Event History Analysis for psychological time-to-event data: A tutorial in R with examples in Bayesian and frequentist workflows

Sven Panis[1] & Richard Ramsey[1]

[1] ETH Zürich

Author Note

Neural Control of Movement lab, Department of Health Sciences and Technology (D-HEST). Social Brain Sciences lab, Department of Humanities, Social and Political Sciences (D-GESS).

Correspondence concerning this article should be addressed to Sven Panis, ETH GLC, room G16.2, Gloriastrasse 37/39, 8006 Zürich. E-mail: sven.panis@hest.ethz.ch

Abstract

Time-to-event data such as response times and saccade latencies form a cornerstone of experimental psychology, and have had a widespread impact on our understanding of human cognition. However, the orthodox method for analyzing such data – comparing means between conditions – is known to conceal valuable information about the timeline of psychological effects, such as their onset time and how they evolve with increasing waiting time. The ability to reveal finer-grained, "temporal states" of cognitive processes can have important consequences for theory development by qualitatively changing the key inferences that are drawn from psychological data. Luckily, well-established analytical approaches, such as event history analysis (EHA), are able to evaluate the detailed shape of time-to-event distributions, and thus characterize the time course of psychological states. One barrier to wider use of EHA, however, is that the analytical workflow is typically more time-consuming and complex than orthodox approaches. To help achieve broader uptake of EHA, in this paper we outline a set of tutorials that detail one distributional method known as discrete-time EHA. We touch upon several key aspects of the workflow, such as how to process raw data and specify regression models, and we also consider the implications for experimental design. We finish the article by considering the benefits of the approach for understanding psychological states, as well as the limitations of this work. Finally, the project is written in R and freely available, which means the approach can easily be adapted to other data sets.

Keywords: response times, event history analysis, Bayesian multilevel regression models, experimental psychology, cognitive psychology

Word count: 10131 (body) + 1709 (references) + 3473 (body supplemental material) + 393 (refs suppl. mat.)

1. Introduction

1.1 Motivation and background context: Comparing means versus distributional shapes

In experimental psychology, it is standard practice to analyse response times (RTs), saccade latencies, and fixation durations by calculating average performance across a series of trials. Such comparisons between means have been the workhorse of experimental psychology over the last century, and have had a substantial impact on theory development as well as our understanding of the structure of cognition and brain function. Indeed, the view that mean values are truth and variations around the mean are error is deeply ingrained in experimental psychology (Bolger, Zee, Rossignac-Milon, & Hassin, 2019). However, differences in mean RT conceal important pieces of information, such as when an experimental effect starts, how it evolves with increasing waiting time, and whether its onset is time-locked to other events (Panis, 2020; Panis, Moran, Wolkersdorfer, & Schmidt, 2020; Panis & Schmidt, 2016, 2022; Panis, Torfs, Gillebert, Wagemans, & Humphreys, 2017; Panis & Wagemans, 2009; Wolkersdorfer, Panis, & Schmidt, 2020). Such absolute timing information is useful not only for the interpretation of experimental effects under investigation, but also for cognitive psychophysiology and computational model selection (Panis, Schmidt, Wolkersdorfer, & Schmidt, 2020).

As a simple illustration, Figure 1 summarises simulated single-subject data (200 trials per condition) that shows how comparing means between two conditions can conceal the shapes of the underlying RT and accuracy distributions. Indeed, compared to the aggregation of data across trials (Figure 1A), a distributional approach offers the possibility to reveal the time course of psychological states (Figure 1B). For example, Figure 1B shows a first state (up to 400 ms after target onset) for which the early upswing in hazard is equal for both conditions, and the emitted responses are always correct in condition 1 and always incorrect in condition 2. In a second state (400 to 500 ms), hazard is higher in condition 1, and conditional accuracies are close to .5 in both conditions. In a third state (>500 ms), the effect disappears in hazard, and all conditional accuracies are equal to 1 (see also Panis & Schmidt, 2016).

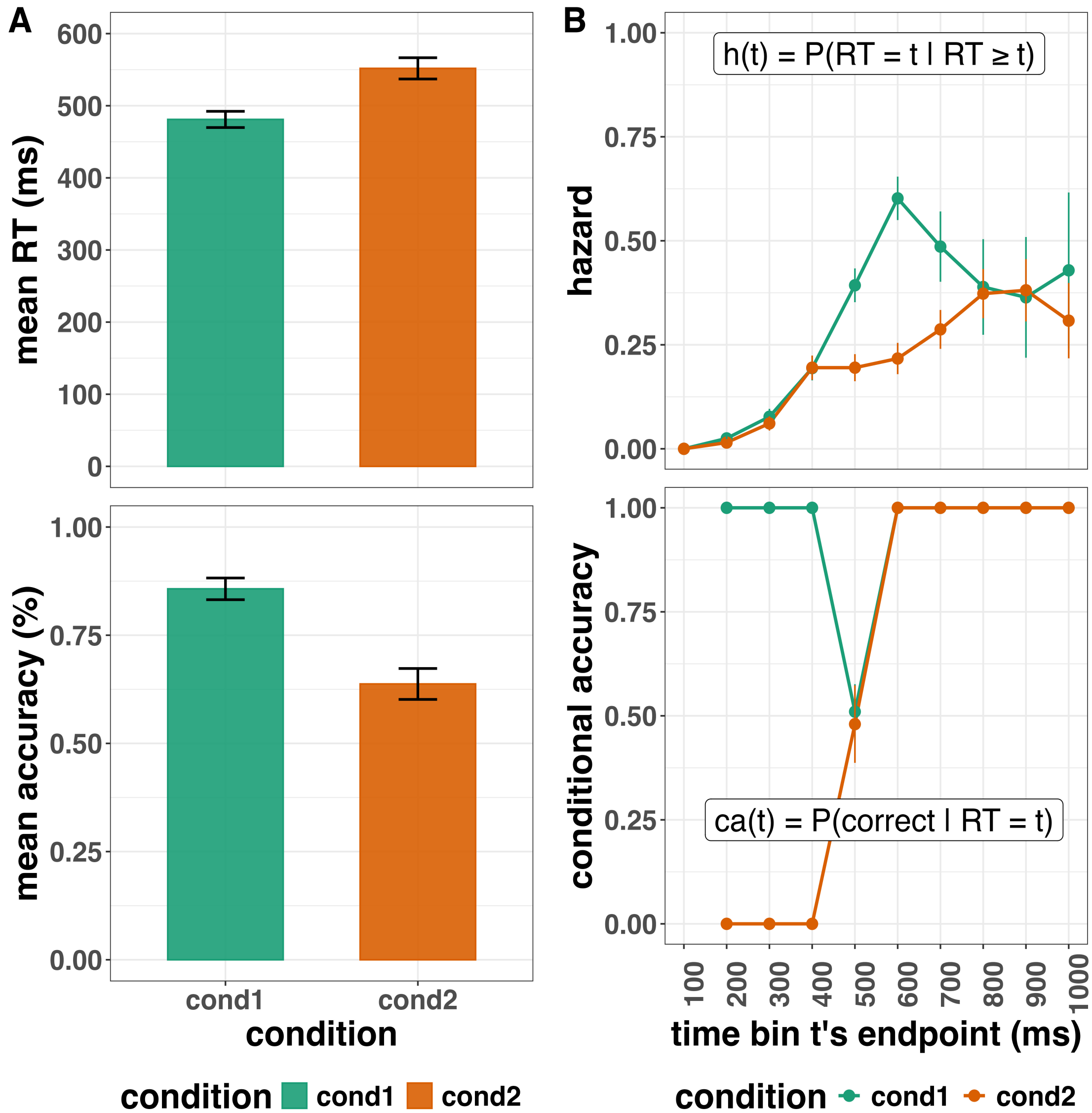

Figure 1. Simulated single-subject data showing mean performance versus distributional (EHA/SAT) analyses. (A) The mean RT (top) and overall accuracy (bottom) for two conditions are plotted. (B) The discrete-time hazard functions (top) and conditional accuracy functions (bottom) are plotted for the same data. The first second after target stimulus onset (time zero) is divided in ten bins of 100 ms. The first bin is (0,100], the last bin is (900,1000]. Two hundred trials were simulated in each condition. Note that the hazard and conditional accuracy estimates are plotted at the endpoint of each time bin. The definitions of discrete-time hazard and conditional accuracy are further explained in section 2.1.2. Error bars represent +/- 1 standard error of the mean (A) or proportion (B).

Why does this matter for research in psychology? For many psychological questions, the estimation of such “temporal states” information can be theoretically meaningful by leading to more fine-grained understanding of psychological processes. Because EHA adds a relatively under-used but ever-present dimension – the passage of time – to the theory building toolkit, it provides one possible answer to the recent call for a temporal science of behavior (Abney, Fausey, Suarez-Rivera, & Tamis-LeMonda, 2025).

1.2 Aims

Our ultimate aim in this paper is twofold: first, we want to convince readers of the many benefits of using EHA when dealing with psychological RT data, and second, we want to provide a set of practical tutorials, which provide step-by-step instructions on how you actually perform a discrete-time EHA on RT data, as well as a complementary discrete-time speed-accuracy tradeoff (SAT) analysis on timed accuracy data in case of choice RT data (Figure 1B).

Even though EHA is a widely used statistical tool and there already exist many excellent reviews (Allison, 1982; Blossfeld & Rohwer, 2002; Box-Steffensmeier, 2004; Hosmer, Lemeshow, & May, 2011; Mills, 2011; Singer & Willett, 2003; Teachman, 1983) and tutorials (Allison, 2010; Elmer, Van Duijn, Ram, & Bringmann, 2023; Landes, Engelhardt, & Pelletier, 2020; Lougheed, Benson, Cole, & Ram, 2019; Stoolmiller, 2015; Stoolmiller & Snyder, 2006), we are not aware of any tutorials that are aimed specifically at psychological RT (+ accuracy) data, and which provide worked examples of the key data processing and Bayesian multilevel regression modelling steps. Set within this context, our overall aim is to introduce a set of tutorials, which explain **how** to do such analyses in the context of experimental psychology, rather than repeat in any detail **why** you may do them. Therefore, we hope that our tutorials will provide a pathway for research avenues in experimental psychology that have the potential to benefit from using EHA in the future.

1.3 Structure

In what follows, the paper is organised in three main sections. In Section 2, we provide a brief overview of EHA to orient the reader to the basic concepts that we will use throughout the paper and why such an approach might be relevant for research in experimental psychology. In Section 3, we outline a series of tutorials, which are written in the R programming language and publicly available on our Github page (https://github.com/sven-panis/Tutorial_Event_History_Analysis), along with all of the other code and material associated with the project. The tutorials provide hands-on, concrete examples of key parts of the analytical process, such as data wrangling, plotting descriptive statistics, model fitting and planning future studies, so that others can apply EHA to their own time-to-event data measured in RT tasks. In Section 4, we discuss the strengths and weaknesses of the approach for researchers in experimental psychology.

2. What is event history analaysis and why is it relevant to research in experimental psychology?

2.1 A brief introduction to event history analysis

EHA is a class of statistical approaches to study the occurrence and timing of events, such as disease onset, marriages, arrests, and job terminations (Allison, 2010). In this section, we want to provide an intuition regarding how EHA works in general, as well as in the context of experimental psychology. For those who want more detailed treatment of EHA and/or regression equations, we refer the reader to several excellent textbooks on these topics (Allison, 2010; Gelman, Hill, & Vehtari, 2020; Mills, 2011; Singer & Willett, 2003; Winter, 2019). We supply a few regression equations in section E of the Supplemental Material.

2.1.1 Terminolgy and minimum requirements for EHA. To avoid possible confusion in terminology used, it is worth noting that EHA is known by various labels, such as survival analysis, hazard analysis, duration analysis, failure-time analysis, and transition analysis (Singer & Willett, 2003). In this paper, we choose to use the term EHA throughout.

In terms of minimum requirements to apply EHA, one must be able to:

1. define an event of interest that represents a qualitative change - a transition from one discrete state to another - that can be situated in time (e.g., a button press, a saccade onset, a fixation offset, etc.);
2. define time point zero in each trial (e.g., target stimulus onset, fixation onset, etc.);
3. measure the passage of time between time point zero and event occurrence in discrete or continuous time units in each trial.

These minimal requirements are fulfilled by the RT data obtained in single-button detection tasks, where the time-to-response is repeatedly measured in different trials in the same individual. In section A of the Supplemental Material we visualize this and other types of time-to-event data which are typically obtained in discrimination and bistable perception tasks.

2.1.2 Types of EHA. There are different types of modeling approaches in EHA. For example, the definition of hazard and the type of models employed depend on whether one is using continuous or discrete time units. As a lab, and mainly for practical reasons, we have much more experience using discrete-time EHA, and that is the approach that we describe and focus on in this paper. This choice may seem counter-intuitive, given that RT is typically treated as a continuous variable. However, continuous forms of EHA require much more data to estimate the continuous-time hazard (rate) function well (Bloxom, 1984; Luce, 1991; Van Zandt, 2000). Thus, by trading a bit of temporal resolution for a lower number of trials, discrete-time methods seem ideal for dealing with typical psychological RT data sets for which there are less than ~200 trials per condition per participant (Panis, Schmidt, et al., 2020). Moreover, as indicated by Allison (2010), learning discrete-time EHA methods first will help in learning continuous-time methods, so it seems like a good starting point.

To apply discrete-time EHA, one divides the within-trial time in discrete, contiguous time bins indexed by t (e.g., t = 1:10 time bins; Figure 1B). Then let RT be a discrete random variable denoting the rank of the time bin in which a particular person's response occurs in a particular trial (i.e., repeated measure). For example, a response in one trial might occur at 546 ms and it would be in time bin 6 (any RTs from 501 ms to 600 ms). One then calculates the sample-based estimate of the discrete-time hazard function of event occurrence for each experimental condition (Figure 1B top). The discrete-time hazard function gives you, for each time bin, the conditional probability that the event occurs (sometime) in bin t, given that the event does not occur in previous bins. In other words, it reflects the instantaneous risk that the event occurs in the current bin t, given that it has not yet occurred in the past, i.e., in one of the prior bins (t-1, t-2, . . . , 1).

In the context of experimental psychology, it is often (but not always), the case that responses can be classified as correct or incorrect. In those cases, one can also calculate the conditional accuracy function (Figure 1B bottom). The conditional accuracy function gives you for each time bin the conditional probability that a response is correct given that it is emitted in time bin t (Allison, 2010; Kantowitz & Pachella, 2021; Wickelgren, 1977). The conditional accuracy function is also known as the micro-level speed-accuracy tradeoff (SAT) function. We refer to this extended (hazard + conditional accuracy) analysis for choice RT data as EHA/SAT. The definitions of these and other discrete-time functions are given in section B of the Supplemental Material.

2.2 Benefits of event history analysis for research in experimental psychology

Statisticians and mathematical psychologists recommend focusing on the hazard function when analyzing time-to-event data for various reasons (Holden, Van Orden, & Turvey, 2009; Luce, 1991; Townsend, 1990). We do not cover these benefits in detail here, as these are more general topics that have been covered elsewhere in textbooks (see also section G of the Supplemental Material). Instead, here we focus on the benefits as we see them for common research programmes in experimental psychology.

We highlight three benefits that we think are relevant to the domain of experimental psychology. First, as illustrated in Figure 1, compared to averaging data across trials, integrating results between hazard functions and their associated conditional accuracy functions for choice RT data can be informative for understanding psychological processes, in terms of inferences about the microgenesis and temporal organization of cognition and theoretical development. As such, the approach permits different kinds of questions to be asked, different inferences to be made, and it holds the potential to discriminate between theoretical accounts of psychological and/or brain-based processes. For example, what kind of theory or set of mechanisms could account for the shape of the functions and the temporally localized effects reported in Figure 1B (Panis & Schmidt, 2016)? Are there new auxiliary assumptions that computational models need to adopt (Panis, Moran, et al., 2020)? Will the temporal effect patterns align nicely with EEG findings (Panis & Schmidt, 2022)? And are there new experiments that need to be performed to test the novel predictions that follow from these analyses?

Second, compared to more conventional analytical approaches, EHA uses more of the data because it deals with missing data differently. It is conventional with RT data to either (a) use a response deadline and discard all trials without a response, or (b) wait in each trial until a response occurs and then apply data trimming techniques, i.e., discarding too short or too long RTs (and perhaps also erroneous responses) before calculating a mean RT

(Berger & Kiefer, 2021). Discarding data can introduce biases, however. Rather than treat non-responses as missing data, EHA treats such trials as *right-censored* observations on the variable RT, because all we know is that RT is greater than some value. Right-censoring is a type of missing data problem and a nearly universal feature of survival data including RT data. For example, if the censoring time was 1 second, then some trials result in observed event times (those with a RT below 1 second), while the other trials result in response times that are right-censored at 1 second. The fact that EHA can deal with right-censoring, therefore, presents a analytical strength of the approach compared to many common approaches in experimental psychology (ANOVA, linear regression, delta plots).

Third, the approach is generalisable and applicable to many tasks that are commonly used in experimental psychology, such as detection, discrimination and bistable perception tasks, and to a range of common experimental manipulations, such as stimulus-onset-asynchrony (see section A of the Supplemental Material). The upshot is that one general analytical approach, which holds several potential advantages, is widely applicable to many substantive use-cases in the domain of experimental psychology, irrespective of the analyst's current view on the nature of cognition (Barack & Krakauer, 2021).

2.3 Implications for research design in experimental psychology

Performing EHA in experimental psychology has implications for how experiments are designed. More specifically, we consider three implications that researchers will need to consider when using discrete-time EHA. First, one can use a response deadline in each trial because EHA deals with right-censored observations.

Second, since the number of trials per condition are spread across bins, it is important to have a relatively large number of trial repetitions per participant and per condition. Accordingly, experimental designs using this approach typically focus on factorial, within-subject designs, in which a large number of observations are made on a relatively small number of participants (so-called small-N designs). This approach emphasizes the precision and reproducibility of data patterns at the individual participant level to increase the inferential validity of the design (Baker et al., 2021; Smith & Little, 2018). Note that because statistical power derives both from the number of participants and from the number of repeated measures per participant and condition, small-N designs can still achieve what are generally considered acceptable levels of statistical power, if they have a sufficient amount of data overall (Baker et al., 2021; Smith & Little, 2018).

Third, the width of each time bin will need to be determined. For instance, in Figure 1B we chose 100 ms in an arbitrary manner. In reality, however, bin width will need to be set by considering a number of factors simultaneously. The optimal bin width will depend on (a) the length of the observation period in each trial, (b) the rarity of event occurrence, (c) the number of repeated measures (or trials) per condition per participant, and (d) the shape of the hazard function. Finding an appropriate bin width in a given user case before fitting models will require testing a number of options, when calculating and plotting the descriptive statistics (see section 3.1). The goal is to find the smallest bin width that is supported by the amount of data available. Based on our experience, a bin width of 50 ms is a good starting value when the number of repeated measures is 100 or less. Too small bin widths will result in erratic hazard functions as many bins will have no events, and thus hazard estimates of zero. Interestingly, the time bins do not need to have the same width. For example, Panis (2020) used larger bins towards the end of the observation period, as fewer events occurred there.

3. Tutorials

Tutorials 1a and 1b show how to calculate and plot the descriptive statistics of EHA/SAT when there are one or two independent variables, respectively. Tutorials 2a and 2b illustrate how to use Bayesian multilevel modeling to fit hazard and conditional accuracy models, respectively. Tutorials 3a and 3b show how to implement, respectively, multilevel models for hazard and conditional accuracy in the frequentist framework. Tutorial 4 shows how to use simulation and power analysis for planning experiments. Additionally, to further simplify the process for other users, the first two tutorials rely on a set of our own custom functions that make sub-processes easier to automate, such as data wrangling and plotting functions (see section C of the Supplemental Material for a list of the custom functions).

The content of the tutorials, in terms of EHA and multilevel regression modelling, is mainly based on Allison (2010), Singer and Willett (2003), McElreath (2020), Heiss (2021), Kurz (2023a), and Kurz (2023b). We used R (Version 4.5.1; R Core Team, 2024) and the R-packages *bayesplot* (Version 1.13.0; Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019), *brms* (Version 2.22.0; Bürkner, 2017, 2018, 2021), *citr* (Version 0.3.2; Aust, 2019), *cmdstanr* (Version 0.9.0.9000; Gabry, Češnovar, Johnson, & Bronder, 2024), *dplyr* (Version 1.1.4; Wickham, François, Henry, Müller, & Vaughan, 2023), *forcats* (Version 1.0.0; Wickham, 2023a), *ggplot2* (Version 3.5.2; Wickham, 2016), *lme4* (Version 1.1.37; Bates, Mächler, Bolker, & Walker, 2015), *lubridate* (Version 1.9.4; Grolemund & Wickham, 2011), *Matrix* (Version 1.7.3; Bates, Maechler, & Jagan, 2024), *nlme* (Version 3.1.168; Pinheiro & Bates, 2000), *papaja* (Version 0.1.3; Aust & Barth, 2024), *patchwork* (Version 1.3.0; Pedersen, 2024), *purrr* (Version 1.0.4; Wickham & Henry, 2023), *RColorBrewer* (Version 1.1.3; Neuwirth, 2022), *Rcpp* (Eddelbuettel & Balamuta, 2018; Version 1.0.14; Eddelbuettel & François, 2011), *readr* (Version 2.1.5; Wickham, Hester, & Bryan, 2024), *RJ-2021-048* (Bengtsson, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021), *rstan* (Version 2.32.7; Stan Development Team, 2024), *standist* (Version 0.0.0.9000; Girard, 2024), *StanHeaders* (Version 2.32.10; Stan Development Team, 2020), *stringr* (Version 1.5.1; Wickham, 2023b), *tibble* (Version 3.3.0; Müller & Wickham, 2023), *tidybayes* (Version 3.0.7; Kay, 2024), *tidyr* (Version 1.3.1; Wickham, Vaughan, & Girlich, 2024), *tidyverse* (Version 2.0.0; Wickham et al., 2019) and *tinylabels* (Version 0.2.5; Barth, 2023) for all reported analyses.

3.1 Tutorial 1a: Calculating descriptive statistics using a life table

3.1.1 Data wrangling aims. Our data wrangling procedures serve two related purposes. First, we want to calculate descriptive statistics for each condition in each individual using a life table. A life table includes for each time bin, the risk set (i.e., the number of trials that are event-free at the start of the bin), the number of observed events, and the estimates of the discrete-time hazard probability h(t), survival probability S(t), probability mass P(t), possibly the conditional accuracy ca(t), and their estimated standard errors (se). The definitions of these quantities are provided in section B of the Supplemental Material.

Second, we want to produce two different data sets that can each be submitted to different types of inferential modelling approaches. The two types of data structure we label as 'person-trial' data and 'person-trial-bin' data. The 'person-trial' data (Table 1) will be familiar to most researchers who record behavioural responses from participants, as it represents the measured RT and accuracy per trial within an experiment. This data set is used when fitting conditional accuracy models (Tutorials 2b and 3b).

Table 1

Data structure for 'person-trial' data

pid	trial	condition	rt	accuracy
1	1	congruent	373.49	1
1	2	incongruent	431.31	1
1	3	congruent	455.43	0
1	4	incongruent	622.41	1
1	5	incongruent	535.98	1
1	6	incongruent	540.08	1
1	7	congruent	511.07	1
1	8	incongruent	444.42	1
1	9	congruent	678.69	1
1	10	congruent	549.79	1

Note. The first 10 trials for participant 1 are shown. These data are simulated and for illustrative purposes only.

In contrast, the 'person-trial-bin' data (Table 2) has a different, more extended structure, which indicates in which bin a response occurred, if at all, in each trial. Therefore, the 'person-trial-bin' data generates a 0 in each bin until an event occurs and then it generates a 1 to signal an event has occurred in that bin. This data set is used when fitting discrete-time hazard models (Tutorials 2a and 3a). It is worth pointing out that there is no requirement for an event to occur at all (in any bin), as maybe there was no response on that trial or the event occurred after the time window of interest. Likewise, when the event occurs in bin 1 there would only be one row of data for that trial in the person-trial-bin data set.

Table 2

Data structure for 'person-trial-bin' data

pid	trial	condition	timebin	event
1	1	congruent	1	0
1	1	congruent	2	0
1	1	congruent	3	0
1	1	congruent	4	1
1	2	incongruent	1	0
1	2	incongruent	2	0
1	2	incongruent	3	0
1	2	incongruent	4	0
1	2	incongruent	5	1

Note. The first 2 trials for participant 1 from Table 1 are shown. The width of the time bins is 100 ms. These data are simulated and for illustrative purposes only.

3.1.2 A real data wrangling example. To illustrate how to quickly set up life tables for calculating the descriptive statistics (functions of discrete time), we use a published data set on masked response priming from Panis and Schmidt (2016), who were interested in the temporal dynamics of the effect of prime-target congruency in RT and accuracy data. In their first experiment, Panis and Schmidt (2016) presented a double arrow for 94 ms that pointed left or right as the target stimulus with an onset at time point zero in each trial. Participants had to indicate the direction in which the double arrow pointed using their corresponding index finger, within 800 ms after target onset. Response time and accuracy were recorded on each trial. Prime type (blank, congruent, incongruent) and mask type were manipulated across trials (i.e., repeated measures of time-to-response). Here we focus for each participant on the subset of 220 trials in which no mask was presented. The 13-ms prime stimulus was a double arrow presented 187 ms before target onset in the congruent (same direction as target) and incongruent (opposite direction as target) prime conditions.

There are several data wrangling steps to be taken. First, we need to load the data before we (a) supply required column names, and (b) specify the factor condition with the correct levels and labels.

The required column names are as follows:

- "pid", indicating unique participant IDs;
- "trial", indicating each unique trial per participant;
- "condition", a factor indicating the levels of the independent variable (1, 2, ...) and the corresponding labels;
- "rt", indicating the response times in ms;
- "acc", indicating the accuracies (1/0).

In the code of Tutorial 1a, this is accomplished as follows.

```
data_wr<-read_csv("../Tutorial_1_descriptive_stats/data/DataExp1_6subjects_wrangled.csv")
data_wr <- data_wr %>%
  rename(pid = vp, condition = prime_type, acc = respac, trial = TrialNr) %>%
  mutate(condition = condition + 1, # original levels were 0, 1, 2.
         condition = factor(condition,
                            levels=c(1,2,3),
                            labels=c("blank","congruent","incongruent")))
```

Next, we can set up the life tables and plot for each condition the discrete-time hazard function h(t), survivor function S(t), probability mass function P(t), and conditional accuracy function ca(t). To do so using a functional programming approach, one has to nest the person-trial data within participants using the group_nest() function, and supply a user-defined censoring time and bin width to our custom function "censor()", as follows.

```
data_nested <- data_wr %>% group_nest(pid)
data_final <- data_nested %>%
  # ! user input: censoring time, and bin width
  mutate(censored = map(data, censor, 600, 40)) %>%
  # create person-trial-bin data set
  mutate(ptb_data = map(censored, ptb)) %>%
  # create life tables without ca(t)
  mutate(lifetable = map(ptb_data, setup_lt)) %>%
  # calculate ca(t)
  mutate(condacc = map(censored, calc_ca)) %>%
  # create life tables with ca(t)
  mutate(lifetable_ca = map2(lifetable, condacc, join_lt_ca)) %>%
  # create plots
  mutate(plot = map2(.x = lifetable_ca, .y = pid, plot_eha,1))
```

Note that the censoring time (here: 600 ms) should be a multiple of the bin width (here: 40 ms). The censoring time should be a time point after which no informative responses are expected anymore, in case one waits for a response in each trial. In experiments that implement a response deadline in each trial the censoring time can equal that deadline time point. Trials with a RT larger than the censoring time, or trials in which no response is emitted during the data observation period, are treated as right-censored observations in EHA. In other words, these trials are not discarded, because they contain the information that the event did not occur before the censoring time. Removing such trials before calculating the mean event time would result in underestimation of the true mean.

The person-trial-bin oriented data set is created by our custom function ptb(), and it has one row for each time bin (of each trial) that is at risk for event occurrence. The variable "event" in the person-trial-bin oriented data set indicates whether a response occurs (1) or not (0) for each bin. The next steps are to set up the life table using our custom function setup_lt(), calculate the conditional accuracies using our custom function calc_ca(), add the ca(t) estimates to the life table using our custom function join_lt_ca(), and then plot the descriptive statistics using our custom function plot_eha(). One can now inspect different aspects, including the life table for a particular condition of a particular subject, and a plot of the different functions for a particular participant.

In general, it is important to visually inspect the functions first for each participant, in order to identify individuals that may not be following task instructions (e.g., a flat conditional accuracy function at .5 indicates that someone is just guessing), outlying individuals, and/or different groups with qualitatively different behavior. Also, to select a suited bin width for model fitting, one can test and compare various bin widths in the censor function, and select the smallest one that is supported by the data. Too small bin widths will result in erratic hazard functions because many bins will have estimates equal to zero.

Table 3 shows the life table for condition "blank" (no prime stimulus presented) for participant 6.

Table 3

The life table for the blank prime condition of participant 6.

bin	risk_set	events	hazard	se_haz	survival	se_surv	ca	se_ca
0	220	NA	NA	NA	1.00	0.00	NA	NA
40	220	0	0.00	0.00	1.00	0.00	NA	NA
80	220	0	0.00	0.00	1.00	0.00	NA	NA
120	220	0	0.00	0.00	1.00	0.00	NA	NA
160	220	0	0.00	0.00	1.00	0.00	NA	NA
200	220	0	0.00	0.00	1.00	0.00	NA	NA
240	220	0	0.00	0.00	1.00	0.00	NA	NA
280	220	7	0.03	0.01	0.97	0.01	0.29	0.17
320	213	13	0.06	0.02	0.91	0.02	0.77	0.12
360	200	26	0.13	0.02	0.79	0.03	0.92	0.05
400	174	40	0.23	0.03	0.61	0.03	1.00	0.00
440	134	48	0.36	0.04	0.39	0.03	0.98	0.02
480	86	37	0.43	0.05	0.22	0.03	1.00	0.00
520	49	32	0.65	0.07	0.08	0.02	1.00	0.00
560	17	9	0.53	0.12	0.04	0.01	1.00	0.00
600	8	4	0.50	0.18	0.02	0.01	1.00	0.00

Note. The column named "bin" indicates the endpoint of each time bin (in ms), and includes time point zero. For example the first bin is (0,40] with the starting point excluded and the endpoint included. At time point zero, no events can occur and therefore h(t=0) and ca(t=0) are undefined. se = standard error. ca = conditional accuracy. NA = undefined.

Figure 2 displays the discrete-time hazard, survivor, conditional accuracy, and probability mass functions for each prime condition for participant 6. By using discrete-time hazard functions of event occurrence – in combination with conditional accuracy functions for two-choice tasks – one can provide an unbiased, time-varying, and probabilistic description of the latency and accuracy of responses based on all trials of any RT data set.

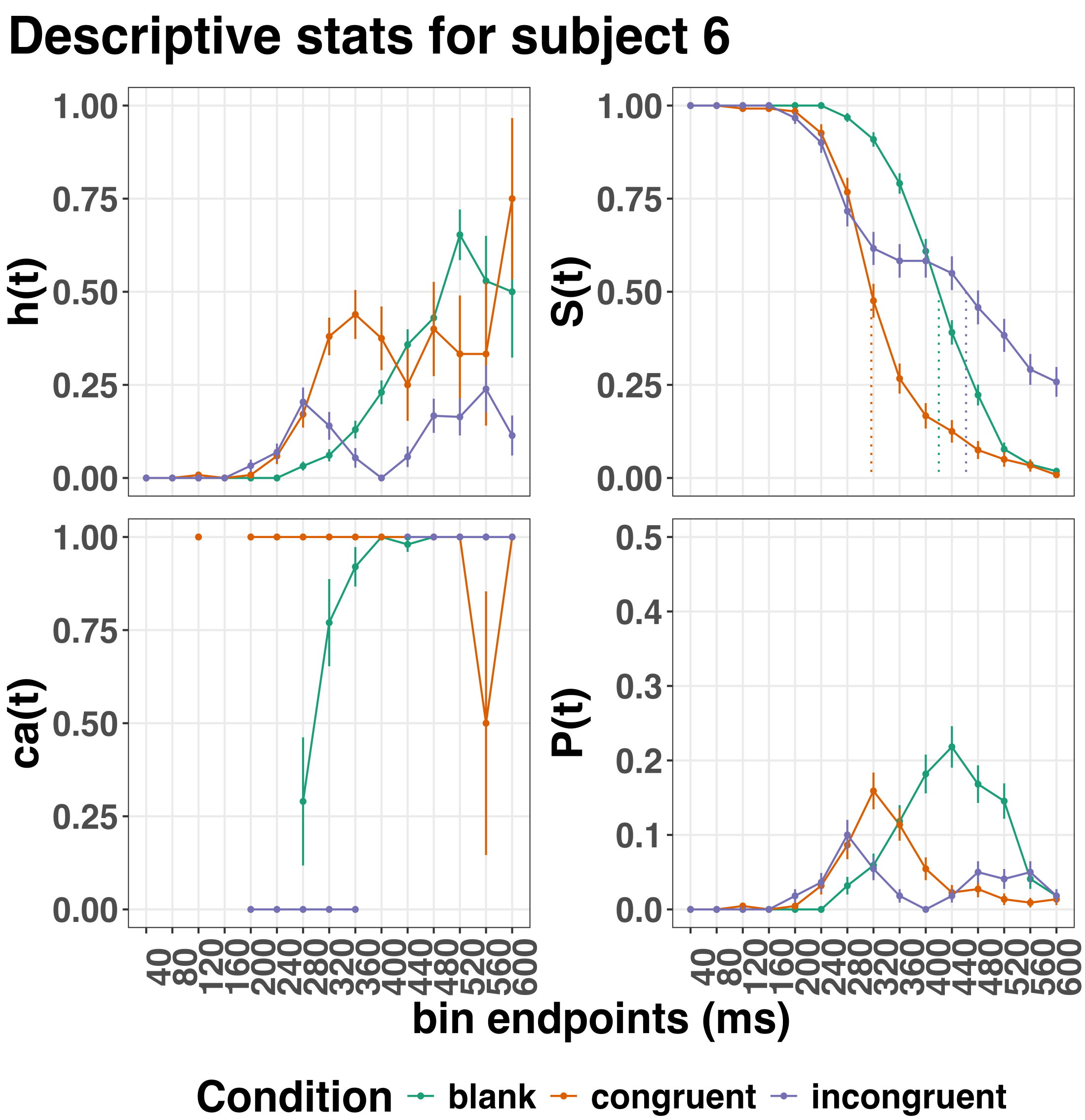

Figure 2. Estimated discrete-time hazard (h), survivor (S), conditional accuracy (ca) and probability mass (P) functions for participant 6. Vertical dotted lines indicate the estimated median RTs. Error bars represent +/- 1 standard error of the respective proportion.

For example, for participant 6, the estimated hazard values in bin (240,280] are 0.03, 0.17, and 0.20 for the blank, congruent, and incongruent prime conditions, respectively. In other words, when the waiting time has increased until *240 ms* after target onset, then the conditional probability of response occurrence in the next 40 ms is more than five times larger for both prime-present conditions, compared to the blank prime condition.

Furthermore, the estimated conditional accuracy values in bin (240,280] are 0.29, 1, and 0 for the blank, congruent, and incongruent prime conditions, respectively. In other words, if a response is emitted in bin (240,280], then the probability that it is correct is estimated to be 0.29, 1, and 0 for the blank, congruent, and incongruent prime conditions, respectively.

However, when the waiting time has increased until *400 ms* after target onset, then the conditional probability of response occurrence in the next 40 ms is estimated to be 0.36, 0.25, and 0.06 for the blank, congruent, and incongruent prime conditions, respectively. And when a response does occur in bin (400,440], then the probability that it is correct is estimated to be 0.98, 1, and 1 for the blank, congruent, and incongruent prime conditions, respectively.

These distributional results suggest that participant 6 is initially responding to the prime even though (s)he was instructed to only respond to the target, that response competition emerges in the incongruent prime condition around 300 ms, and that only slower responses are fully controlled by the target stimulus. Qualitatively similar results were obtained for the other five participants. When participants show qualitatively similar distributional patterns, one might consider aggregating their data and plotting the group-average distribution per condition (see Tutorial_1a.Rmd). More generally, these results go against the (often implicit) assumption in research on priming that all observed responses are primed responses to the target stimulus. Instead, the distributional data show that fast responses are triggered exclusively by the prime stimulus, while only the slower responses reflect primed responses to the target stimulus.

At this point, we have calculated and plotted the descriptive statistics for each type of prime stimulus. As we will show in later Tutorials, statistical models for hazard and conditional accuracy functions can be implemented as generalized linear mixed regression models predicting event occurrence (1/0) and conditional accuracy (1/0) in each bin of a selected time window for analysis. But first we consider calculating the descriptive statistics for within-subject designs with two independent variables.

3.2 Tutorial 1b: Generalising to a more complex design

So far in this paper, we have used a simple experimental design, which involved one condition with three levels. But psychological experiments are often more complex, with crossed factorial designs and/or conditions with more than three levels. The purpose of Tutorial 1b, therefore, is to provide a generalisation of the basic approach, which extends to a more complicated design. We feel that this might be useful for researchers in experimental psychology that typically use crossed factorial designs.

To this end, Tutorial 1b illustrates how to calculate and plot the descriptive statistics for the full data set of Experiment 1 of Panis and Schmidt (2016), which includes two independent variables: mask type and prime type. As we use the same functional programming approach as in Tutorial 1a, we simply present the sample-based functions for each participant as part of Tutorial_1b.Rmd for those that are interested.

3.3 Tutorial 2a: Fitting Bayesian hazard models to interval-censored RT data

In this third tutorial, we illustrate how to fit Bayesian multilevel regression models to the RT data of the masked response priming data used in Tutorial 1a. Fitting (Bayesian or non-Bayesian) regression models to time-to-event data is important when you want to study how the shape of the hazard function depends on various predictors (Singer &

Willett, 2003).

In general, when fitting regression models, our lab adopts an estimation approach to multilevel regression (Kruschke & Liddell, 2018; Winter, 2019), which is heavily influenced by the Bayesian framework as suggested by Richard McElreath (Kurz, 2023b; McElreath, 2020). We also use a "keep it maximal" approach by specifying a full varying (or random) effects structure (Barr, Levy, Scheepers, & Tily, 2013). This means that wherever possible we include varying intercepts and slopes per participant. To make inferences, we use two main approaches. We compare models of different complexity using information criteria and cross-validation, to evaluate out-of-sample predictive accuracy (McElreath, 2020). We also take the most complex model and evaluate key parameters of interest using point and interval estimates.

3.3.1 Hazard model considerations. There are several analytic decisions one has to make when fitting a discrete-time hazard model. First, because the first few bins typically contain no responses, one has to select an analysis time window, i.e., a contiguous set of bins for which there is data for each participant. Second, given that the dependent variable (event occurrence) is binary, one has to select a link function (see section D of the Supplemental Material). The cloglog link is preferred over the logit link when events can occur in principle at any time point within a bin, which is the case for RT data (Singer & Willett, 2003). Third, one has to choose whether to treat TIME (i.e., the time bin index t) as a categorical or continuous predictor (see also section E of the Supplemental Material). For example, when you want to know if cloglog-hazard is changing linearly or quadraticly over time, you should treat TIME as a continuous predictor. When you are only interested in the effect of covariates on hazard, you can treat TIME as a categorical predictor (i.e., fit an intercept for each bin), in which case you can choose between reference coding and index coding. With reference coding, one defines the variable as a factor and selects one of the k categories as the reference level. Brm() will then construct k-1 indicator variables (see model M1d in Tutorial_2a.Rmd for an example). With index coding, one constructs an index variable that contains integers that correspond to different categories (see models M0i and M1i below). As explained by McElreath (2020), the advantage of index coding is that the same prior can be assigned to each level of the index variable, so that each category has the same prior uncertainty.

In the case of a large-N design without repeated measurements, the parameters of a discrete-time hazard model can be estimated using standard logistic regression software after expanding the typical person-trial data set into a person-trial-bin data set (Allison, 2010). When there is clustering in the data, as in the case of a small-N design with repeated measurements, the parameters of a discrete-time hazard model can be estimated using population-averaged methods (e.g., Generalized Estimating Equations), and Bayesian or frequentist generalized linear mixed models (Allison, 2010).

In general, there are three assumptions one can make or relax when adding experimental predictor variables and other covariates: The linearity assumption for continuous predictors (the effect of a 1 unit change is the same anywhere on the scale), the additivity assumption (predictors do not interact), and the proportionality assumption (predictors do not interact with TIME).

In tutorial_2a.Rmd we fit several Bayesian multilevel models (i.e., generalized linear mixed models) that differ in complexity to the person-trial-bin oriented data set that we created in Tutorial 1a. We decided to select the analysis time window (200,600] and the cloglog link. Below, we shortly discuss two of these models. The person-trial-bin data set is prepared as follows.

```
# read in the file we saved in tutorial 1a
ptb_data <- read_csv("Tutorial_1_descriptive_stats/data/inputfile_hazard_modeling.csv")


ptb_data <- ptb_data %>%
# select analysis time range: (200,600] with 10 bins (time bin ranks 6 to 15)
filter(period > 5) %>%
```

```
        # define categorical predictor TIME as index variable named timebin
mutate(timebin = factor(period, levels = c(6:15)),
        # factor "condition" using reference coding, with "blank" as the reference level
        condition = factor(condition, labels = c("blank", "congruent","incongruent")),
        # categorical predictor "prime" with index coding
        prime = ifelse(condition=="blank", 1, ifelse(condition=="congruent", 2, 3)),
        prime = factor(prime, levels = c(1,2,3)))
```

3.3.2 Prior distributions. To get the posterior distribution of each model parameter given the data, we need to specify prior distributions for the model parameters which reflect our prior beliefs. In Tutorial_2a.Rmd we perform a few prior predictive checks to make sure our selected prior distributions reflect our prior beliefs (Gelman, Vehtari, et al., 2020).

The middle column of Supplementary Figure 3 (section F of the Supplemental Material) shows six examples of prior distributions for an intercept on the logit and/or cloglog scales. While a normal distribution with relatively large variance is often used as a weakly informative prior for continuous dependent variables, rows A and B of Supplementary Figure 3 show that specifying such distributions on the logit and cloglog scales actually leads to rather informative distributions on the original probability scale, as most mass is pushed to probabilities of 0 and 1.

3.3.3 Model M0i: A null model with index coding. When you do not want to make assumptions about the shape of the hazard function, or its shape is not smooth but irregular, then you can use a general specification of TIME, i.e., fit one grand intercept per time bin. In this first baseline or reference model, we use a general specification of TIME using index coding, and do not include experimental predictors. We call this model "M0i". The other model (see section 3.3.4) extends model M0i by including our experimental predictor prime type.

Before we fit model M0i, we select the necessary columns from the data, and specify our priors. In the code of Tutorial 2a, model M0i is specified as follows.

```
model_M0i <-
  brm(data = data_M0i,
      family = bernoulli(link="cloglog"),
      formula = event ~ 0 + timebin + (0 + timebin | pid),
      prior = priors_M0i,
      chains = 4, cores = 4,
      iter = 3000, warmup = 1000,
      control = list(adapt_delta = 0.999,
                     step_size = 0.04,
                     max_treedepth = 12),
      seed = 12, init = "0",
      file = "Tutorial_2_Bayesian/models/model_M0i")
```

After selecting the bernoulli family and the cloglog link, the model formula is specified. The specification "0 + ..." removes the default intercept in brm(). The fixed effects include an intercept for each level of timebin. Each of these intercepts is allowed to vary across individuals (variable pid). We request 2000 samples from the posterior distribution for each of four chains. Estimating model M0i took about 30 minutes on a MacBook Pro (Sonoma 14.6.1 OS, 18GB Memory, M3 Pro Chip).

3.3.4 Model M1i: Adding the effects of prime-target congruency. Previous research has shown that psychological effects typically change over time (Panis, 2020; Panis, Moran, et al., 2020; Panis & Schmidt, 2022; Panis et al., 2017; Panis & Wagemans, 2009). In the next model, therefore, we use index coding for both TIME (variable "timebin") and the categorical predictor prime-target-congruency (variable "prime"), so that we get 30 grand intercepts, one for each combination of timebin level and prime level. Here is the model formula of this model that we call "M1i".

```
event ~ 0 + timebin:prime + (0 + timebin:prime | pid)
```

Estimating model M1i took about 124 minutes.

3.3.5 Compare the models. There are two popular strategies to evaluate how well models will perform in predicting new data on average: Leave-One-Out (LOO) cross-validation and the Widely Applicable Information Criterion or WAIC (McElreath, 2020). LOO-weights represent the optimal linear combination of models for predictive performance, with higher weights for models with better out-of-sample predictive performance. WAIC-weights represent the relative evidence for each model, with higher weights for models with a better fit while accounting for model complexity (Kurz, 2023a; McElreath, 2020).

```
model_weights(model_M0i, model_M1i, weights = "loo") %>% round(digits = 2) %>% format(nsmall=2)
```

```
## model_M0i model_M1i
##    "0.00"    "1.00"
```

```
model_weights(model_M0i, model_M1i, weights = "waic") %>% round(digits = 1) %>% format(nsmall=2)
```

```
## model_M0i model_M1i
##    "0.00"    "1.00"
```

Clearly, both the loo and waic weighting schemes assign a weight of 1 to model M1i, and a weight of 0 to model M0i.

3.3.6 Evaluating parameter estimates in model M1i. To make causal inferences from the parameter estimates in model M1i, we first plot the densities of the draws from the posterior distributions of its population-level parameters in Figure 3, together with point (median) and interval estimates (80% and 95% credible intervals). A credible interval is a range of values that contains a parameter's true value with a specified probability, given the observed data and model.

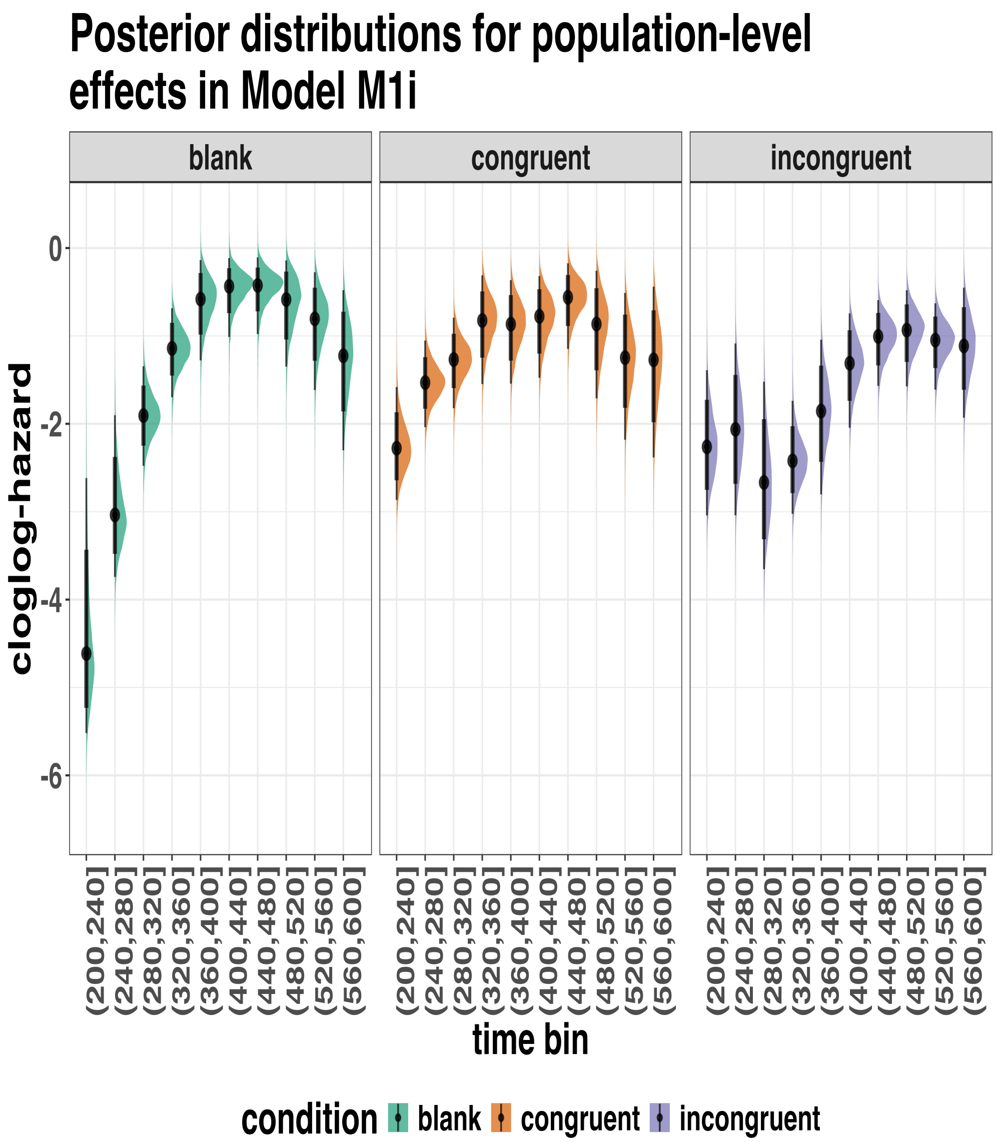

Figure 3. Medians and 80/95% credible intervals of the posterior distributions of the population-level parameters of model M1i.

Because the parameter estimates are on the cloglog-hazard scale, we can ease our interpretation by plotting the expected value of the posterior predictive distribution – the predicted hazard values – at the population level (Figure 4A), and for each participant in the data set (Figure 4B).

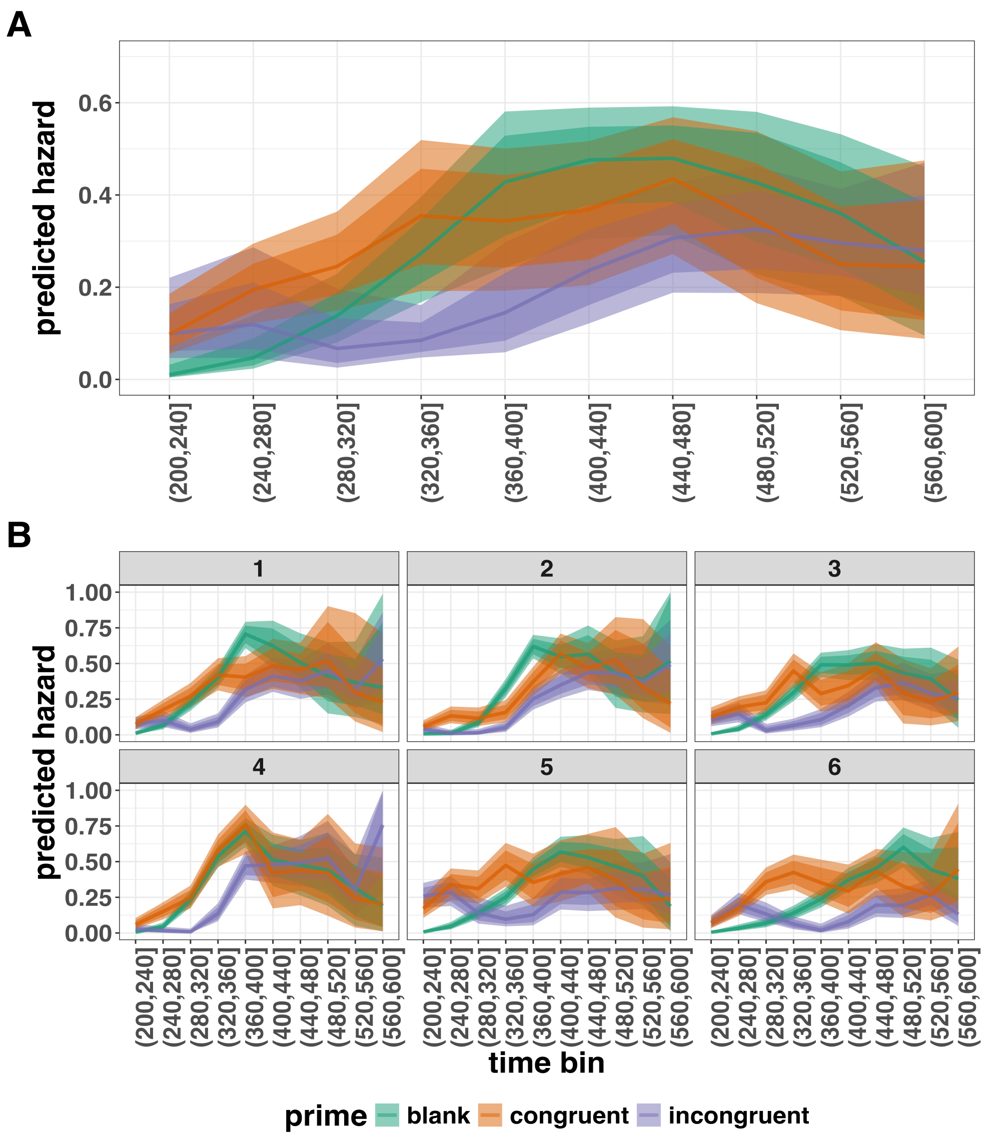

Figure 4. Point (median) and 80/95% credible interval summaries of the hazard estimates (expected values of the draws from the posterior predictive distributions) in each time bin at the population level (A), and for each participant (B).

As we are actually interested in the effects of congruent and incongruent primes, relative to the blank prime condition, we can construct two contrasts (congruent-blank, incongruent-blank), and plot the posterior distributions of these contrast effects, both at the population level (Figure 5A) and at the participant level (Figure 5B).

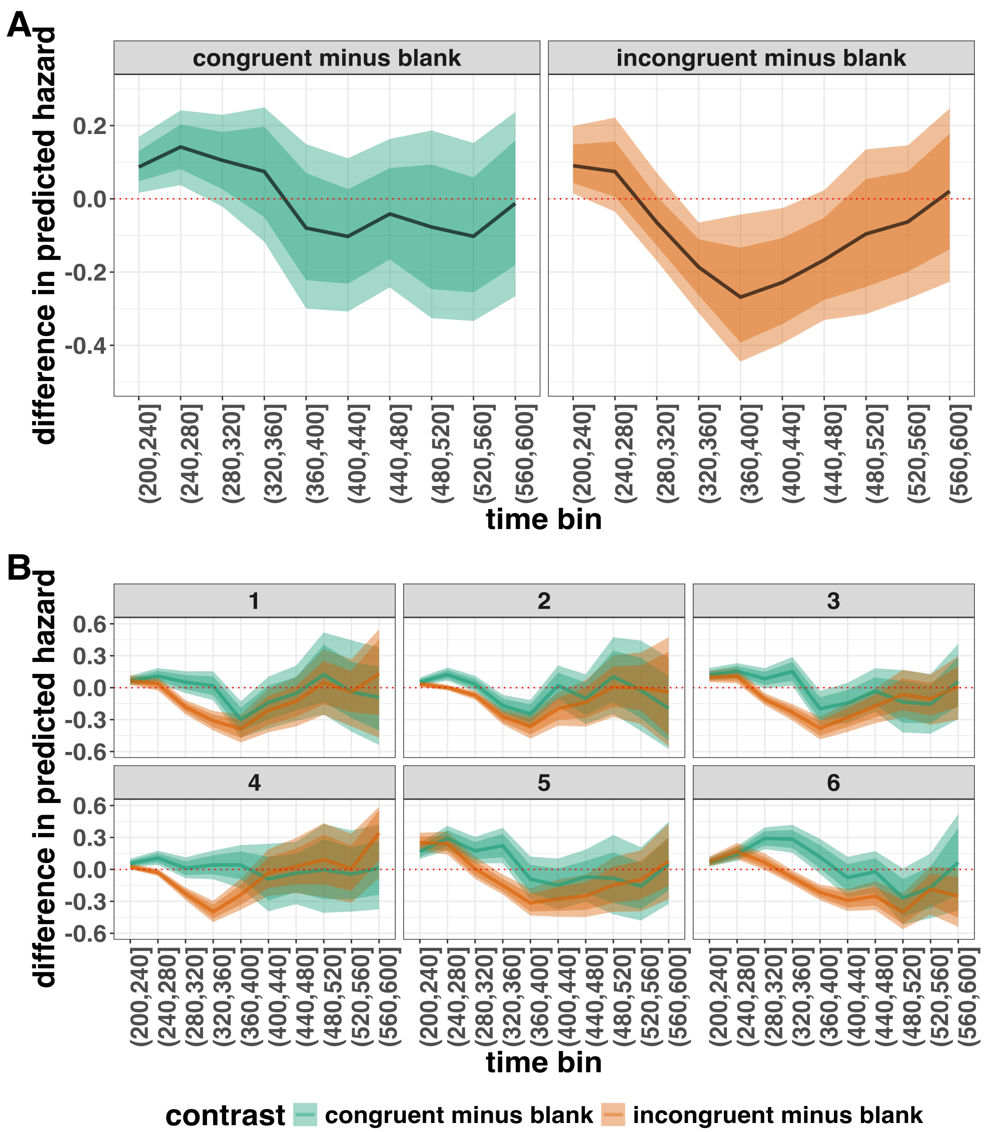

Figure 5. Point (mean) and 80/95% credible interval summaries of estimated differences in hazard in each time bin at the population level (A), and for each participant (B).

The point estimates and quantile intervals can also be reported in a table (see Tutorial_2a.Rmd for details).

Example conclusions for M1i. What can we conclude from model M1i about our research question, i.e., the temporal dynamics of the effect of prime-target congruency on RT? In other words, in which of the 40-ms time bins between 200 and 600 ms after target onset does changing the prime from blank to congruent or incongruent affect the hazard of response occurrence (for a prime-target stimulus-onset-asynchrony of 187 ms)?

If we want to estimate the population-level effect of prime type on hazard, we can base our conclusion on the credible Intervals (CrIs) in Figure 5A. The contrast "congruent minus blank" was estimated to be 0.09 hazard units in bin (200,240] (95% CrI = [0.02, 0.17]), and 0.14 hazard units in bin (240,280]) (95% CrI = [0.04, 0.25]). For the other bins, the 95% credible interval contained zero. The contrast "incongruent minus blank" was estimated to be 0.09 hazard units in bin (200,240] (95% CrI = [0.01, 0.21]), -0.19 hazard units in bin (320,360] (95% CrI = [-0.31, -0.06]), -0.27 hazard units in bin (360,400] (95% CrI = [-0.45, -0.04]), and -0.23 hazard units in bin (400,440] (95% CrI = [-0.40, -0.03]). For the other bins, the 95% credible interval contained zero.

There are thus two phases of performance for the average person between 200 and 600 ms after target onset. In the first phase, the addition of a congruent or incongruent prime stimulus increases the hazard of response occurrence compared to blank prime trials in the time period (200, 240]. In the second phase, only the incongruent prime decreases the hazard of response occurrence compared to blank primes, in the time period (320,440]. The sign of the effect of incongruent primes on the hazard of response occurrrence thus depends on how much waiting time has passed since target onset.

If we want to focus more on inter-individual differences, we can study the subject-specific hazard functions in Figure 5B. Note that three participants (1, 2, and 3) show a negative difference for the contrast "congruent minus incongruent" in bin (360,400] – subject 2 also in bin (320,360]. Interestingly, all subjects show a tendency in their mean difference (congruent minus blank) to "dip" around that time. Therefore, future modeling efforts could incorporate the trial number into the model formula, in order to also study how the effects of prime type on hazard change on the long experiment-wide time scale, next to the short trial-wide time scale. In Tutorial_2a.Rmd we provide a number of model formulae that should get you going.

3.4 Tutorial 2b: Fitting Bayesian conditional accuracy models

In this fourth tutorial, we illustrate how to fit a Bayesian multilevel regression model to the timed accuracy data from the masked response priming data used in Tutorial 1a. The general process is similar to Tutorial 2a, except that (a) we use the person-trial data, (b) we use the symmetric logit link function, and (c) we change the priors (our prior belief is that conditional accuracy values between 0 and 1 are equally likely). To keep the tutorial short, we only fit one conditional accuracy model, which was based on model M1i from Tutorial 2a and labelled M1i_ca.

To make inferences from the parameter estimates in model M1i_ca, we first plot the densities of the draws from the posterior distributions of its population-level parameters in Figure 6, together with point (median) and interval estimates (80% and 95% credible intervals).

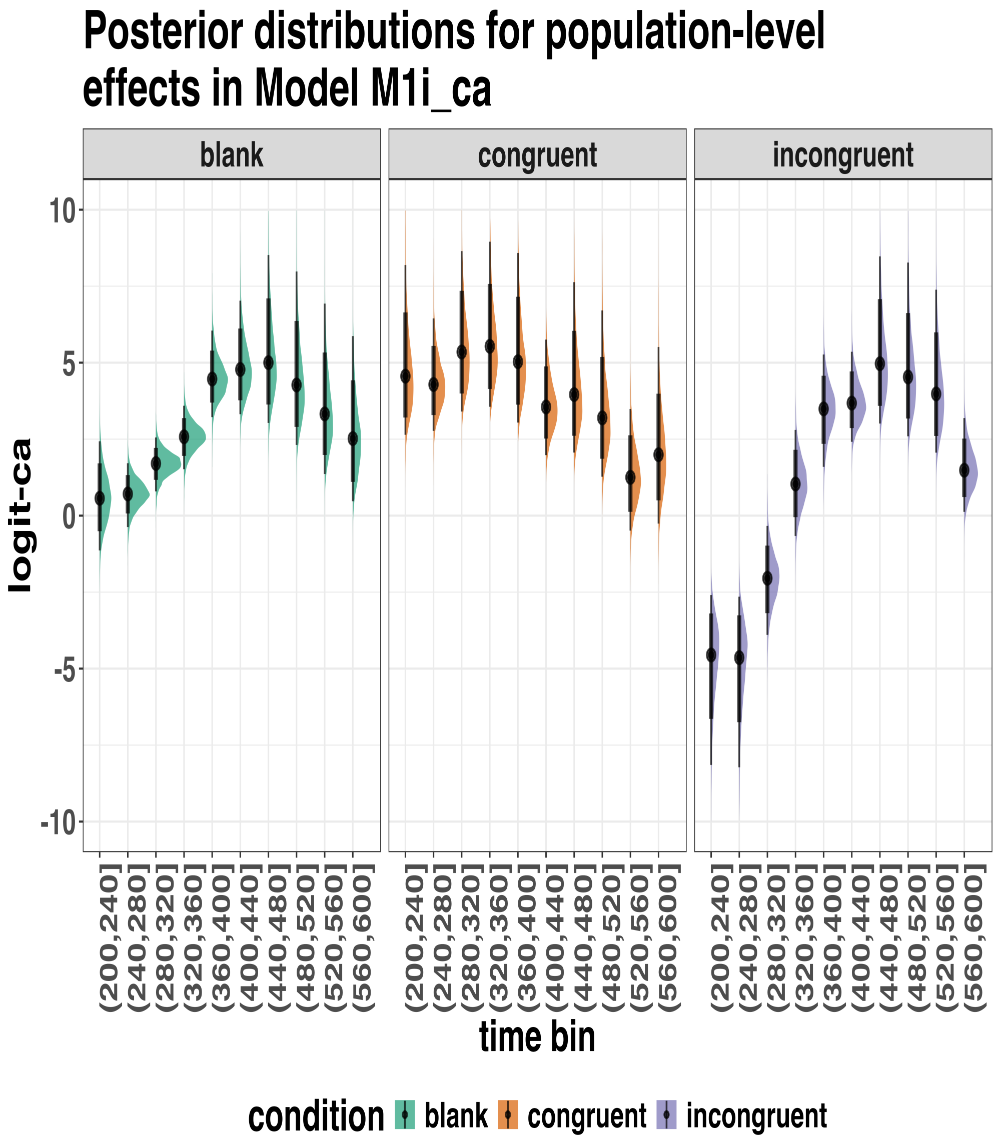

Figure 6. Medians and 80/95% credible intervals of the posterior distributions of the population-level parameters of model M1i_ca. ca = conditional accuracy.

Because the parameter estimates are on the logit-ca scale, we can ease our interpretation by plotting the expected value of the posterior predictive distribution – the predicted conditional accurcies – at the population level (Figure 7A), and for each participant in the data set (Figure 7B).

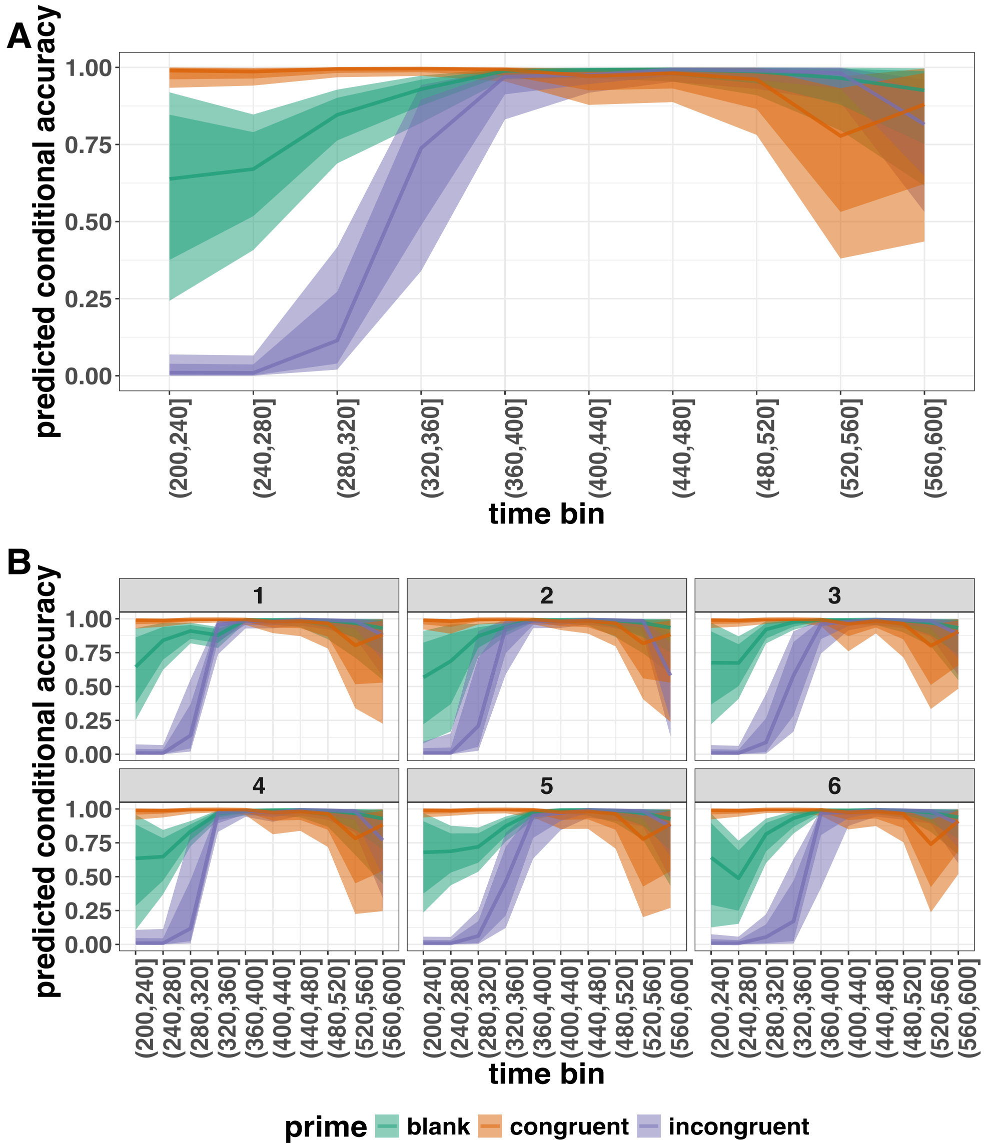

Figure 7. Point (median) and 80/95% credible interval summaries of the conditional accuracy estimates (expected values of the draws from the posterior predictive distributions) in each time bin at the population level (A), and for each participant (B).

As we are actually interested in the effects of congruent and incongruent primes, relative to the blank prime condition, we can construct two contrasts (congruent-blank, incongruent-blank), and plot the posterior distributions of these contrast effects at the population level (Figure 8A) and for each participant (Figure 8B).

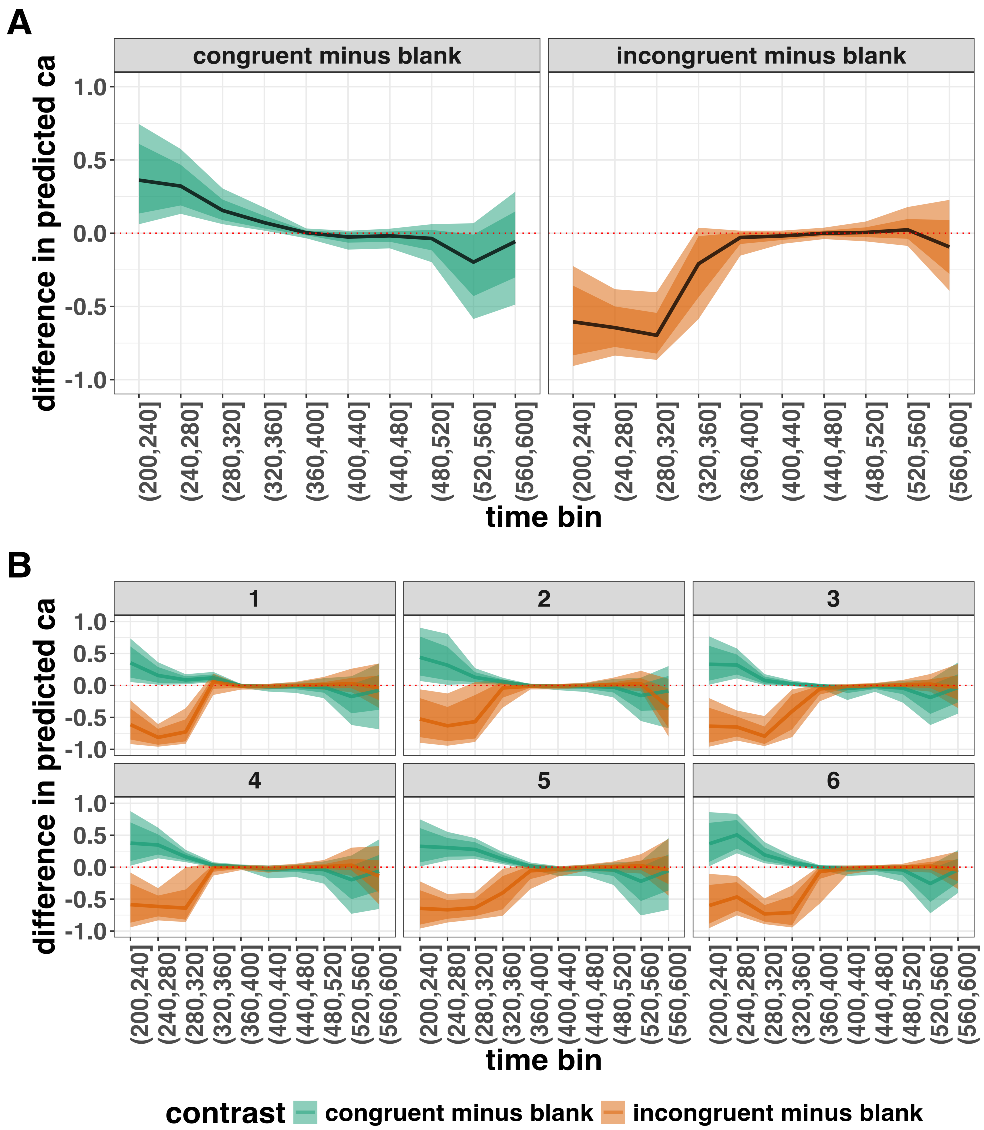

Figure 8. Point (mean) and 80/95% credible interval summaries of estimated differences in conditional accuracy in each time bin at the population level (A), and for each participant (B).

Based on Figure 8A we see that on the population level congruent primes have a positive effect on the conditional accuracy of emitted responses in time bins (200,240], (240,280], (280,320], and (320,360], relative to the estimates in the baseline condition (blank prime; red dashed lines in Figure 8A). Incongruent primes have a negative effect on the conditional accuracy of emitted responses in the first time bins, relative to the estimates in the baseline condition.

Before we move to our Tutorial on planning experiments, we also provide code to fit hazard and conditional accuracy models in the frequentist framework (see Tutorial_3a.Rmd and Tutorial_3b.Rmd). However, because multilevel generalized linear regression models often do not converge with complex random-coefficient structures, we do not discuss them here.

3.5 Tutorial 4: Planning

In the final tutorial, we look at planning a future experiment, which uses EHA.

3.5.1 Background. The general approach to planning that we adopt here involves simulating reasonably structured data to help guide what you might be able to expect from your data once you collect it (Gelman, Vehtari, et al., 2020). The basic structure and code follows the examples outlined by Solomon Kurz in his 'power' blog posts (https://solomonkurz.netlify.app/blog/bayesian-power-analysis-part-i/) and Lisa Debruine's R package faux{} (https://debruine.github.io/faux/) as well as these related papers (DeBruine & Barr, 2021; Pargent, Koch, Kleine, Lermer, & Gaube, 2024).

3.5.2 Basic workflow. The basic workflow is as follows:

1. Fit a regression model to existing data.
2. Use the regression model parameters to simulate new data.
3. Write a function to create 1000s of datasets and vary parameters of interest (e.g., sample size, trial count, effect size).
4. Summarise the simulated data to estimate likely power or precision of the research design options.

Ideally, in the above workflow, we would also fit a model to each dataset and summarise the model output, rather than the raw data. However, when each model takes several hours to build, and we may want to simulate many 1000s of datasets, it can be computationally demanding for desktop machines. So, for ease, here we just use the raw simulated datasets to guide future expectations.

In the below, we only provide a high-level summary of the process and let readers dive into the details within the tutorial should they feel so inclined.

3.5.3 Fit a regression model and simulate one dataset. We again use the data from Panis and Schmidt (2016) to provide a worked example. We fit an index coding model on a subset of time bins (six time bins in total) and for two prime conditions (congruent and incongruent). We chose to focus on a subsample of the data to ease the computational burden. We also used a full varying effects structure, with the model formula as follows:

```
event ~ 0 + timebin:prime + (0 + timebin:prime | pid)
```

We then took parameters from this model and used them to create a single dataset with 200 trials per condition for 10 individual participants. The raw data and the simulated data are plotted in Figure 9 and show quite close correspondence, which is re-assuring. But, this is only one dataset. What we really want to do is simulate many datasets and vary parameters of interest, which is what we turn to in the next section.

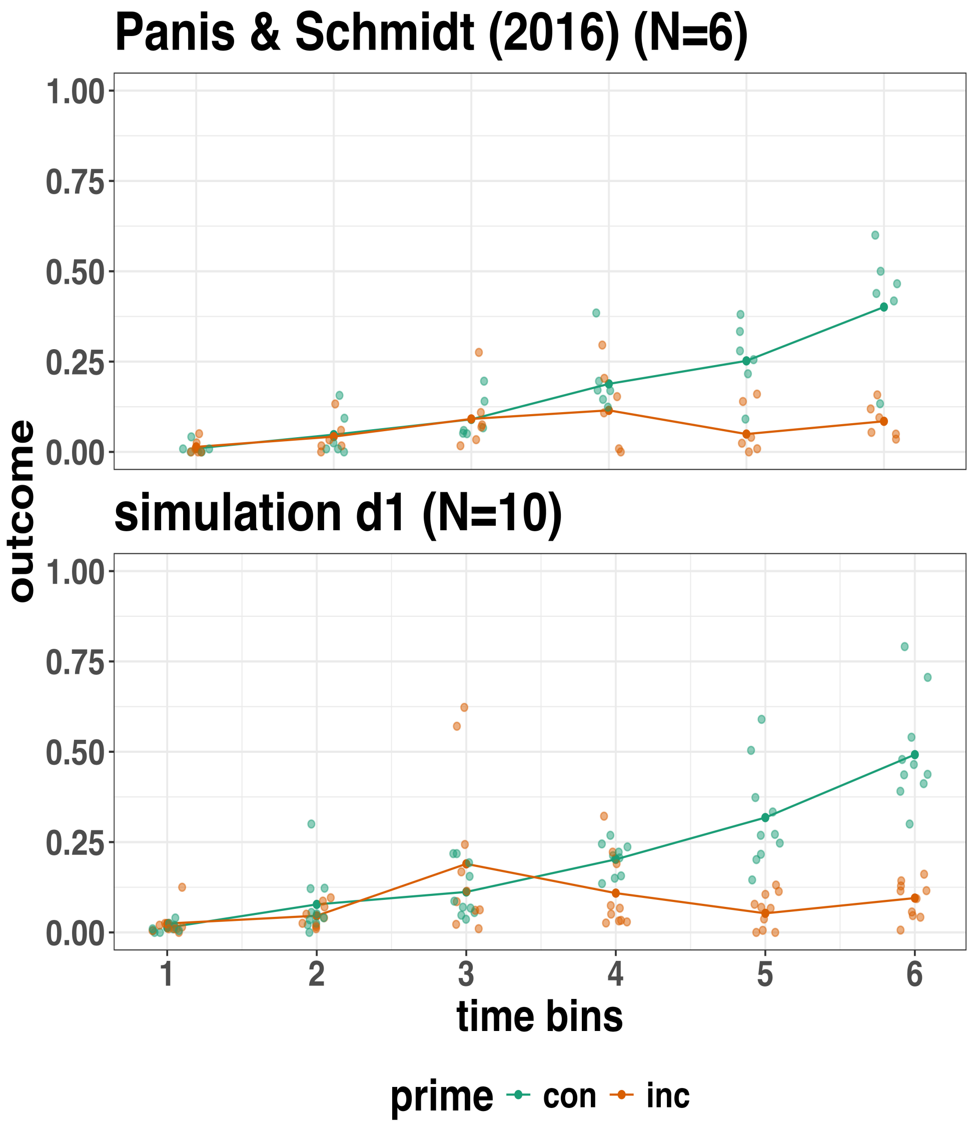

Figure 9. Raw data from Panis and Schmidt (2016) and simulated data from 10 participants.

3.5.4 Simulate and summarise data across a range of parameter values.

Here we use the same data simulation process as used above, but instead of simulating one dataset, we simulate 1000 datasets per variation in parameter values. Specifically, in Simulation 1, we vary the number of trials per condition (100, 200, and 400), as well as the effect size in bin 6. We focus on bin 6 only, in terms of varying the effect size, just to make things simpler and easier to understand. The effect size observed in bin 6 in this subsample of data was a 79% reduction in hazard value from the congruent prime (0.401 hazard value) to the incongruent prime condition (0.085 hazard value). In other words, a hazard ratio of 0.21 (e.g., 0.085/0.401 = 0.21). As a starting point, we chose three effect sizes, which covered a fairly broad range of hazard ratios (0.25, 0.5, 0.75), which correspond to a 75%, 50% and 25% reduction in hazard value as a function of prime condition.

Summary results from Simulation 1 are shown in Figure 10A. Figure 10A depicts statistical "power" as calculated by the percentage of lower-bound 95% confidence intervals that exclude zero when the difference between prime condition is calculated (congruent - incongruent). In other words, what fraction of the simulated datasets generated an effect of prime that excludes the criterion mark of zero. We are aware that "power" is not part of a Bayesian analytical workflow, but we choose to include it here, as it is familiar to most researchers in experimental psychology.

The results of Simulation 1 show that if we were targeting an effect size similar to the one reported in the original study, then testing 10 participants and collecting 100 trials per condition would be enough to provide over 95% power. However, we could not be as confident about smaller effects, such as a hazard ratio of 50% or 25%. From this simulation, we can see that somewhere between an effect size of a 50% and 75% reduction in hazard value, power increases to a range that most researchers would consider acceptable (i.e., >95% power). To probe this space a little further, we decided to run a second simulation, which varied different parameters.

In Simulation 2, we varied the effect size between a different range of values (0.5, 0.4, 0.3), which correspond to a 50%, 60% and 70% reduction in hazard value as a function of prime condition. In addition, we varied the number of participants per experiment between 10, 15, and 20 participants. Given that trial count per condition made little difference to power in Simulation 1, we fixed trial count at 200 trials per condition in Simulation 2. Summary results from Simulation 2 are shown in Figure 10B. A summary of these power calculations might be as follows (trial count = 200 per condition in all cases):

- For a 70% reduction (0.3 hazard ratio), N=10 would give nearly 100% power.
- For a 60% reduction (0.4 hazard ratio), N=10 would give nearly 90% power.
- For a 50% reduction (0.5 hazard ratio), N=15 would give over 80% power.

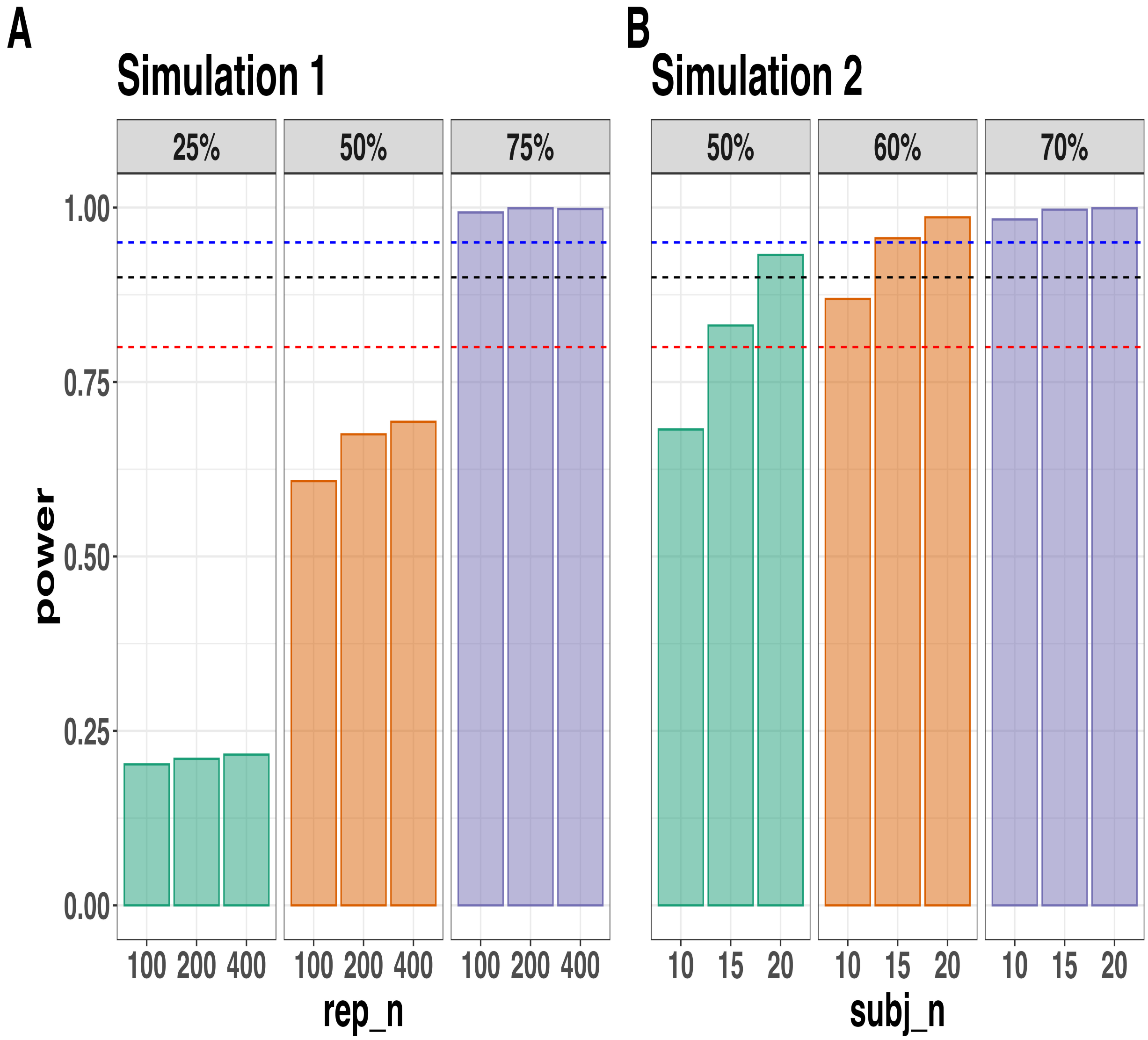

Figure 10. Statistical power across data Simulation 1 (A) and Simulation 2 (B). Power was calculated as the percentage of lower-bound 95% confidence intervals that exclude zero when the difference between prime condition is calculated (congruent - incongruent). In Simulation 1, the effect size was varied between a 25%, 50% and 75% reduction in hazard value, whereas the trial count was varied between 100, 200 and 400 trials per condition (the number of participants was fixed at N=10). In Simulation 2, the effect size was varied between a 50%, 60% and 70% reduction in hazard value, whereas the number of participants was varied between N=10, 15 and 20 (the number of trials per condition was fixed at 200). The dashed lines represent 80% (red), 90% (black) and 95% (blue) power. Abbreviations: rep_n = the number of trials per experimental condition; subj_n = the number of participants per simulated experiment.

3.5.5 Planning decisions. Now that we have summarised our simulated data, what planning decisions could we make about a future study? More concretely, how many trials per condition should we collect and how many participants should we test? Like almost always when planning future studies, the answer depends on your objectives, as well as the available resources (Lakens, 2022). There is no straigtforward and clear-cut answer. Some considerations might be as follows:

- How much power or precision are you looking to obtain in this particular study?
- Are you running multiple studies that have some form of replication built in?
- What level of resources do you have at your disposal, such as time, money and personnel?
- How easy or difficult is it to obtain the specific type of sample?

If we were running this kind of study in our lab, what would we do? We might pick a hazard ratio of 0.4 or 0.5 as a target effect size since this is much smaller than that observed previously (Panis & Schmidt, 2016). Then we might pick the corresponding combination of trial count per condition (e.g., 200) and participant sample size (e.g., N=10 or N=15) that takes you over the 80% power mark. If we wanted to maximise power based on these simulations, and we had the time and resources available, then we would test N=20 participants, which would provide >90% power for an effect size of 0.5.

But, and this is an important "but", unless there are unavoidable reasons, no matter what planning choices we made based on these data simulations, we would not solely rely on data collected from one single study. Instead, we would run a follow-up experiment that replicates and extends the initial result. By doing so, we would aim to avoid the Cult of the Isolated Single Study (Nelder, 1999; Tong, 2019), and thus reduce the reliance on any one type of planning tool, such as a power analysis. Then, we would look for common patterns across two or more experiments, rather than trying to make the case that a single study on its own has sufficient evidential value to hit some criterion mark.

4. Discussion

This main motivation for writing this paper is the observation that EHA and SAT analysis remain under-used in psychological research. As a consequence, the field of psychological research is not taking full advantage of the many benefits EHA/SAT provides compared to more conventional analyses. By providing a freely available set of tutorials, which provide step-by-step guidelines and ready-to-use R code, we hope that researchers will feel more comfortable using EHA/SAT in the future. Indeed, we hope that our tutorials may help to overcome a barrier to entry with EHA/SAT, which is that such approaches require more analytical complexity compared to standard approaches. While we have focused here on within-subject, factorial, small-N designs, it is important to realize that EHA/SAT can be applied to other designs as well (large-N designs with only one measurement per subject, between-subject designs, etc.). As such, the general workflow and associated code can be modified and applied more broadly to other contexts and research questions. In the following, we discuss the main use-cases, issues relating to model complexity and interpretability, as well as limitations of the approach.

4.1 What are the main use-cases of EHA for understanding cognition and brain function?

For those researchers, like ourselves, who are primarily interested in understanding human cognitive and brain systems, we consider two broadly-defined, main use-cases of EHA. First, as we hope to have made clear by this point, EHA is one way to investigating a "temporal states" approach to cognitive processes, by tracking behavior as a function of step-wise increases in absolute waiting time. EHA thus provides a way to uncover the microgenesis of cognitive effects, by revealing when cognitive states may start and stop, how states are replaced with others, as well as what they may be tied to or interact with. Therefore, if your research questions concern **when psychological states occur, and**

how they are temporally organized, our EHA tutorials could be useful tools for you to use.

Second, even if you are not primarily interested in studying the temporal organization of cognitive states, EHA could still be a useful tool to consider using, in order to qualify inferences that are being made based on comparisons between means. Given that distinctly different inferences can be made from the same data based on whether one computes a mean across trials or a RT distribution of events (Figure 1), it may be important for researchers to supplement comparisons between means with EHA. Therefore, if you have a lot of right-censored observations in your RT data set, and/or your research question concerns whether and when responses occur, and whether and when experimental manipulations affect the instantaneous risk of response occurrence, then EHA should be your method of choice.

4.2 Model complexity versus interpretability

Hazard and conditional accuracy models can quickly become very complex when adding more than one time scale, due to the many possible higher-order interactions. For example, some of the models discussed in Tutorial 2a, which we did not focus on in the main text, contain two time scales as covariates: the passage of time on the within-trial time scale, and the passage of time on the across-trial (or within-experiment) time scale. However, when trials are presented in blocks, and blocks of trials within sessions, and when the experiment comprises a number of sessions, then four time scales can be defined (within-trial, within-block, within-session, and within-experiment). From a theoretical perspective, adding more than one time scale – and their interactions – can be important to capture plasticity and other learning effects that may play out on such longer time scales, and that are probably present in each experiment in general (Schöner & Spencer, 2016). From a practical perspective, therefore, some choices need to be made to balance the amount of data that is being collected per participant, condition and across the varying timescales. As one example, if there are several times cales of relevance, then it might be prudent for interpretational purposes to limit the number of experimental predictor variables (conditions). This is of course where planning and data simulation efforts would be important to provide a guide to experimental design choices (see Tutorial 4 and section 2.3).

4.3 Limitations

Compared to the orthodox method – comparing means between conditions – the most important limitation of multilevel hazard and conditional accuracy modeling is that it might take a long time to estimate the parameters using Bayesian methods or the model might have to be simplified significantly to use frequentist methods.

Another issue is that you need a relatively large number of trials per condition to estimate the discrete-time hazard function with relatively high temporal resolution (e.g., 20 ms), which is required when testing predictions of process models of cognition. Indeed, in general, there is a trade-off between the number of trials per condition and the temporal resolution (i.e., bin width) of the discete-time hazard function. Therefore, we recommend researchers to collect as many trials as possible per experimental condition, given the available resources and considering the participant experience (e.g., fatigue and boredom). For instance, if the maximum session length deemed reasonable is between 1 and 2 hours, what is the maximum number of trials per condition that you could reasonably collect? After consideration, it might be worth conducting multiple testing sessions per participant and/or reducing the number of experimental conditions. There is a user-friendly online tool for calculating statistical power as a function of the number of trials as well as the number of participants, and this might be worth consulting to guide the research design process (Baker et al., 2021). Finally, if you have a lot of repeated measurements per condition per participant, you can of course also try continuous-time methods (Allison, 2010; Elmer et al., 2023).

5. Conclusions

Estimating the temporal distributions of RT and accuracy provide a rich source of information on the time course of cognitive processing, which have been largely undervalued in the history of experimental psychology and cognitive neuroscience. We hope that by providing a set of hands-on, step-by-step tutorials, which come with custom-built and freely available code, researchers will feel more comfortable embracing EHA and investigating the shape of empirical hazard functions and the temporal profile of cognitive states. On a broader level, we think that wider adoption of such approaches will have a meaningful impact on the inferences drawn from data, as well as the development of theories regarding the structure of cognition.

Author contributions

Conceptualization: S. Panis and R. Ramsey; Software: S. Panis and R. Ramsey; Writing - Original Draft Preparation: S. Panis; Writing - Review & Editing: S. Panis and R. Ramsey; Supervision: R. Ramsey.

Conflicts of Interest

The author(s) declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Prior versions

All of the submitted manuscript and Supplemental Material was previously posted to a preprint archive: https://doi.org/10.31234/osf.io/57bh6

Supplemental Material

Disclosures

Data, materials, and online resources

Link to public archive:
https://github.com/sven-panis/Tutorial_Event_History_Analysis

Supplemental Material: Panis_Ramsey_suppl_material.pdf

Ethical approval

Ethical approval was not required for this tutorial in which we reanalyze existing data sets.

References

Abney, D. H., Fausey, C. M., Suarez-Rivera, C., & Tamis-LeMonda, C. S. (2025). Advancing a temporal science of behavior. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2025.05.010

Allison, P. D. (1982). Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology*, *13*, 61. https://doi.org/10.2307/270718

Allison, P. D. (2010). *Survival analysis using SAS: A practical guide* (2. ed). Cary, NC: SAS Press.

Aust, F. (2019). *Citr: 'RStudio' add-in to insert markdown citations*. Retrieved from https://github.com/crsh/citr

Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://doi.org/10.32614/CRAN.package.papaja

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295–314. https://doi.org/10.1037/met0000337

Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, *22*(6), 359–371. https://doi.org/10.1038/s41583-021-00448-6

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 10.1016/j.jml.2012.11.001. https://doi.org/10.1016/j.jml.2012.11.001

Barth, M. (2023). *tinylabels: Lightweight variable labels*. Retrieved from https://cran.r-project.org/package=tinylabels

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, D., Maechler, M., & Jagan, M. (2024). *Matrix: Sparse and dense matrix classes and* *methods*. Retrieved from https://Matrix.R-forge.R-project.org

Bengtsson, H. (2021). A unifying framework for parallel and distributed processing in r using futures. *The R Journal*, *13*(2), 208–227. https://doi.org/10.32614/RJ-2021-048

Berger, A., & Kiefer, M. (2021). Comparison of Different Response Time Outlier Exclusion Methods: A Simulation Study. *Frontiers in Psychology*, *12*, 675558. https://doi.org/10.3389/fpsyg.2021.675558

Blossfeld, H.-P., & Rohwer, G. (2002). *Techniques of event history modeling: New approaches to causal analysis, 2nd ed* (pp. x, 310). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Bloxom, B. (1984). Estimating response time hazard functions: An exposition and extension. *Journal of Mathematical Psychology*, *28*(4), 401–420. https://doi.org/10.1016/0022-2496(84)90008-7

Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General*, *148*(4), 601–618. https://doi.org/10.1037/xge0000558

Box-Steffensmeier, J. M. (2004). Event history modeling: A guide for social scientists. Cambridge: University Press.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920965119. https://doi.org/10.1177/2515245920965119

Eddelbuettel, D., & Balamuta, J. J. (2018). Extending R with C++: A Brief Introduction to Rcpp. *The American Statistician*, *72*(1), 28–36. https://doi.org/10.1080/00031305.2017.1375990

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

Elmer, T., Van Duijn, M. A. J., Ram, N., & Bringmann, L. F. (2023). Modeling categorical time-to-event data: The example of social interaction dynamics captured with event-contingent experience sampling methods. *Psychological Methods*. https://doi.org/10.1037/met0000598

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). *Cmdstanr: R interface to 'CmdStan'*. Retrieved from https://github.com/stan-dev/cmdstanr

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in bayesian workflow. *J. R. Stat. Soc. A*, *182*, 389–402. https://doi.org/10.1111/rssa.12378

Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and Other Stories. https://www.cambridge.org/highereducation/books/regression-and-other-stories/DD20DD6C9057118581076E54E40C372C; Cambridge University Press. https://doi.org/10.1017/9781139161879

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., … Modrák, M. (2020). *Bayesian Workflow.* arXiv. https://doi.org/10.48550/arXiv.2011.01808

Girard, J. (2024). *Standist: What the package does (one line, title case)*. Retrieved from https://github.com/jmgirard/standist

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from https://www.jstatsoft.org/v40/i03/

Heiss, A. (2021, November 10). A Guide to Correctly Calculating Posterior Predictions and Average Marginal Effects with Multilievel Bayesian Models.

https://doi.org/10.59350/wbn93-edb02

Holden, J. G., Van Orden, G. C., & Turvey, M. T. (2009). Dispersion of response times reveals cognitive dynamics. *Psychological Review*, *116*(2), 318–342. https://doi.org/10.1037/a0014849

Hosmer, D. W., Lemeshow, S., & May, S. (2011). *Applied Survival Analysis: Regression Modeling of Time to Event Data* (2nd ed). Hoboken: John Wiley & Sons.

Kantowitz, B. H., & Pachella, R. G. (2021). The Interpretation of Reaction Time in Information-Processing Research 1. *Human Information Processing*, 41–82. https://doi.org/10.4324/9781003176688-2

Kay, M. (2024). *tidybayes: Tidy data and geoms for Bayesian models.* https://doi.org/10.5281/zenodo.1308151

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kurz, A. S. (2023a). *Applied longitudinal data analysis in brms and the tidyverse* (version 0.0.3). Retrieved from https://bookdown.org/content/4253/

Kurz, A. S. (2023b). *Statistical rethinking with brms, ggplot2, and the tidyverse: Second edition* (version 0.4.0). Retrieved from https://bookdown.org/content/4857/

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Landes, J., Engelhardt, S. C., & Pelletier, F. (2020). An introduction to event history analyses for ecologists. *Ecosphere*, *11*(10), e03238. https://doi.org/10.1002/ecs2.3238

Lougheed, J. P., Benson, L., Cole, P. M., & Ram, N. (2019). Multilevel survival analysis: Studying the timing of children's recurring behaviors. *Developmental Psychology*, *55*(1), 53–65. https://doi.org/10.1037/dev0000619

Luce, R. D. (1991). *Response times: Their role in inferring elementary mental organization*

(1. issued as paperback). Oxford: Univ. Press.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN* (2nd ed.). New York: Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

Mills, M. (2011). *Introducing Survival and Event History Analysis*. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd. https://doi.org/10.4135/9781446268360

Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Nelder, J. A. (1999). From Statistics to Statistical Science. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *48*(2), 257–269. Retrieved from https://www.jstor.org/stable/2681191

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer palettes*. Retrieved from https://CRAN.R-project.org/package=RColorBrewer

Panis, S. (2020). How can we learn what attention is? Response gating via multiple direct routes kept in check by inhibitory control processes. *Open Psychology*, *2*(1), 238–279. https://doi.org/10.1515/psych-2020-0107

Panis, S., Moran, R., Wolkersdorfer, M. P., & Schmidt, T. (2020). Studying the dynamics of visual search behavior using RT hazard and micro-level speed–accuracy tradeoff functions: A role for recurrent object recognition and cognitive control processes. *Attention, Perception, & Psychophysics*, *82*(2), 689–714. https://doi.org/10.3758/s13414-019-01897-z

Panis, S., Schmidt, F., Wolkersdorfer, M. P., & Schmidt, T. (2020). Analyzing Response Times and Other Types of Time-to-Event Data Using Event History Analysis: A Tool for Mental Chronometry and Cognitive Psychophysiology. *I-Perception*, *11*(6), 2041669520978673. https://doi.org/10.1177/2041669520978673

Panis, S., & Schmidt, T. (2016). What Is Shaping RT and Accuracy Distributions? Active and Selective Response Inhibition Causes the Negative Compatibility Effect. *Journal of Cognitive Neuroscience*, *28*(11), 1651–1671. https://doi.org/10.1162/jocn_a_00998

Panis, S., & Schmidt, T. (2022). When does "inhibition of return" occur in spatial cueing tasks? Temporally disentangling multiple cue-triggered effects using response history and conditional accuracy analyses. *Open Psychology*, *4*(1), 84–114. https://doi.org/10.1515/psych-2022-0005

Panis, S., Torfs, K., Gillebert, C. R., Wagemans, J., & Humphreys, G. W. (2017). Neuropsychological evidence for the temporal dynamics of category-specific naming. *Visual Cognition*, *25*(1-3), 79–99. https://doi.org/10.1080/13506285.2017.1330790

Panis, S., & Wagemans, J. (2009). Time-course contingencies in perceptual organization and identification of fragmented object outlines. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 661–687. https://doi.org/10.1037/a0013547

Pargent, F., Koch, T. K., Kleine, A.-K., Lermer, E., & Gaube, S. (2024). A Tutorial on Tailored Simulation-Based Sample-Size Planning for Experimental Designs With Generalized Linear Mixed Models. *Advances in Methods and Practices in Psychological Science*, *7*(4), 25152459241287132. https://doi.org/10.1177/25152459241287132

Pedersen, T. L. (2024). *Patchwork: The composer of plots*. Retrieved from https://patchwork.data-imaginist.com

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-PLUS*. New York: Springer. https://doi.org/10.1007/b98882

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Schöner, G., & Spencer, J. P. (2016). *Dynamic thinking: A primer on dynamic field theory*. New York, NY: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199300563.001.0001

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford, New York: Oxford University Press.

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Stan Development Team. (2020). *StanHeaders: Headers for the R interface to Stan*. Retrieved from https://mc-stan.org/

Stan Development Team. (2024). *RStan: The R interface to Stan*. Retrieved from https://mc-stan.org/

Stoolmiller, M. (2015). *An Introduction to Using Multivariate Multilevel Survival Analysis to Study Coercive Family Process* (Vol. 1; T. J. Dishion & J. Snyder, Eds.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199324552.013.27

Stoolmiller, M., & Snyder, J. (2006). Modeling heterogeneity in social interaction processes using multilevel survival analysis. *Psychological Methods*, *11*(2), 164–177. https://doi.org/10.1037/1082-989X.11.2.164

Teachman, J. D. (1983). Analyzing social processes: Life tables and proportional hazards models. *Social Science Research*, *12*(3), 263–301. https://doi.org/10.1016/0049-089X(83)90015-7

Tong, C. (2019). Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science. *The American Statistician*, *73*(sup1), 246–261. https://doi.org/10.1080/00031305.2018.1518264

Townsend, J. T. (1990). Truth and consequences of ordinal differences in statistical distributions: Toward a theory of hierarchical inference. *Psychological Bulletin*, *108*(3), 551–567. https://doi.org/10.1037/0033-2909.108.3.551

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*(3), 424–465. https://doi.org/10.3758/BF03214357

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics.

Acta Psychologica, *41*(1), 67–85. https://doi.org/10.1016/0001-6918(77)90012-9

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors).* Retrieved from https://forcats.tidyverse.org/

Wickham, H. (2023b). *Stringr: Simple, consistent wrappers for common string operations.* Retrieved from https://stringr.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation.* Retrieved from https://dplyr.tidyverse.org

Wickham, H., & Henry, L. (2023). *Purrr: Functional programming tools.* Retrieved from https://purrr.tidyverse.org/

Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data.* Retrieved from https://readr.tidyverse.org

Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data.* Retrieved from https://tidyr.tidyverse.org

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R.* New York: Routledge. https://doi.org/10.4324/9781315165547

Wolkersdorfer, M. P., Panis, S., & Schmidt, T. (2020). Temporal dynamics of sequential motor activation in a dual-prime paradigm: Insights from conditional accuracy and hazard functions. *Attention, Perception, & Psychophysics*, *82*(5), 2581–2602. https://doi.org/10.3758/s13414-020-02010-5