Tutorial

Advances in Methods and Practices in Psychological Science
July-September 2026, Vol. 9, No. 3, pp. 1–24

DOI: 10.1177/25152459261416470
www.psychologicalscience.org/AMPPS

Event-History Analysis for Psychological Time-to-Event Data: A Tutorial in R With Examples in Bayesian and Frequentist Workflows

Sven Panis[1] and Richard Ramsey[1,2]
[1]Neural Control of Movement Lab, Department of Health Sciences and Technology, ETH Zürich, Zürich, Switzerland, and [2]Social Brain Sciences Lab, Department of Humanities, Social and Political Sciences, ETH Zürich, Zürich, Switzerland

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 the 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 over 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. Well-established analytical approaches, such as event-history analysis (EHA), can 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 article, we outline a set of tutorials that detail one distributional method, known as discrete-time EHA. We touch on 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 and its limitations. 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, open materials

Received 12/17/24; revision accepted 12/8/25

In experimental psychology, it is standard practice to analyze 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 and the understanding of the structure of cognition and brain function. Indeed, the view that mean values represent truth and variations around the mean are error is deeply ingrained in experimental psychology (Bolger et al., 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 et al., 2017; Panis, Moran, et al., 2020; Panis & Schmidt, 2016, 2022; Panis & Wagemans, 2009; Wolkersdorfer et al., 2020).

As a simple illustration, Figure 1 summarizes simulated data for one subject that shows how comparing

Corresponding Author:
Sven Panis, Neural Control of Movement Lab, Department of Health Sciences and Technology, ETH Zürich, Zürich, Switzerland
Email: sven.panis@hest.ethz.ch

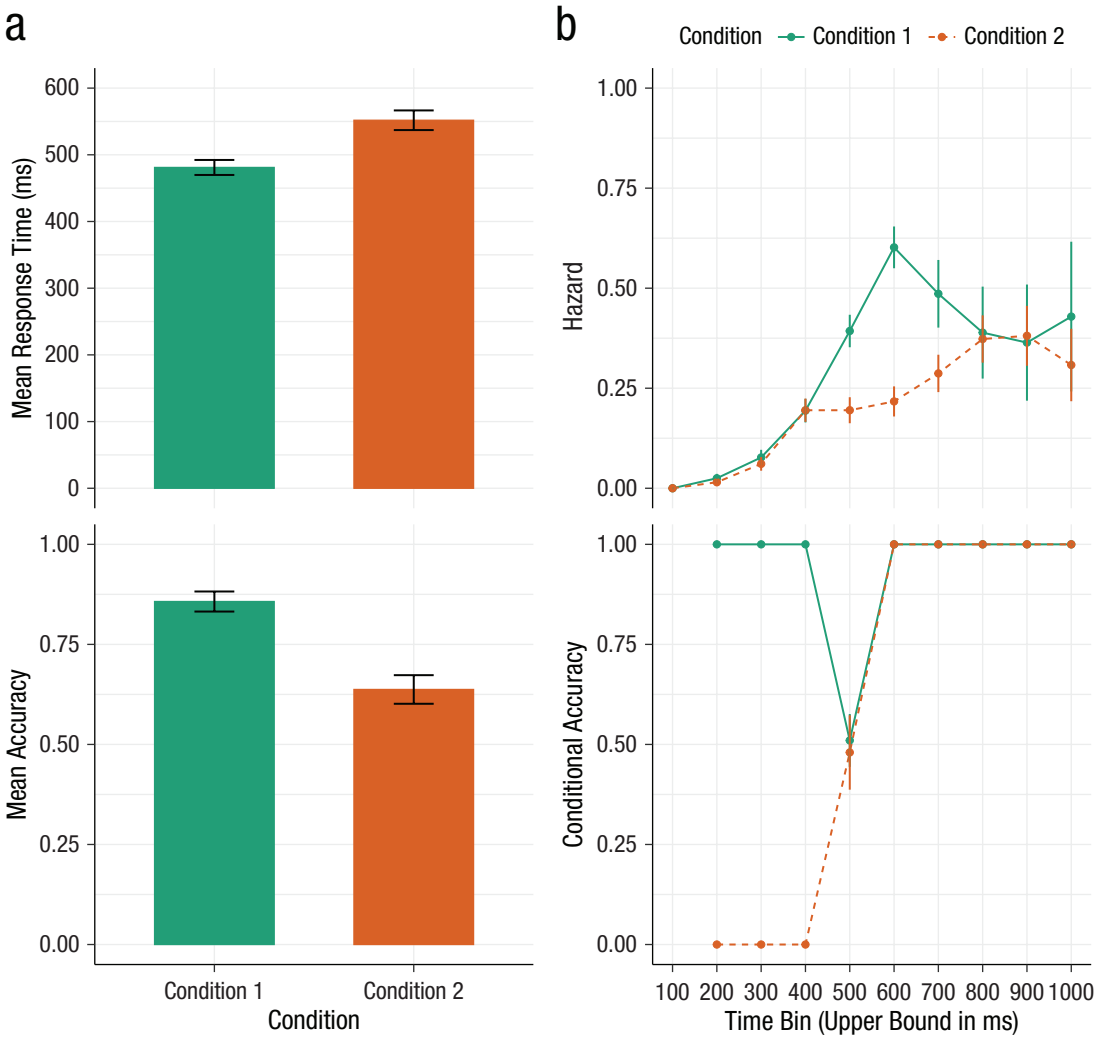

Fig. 1. Simulated single-subject data showing mean performance versus a distributional analysis. (a) The (top) mean reaction time and (bottom) overall accuracy for two conditions are plotted. Two hundred trials are simulated in each condition. (b) For the same data, we perform a descriptive discrete-time event-history analysis by plotting (top) the discrete-time hazard functions and a speed-accuracy trade-off analysis by plotting the (bottom) conditional-accuracy functions. The first second after target stimulus onset (time zero) is divided in 10 time bins of 100 ms (indexed by t = 1–10). Hazard for bin t is the conditional probability of response occurrence in bin t given no response in any of the earlier bins. Conditional accuracy for bin t is the conditional probability of a correct response given response occurrence in bin t. The hazard and conditional-accuracy estimates are plotted at the upper bound of each time bin. For the mathematical definitions of discrete-time hazard and conditional accuracy, see Section B of the Supplemental Material available online. Error bars represent ±1 *SE* of the (a) mean or (b) proportion.

means between two conditions conceals the shapes of the underlying RT and accuracy distributions. Furthermore, compared with the aggregation of data across trials (Fig. 1a), a distributional approach offers the possibility to reveal the temporal organization of psychological states (Fig. 1b). The distributional method we apply is known as discrete-time event-history analysis (EHA), which is extended with a speed-accuracy trade-off (SAT) analysis.

For example, Figure 1b shows a first state (up to 400 ms after target onset) for which the early upswing in the hazard functions of response occurrence 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 (400–500 ms), the hazard of response occurrence 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.[1] Importantly from a face-validity perspective, this pattern of simulated data can be seen in the experimental-psychology literature (Panis, 2020; Panis & Schmidt, 2016, 2022), as illustrated in Figure S6 in the Supplemental Material available online.

Why does this matter for research in psychology? First, comparing the shapes of the RT distributions between experimental conditions can be theoretically meaningful by leading to more fine-grained understanding of the time course of psychological processes. Second, extracting 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, et al., 2020). Third, EHA adds a relatively underused but ever-present dimension—the passage of time—to the theory-building tool kit and thus constitutes one possible response to the recent call for a temporal science of behavior (Abney et al., 2025).

Aims

Our aim in this article is twofold. First, we want to inform readers of the many benefits of using EHA when dealing with psychological RT data and other types of time-to-event data. Second, we want to provide a set of practical tutorials, which include step-by-step instructions on how you actually perform a (repeated event) discrete-time EHA on RT data and a complementary SAT analysis on timed accuracy data in case of choice RT data (Fig. 1b).

Even though EHA is a widely used statistical tool and there already exist many excellent textbooks (Allison, 1982; Blossfeld & Rohwer, 2002; Box-Steffensmeier, 2004; Hosmer et al., 2011; Mills, 2011; Singer & Willett, 2003; Teachman, 1983) and tutorials (Allison, 2010; Elmer et al., 2025; Landes et al., 2020; Lougheed et al., 2019; Stoolmiller, 2015; Stoolmiller & Snyder, 2006), we are not aware of any tutorials that are aimed specifically at psychological RT (+ accuracy) data and that provide worked examples of the key data-processing and Bayesian multilevel-regression-modeling steps.

Set within this context, our overall aim is to introduce a set of tutorials that explain how to do such analyses in the context of experimental psychology rather than repeat in any detail why you may do them. For more background context on EHA, however, see the Research Questions section and Section I of the Supplemental Material. 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.

Structure

The article is organized in four main sections. In the following section, we provide a brief overview of EHA, including basic concepts, modeling approaches, and the new types of research questions that EHA allows to answer about behavioral dynamics. Then, 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 discrete-time EHA to their own time-to-event data measured in RT tasks. Next, we discuss the overall strengths and weaknesses of the approach for researchers in experimental psychology. In the final section, we provide a glossary containing all key terms.

What Is EHA, and Why Is It Relevant to Research in Experimental Psychology?

In this section, we want to provide an intuition regarding how EHA works in general and in the context of experimental psychology. We assume readers are familiar with regression as a statistical-modeling technique. For individuals who want more detailed treatment of regression, we refer the reader to several excellent textbooks (Gelman, Hill, & Vehtari, 2020; Winter, 2019). To make the article and tutorials more widely accessible, we specifically aimed to keep this article equation-light. Therefore, we provide additional information and mathematical definitions in the Supplemental Material (for regression equations, see Section G in the Supplemental Material).

A brief introduction to EHA

EHA is a class of statistical approaches to study the occurrence and timing of events, such as death, disease onset, marriages, arrests, and job terminations (Allison, 2010). EHA is known by various labels, including "survival analysis", "hazard analysis", "duration analysis", "failure-time analysis", and "transition analysis" (Singer & Willett, 2003). In this article, we choose to use the term "EHA" throughout. An event is defined as a qualitative change—a transition from one discrete state to another—that can be situated in time. Typical events of interest in cognitive science include manual button presses, vocal-response onsets, saccadic eye-movement onsets, fixation offsets, and spatial boundary crossings

when measuring movement trajectories, for example, in mouse-tracking studies (Schoemann et al., 2021; Spivey & Dale, 2006). Although some events can happen only once (e.g., death, chronic-disease onset), most events of interest for cognitive scientists can happen repeatedly to the same individual.

EHA concepts applied to RT tasks. To apply EHA, one also must define time-point zero (e.g., imperative stimulus onset) and measure the passage of time between time-point zero and event occurrence in discrete or continuous time units. In RT tasks, RTs are typically measured with a temporal resolution of 1 ms or higher. Because 1 ms is typically short compared with the range of empirical RT distributions (e.g., a half second), RTs are generally treated as continuous time-to-event data.

Risk episodes and observation period. In standard applications, each individual can experience the event only once. In experimental psychology and cognitive neuroscience, however, events are typically repeatedly measured across experimental trials in which each individual completes multiple independent trials and each trial yields one time-to-event observation. For such repeated measurements or trial-based events, the observation period refers to the entire experimental session during which multiple independent trials are completed. Within each trial, as soon as the timer starts, the individual becomes "at risk" for event occurrence and stays at risk for event occurrence as long as the participant is eligible to experience the event. Thus, the participant is no longer at risk when the event occurs or the response deadline is reached. The stretch of time within a trial during which the participant is at risk for event occurrence is called the "risk episode." In Section A of the Supplemental Material, we visualize various types of time-to-event data sets obtained in detection, discrimination, and bistable perception tasks.

Censoring. Censoring is a situation in which the exact time to an event is unknown for some trials, resulting in incomplete information. 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, that is, 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 nonresponses as missing data, EHA treats trials without a response as "right-censored" observations on the variable RT because all that is known 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. The fact that EHA can deal with right censoring, therefore, presents an analytical strength of the approach compared with many common approaches in experimental psychology (e.g., analysis of variance, linear regression, delta plots). Other types of censoring include left censoring and interval censoring.

Competing risks. In many applications, there are multiple events of interest, such as correct versus error responses, for example. In EHA parlance, this is known as a "competing-risks" situation because from time zero onward, both correct and incorrect responses are competing with another in each trial until one occurs that determines the event type and latency that are recorded. In other words, in each trial, the participant can transition from an idle state to either a "correct-response" state or an "incorrect-response" state. There are several approaches to deal with a competing-risks situation (see Section A of the Supplemental Material). Here, we estimate and model the hazard of (any) response occurrence and extend it with an SAT analysis (Heitz, 2014) by plotting and modeling the conditional-accuracy function (see Fig. 1b and the next section).

Modeling approaches. There are several different modeling approaches available to analyze time-to-event data. The type of models employed and the definition of hazard depend on whether one is using continuous- or discrete-time units. Even though RT is typically treated as a continuous variable, here, we focus on discrete-time methods that can be applied after grouping RTs into time periods (i.e., time intervals or time bins). Discrete-time methods are particularly useful when (a) the shape of the hazard function is unknown and the analyst wants to estimate it flexibly without parametric assumptions and (b) the analyst is interested in examining and interpreting the shape of the baseline hazard function. In Section D of the Supplemental Material, we provide some examples to justify our choice of discrete-time modeling. In addition, because continuous forms of EHA require a lot of data to reliably estimate the shape of the continuous-time hazard (rate) function (Bloxom, 1984; Luce, 1991; Van Zandt, 2000), discrete-time methods seem ideal for dealing with psychological RT data sets, for which there are typically fewer than ≈200 trials per condition per participant. Thus, discrete-time methods allow a suitable trade-off between temporal resolution (i.e., bin width) and the number of repeated measurements per condition per participant (Panis, Schmidt, et al., 2020). Moreover, as indicated by Singer and Willett (2003), learning discrete-time methods first will help in learning continuous-time methods, so it seems like a good entry point into EHA. Note that there are no strict cutoffs for when time-to-event data should be treated as discrete versus continuous.

To apply discrete-time EHA, one divides the risk episode in each trial in discrete, contiguous time bins

indexed by t (e.g., t = 1–10; Fig. 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. For example, with bin widths of 100 ms, 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 (Fig. 1b, upper panel). The discrete-time hazard function gives you for each time bin the conditional probability that the response occurs (sometime) in bin t given that the response does not occur in previous bins. In other words, because the waiting time in a trial is increasing incrementally in small steps, hazard reflects the risk that the response occurs in the current bin t given that it has not yet occurred in the past, that is, 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 (Fig. 1b, lower panel). 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). This type of conditional-accuracy function is also known as the micro-level SAT function. We refer to this extended (hazard + conditional accuracy) analysis for choice RT data as EHA/SAT. For the mathematical definitions of both these and other discrete-time functions, see Section B in the Supplemental Material.

Research questions. EHA/SAT enables a range of new research questions to be addressed, which are not possible with a more conventional comparison of means analysis, such as:

- What is the shape of the hazard function of response occurrence? In other words, what is the effect of the passage of time on the within-trials timescale on the hazard estimate? Does the hazard function increase to a plateau, or does it increase to a peak and then decrease? Is it flat in the right tail (Panis, Moran, et al., 2020)?
- Does an experimental manipulation affect the early, middle, and/or late part of the distribution? Does an experimental effect increase, decrease, or remain stable over time within a trial? Is an experimental effect restricted only to a short time segment within a trial? What is the onset time of an effect?
- Can one identify several successive cognitive states when considering the hazard and conditional-accuracy functions together (see Fig. 1b and Fig. S6 in the Supplemental Material)?
- Do timescales interact? In other words, is the shape of the hazard (or conditional accuracy) function within a trial changing as time passes by on longer timescales? For example, some effects might increase or decrease in size and disappear over the course of an experiment (the across-trials timescale). Another example is that people typically speed up during the course of an experiment, but is this reflected in hazard changes in all time periods or only in the left or right tail of the distribution?
- Do predictor variables interact in their effect on hazard? Do timescales interact with predictor variables? Is the effect of a continuous predictor on hazard linear or nonlinear?
- Are experimental effects time-locked to certain stimuli (Panis & Schmidt, 2016)?
- Are there interindividual differences in the shape of the hazard function (Panis, Moran, et al., 2020)?
- How and when do time-varying predictors (e.g., eye-gaze position, eye dilation, heart rate, parietal-frontal synchronization) affect the risk of event occurrence?

Benefits of EHA 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 et al., 2009; Luce, 1991; Townsend, 1990), as explained in Section I of the Supplemental Material. Here, we highlight three benefits that we think are relevant to the domain of experimental psychology.

First, as illustrated in Figure 1, compared with 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 cognitive states and theoretical development. Thus, the approach permits different kinds of questions to be asked and 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 electroencephalography 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, the approach is generalizable and applicable to many tasks that are commonly used in experimental psychology, such as detection, comparison, discrimination, recognition, 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 RT domain of experimental psychology irrespective of the analyst's current view on the nature of cognition (Barack & Krakauer, 2021).

Third, in contrast to orthodox approaches, EHA can deal with right-censored observations when estimating the hazard function of event occurrence. Although a few right-censored observations can be expected in any RT task, a lot of right-censored observations are expected in studies on visual masking, the attentional blink, and so on, that is, in studies in which RT is typically not analyzed. EHA thus has a wider scope than orthodox methods.

Implications for research in experimental psychology

We consider three implications that researchers will need to consider when using EHA. First, because EHA deals with right-censored observations, one can design a fixed response deadline in each trial. This will increase time efficiency because one does not need to wait for very long RTs that would be trimmed anyway.

Second, because the number of trials per condition are spread across time 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-subjects designs in which a relatively 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 from both the number of participants and 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. The optimal bin width generally depends 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 before fitting models will require testing a number of options when calculating and plotting the descriptive statistics (see Tutorial 1a). The goal is to find an appropriate 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 fewer. Careful consideration of bin width is important because overly small bin widths will result in erratically shaped hazard functions, and some bins will have no events. In contrast, overly large bin widths will result in a loss of temporal precision (see Fig. S3 in the Supplemental Material). Note, however, that time bins do not need to have the same width. For example, Panis (2020) used larger bins toward the end of each trial because fewer events occur there.

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, to deal with unobserved heterogeneity. 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. In addition, to further simplify the process for other users, the first two tutorials rely on a set of our own custom functions that make subprocesses easier to automate, such as data-wrangling and -plotting functions (for a list of the custom functions, see Section E of the Supplemental Material).

The content of the tutorials, in terms of EHA and multilevel regression modeling, 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)[2] for all reported analyses.

Tutorial 1a: calculating descriptive statistics using a life table

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 (see Table 3) includes for each time bin indexed by t 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. For the mathematical definitions

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. PID = participant ID; RT = response time.

of these quantities, see 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 modeling 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 behavioral responses from participants because it represents the measured RT and accuracy per trial in an experiment. This data set is used when fitting conditional-accuracy models (Tutorials 2b and 3b).

In contrast, the person-trial-bin data (Table 2) have a different, more extended structure that indicates in which bin a response occurred, if at all, in each trial. Therefore, the person-trial-bin data generate a 0 in each bin until an event occurs and then 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) because maybe there was no response on that trial or the event occurred after the censoring time. Likewise, when the event occurs in Bin 1, there would be only 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	Time bin	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 two 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. PID = participant ID.

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 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. RT 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-prime (same direction as target) and incongruent-prime (opposite direction as target) 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; and "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 participant and 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 in milliseconds
  mutate(censored = map(data, censor,
    600, 40)) |>
  # create person-trial-bin data set
  mutate(ptb_data = map(censored, ptb))
    |>
  # create life tables without
    conditional accuracies
  mutate(lifetable = map(ptb_data,
    setup_lt)) |>
  # calculate conditional accuracies
  mutate(condacc = map(censored, calc_
    ca)) |>
  # create life tables with conditional
    accuracies
  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 an RT larger than the censoring time or trials in which no response is emitted during the follow-up 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.

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 of these bins. 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 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. In addition, to select a suited bin width for model fitting, one can test and compare various bin widths in the censor function and select an appropriate one—in terms of visualizing the effect—that is supported by the data (see Fig. S3 in the Supplemental Material).

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

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.

For example, for Participant 6, the estimated hazard values in Bin 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 5 times larger for both prime-present conditions compared with the blank-prime condition.

Furthermore, the estimated conditional-accuracy values in Bin 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 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 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 the

Table 3. The Life Table for the Blank-Prime Condition of Participant 6

Bin	Index *t*	RS	*n* events	h(*t*)	*SE*	S(*t*)	*SE*	ca(*t*)	*SE*	P(*t*)	*SE*
0	0	220	NA	NA	NA	1.00	0.00	NA	NA	0.00	0.00
40	1	220	0	0.00	0.00	1.00	0.00	NA	NA	0.00	0.00
80	2	220	0	0.00	0.00	1.00	0.00	NA	NA	0.00	0.00
120	3	220	0	0.00	0.00	1.00	0.00	NA	NA	0.00	0.00
160	4	220	0	0.00	0.00	1.00	0.00	NA	NA	0.00	0.00
200	5	220	0	0.00	0.00	1.00	0.00	NA	NA	0.00	0.00
240	6	220	0	0.00	0.00	1.00	0.00	NA	NA	0.00	0.00
280	7	220	7	0.03	0.01	0.97	0.01	0.29	0.17	0.03	0.01
320	8	213	13	0.06	0.02	0.91	0.02	0.77	0.12	0.06	0.02
360	9	200	26	0.13	0.02	0.79	0.03	0.92	0.05	0.12	0.02
400	10	174	40	0.23	0.03	0.61	0.03	1.00	0.00	0.18	0.03
440	11	134	48	0.36	0.04	0.39	0.03	0.98	0.02	0.22	0.03
480	12	86	37	0.43	0.05	0.22	0.03	1.00	0.00	0.17	0.03
520	13	49	32	0.65	0.07	0.08	0.02	1.00	0.00	0.15	0.02
560	14	17	9	0.53	0.12	0.04	0.01	1.00	0.00	0.04	0.01
600	15	8	4	0.50	0.18	0.02	0.01	1.00	0.00	0.02	0.01

Note: The column named "bin" indicates the upper bound of each time bin (in milliseconds) and includes time-point zero. At time-point zero, no events can occur, and therefore, both the discrete-time hazard h($t = 0$) and the conditional-accuracy ca($t = 0$) are undefined. Conditional accuracy is also undefined for bins without events. RS = risk set; NA = undefined; S(*t*) = survivor function; P(*t*) = probability-mass function.

participant was instructed to respond only 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, Github). 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, whereas 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 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) or conditional accuracy (1/0) in each bin of a selected time window for analysis. But first, we consider calculating the descriptive statistics for within-subjects designs with two independent variables.

Tutorial 1b: generalizing to a more complex design

So far in this article, 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 generalization 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, in Tutorial 1b, we illustrate 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. Because we use the same functional programming approach as in Tutorial 1a, we simply refer the reader to Tutorial_1b.Rmd, Github.

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).

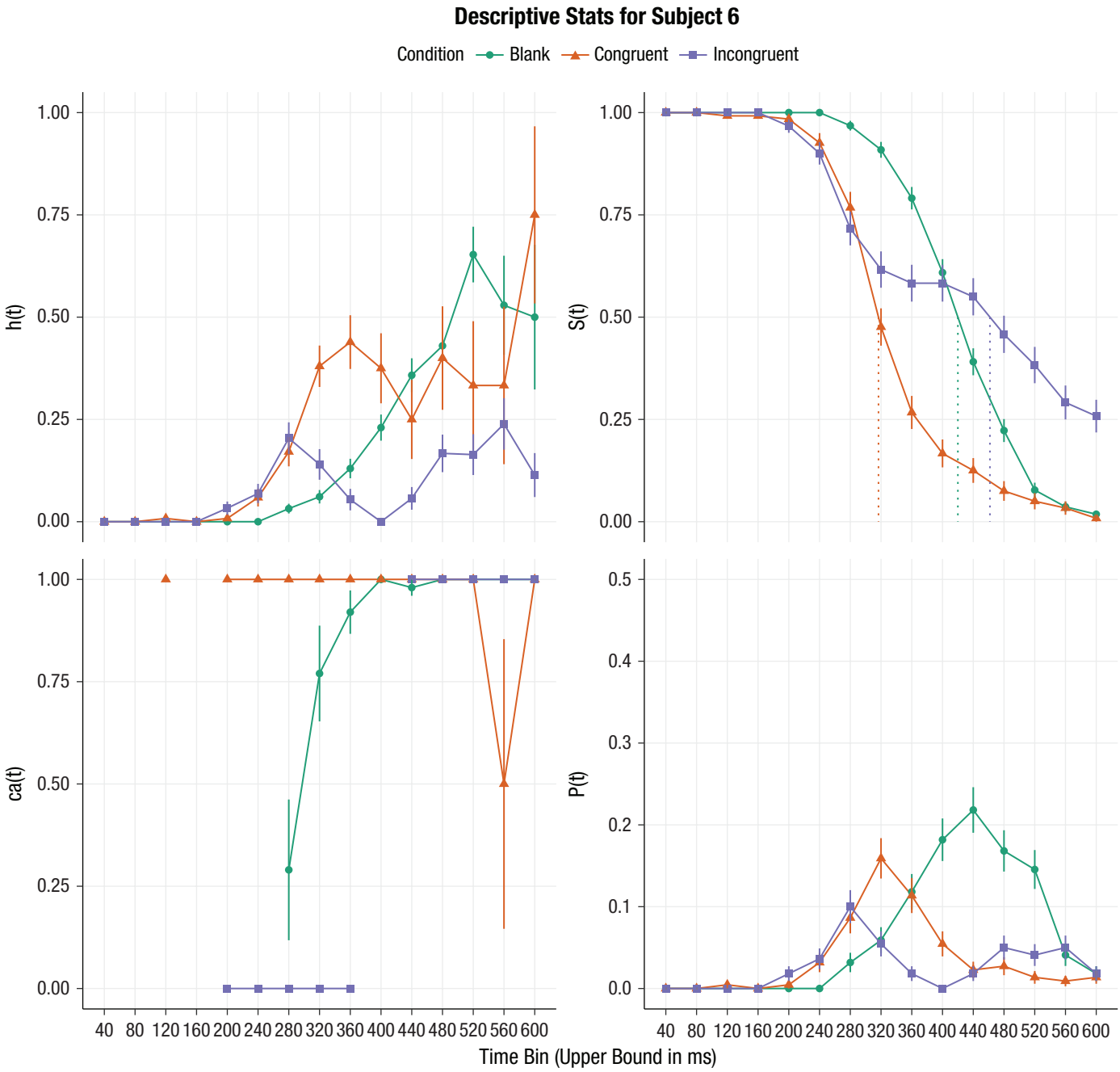

Fig. 2. Descriptive statistics. Estimated sample-based discrete-time hazard (h), survivor (S), conditional-accuracy (ca), and probability-mass (P) functions for Participant 6. Vertical dotted lines indicate the estimated median response times. Error bars represent ±*SE* of the respective proportion.

We also use a "keep it maximal" approach by specifying a full varying-effects (or random-effects) structure (Barr et al., 2013). This means that whenever possible, we include varying intercepts and slopes per participant. To make inferences, we typically use two main approaches (McElreath, 2020). We compare models of different complexity using information criteria and cross-validation to evaluate out-of-sample predictive accuracy. We also take the most complex model and evaluate key parameters of interest using point and interval estimates.

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 often contain no responses in RT tasks, one may want to select a smaller time window for analysis compared with the

follow-up period in each trial (see Section A in the Supplemental Material). Second, given that the dependent variable (event occurrence) is binary, one has to select a link function (see Section F in the Supplemental Material). The complementary log-log (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 Section G in the Supplemental Material). For example, when the shape of the empirical hazard function is smooth and/or you want to test if cloglog-hazard is changing linearly or quadratically over time, then you should treat TIME as a continuous predictor and center it on one of the bins. 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 treat TIME as a categorical predictor and use a general specification of TIME, that is, fit one grand intercept per time 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 (for an example, see Model M1d in Tutorial_2a.Rmd). 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 200-ms to 600-ms time window for inferential analysis 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
  ms, 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)))
```

Prior distributions. To get the posterior distribution of each model parameter given the data, we need to specify prior distributions for the model parameters that 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 Figure S8 in the Supplemental Material (Section H in the Supplemental Material) shows six examples of prior distributions for an intercept on the logit and/or cloglog scales. Although a normal distribution with relatively large variance is often used as a weakly informative prior for continuous dependent variables, Rows A and B of Figure S8 show that specifying such distributions on the logit and cloglog scales actually leads to rather informative distributions on the original probability scale because most mass is pushed to probabilities of 0 and 1. Thus, we modify the prior formulation to make sure that it remains consistent with a weakly informative approach.

Model M0i: a null model with index coding. 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 the following section) 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 time bin. Each of these intercepts is allowed to vary across individuals (variable pid). We request 2,000 samples from the posterior distribution for each of four chains. Estimating Model M0i took about 30 min on a MacBook Pro (Sonoma 14.6.1 OS, 18 GB memory, M3 Pro Chip).

Model M1i: adding the effects of prime-target congruency. Previous research has shown that psychological effects typically change over time (Panis, 2020; Panis et al., 2017; Panis, Moran, et al., 2020; Panis & Schmidt, 2022; 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 time-bin 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 min using the same MacBook Pro.

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 (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 (see Kurz, 2023a, Section 4.6.4; 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 = 2) |>
  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.

Evaluating parameter estimates in Model M1i. To make causal inferences from the parameter estimates in Model M1i (Frank et al., 2025), we first plot the densities of the draws from the posterior distributions of its population-level parameters and point (median) and interval estimates (80% and 95% credible intervals [CrIs]) (Fig. 3a). A CrI is a range of values that contains a parameter's true value with a specified probability given the observed data and model.

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 (Fig. 3b). Because 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 (Fig. 3c). The point estimates and quantile intervals can also be reported in a table (for details, see Tutorial_2a.Rmd).

Example conclusions for M1i. What can we conclude from Model M1i about our research question, that is, the temporal dynamics of the effect of prime-target congruency on RT? In other words, in which of the 40-ms time

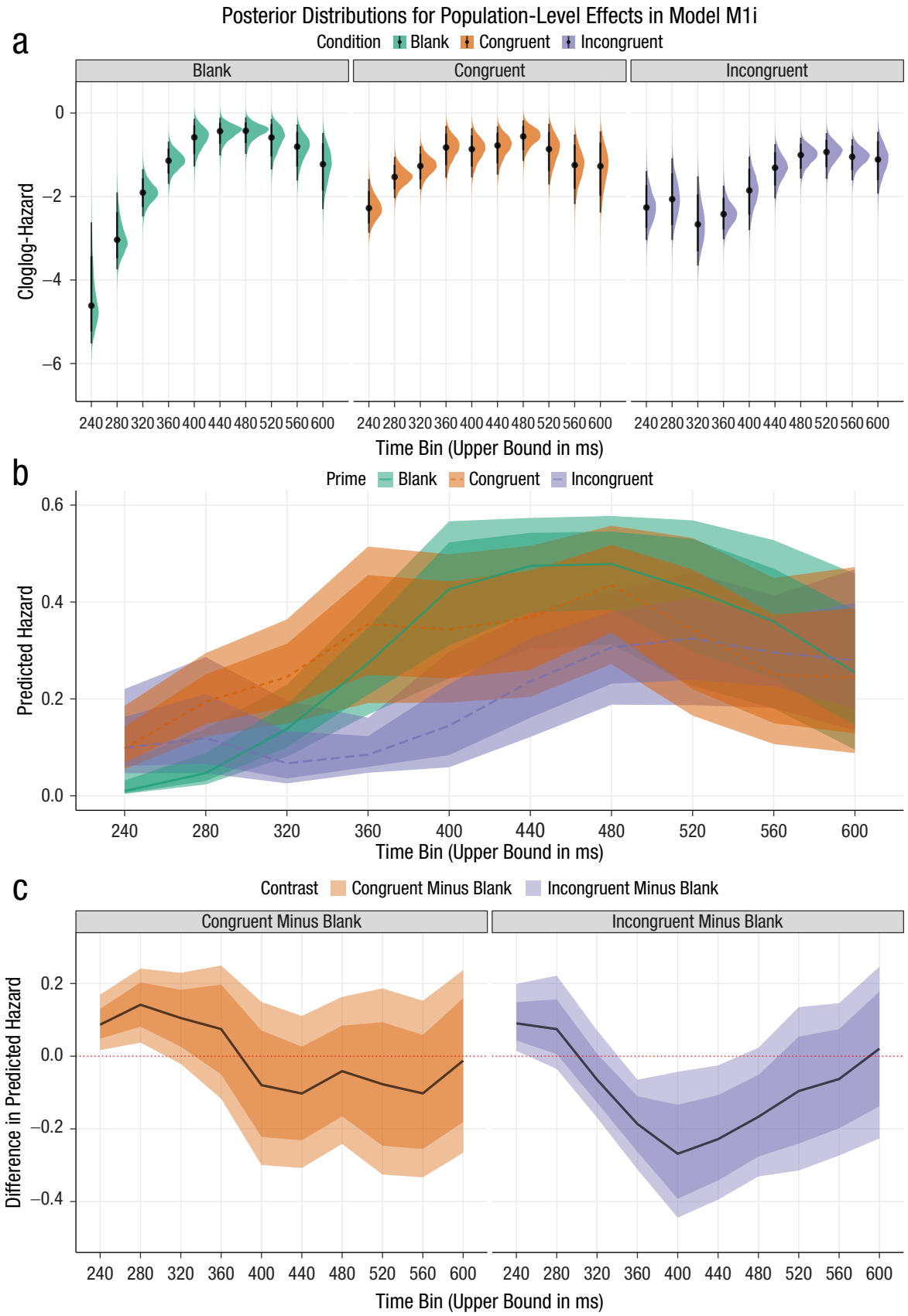

Fig. 3. Discrete-time hazard-modeling results at the population level. (a) Medians and 80%/95% credible intervals of the posterior distributions of the population-level parameters of Model M1i. (b) Point (median) and 80%/95%-credible-interval summaries of the hazard estimates (expected values of the draws from the posterior predictive distributions) in each time bin. (c) Point (mean) and 80%/95%-credible-interval summaries of estimated differences in hazard in each time bin.

bins between 200 ms 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 CrIs in Figure 3c. The contrast "congruent minus blank" was estimated to be 0.09 hazard units in Bin 240 (95% CrI = [0.02, 0.17]) and 0.14 hazard units in Bin 280 (95% CrI = [0.04, 0.25]). For the other bins, the 95% CrI contained zero. The contrast "incongruent minus blank" was estimated to be 0.09 hazard units in Bin 240 (95% CrI = [0.01, 0.21]), −0.19 hazard units in Bin 360 (95% CrI = [−0.31, −0.06]), −0.27 hazard units in Bin 400 (95% CrI = [−0.45, −0.04]), and −0.23 hazard units in Bin 440 (95% CrI = [−0.40, −0.03]). For the other bins, the 95% CrI contained zero.

There are thus two phases of performance for the average person between 200 ms 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 with blank-prime trials in Time Bin 240. In the second phase, only the incongruent prime decreases the hazard of response occurrence compared with blank primes in the time period 320 ms to 440 ms. The sign of the effect of incongruent primes on the hazard of response occurrence thus depends on how much waiting time has passed since target onset. Future modeling efforts could incorporate the trial number into the model formula to also study how the effects of prime type on hazard change on the longer experiment-wide timescale next to the short trial-wide timescale. In Tutorial_2a.Rmd, we provide a number of model formulae that should get you going.

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 fit only one conditional-accuracy model, which was based on Model M1i from Tutorial 2a and labeled "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 4a and point (median) and interval estimates (80% and 95% CrIs) (Fig. 4a).

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 accuracies—at the population level (Fig. 4b). Because 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 (Fig. 4c). Based on Figure 4c, we see that on the population level, congruent primes have a positive effect on the conditional accuracy of emitted responses in Time Bins 240, 280, 320, and 360 relative to the estimates in the baseline condition (blank prime; red dashed lines in Fig. 4c). Incongruent primes have a negative effect

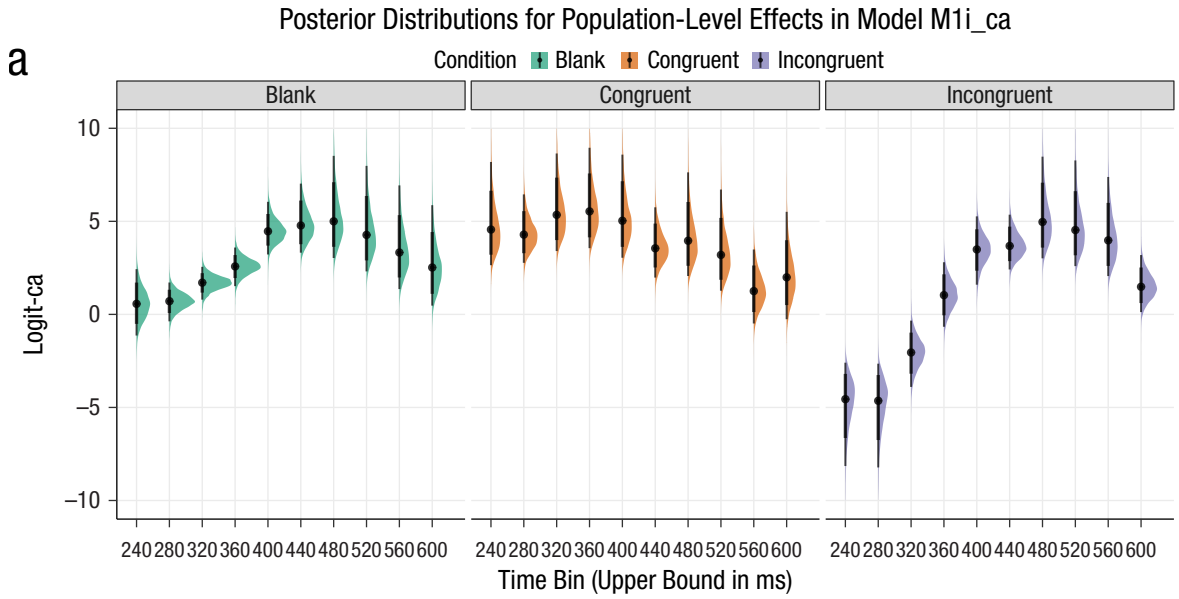

Fig. 4. *(continued on next page)*

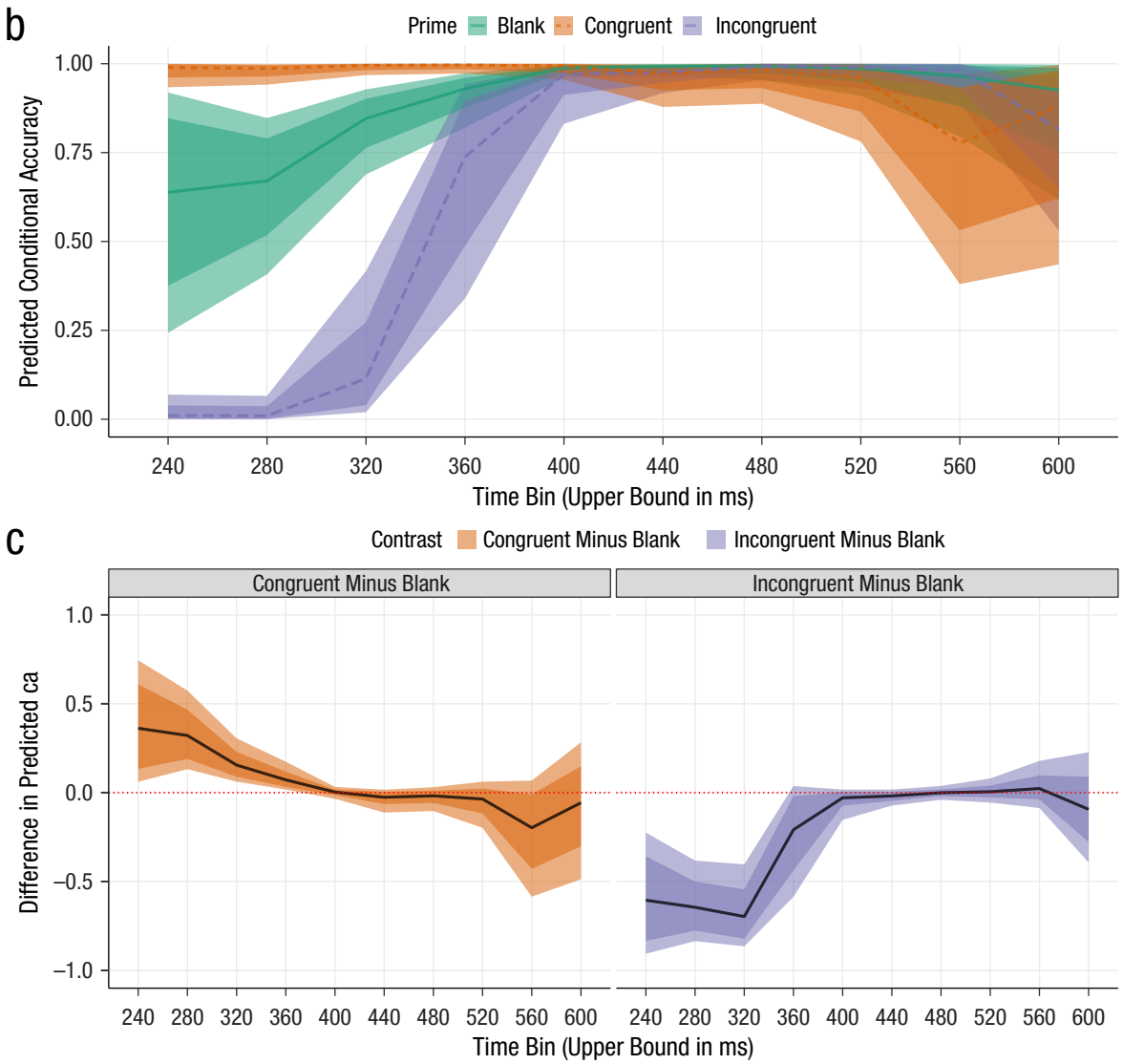

Fig. 4. Conditional-accuracy-modeling results at the population level. (a) Medians and 80/95% credible intervals of the posterior distributions of the population-level parameters of Model M1i_ca. (b) Point (median) and 80/95%-credible-interval summaries of the conditional accuracy (ca) estimates (expected values of the draws from the posterior predictive distributions) in each time bin. (c) Point (mean) and 80/95%-credible-interval summaries of estimated differences in conditional accuracy in each time bin.

on the conditional accuracy of emitted responses in the first three time bins relative to blank primes.

Finally, because many researchers will be more familiar with frequentist statistics, we also provide code to fit hazard and conditional-accuracy models in the frequentist framework in Tutorial_3a.Rmd and Tutorial_3b.Rmd using the R package *lme4* (Bates et al., 2015).

Tutorial 4: planning

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

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 them (Gelman, Vehtari, et al., 2020). The basic structure and code follows the examples outlined by Solomon Kurz (2019) in his "power" blog posts, Lisa Debruine's (n.d.) R package *faux*, and related articles from DeBruine and Barr (2021) and Pargent et al. (2024).

Basic workflow. The basic workflow is as follows: (a) fit a regression model to existing data, (b) use the regression-model parameters to simulate new data, (c) write a function

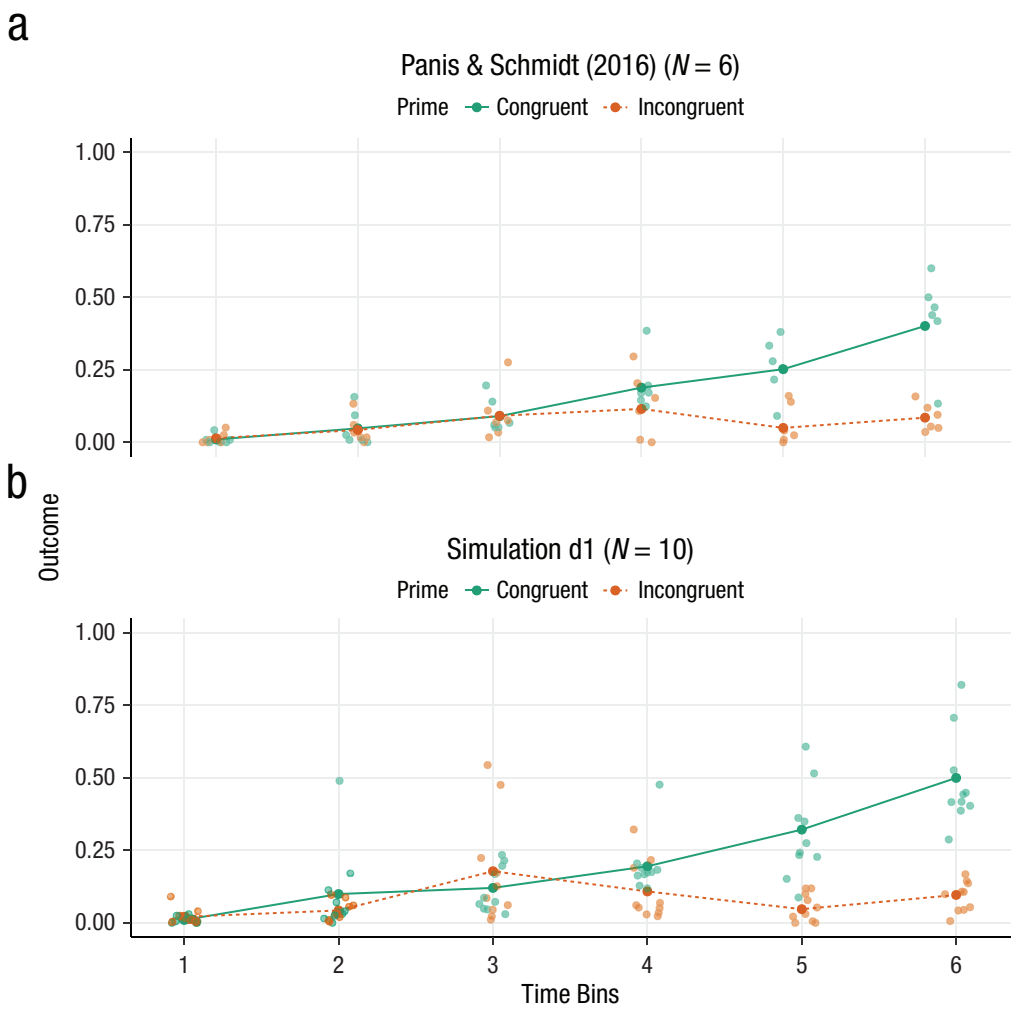

Fig. 5. (a) Raw data from Panis and Schmidt (2016). (b) Simulated data for 10 participants.

to create thousands of data sets and vary parameters of interest (e.g., sample size, trial count, effect size), and (d) summarize 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 data set and summarize the model output rather than the raw data. However, when each model takes several hours to build and we may want to simulate many thousands of data sets, it can be computationally demanding for desktop machines. So for ease, here, we just use the raw simulated data sets to guide future expectations.

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

Fit a regression model and simulate one data set. 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 data set with 200 trials per condition for 10 individual participants. The raw data and the simulated data are plotted in Figure 5 and show quite close correspondence, which is reassuring. But this is only one data set. What we really want to do is simulate many data sets and vary parameters of interest, which is what we turn to in the next section.

Simulate and summarize data across a range of parameter values. Here, we use the same data-simulation

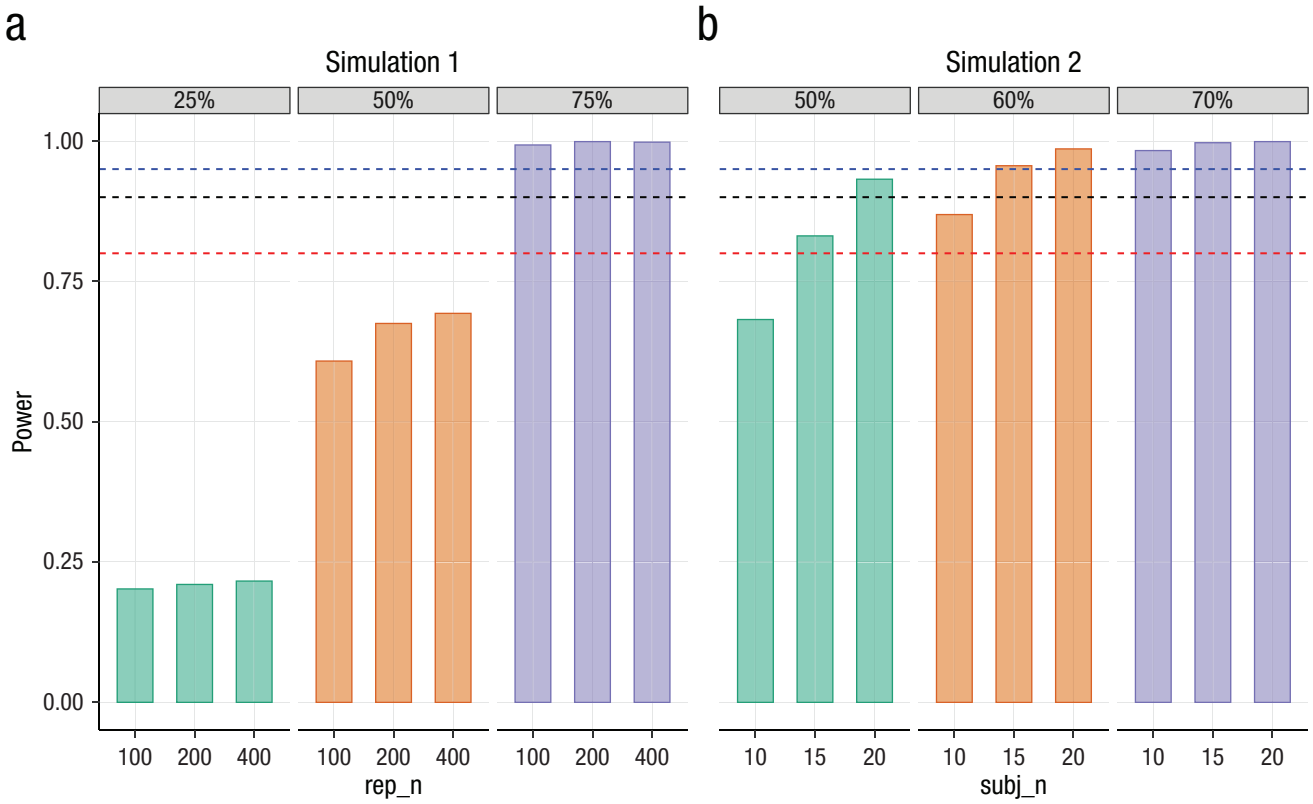

Fig. 6. Statistical power across (a) Data Simulation 1 and (b) Data Simulation 2. Power was calculated as the percentage of lower-bound 95% confidence intervals that exclude zero when the difference between prime condition is calculated (congruent – incongruent). In Simulation 1, the effect size was varied between a 25%, 50%, and 75% reduction in hazard value, and 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, and the number of participants was varied between *N*s = 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. rep_n = the number of trials per experimental condition; subj_n = the number of participants per simulated experiment.

process as used above, but instead of simulating one data set, we simulate 1,000 data sets per variation in parameter values. Specifically, in Simulation 1, we vary the number of trials per condition (100, 200, and 400) and 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 condition (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 that 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 6a. Figure 6a 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, we calculate the fraction of simulated data sets that 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 because it is familiar to most researchers in experimental psychology.

The results of Simulation 1 (Fig. 6) 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 0.5 or 0.75. 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 (Fig. 6b), we varied the effect size between a different range of values (0.5, 0.4, 0.3) that 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 6b. 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; and for a 50% reduction (0.5 hazard ratio), $N = 15$ would give over 80% power.

Planning decisions. Now that we have summarized 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 and the available resources (Lakens, 2022). There is no straightforward 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 that is weaker than previously reported, such as 0.4 or 0.5, as an initial target effect size because this embodies a sense of caution in terms of likely effect sizes (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 us over the 80% power mark. If we wanted to maximize 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 caveat, unless there are unavoidable reasons, no matter what kind of planning choices we made based on these data simulations, we would not rely solely 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.

Discussion

The main motivation for writing this article is the observation that EHA and SAT analysis remain underused in psychological research. As a consequence, the field of psychological research is not taking full advantage of the many benefits EHA/SAT provides compared with 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 with standard approaches. Although we have focused here on within-subjects, 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-subjects designs, etc.). Thus, 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, and limitations of the approach.

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 investigate a "temporal states" approach to cognitive processes by tracking behavior as a function of stepwise 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, and 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 to use for basic knowledge development and theory building.

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 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 (Fig. 1), it may be important for

researchers to supplement comparisons between means with EHA. For instance, EHA might reveal that the conclusion of interest based on averaging across trials does not apply to all responses but is instead restricted to certain periods of within-trials time.

Model complexity versus interpretability

Hazard and conditional-accuracy models can quickly become very complex when adding more than one timescale because of 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 timescales as covariates: the passage of time on the within-trials timescale and the passage of time on the across-trials (or within-experiments) timescale. However, when trials are presented in blocks and blocks of trials within sessions and when the experiment comprises a number of sessions, then four timescales can be defined (within-trials, within-blocks, within-sessions, and within-experiments). From a theoretical perspective, adding more than one timescale—and their interactions—can be important to capture plasticity and other learning effects that may play out on such longer timescales 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 are being collected per participant and condition and across the varying timescales. As one example, if there are several timescales of relevance, then it might be prudent for interpretational purposes to limit the number of experimental predictor variables (conditions). This is, of course, a case in which planning and data-simulation efforts would be important to provide a guide to experimental-design choices (see Tutorial 4 and Implications for Research in Experimental Psychology section).

Limitations

Compared with 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. Relatedly, because these models can be quite complex in terms of the number of possible parameters, more thought is required at the model-specification and model-building stages.

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 discrete-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 hr and 2 hr, 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 and 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., 2025).

Conclusions

Estimating the temporal distributions of RT and accuracy provides a rich source of information on the time course of cognitive processing, which has 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 and the development of theories regarding the structure of cognition.

Glossary

Discrete-time event-history analysis: a statistical method for analyzing the occurrence and timing of events when time is measured using (or grouped into) discrete intervals (e.g., years, months, survey waves). Commonly used when events are naturally recorded at specific measurement occasions but also applicable to continuous time-to-event data after grouping into intervals. Discrete-time methods are particularly useful when (a) the shape of the hazard function is unknown and the analyst wants to estimate it flexibly without parametric assumptions and (b) the analyst is interested in examining and interpreting the shape of the baseline hazard function (in contrast to Cox regression, which leaves the baseline hazard unspecified).

Speed-accuracy trade-off analysis: A speed-accuracy trade-off analysis plots a conditional-accuracy function to examine how an individual's accuracy changes at different speeds for specific response-time intervals, often under different conditions, such as varying deadlines or payoffs. This approach moves beyond simple average error rates to show how accuracy is not uniform across all speeds.

Discrete-time hazard function: the pattern of hazard probabilities over time, showing how the risk of event occurrence changes across time periods.

Conditional-accuracy function: a plot that shows the probability of an accurate response as a function of the response time or a specific time period.

Event: the outcome of interest whose occurrence and timing are being studied (e.g., button press, job loss, marriage, death, graduation). Events can be repeatable or nonrepeatable.

Time period (bin, interval): the discrete unit of time used in the analysis (e.g., year, month, week, second). Each observation contributes one record per time period at risk.

Hazard (hazard probability): the conditional probability that an individual will experience the event in a particular time bin or time period (of a trial or risk episode) given that the individual has not yet experienced it by the beginning of that period. Denoted as h(t) or $P(T = t \mid T \geq t)$.

Continuous-time hazard (hazard rate, instantaneous hazard): the instantaneous rate at which events occur at time point t given survival to time t. Formally defined as $\lim(\Delta t \to 0)\ P(t \leq T < t + \Delta t \mid T \geq t) / \Delta t$. Unlike discrete-time hazard (a conditional probability), continuous-time hazard is a rate that can exceed 1. Denoted as $\lambda(t)$.

Observation period: the span of time during which an individual is under observation or follow-up in the study. The observation period may encompass one or more risk episodes. For nonrepeatable events, the observation period and risk episode are typically the same. For repeatable events in natural history (e.g., multiple job changes), the observation period can contain multiple sequential risk episodes. For trial-based events, the observation period refers to the entire experimental session during which multiple independent trials (risk episodes) are completed. The observation period ends when the event occurs (for nonrepeatable events), the study ends, or the individual is lost to follow-up.

Risk episode (spell): the duration from the start of observation or from a previous event until either the event occurs or observation ends. Also called a "spell" (particularly in labor economics). A risk episode represents one complete at-risk period being studied. Within each risk episode, data may be collected at multiple discrete time points (e.g., monthly or yearly intervals)—these are the time periods or measurement occasions. In discrete-time analysis, each risk episode is typically divided into multiple person-period observations in which each person-period is the actual unit of analysis for model estimation. In repeatable-event contexts, individuals may contribute multiple risk episodes.

Repeated measurements (trial-based events): events that are repeatedly measured across experimental trials, in which each individual completes multiple independent trials and each trial yields one time-to-event observation (e.g., reaction-time tasks). Unlike recurrent events in natural history, these trials are typically designed to be independent realizations of the same process, although individual-level random effects may be needed to account for within-subjects correlation. In this context, each trial constitutes a separate risk episode, and discrete time periods within each trial represent measurement occasions.

Baseline hazard: the pattern of hazard over time when all covariates equal zero (or their reference categories). Often modeled with time dummy variables or polynomials (see Section G in the Supplemental Material).

Time-varying covariates (time-dependent covariates): predictors whose values change over time periods within the same individual's episode (e.g., employment status, gaze position, age, income).

Risk set: for nonrepeatable events with continuous-time units: the set of individuals who are eligible to experience the event at a given time point. In discrete-time analysis, the risk set at time period t consists of all individuals who have not yet experienced the (nonrepeatable) event by the beginning of period t and are still under observation. An individual exits the risk set once the individual experiences the (nonrepeatable) event or is censored. In trial-based or repeated-measurements designs with discrete time, the risk set at each discrete time period within a trial consists of all observations (trials) in which the event has not yet occurred by that time period, even when multiple trials come from the same individual.

Right censoring: occurs when an individual's observation period ends before the event occurs (e.g., study ends, lost to follow-up). Researchers know the event did not occur by time t but do not know if/when it occurs afterward.

Left censoring: occurs when the event may have occurred before the observation period began, but exactly when is not known.

Interval censoring: occurs when researchers know the event happened within a specific time interval but do not know the exact timing within that interval.

Life table: a tabular method for describing the distribution of discrete-event times, showing at least the number

at risk, number of events, hazard, and survival probability for each time period.

Competing risks: situations in which multiple types of events can occur and the occurrence of one event prevents the others from occurring (e.g., death prevents marriage, different types of job transitions).

Unobserved heterogeneity (frailty): individual-level differences in event risk that are not captured by observed covariates. Can lead to spurious duration dependence if not addressed. Commonly modeled using hierarchical or multilevel models with random effects (also called "frailty models"), which account for unobserved individual-level variation in baseline risk.

Transparency

Action Editor: David A. Sbarra
Editor: David A. Sbarra
Author Contributions

Sven Panis: Conceptualization; Software; Writing – original draft; Writing – review & editing.

Richard Ramsey: Conceptualization; Software; Supervision; Writing – review & editing.

Declaration of Conflicting Interests

The authors declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Sven Panis https://orcid.org/0000-0002-6321-583X
Richard Ramsey https://orcid.org/0000-0002-0329-2112

Acknowledgments

Ethical approval was not required for this tutorial, in which we reanalyze existing data sets. Link to public archive (containing code, manuscript, and supplemental material): https://github.com/sven-panis/Tutorial_Event_History_Analysis.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152459261416470

Notes

1. We always refer to a time bin by its upper bound. For example, Time Bin "500" in Figure 1b refers to the time period running from 400 ms to 500 ms; the lower bound of 400 ms is excluded, and the upper bound of 500 ms is included.

2. We, furthermore, used the R packages *bayesplot* (Version 1.13.0; Gabry et al., 2019), *brms* (Version 2.22.0; Bürkner, 2017, 2018, 2021), *citr* (Version 0.3.2; Aust, 2019), *cmdstanr* (Version 0.9.0.9000; Gabry et al., 2024), *dplyr* (Version 1.1.4; Wickham et al., 2023), *forcats* (Version 1.0.0; Wickham, 2023a), *futures* (Bengtsson, 2021), *ggplot2* (Version 3.5.2; Wickham, 2016), *lme4* (Version 1.1.37; Bates et al, 2015), *lubridate* (Version 1.9.4; Grolemund & Wickham, 2011), *Matrix* (Version 1.7.3; Bates et al., 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), *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).

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*, *29*(12), P1109–P1119. 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*, 61–98. https://doi.org/10.2307/270718

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

Aust, F. (2019). *Citr: 'RStudio' add-in to insert markdown citations*. 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), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barth, M. (2023). *tinylabels: Lightweight variable labels*. 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*. 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*, Article 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.). 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. (n.d.). *faux*. https://debruine.github.io/faux/

DeBruine, L. M., & Barr, D. J. (2021). Understanding mixed-effects models through data simulation. *Advances in Methods and Practices in Psychological Science*, *4*(1). 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. (2025). Modeling categorical time-to-event data: The example of social interaction dynamics captured with event-contingent experience sampling methods. *Psychological Methods*, *30*(6), 1345–1363. https://doi.org/10.1037/met0000598

Frank, M. C., Braginsky, M., Cachia, J., Coles, N. A., Hardwicke, T. E., Hawkins, R. D., Mathur, M. B., & Williams, R. (2025). *Experimentology: An open science approach to experimental psychology methods*. Stanford University. https://doi.org/10.25936/3JP6-5M50

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

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*, 389–402. https://doi.org/10.1111/rssa.12378

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press. https://doi.org/10.1017/9781139161879

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & 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)*. 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. https://www.jstatsoft.org/v40/i03/

Heiss, A. (2021). *A guide to correctly calculating posterior predictions and average marginal effects with multilievel Bayesian models*. https://doi.org/10.59350/wbn93-edb02

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, Article 150. https://doi.org/10.3389/fnins.2014.00150

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.). John Wiley & Sons.

Kantowitz, B. H., & Pachella, R. G. (2021). The interpretation of reaction time in information-processing research. In B. H. Kantowitz (Ed.), *Human information processing* (pp. 41–82). Routledge. https://doi.org/10.4324/9781003176688-2

Kay, M. (2024). *tidybayes: Tidy data and geoms for Bayesian models*. Zenodo. 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. (2019, July 18). *Bayesian power analysis: Part I. Prepare to reject '$\mathbf{H_0}$' with simulation*. https://solomonkurz.netlify.app/blog/bayesian-power-analysis-part-i/

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

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

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), Article 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), Article 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*. Oxford University Press.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and STAN* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

Mills, M. (2011). *Introducing survival and event history analysis*. Sage. https://doi.org/10.4135/9781446268360

Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. 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. https://www.jstor.org/stable/2681191

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer palettes*. 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). 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). https://doi.org/10.1177/25152459241287132

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

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

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

Schoemann, M., O'Hora, D., Dale, R., & Scherbaum, S. (2021). Using mouse cursor tracking to investigate online cognition: Preserving methodological ingenuity while moving toward reproducible science. *Psychonomic Bulletin & Review*, *28*(3), 766–787.

Schöner, G., & Spencer, J. P. (2016). *Dynamic thinking: A primer on dynamic field theory*. 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 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

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*(5), 207–211.

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

Stan Development Team. (2024). *RStan: The R interface to Stan*. 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*(Suppl. 1), 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. https://ggplot2.tidyverse.org

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

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

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse.

Journal of Open Source Software, *4*(43), Article 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*. https://dplyr.tidyverse.org

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

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

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

Winter, B. (2019). *Statistics for linguists: An introduction using R*. 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