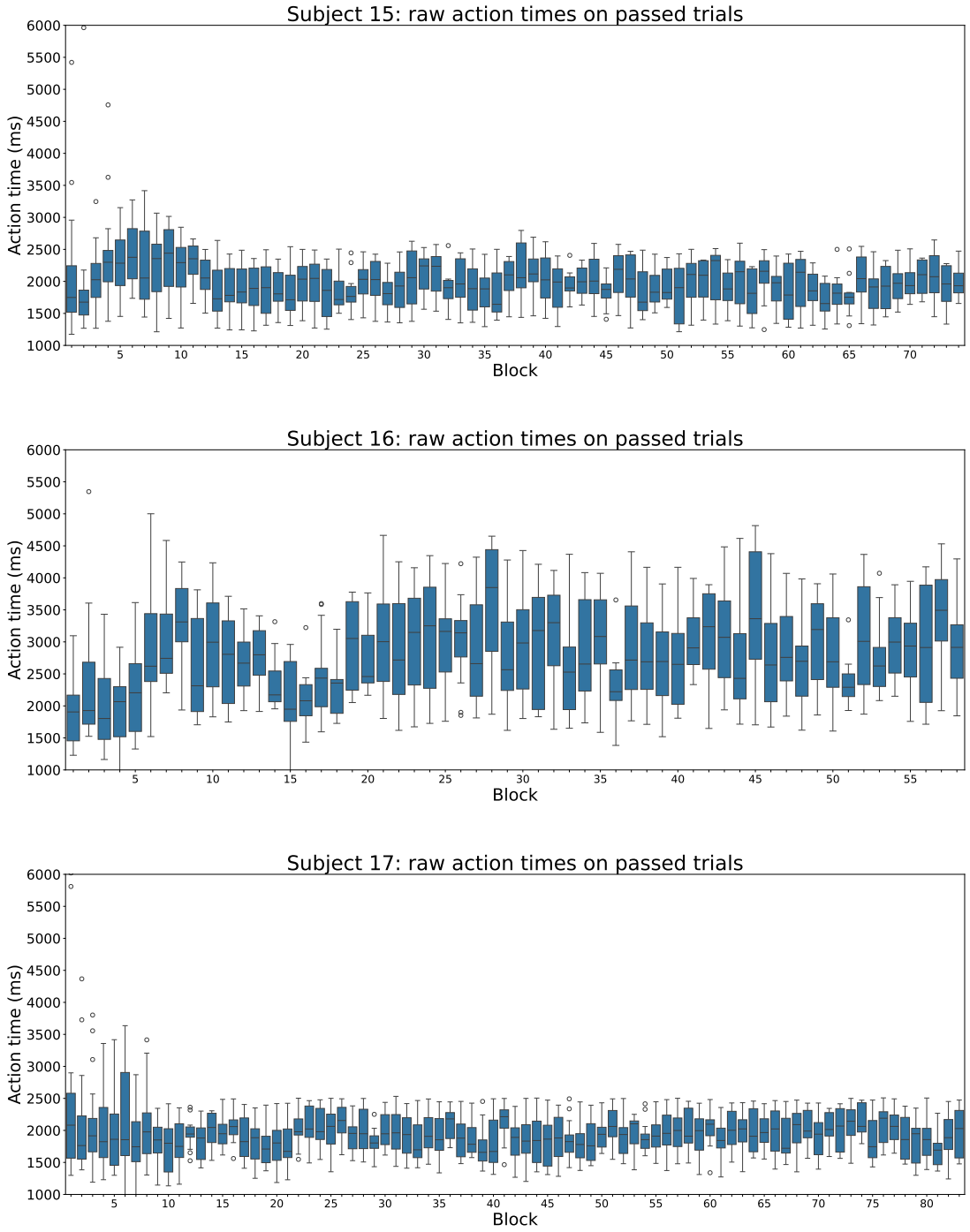

Figure A.6: AT on passed trials for subjects 15, 16, and 17.

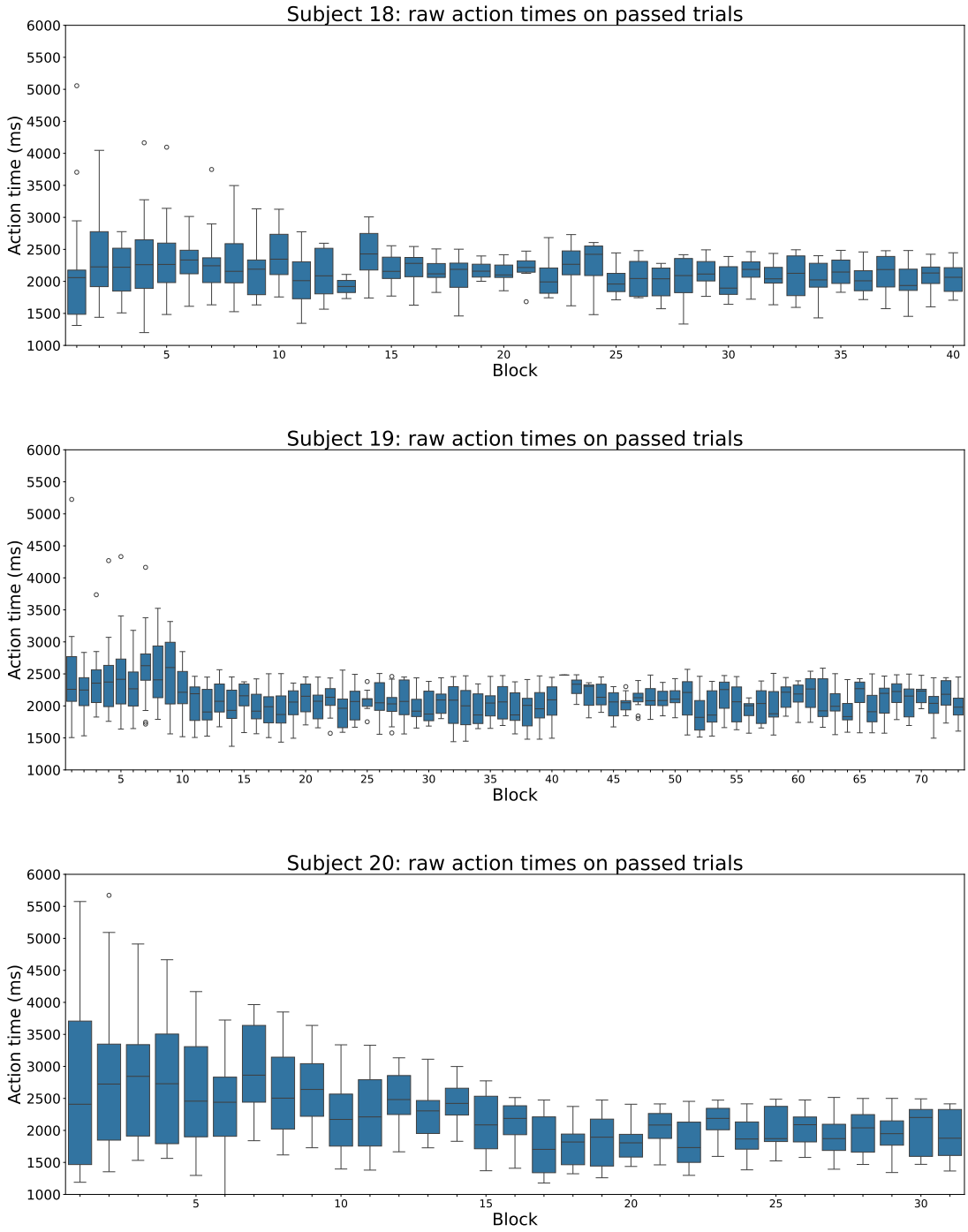

Figure A.7: AT on passed trials for subjects 18, 19, and 20.

Appendix B

Raw Results – Other Subjects

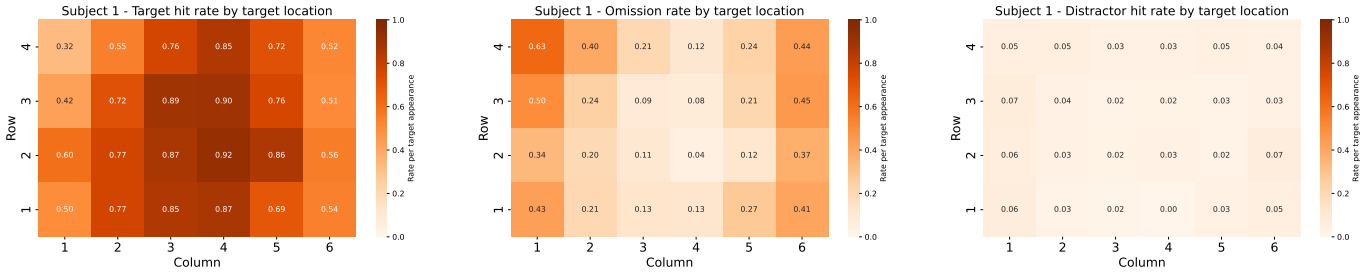

Figure B.1: Trial outcome rates by target location for subject 1.

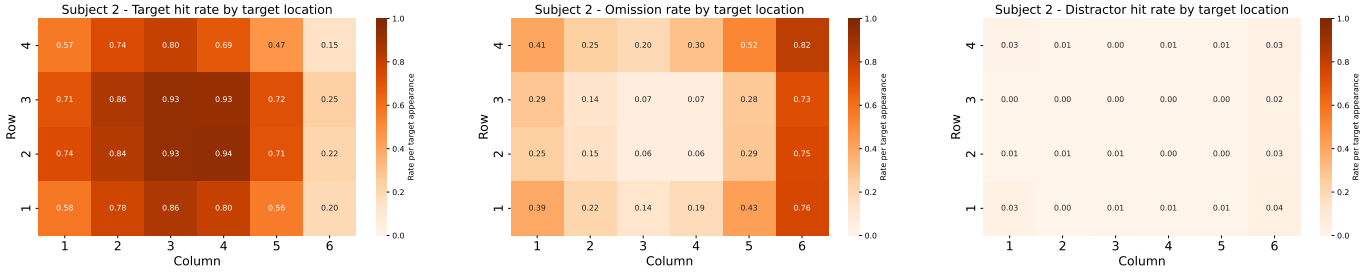

Figure B.2: Trial outcome rates by target location for subject 2.

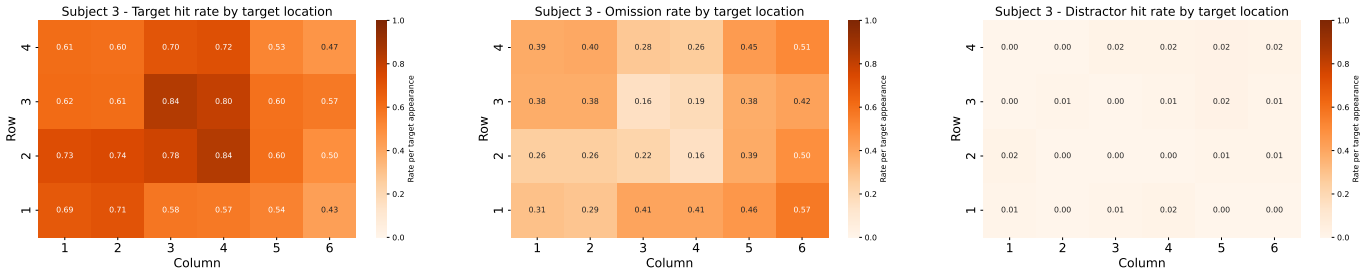

Figure B.3: Trial outcome rates by target location for subject 3.

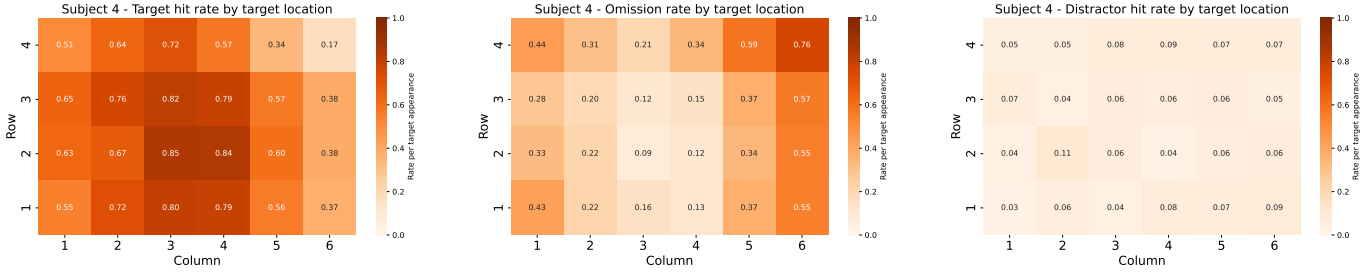

Figure B.4: Trial outcome rates by target location for subject 4.

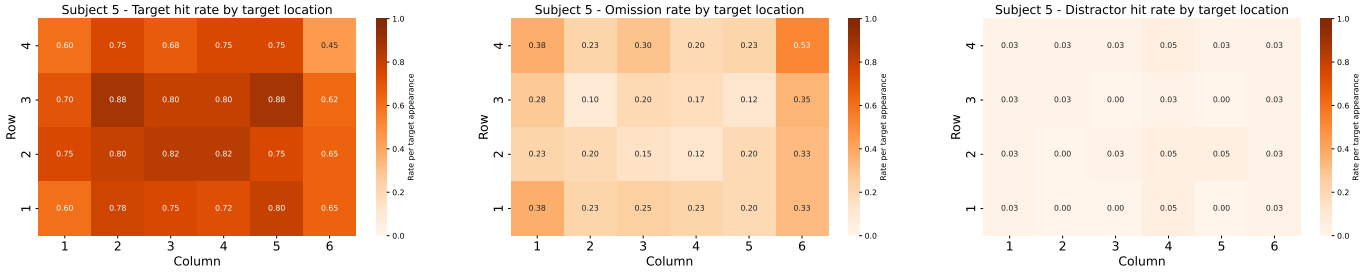

Figure B.5: Trial outcome rates by target location for subject 5.

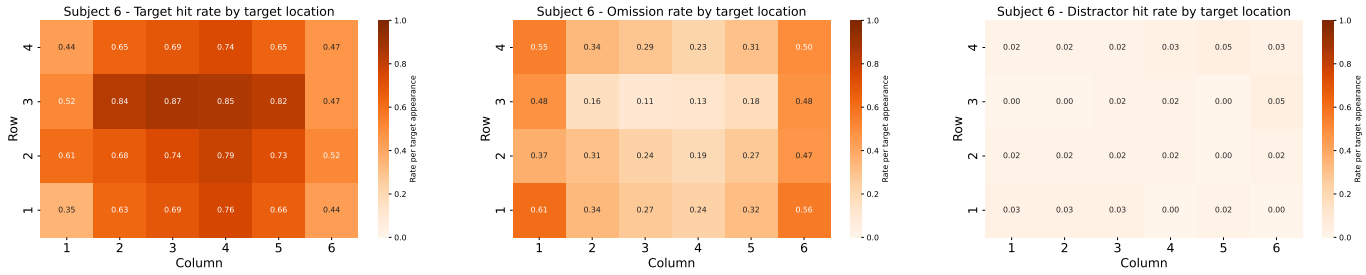

Figure B.6: Trial outcome rates by target location for subject 6.

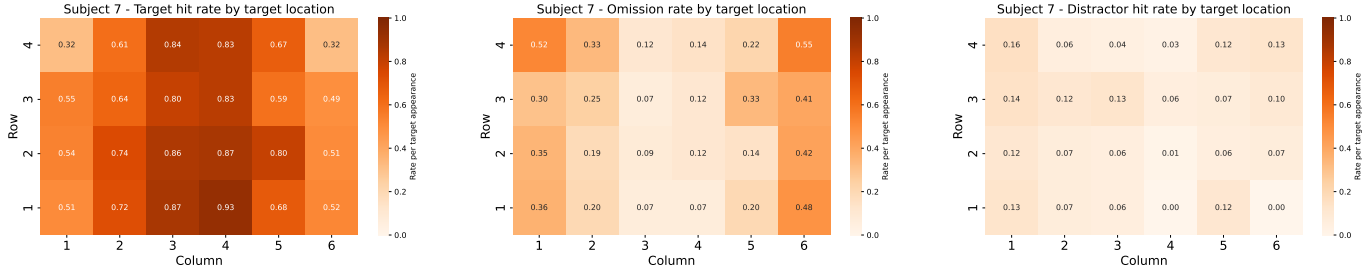

Figure B.7: Trial outcome rates by target location for subject 7.

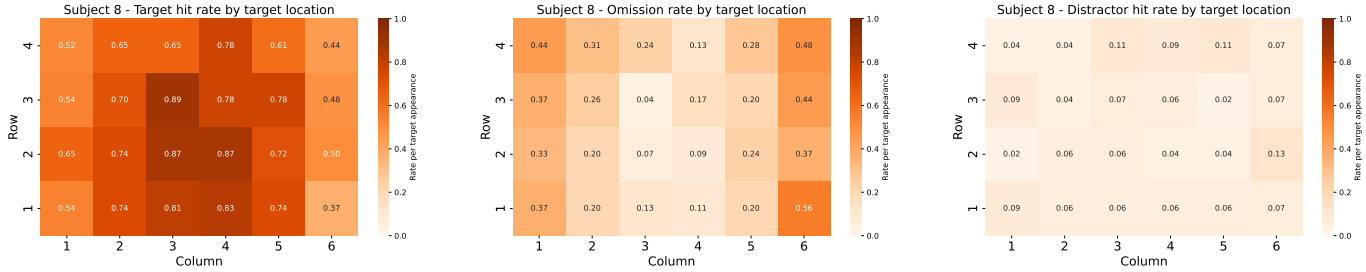

Figure B.8: Trial outcome rates by target location for subject 8.

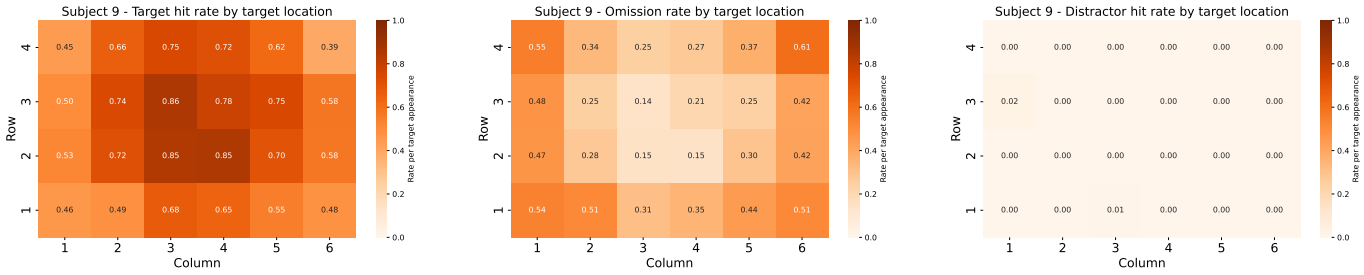

Figure B.9: Trial outcome rates by target location for subject 9.

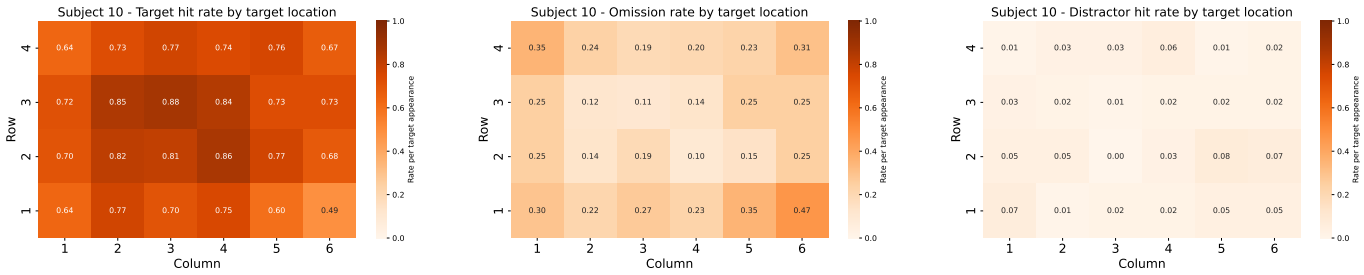

Figure B.10: Trial outcome rates by target location for subject 10.

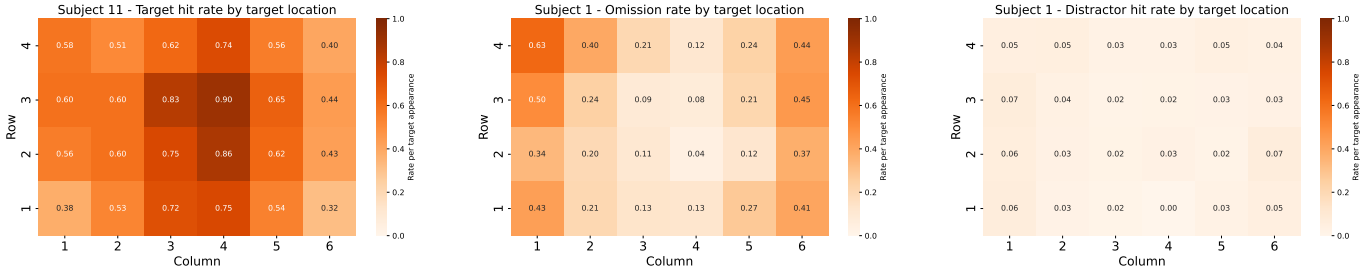

Figure B.11: Trial outcome rates by target location for subject 11.

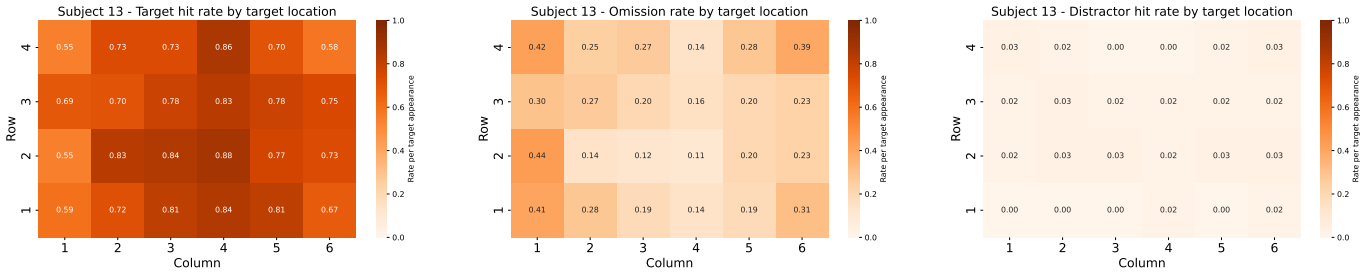

Figure B.12: Trial outcome rates by target location for subject 13.

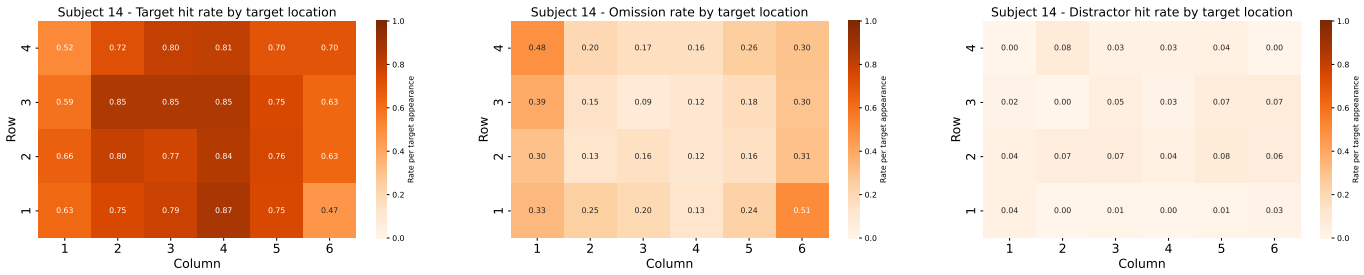

Figure B.13: Trial outcome rates by target location for subject 14.

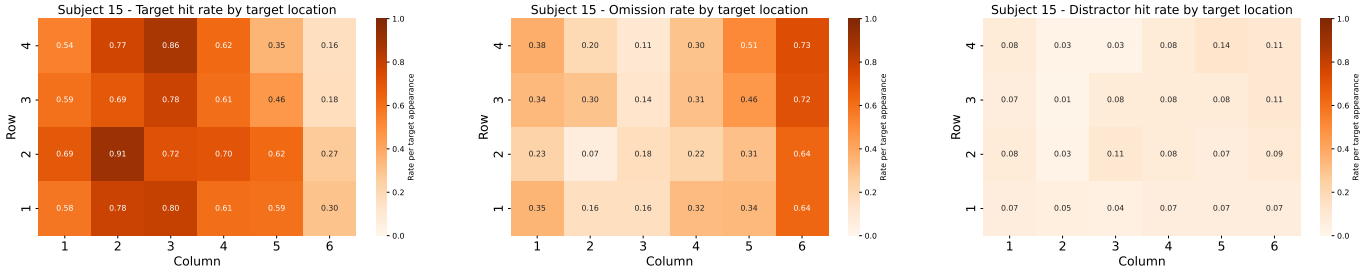

Figure B.14: Trial outcome rates by target location for subject 15.

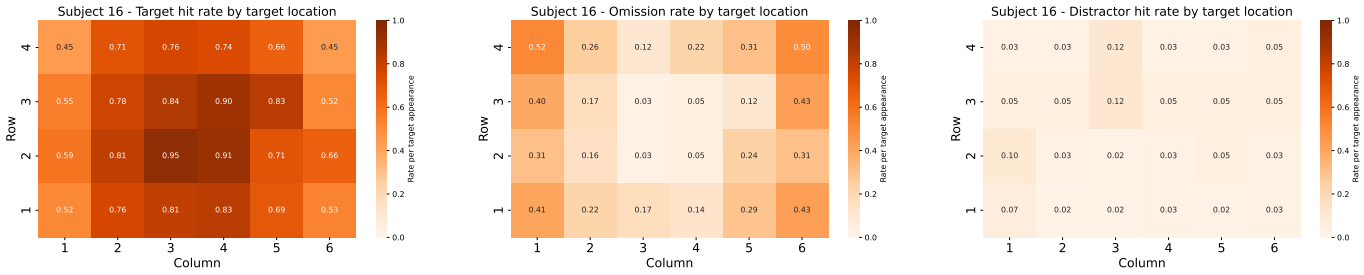

Figure B.15: Trial outcome rates by target location for subject 16.

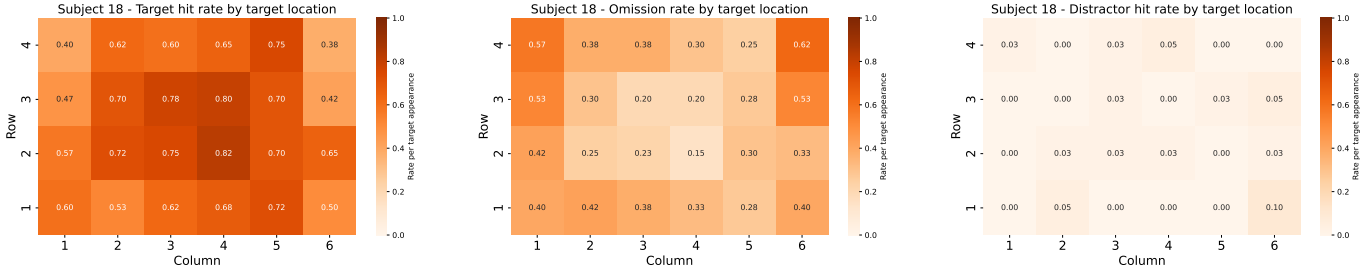

Figure B.16: Trial outcome rates by target location for subject 18.

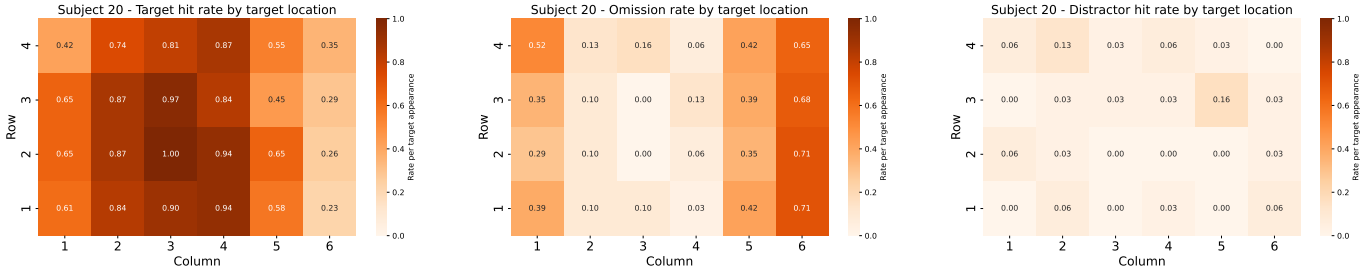

Figure B.17: Trial outcome rates by target location for subject 20.

Appendix C

Action Time Models – Other Subjects

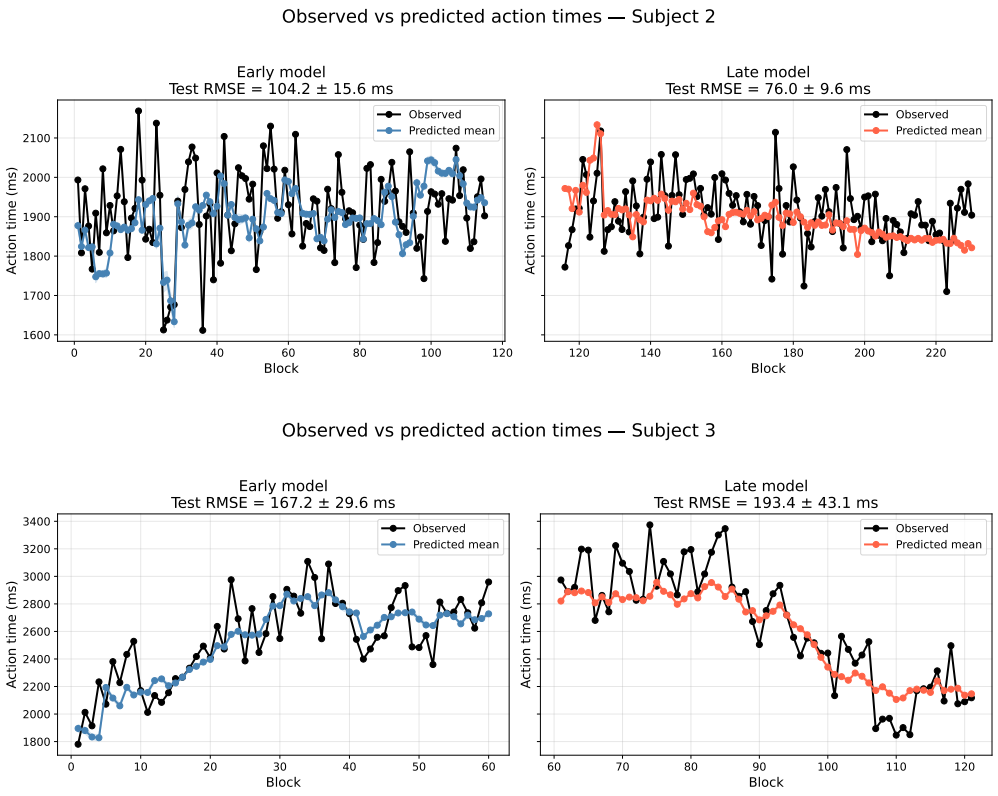

Figure C.1: Observed and predicted AT on the test set for subjects 2 and 3 in the repeated 80–20 test procedure.

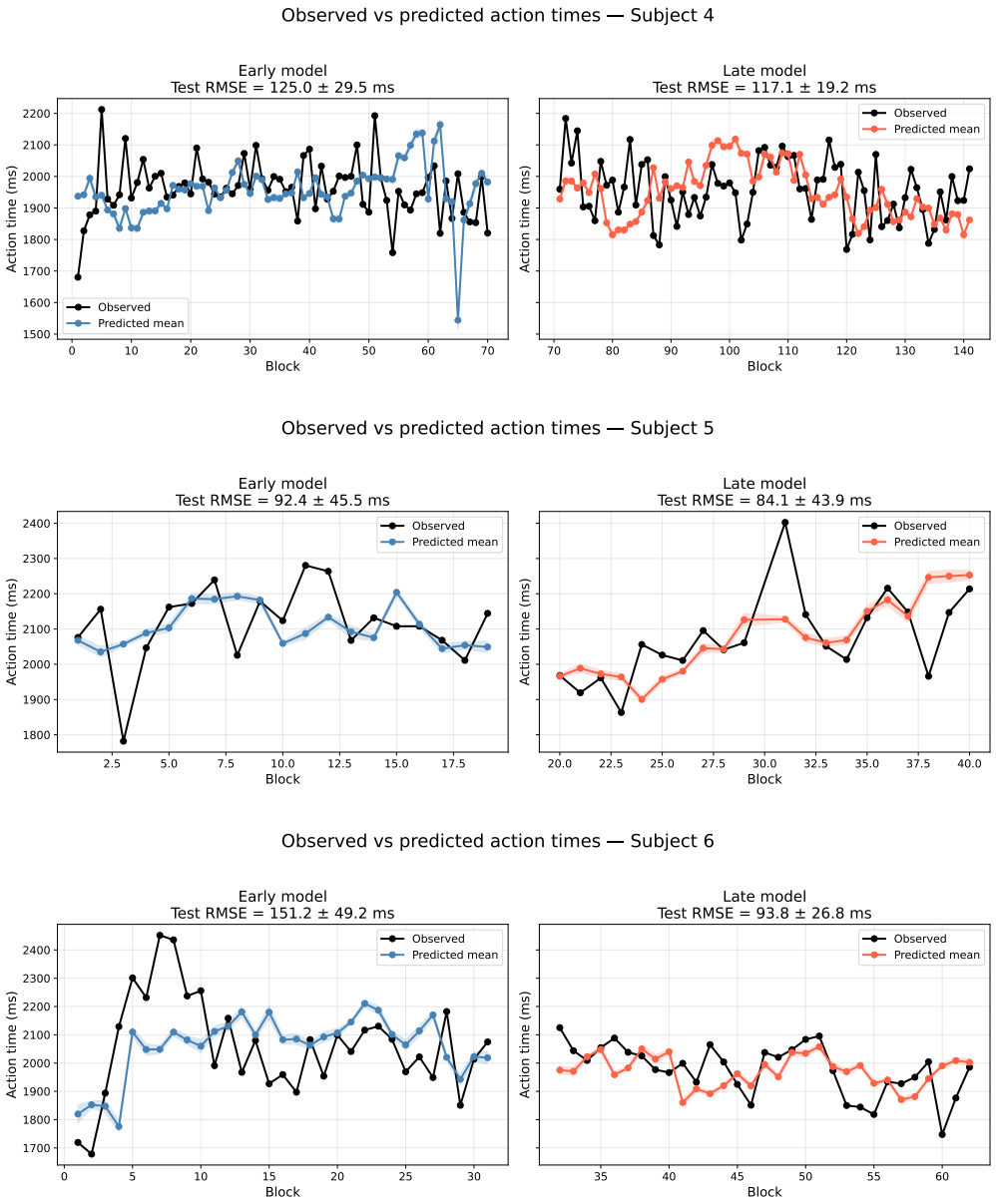

Figure C.2: Observed and predicted AT on the test set for subjects 4, 5, and 6 in the repeated 80–20 test procedure.

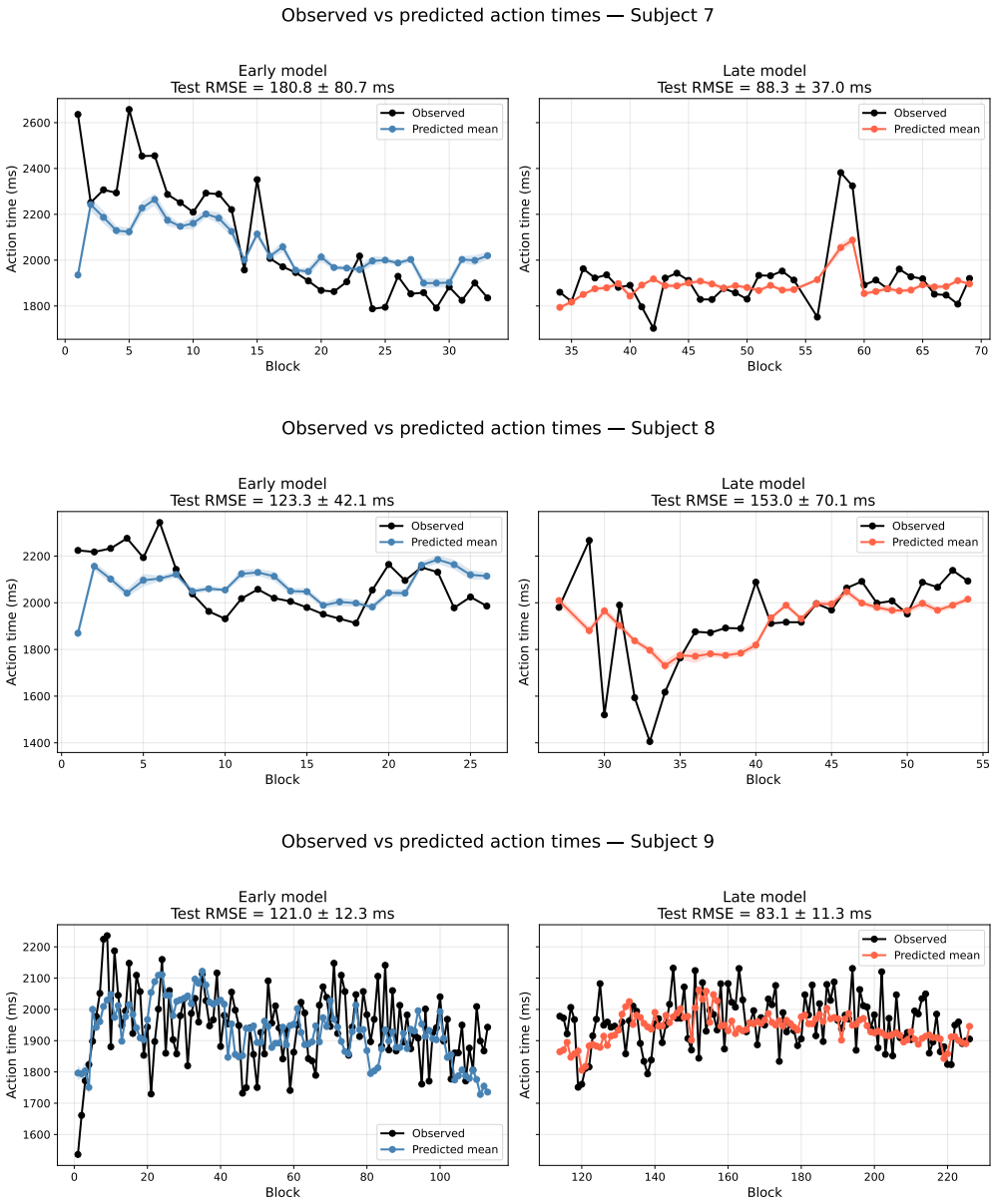

Figure C.3: Observed and predicted AT on the test set for subjects 7, 8, and 9 in the repeated 80–20 test procedure.

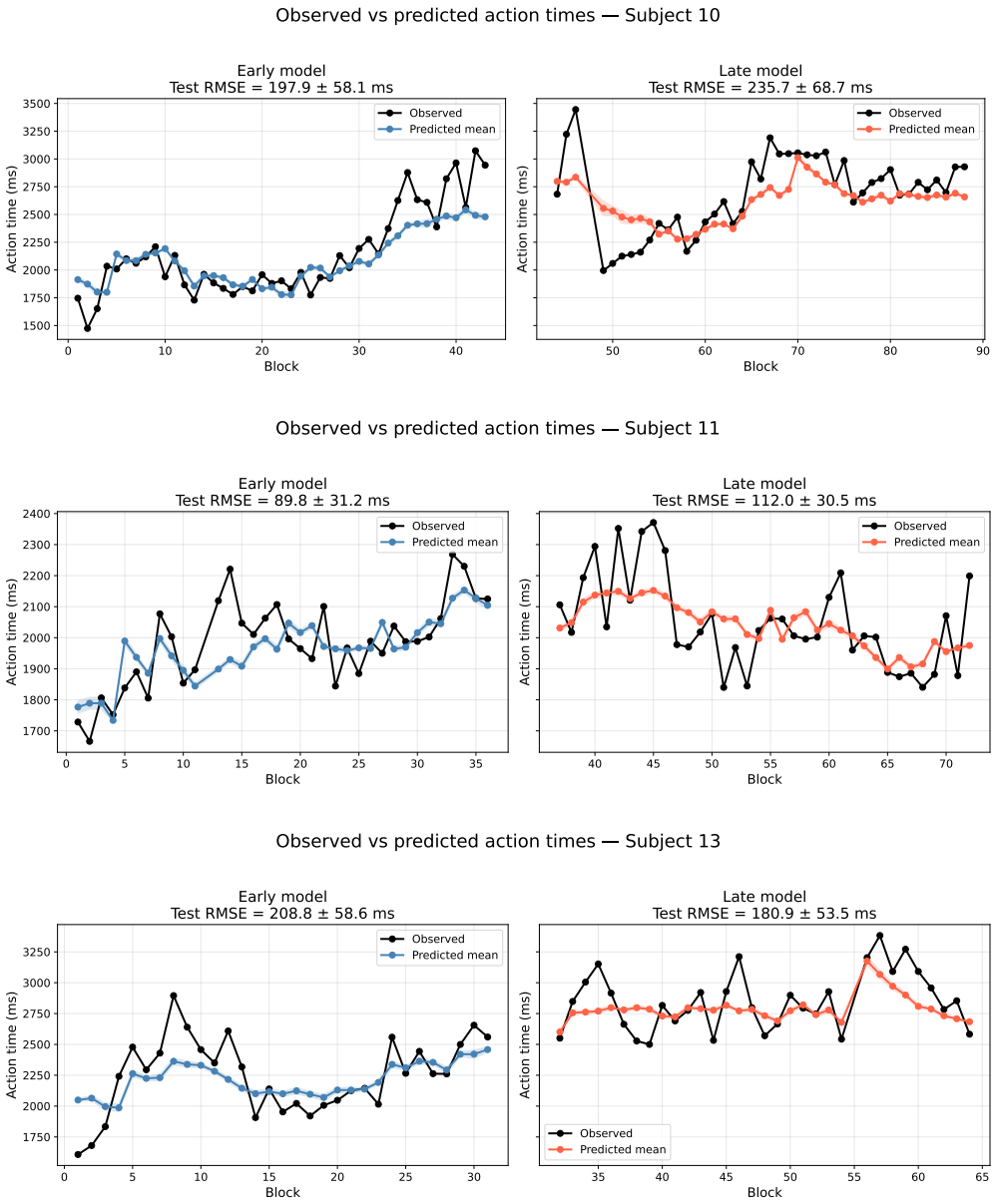

Figure C.4: Observed and predicted AT on the test set for subjects 10, 11, and 13 in the repeated 80–20 test procedure.

Figure C.5: Observed and predicted AT on the test set for subjects 14, 15, and 16 in the repeated 80–20 test procedure.

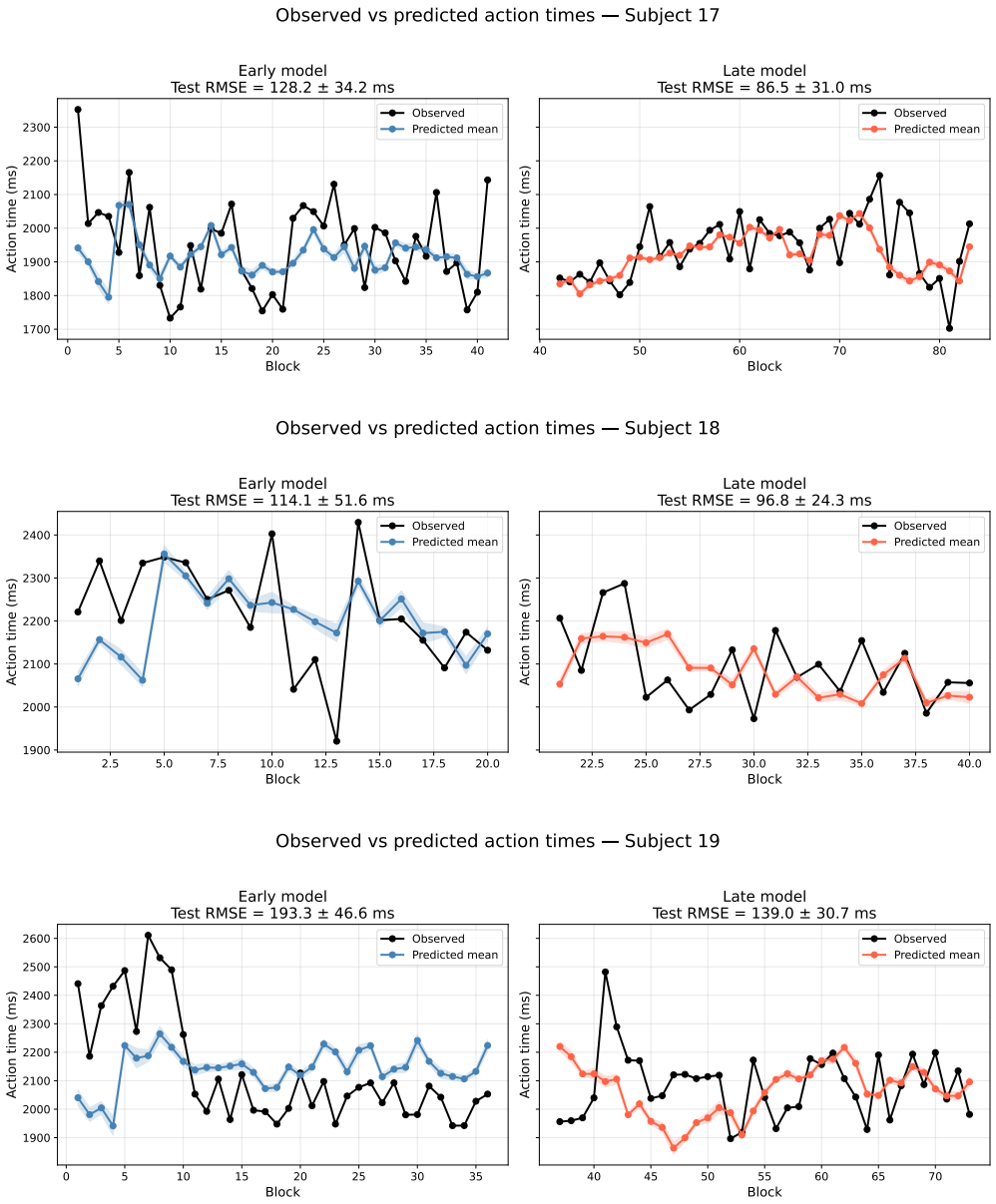

Figure C.6: Observed and predicted AT on the test set for subjects 17, 18, and 19 in the repeated 80–20 test procedure.

Appendix D

Binary Hit Models – Other Subjects

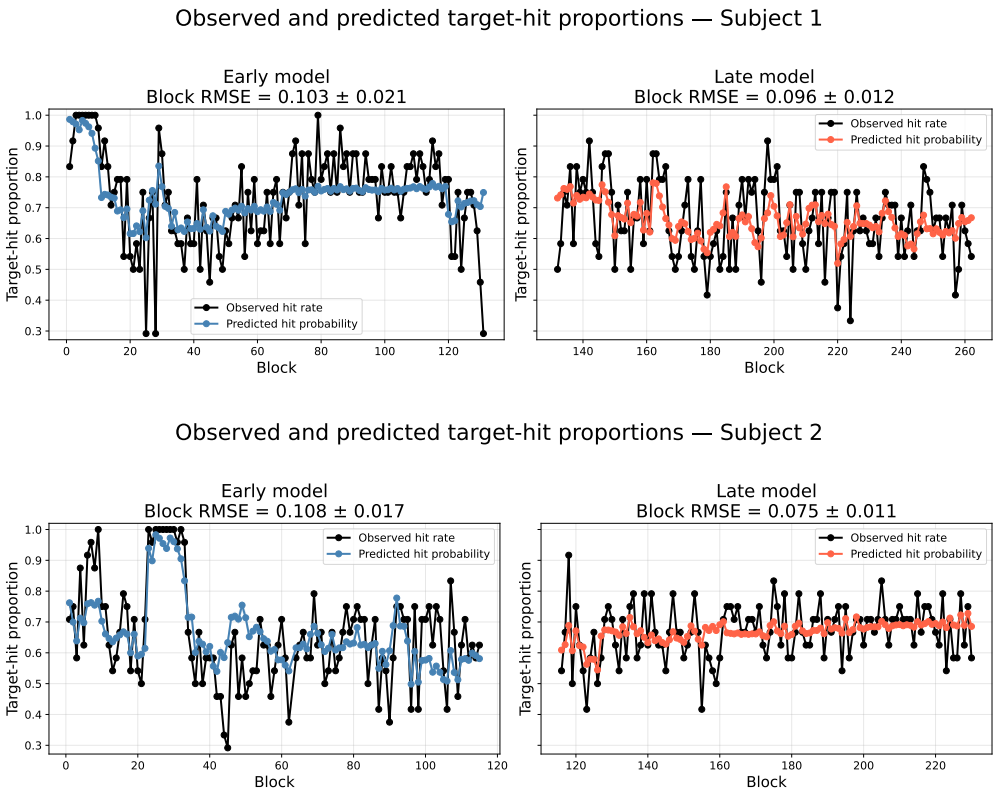

Figure D.1: Observed hit rates and predicted hit probabilities on the test set for subjects 1 and 2 in the repeated 80–20 test procedure.

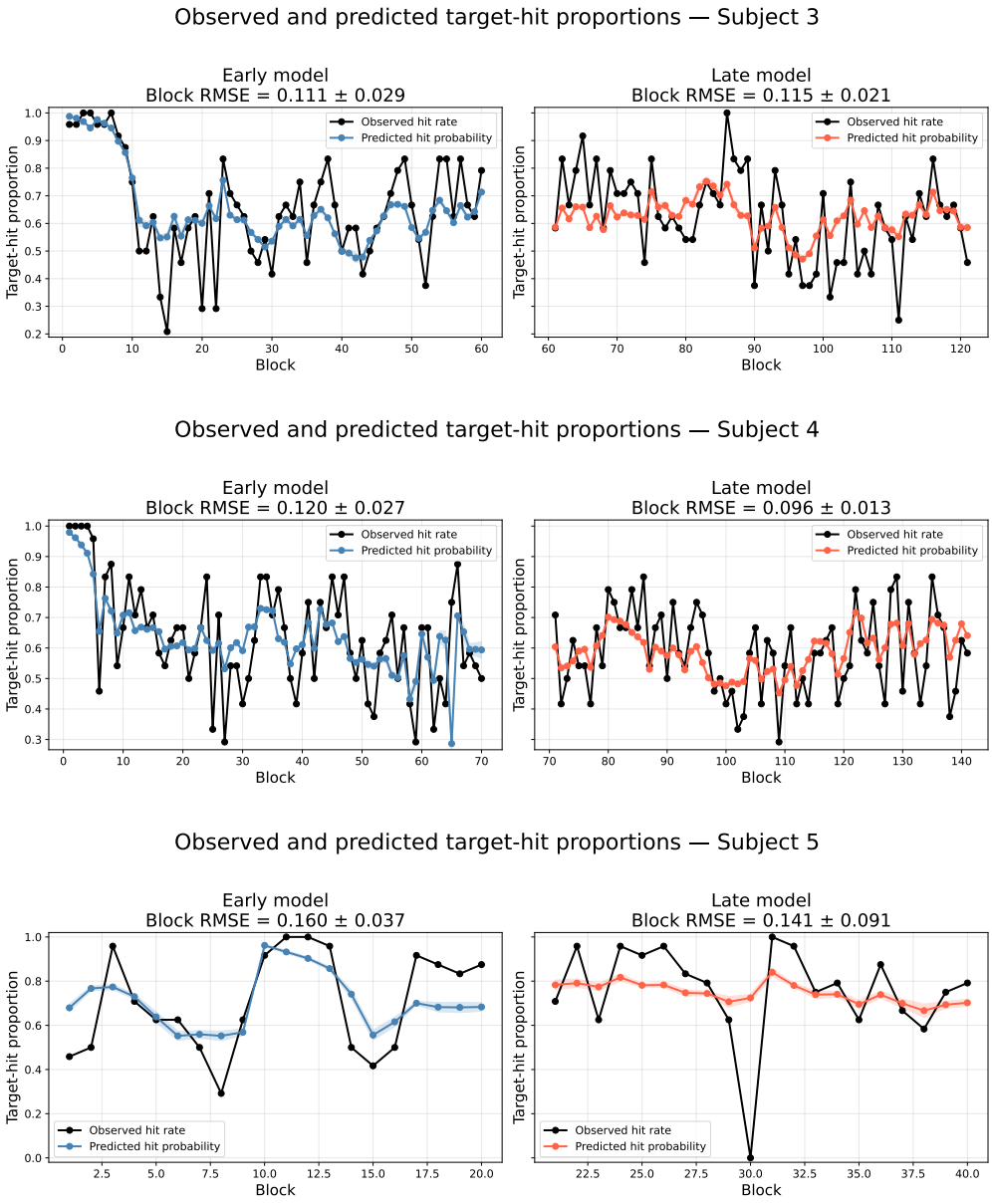

Figure D.2: Observed hit rates and predicted hit probabilities on the test set for subjects 3, 4, and 5 in the repeated 80–20 test procedure.

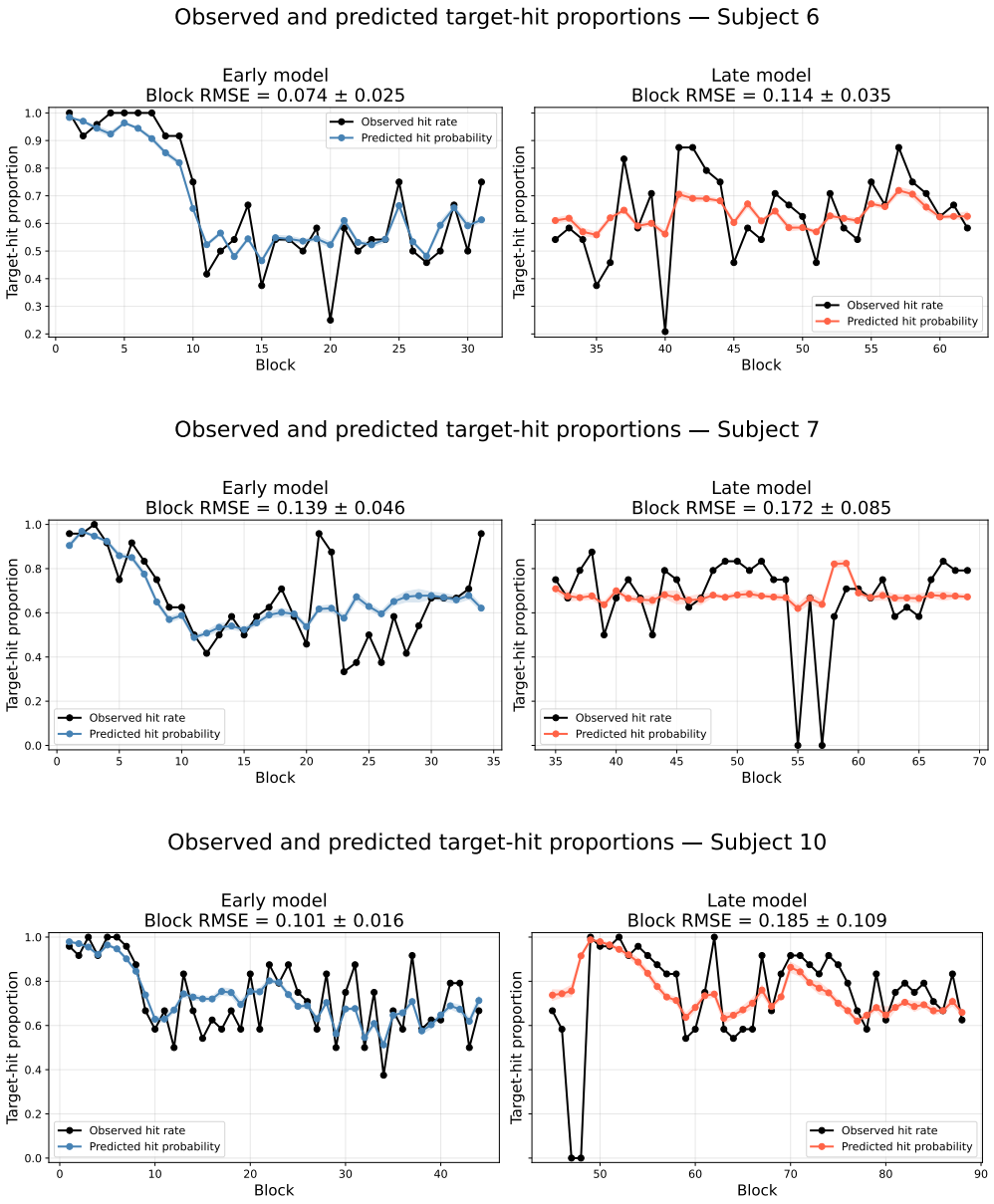

Figure D.3: Observed hit rates and predicted hit probabilities on the test set for subjects 6, 7, and 10 in the repeated 80–20 test procedure.

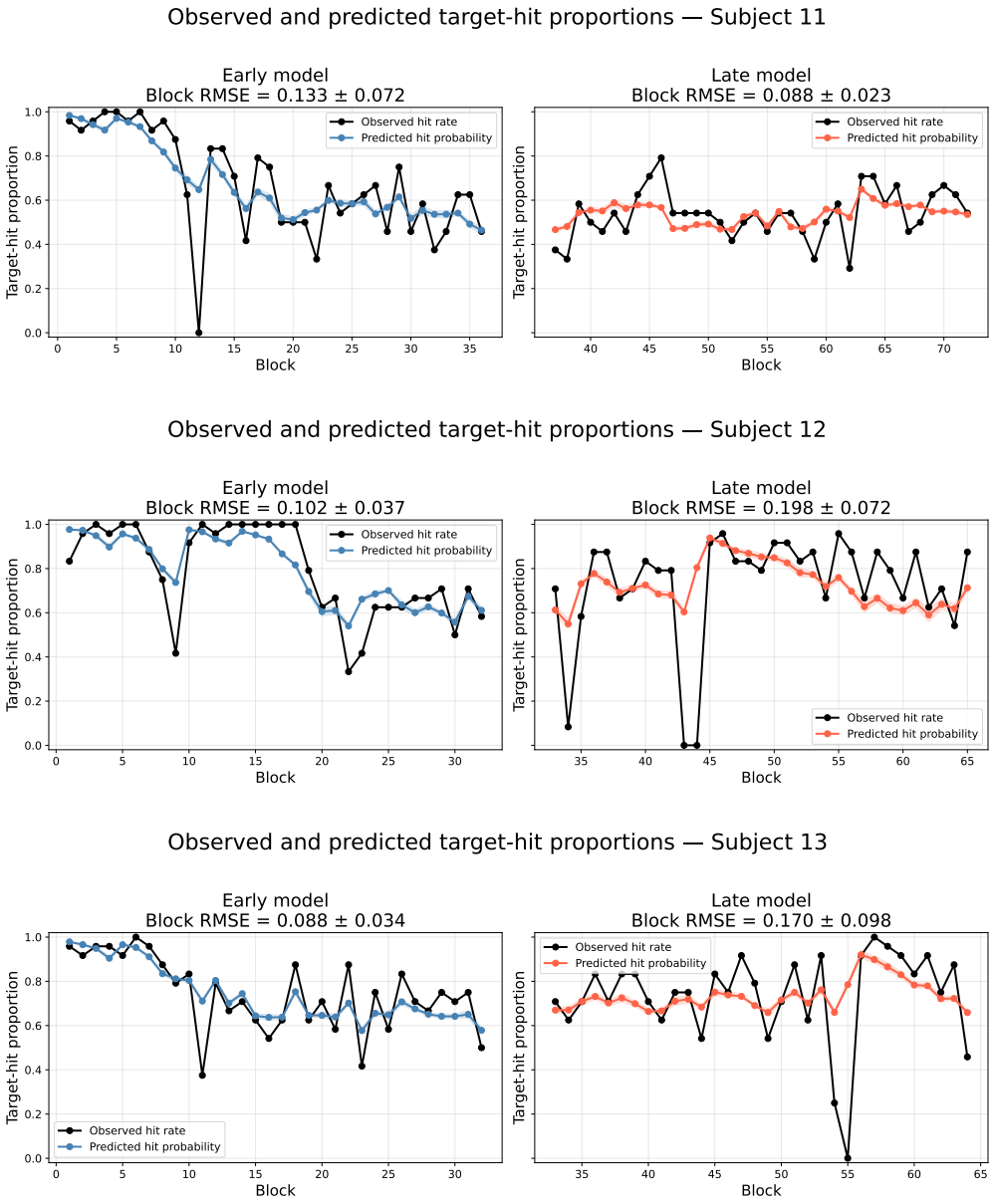

Figure D.4: Observed hit rates and predicted hit probabilities on the test set for subjects 11, 12, and 13 in the repeated 80–20 test procedure.

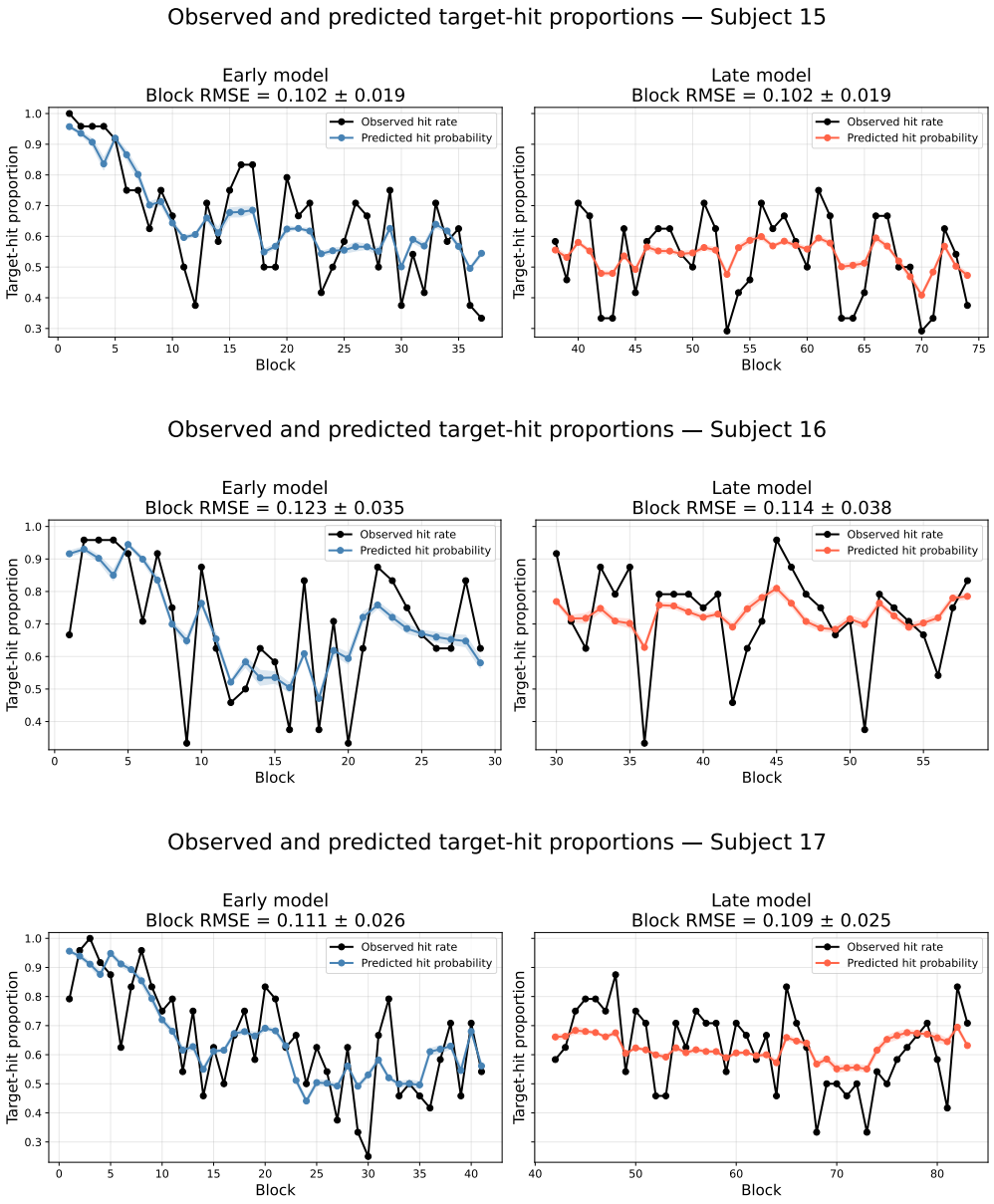

Figure D.5: Observed hit rates and predicted hit probabilities on the test set for subjects 15, 16, and 17 in the repeated 80–20 test procedure.

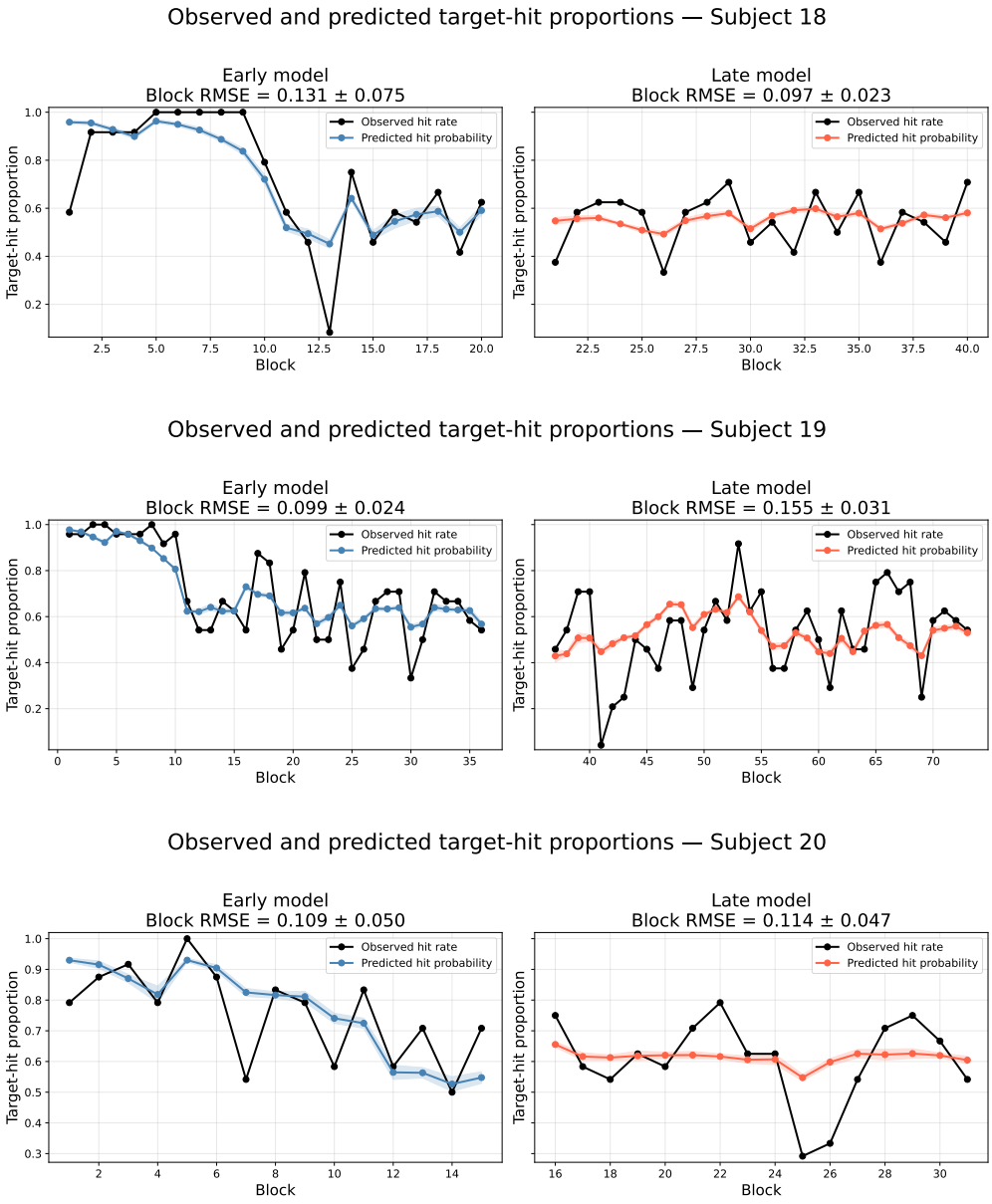

Figure D.6: Observed hit rates and predicted hit probabilities on the test set for subjects 18, 19, and 20 in the repeated 80–20 test procedure.

Appendix E

Random Slope

Model	Train RMSE (ms)	Test RMSE (ms)	Test RMSE (log)	Test MAE (ms)
Early model	289.90	285.75	0.129	210.14
Late model	252.04	250.82	0.100	176.29

Table E.1: Mean repeated 80–20 test performance of the AT LMEMs across 50 random splits, with an added random slope per subject.

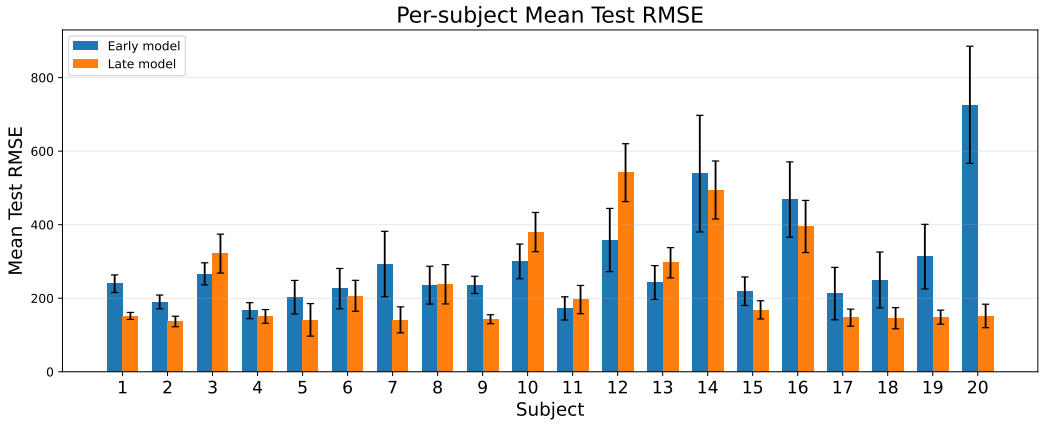

Figure E.1: Test RMSE computed separately for each subject for the early and late AT models, with an added random slope per subject.

	Early model			Late model		
Subject	Mean RMSE (ms)	Mean MAE (ms)	Mean Test obs.	Mean RMSE (ms)	Mean MAE (ms)	Mean Test obs.
1	239.4 ± 24.0	197.5 ± 21.8	91.6	151.8 ± 9.5	121.0 ± 8.0	83.6
2	189.9 ± 18.7	153.7 ± 16.2	66.4	136.7 ± 14.3	109.7 ± 11.7	68.7
3	266.4 ± 30.0	208.7 ± 19.8	36.0	321.3 ± 52.9	249.8 ± 40.5	35.9
4	166.2 ± 21.8	133.0 ± 16.5	41.9	150.6 ± 18.6	120.7 ± 15.1	40.3
5	202.8 ± 45.6	179.8 ± 45.4	12.7	141.2 ± 44.3	111.4 ± 35.5	16.1
6	226.1 ± 54.8	180.2 ± 47.7	18.1	206.6 ± 42.1	176.5 ± 39.9	18.2
7	293.0 ± 89.0	230.8 ± 66.3	18.3	141.2 ± 35.4	113.5 ± 23.6	21.9
8	235.6 ± 51.5	182.1 ± 41.8	17.1	237.9 ± 53.4	197.4 ± 49.3	15.9
9	236.4 ± 23.4	191.5 ± 21.0	66.2	142.9 ± 12.2	113.8 ± 10.7	68.4
10	300.1 ± 47.2	243.4 ± 43.0	27.0	379.9 ± 53.4	309.3 ± 44.0	29.4
11	172.4 ± 31.6	138.9 ± 30.2	20.6	196.3 ± 38.5	163.4 ± 33.0	18.1
12	358.3 ± 85.9	297.9 ± 76.7	20.7	541.5 ± 78.7	442.3 ± 78.0	23.5
13	242.9 ± 45.8	197.6 ± 42.1	22.2	296.5 ± 41.1	244.9 ± 35.2	21.0
14	538.8 ± 158.6	397.4 ± 97.1	18.6	494.4 ± 78.9	389.1 ± 72.1	24.4
15	219.1 ± 38.8	172.3 ± 32.6	22.4	168.4 ± 24.8	133.7 ± 20.2	19.2
16	468.4 ± 102.5	386.1 ± 90.9	17.8	395.1 ± 70.9	325.2 ± 62.5	18.2
17	212.6 ± 71.3	159.3 ± 37.0	23.4	147.1 ± 23.2	120.6 ± 20.8	23.3
18	249.7 ± 76.0	198.6 ± 56.0	12.8	145.7 ± 28.8	118.9 ± 26.6	10.3
19	313.2 ± 87.9	228.1 ± 69.1	22.7	148.5 ± 19.1	121.1 ± 16.5	19.5
20	726.1 ± 159.2	650.8 ± 144.3	11.0	151.8 ± 31.9	126.3 ± 27.6	10.1

Table E.2: Per-subject mean test RMSE and MAE for the AT models at the chunk level across 50 train-test splits when adding a random slope per subject.

Model	Set	AUC	Brier score	Accuracy
Early model	Train	0.652 ± 0.005	0.211 ± 0.003	0.657 ± 0.006
Early model	Test	0.650 ± 0.015	0.211 ± 0.007	0.656 ± 0.014
Late model	Train	0.563 ± 0.011	0.242 ± 0.008	0.622 ± 0.008
Late model	Test	0.565 ± 0.014	0.241 ± 0.007	0.621 ± 0.011

Table E.3: Mean performance of the binary hit mixed-effects models with a random slope for block across the 50 random 80–20 train–test splits.

Model	Set	Block RMSE	Block MAE	Bias
Early model	Train	0.172 ± 0.007	0.134 ± 0.006	0.039 ± 0.004
Early model	Test	0.176 ± 0.015	0.137 ± 0.010	0.039 ± 0.010
Late model	Train	0.208 ± 0.020	0.154 ± 0.014	0.007 ± 0.007
Late model	Test	0.206 ± 0.016	0.153 ± 0.013	0.005 ± 0.011

Table E.4: Block-level prediction performance of the binary hit mixed-effects models with a random slope for block across the 50 random 80–20 train–test splits.

	Early model			Late model		
Subject	Block RMSE	Block MAE	Bias	Block RMSE	Block MAE	Bias
1	0.164 ± 0.019	0.137 ± 0.019	0.104 ± 0.027	0.108 ± 0.016	0.088 ± 0.013	0.038 ± 0.032
2	0.154 ± 0.019	0.126 ± 0.016	0.103 ± 0.024	0.103 ± 0.016	0.084 ± 0.015	0.064 ± 0.024
3	0.188 ± 0.034	0.166 ± 0.034	0.146 ± 0.036	0.361 ± 0.088	0.336 ± 0.090	0.335 ± 0.094
4	0.173 ± 0.025	0.147 ± 0.021	0.127 ± 0.031	0.124 ± 0.020	0.103 ± 0.020	0.072 ± 0.037
5	0.148 ± 0.034	0.133 ± 0.035	0.018 ± 0.074	0.346 ± 0.048	0.339 ± 0.050	0.319 ± 0.066
6	0.241 ± 0.049	0.224 ± 0.051	0.222 ± 0.054	0.124 ± 0.031	0.105 ± 0.029	0.063 ± 0.043
7	0.149 ± 0.050	0.115 ± 0.044	-0.012 ± 0.075	0.177 ± 0.074	0.133 ± 0.044	0.012 ± 0.084
8	0.127 ± 0.036	0.108 ± 0.034	0.050 ± 0.054	0.316 ± 0.081	0.248 ± 0.066	-0.214 ± 0.082
9	0.147 ± 0.021	0.124 ± 0.020	0.054 ± 0.027	0.150 ± 0.024	0.127 ± 0.022	-0.113 ± 0.030
10	0.107 ± 0.020	0.087 ± 0.019	-0.028 ± 0.040	0.255 ± 0.076	0.178 ± 0.045	-0.162 ± 0.045
11	0.158 ± 0.058	0.135 ± 0.050	0.091 ± 0.034	0.132 ± 0.031	0.115 ± 0.030	0.096 ± 0.047
12	0.095 ± 0.034	0.076 ± 0.028	0.003 ± 0.033	0.195 ± 0.093	0.122 ± 0.044	-0.071 ± 0.056
13	0.088 ± 0.030	0.070 ± 0.023	0.003 ± 0.039	0.317 ± 0.074	0.301 ± 0.076	0.247 ± 0.111
14	0.316 ± 0.149	0.225 ± 0.104	-0.205 ± 0.115	0.170 ± 0.078	0.131 ± 0.061	0.045 ± 0.074
15	0.131 ± 0.040	0.102 ± 0.031	-0.074 ± 0.043	0.296 ± 0.041	0.277 ± 0.042	-0.277 ± 0.042
16	0.272 ± 0.053	0.227 ± 0.051	-0.227 ± 0.051	0.123 ± 0.047	0.096 ± 0.037	-0.042 ± 0.058
17	0.158 ± 0.040	0.126 ± 0.035	-0.109 ± 0.040	0.114 ± 0.025	0.098 ± 0.023	0.007 ± 0.052
18	0.204 ± 0.077	0.158 ± 0.057	-0.137 ± 0.055	0.175 ± 0.044	0.153 ± 0.044	-0.152 ± 0.045
19	0.153 ± 0.037	0.127 ± 0.035	0.119 ± 0.041	0.320 ± 0.061	0.283 ± 0.060	-0.282 ± 0.062
20	0.219 ± 0.072	0.197 ± 0.068	-0.196 ± 0.069	0.199 ± 0.078	0.177 ± 0.074	-0.176 ± 0.075

Table E.5: Per-subject block-level prediction performance of the binary hit mixed-effects models with a random slope on block per subject, evaluated on the test set and averaged across the 50 random 80–20 train–test splits.

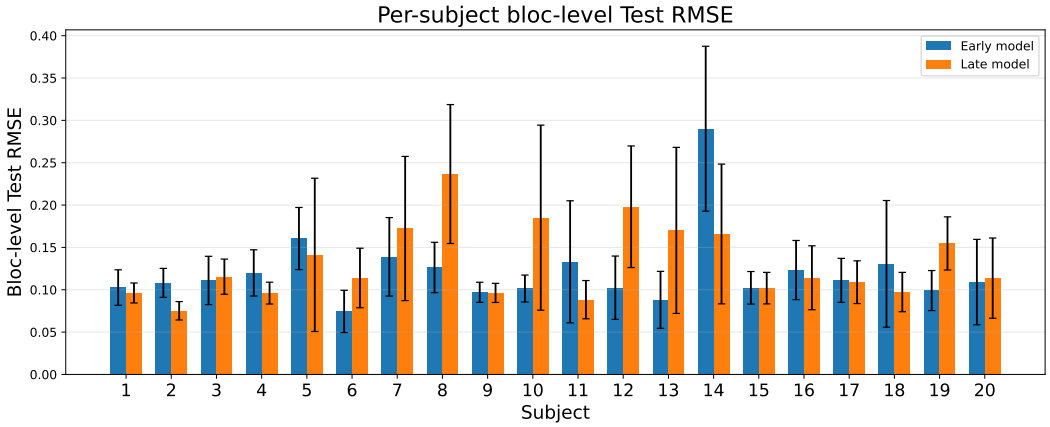

Figure E.2: Test RMSE computed separately for each subject for the early and late binary hit models, with an added random slope per subject.

Appendix F

LOSO Models – Other Subjects

1 Action Time Models

	Early model			Late model		
Variable	Mean coef.	SD	Range	Mean coef.	SD	Range
Intercept	7.5820***	0.0077	$^{+0.0173}_{-0.0132}$	7.5753***	0.0128	$^{+0.0267}_{-0.0253}$
Block	−0.0124**	0.0018	$^{+0.0044}_{-0.0056}$	−0.0465***	0.0055	$^{+0.0159}_{-0.0083}$
Blue horn distractors	0.0038	0.0028	$^{+0.0070}_{-0.0067}$	0.0347***	0.0045	$^{+0.0061}_{-0.0131}$
Red horn distractors	0.0775***	0.0056	$^{+0.0109}_{-0.0132}$	0.0929***	0.0097	$^{+0.0174}_{-0.0305}$
Animation cue	−0.0333**	0.0043	$^{+0.0078}_{-0.0076}$	−0.0102	0.0045	$^{+0.0133}_{-0.0066}$
Arrow cue	−0.2070***	0.0099	$^{+0.0221}_{-0.0180}$	−0.0124	0.0058	$^{+0.0088}_{-0.0163}$
Sound cue	0.0246***	0.0032	$^{+0.0056}_{-0.0069}$	0.0034	0.0029	$^{+0.0104}_{-0.0031}$
Lifetime	0.0698***	0.0036	$^{+0.0085}_{-0.0071}$	0.0612***	0.0068	$^{+0.0243}_{-0.0113}$
Target column	0.0053	0.0016	$^{+0.0052}_{-0.0021}$	0.0122**	0.0023	$^{+0.0057}_{-0.0066}$
Target row	0.0257***	0.0010	$^{+0.0027}_{-0.0018}$	0.0265***	0.0010	$^{+0.0022}_{-0.0019}$

Table F.1: Average coefficient estimates of the LOSO AT LMEMs across the 20 folds. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

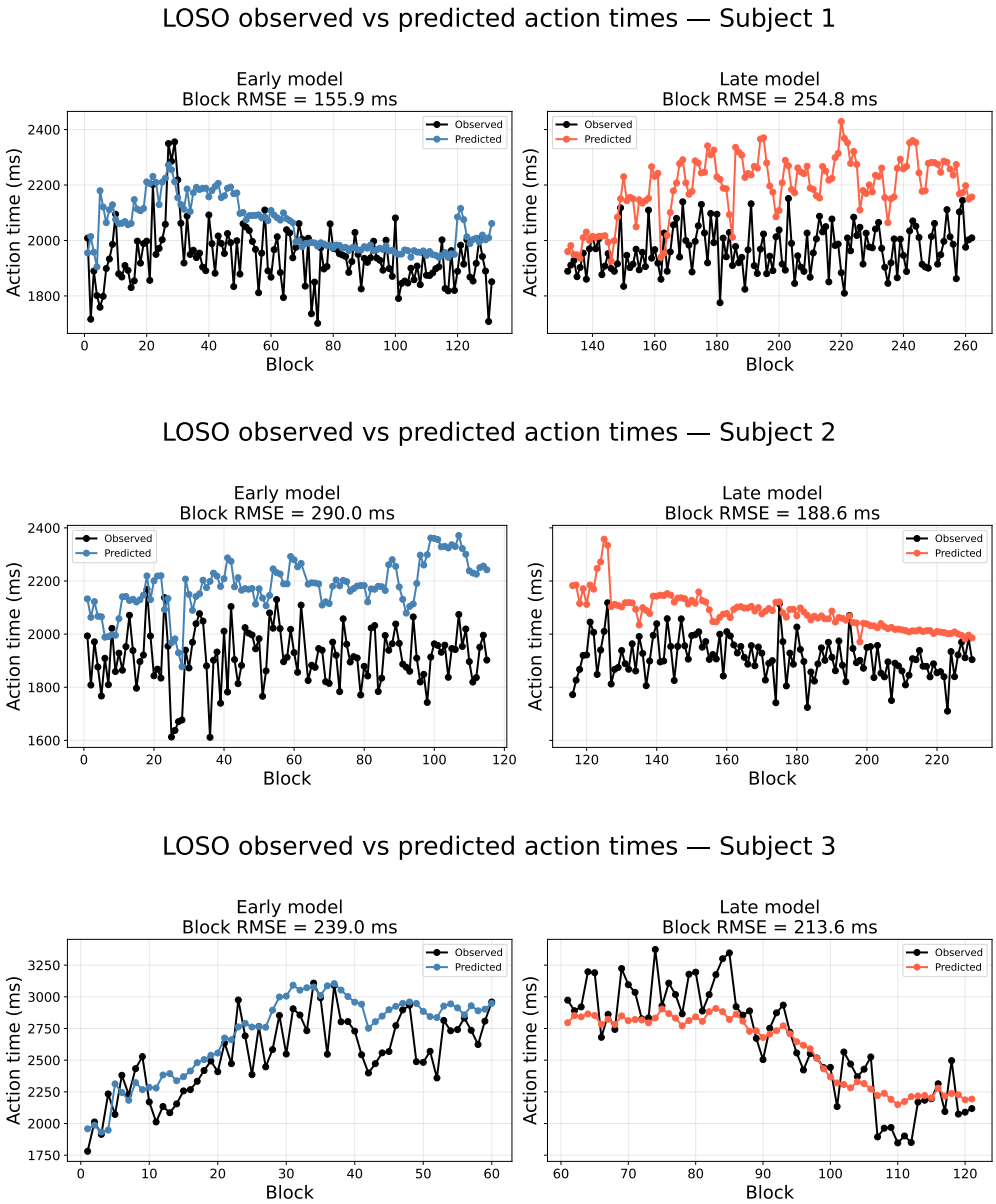

Figure F.1: LOSO AT predictions for subjects 1, 2, and 3.

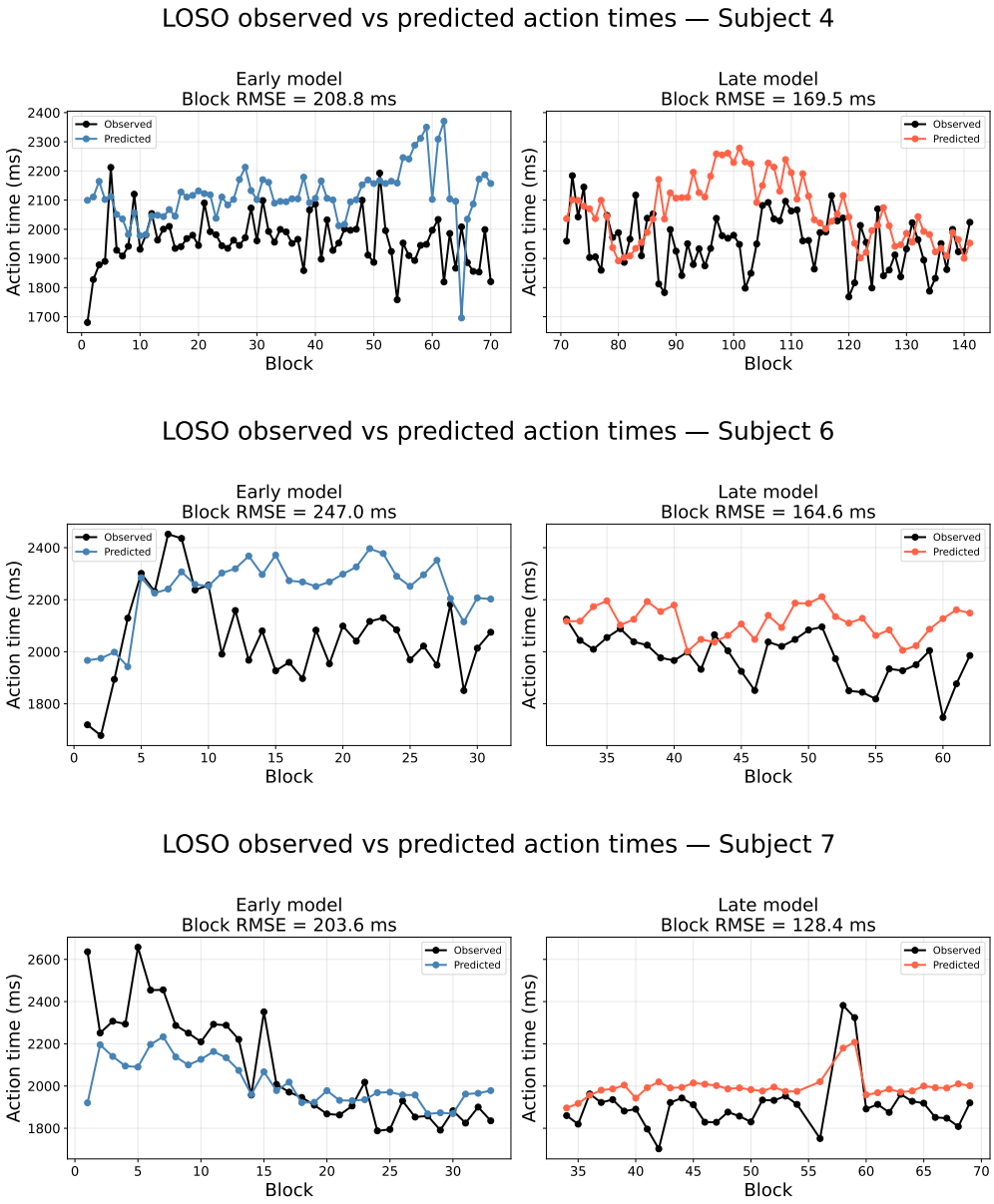

Figure F.2: LOSO AT predictions for subjects 4, 6, and 7.

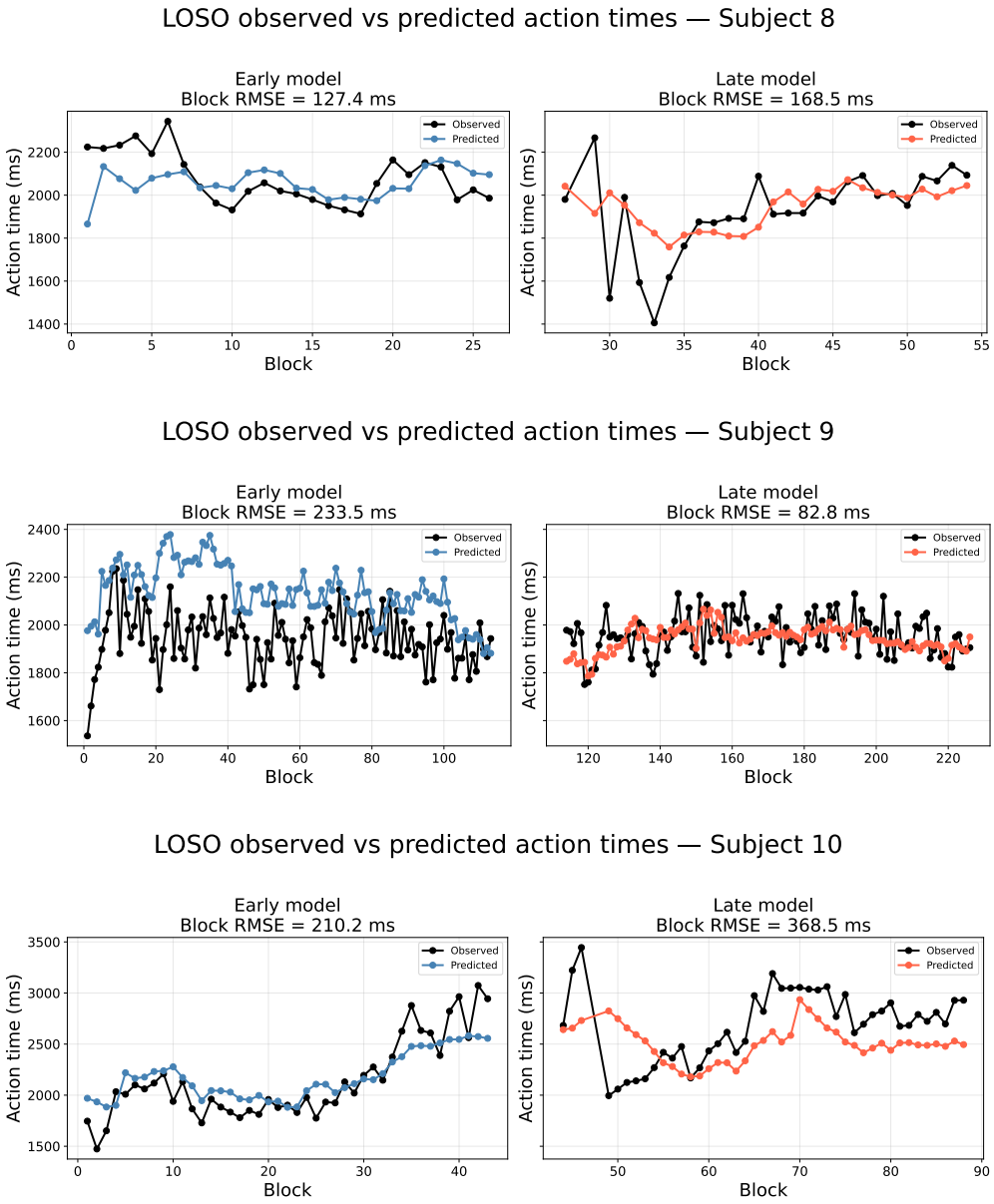

Figure F.3: LOSO AT predictions for subjects 8, 9, and 10.

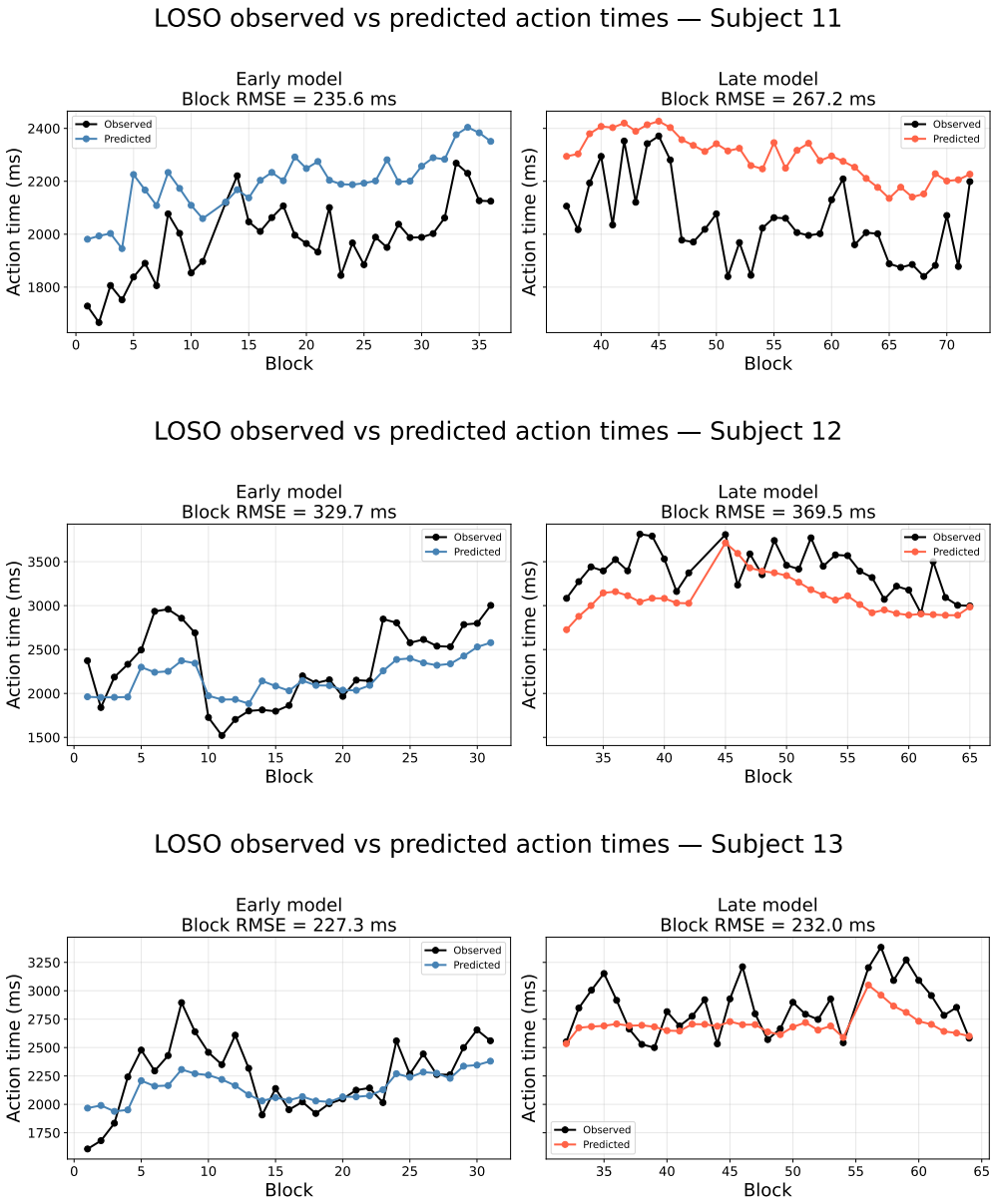

Figure F.4: LOSO AT predictions for subjects 11, 12, and 13.

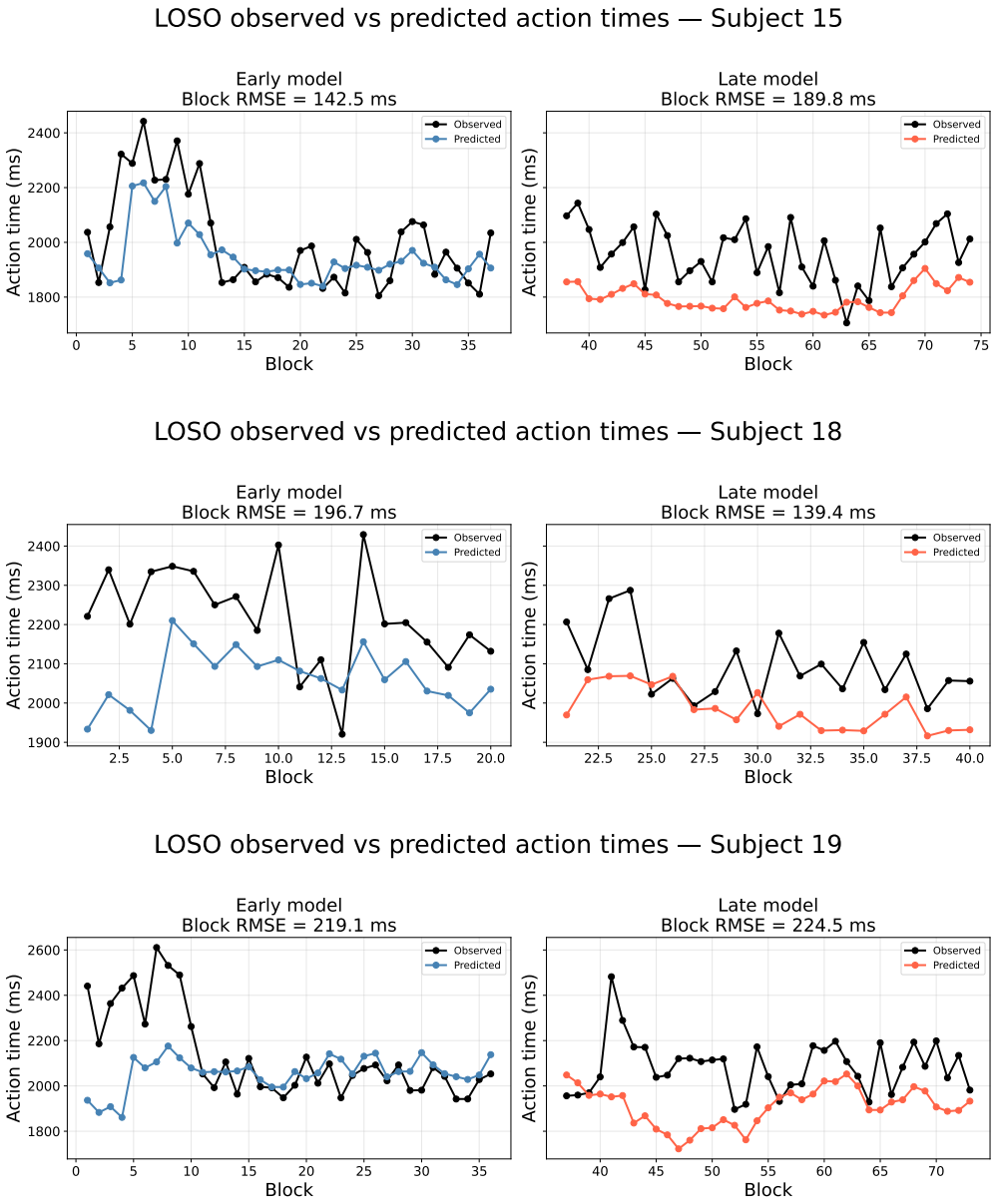

Figure F.5: LOSO AT predictions for subjects 15, 18, and 19.

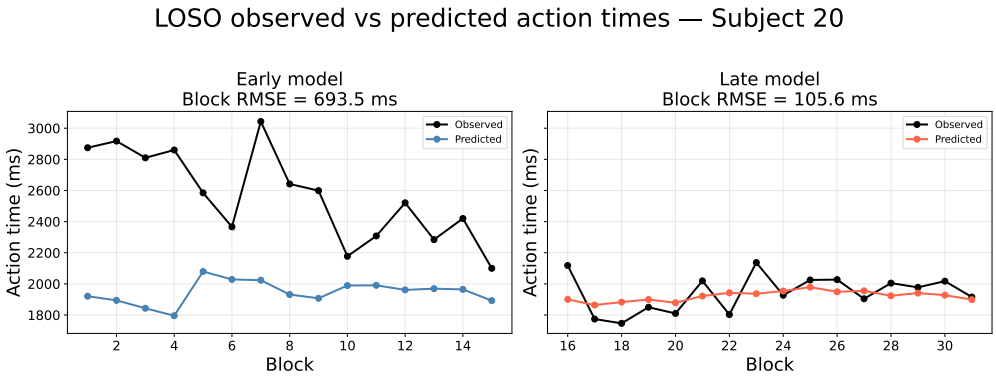

Figure F.6: LOSO AT predictions for subject 20.

2 Binary Hit Models

	Early model			Late model		
Variable	Mean coef.	SD	Range	Mean coef.	SD	Range
Intercept	3.2123***	0.0629	$^{+0.1214}_{-0.0979}$	3.7094***	0.0882	$^{+0.2474}_{-0.1701}$
Block	−0.3586**	0.0646	$^{+0.2520}_{-0.0921}$	0.4748*	0.1175	$^{+0.2159}_{-0.3997}$
Trial	0.0447	0.0086	$^{+0.0192}_{-0.0161}$	−0.1235**	0.0107	$^{+0.0215}_{-0.0200}$
Previous hit	0.1873***	0.0234	$^{+0.0827}_{-0.0319}$	0.2999***	0.0270	$^{+0.0560}_{-0.0839}$
Lifetime	0.3253***	0.0229	$^{+0.0272}_{-0.0625}$	0.0834	0.0639	$^{+0.0744}_{-0.2293}$
Red horn distractors	−0.2128***	0.0470	$^{+0.0751}_{-0.1123}$	−0.1568**	0.0783	$^{+0.0841}_{-0.2892}$
Blue horn distractors	0.3731***	0.0687	$^{+0.1771}_{-0.2063}$	−0.1287*	0.0518	$^{+0.1979}_{-0.0551}$
Target row	−0.0915*	0.0104	$^{+0.0191}_{-0.0256}$	−0.1268**	0.0113	$^{+0.0212}_{-0.0305}$
Target column	−0.0084	0.0229	$^{+0.0486}_{-0.0542}$	−0.1845***	0.0216	$^{+0.0495}_{-0.0504}$
Sound cue	0.0192	0.0223	$^{+0.0451}_{-0.0397}$	−0.2394*	0.0397	$^{+0.1225}_{-0.1001}$

Table F.2: Average fixed-effect coefficient estimates of the LOSO binary hit mixed-effects models across the 20 held-out-subject folds. Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

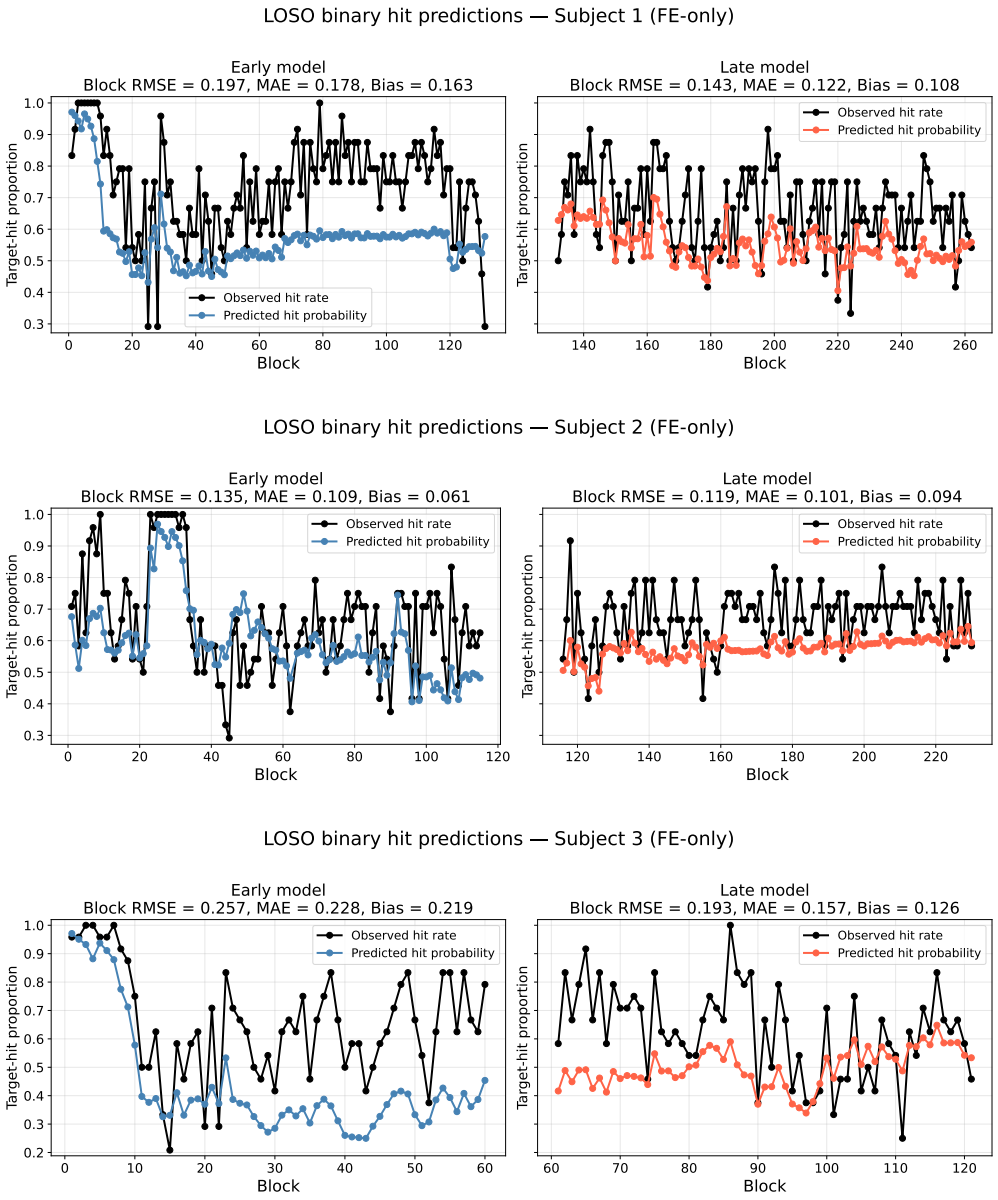

Figure F.7: LOSO binary hit predictions for subjects 1, 2, and 3.

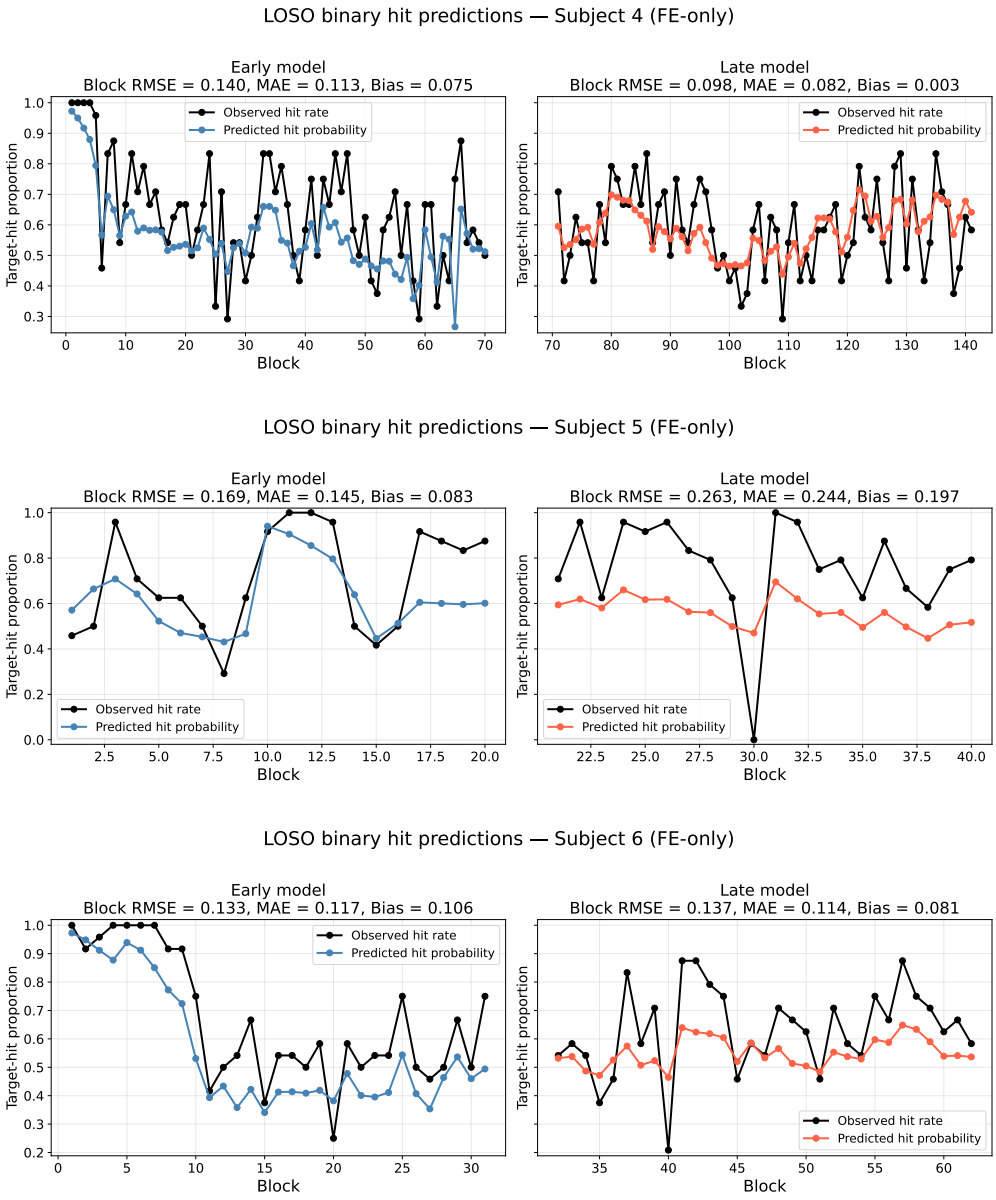

Figure F.8: LOSO binary hit predictions for subjects 4, 5, and 6.

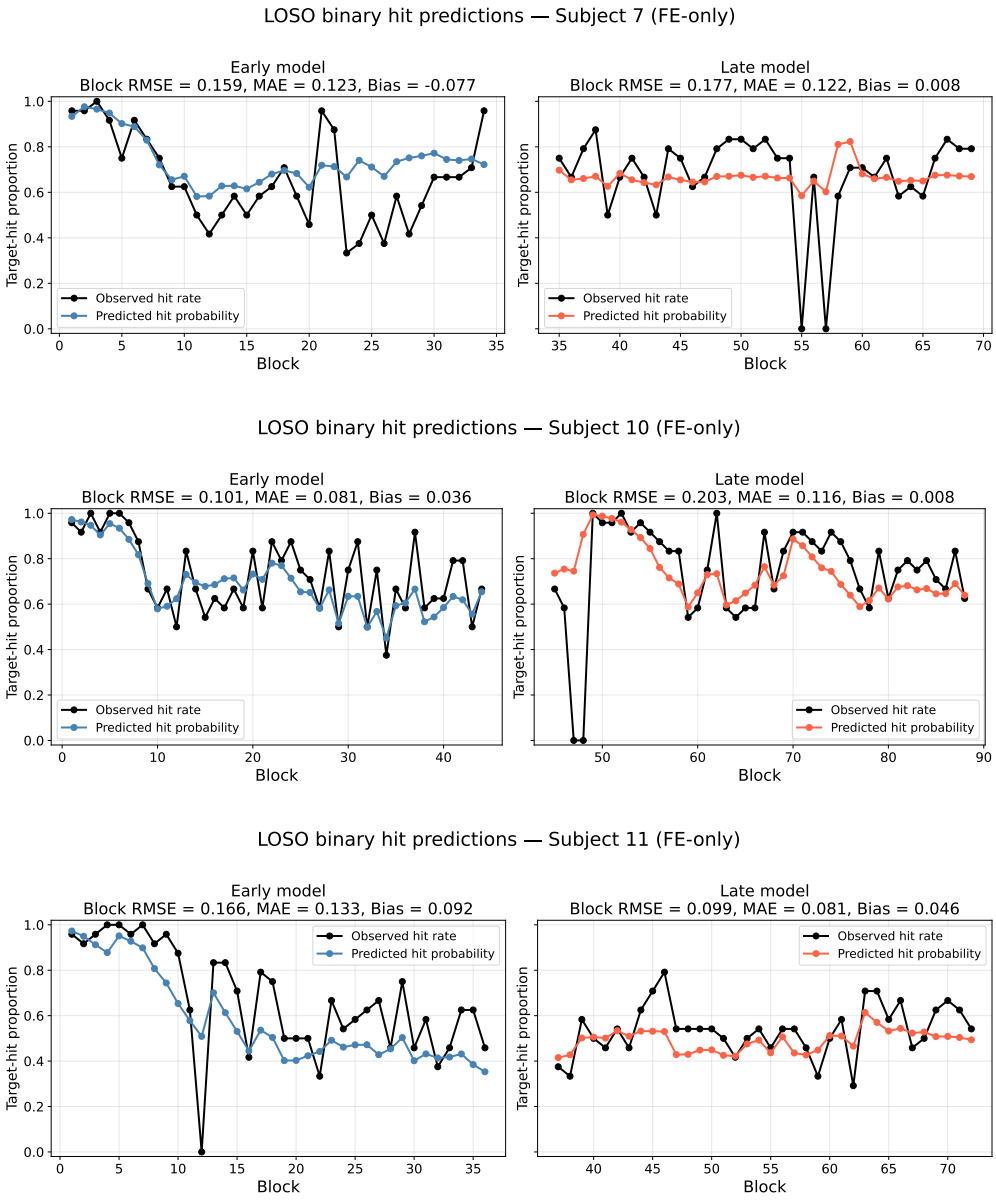

Figure F.9: LOSO binary hit predictions for subjects 7, 10, and 11.

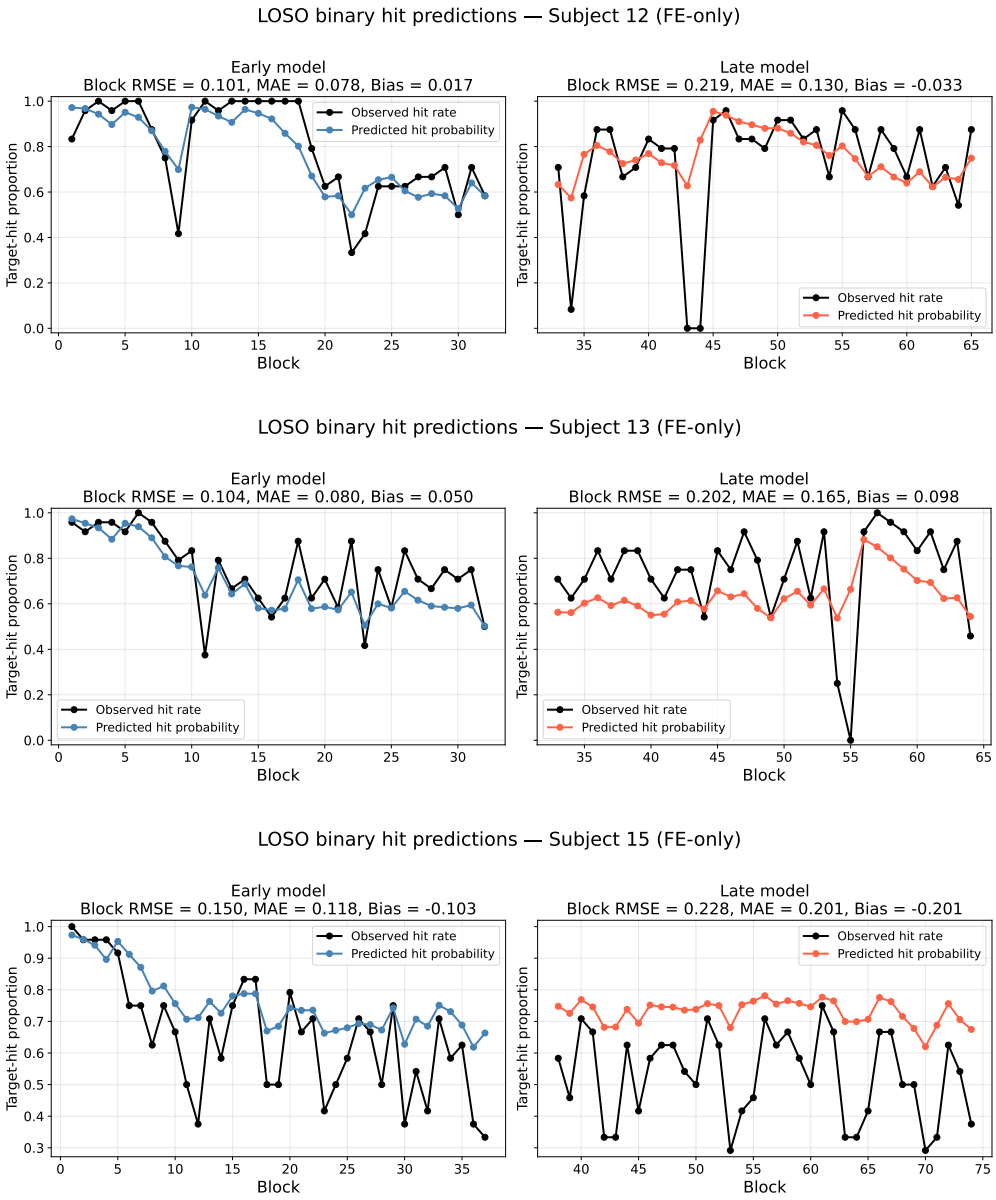

Figure F.10: LOSO binary hit predictions for subjects 12, 13, and 15.

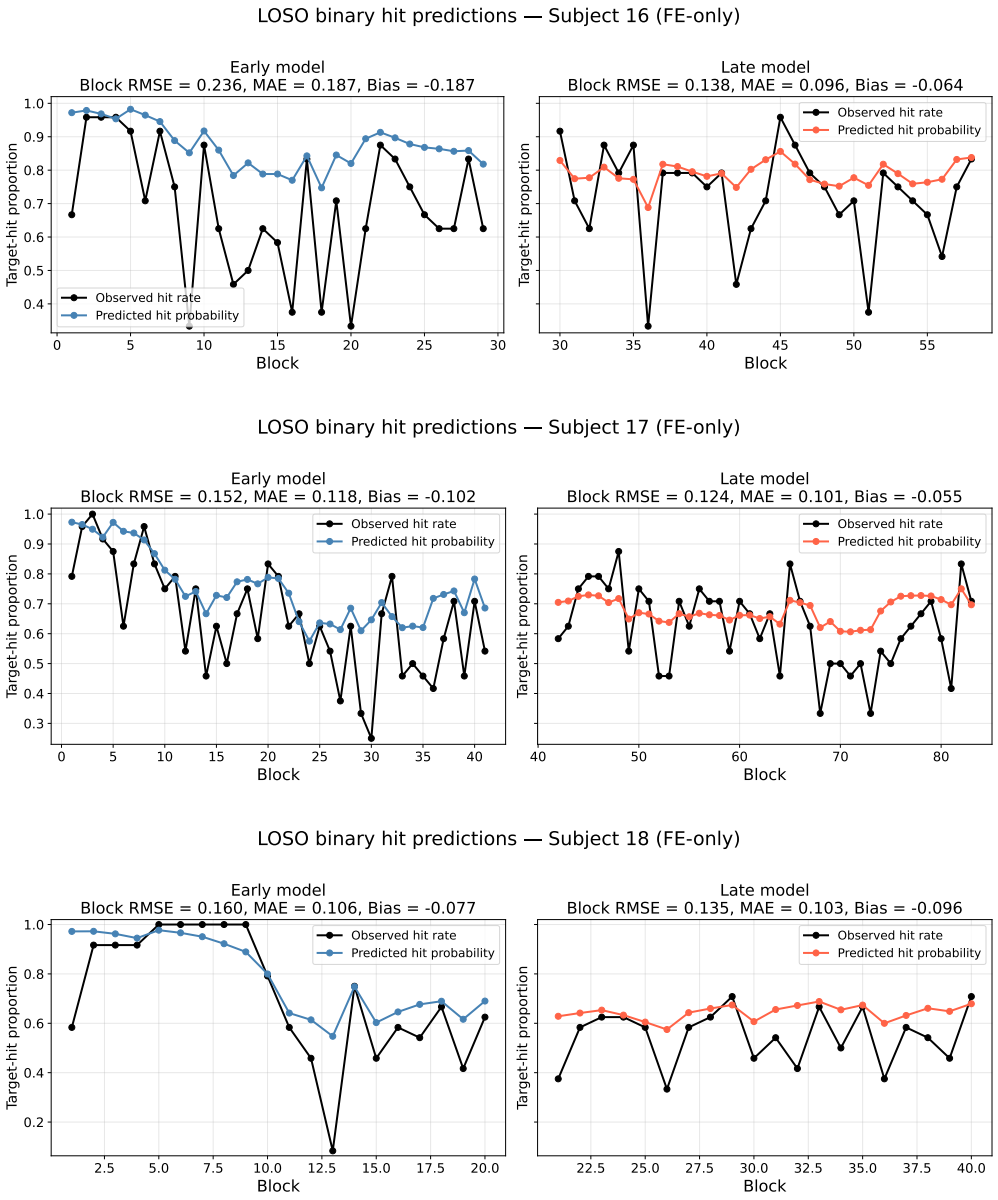

Figure F.11: LOSO binary hit predictions for subjects 16, 17, and 18.

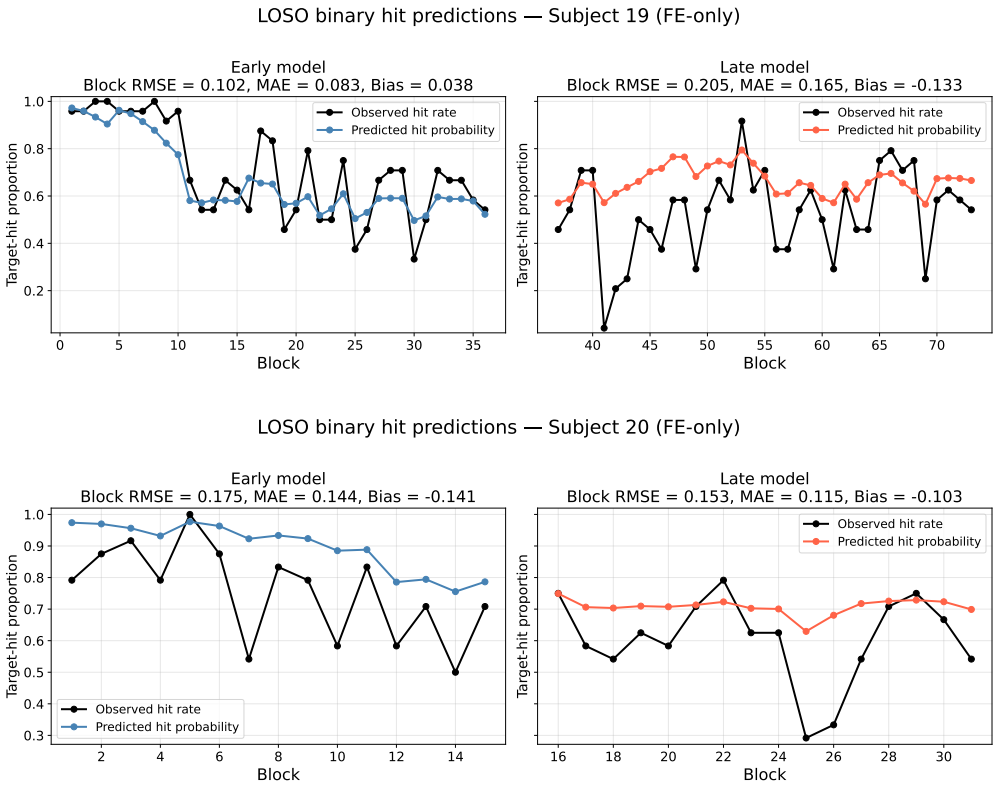

Figure F.12: LOSO binary hit predictions for subjects 19 and 20.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**