

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

3 Sven Panis¹ & Richard Ramsey¹

4 ¹ ETH Zürich

5 Author Note

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

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

11

Abstract

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

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

33 Word count: 10102 (body) + 1742 (references) + 3442 (body supplemental material)
34 + 393 (refs suppl. mat.)

35

1. Introduction

36 1.1 Motivation and background context: Comparing means versus 37 distributional shapes

38 In experimental psychology, it is standard practice to analyse response times (RTs),
39 saccade latencies, and fixation durations by calculating average performance across a series
40 of trials. Such comparisons between means have been the workhorse of experimental
41 psychology over the last century, and have had a substantial impact on theory development
42 as well as our understanding of the structure of cognition and brain function. Indeed, the
43 view that mean values represent truth and variations around the mean are error is deeply
44 ingrained in experimental psychology (Bolger, Zee, Rossignac-Milon, & Hassin, 2019).

45 However, differences in mean RT conceal important pieces of information, such as when an
46 experimental effect starts, how it evolves with increasing waiting time, and whether its
47 onset is time-locked to other events (Panis, 2020; Panis, Moran, Wolkersdorfer, & Schmidt,
48 2020; Panis & Schmidt, 2016, 2022; Panis, Torfs, Gillebert, Wagemans, & Humphreys,
49 2017; Panis & Wagemans, 2009; Wolkersdorfer, Panis, & Schmidt, 2020). Such absolute
50 timing information is useful not only for the interpretation of experimental effects under
51 investigation, but also for cognitive psychophysiology and computational model selection
52 (Panis, Schmidt, Wolkersdorfer, & Schmidt, 2020).

53 As a simple illustration, Figure 1 summarises simulated data for one subject that
54 shows how comparing means between two conditions can conceal the shapes of the
55 underlying RT and accuracy distributions. Indeed, compared to the aggregation of data
56 across trials (Figure 1A), a distributional approach offers the possibility to reveal the time
57 course of psychological states (Figure 1B). Here we apply a distributional method known as
58 event history analysis (EHA) extended with speed-accuracy tradeoff (SAT) analysis. For
59 example, Figure 1B shows a first state (up to 400 ms after target onset) for which the early
60 upswing in hazard is equal for both conditions, and the emitted responses are always

61 correct in condition 1 and always incorrect in condition 2. In a second state (400 to 500
62 ms), hazard is higher in condition 1, and conditional accuracies are close to .5 in both
63 conditions. In a third state (>500 ms), the effect disappears in hazard, and all conditional
64 accuracies are equal to 1. Importantly from a face-validity perspective, this pattern of
65 simulated data can be seen in the experimental psychology literature (Panis & Schmidt,
66 2016).

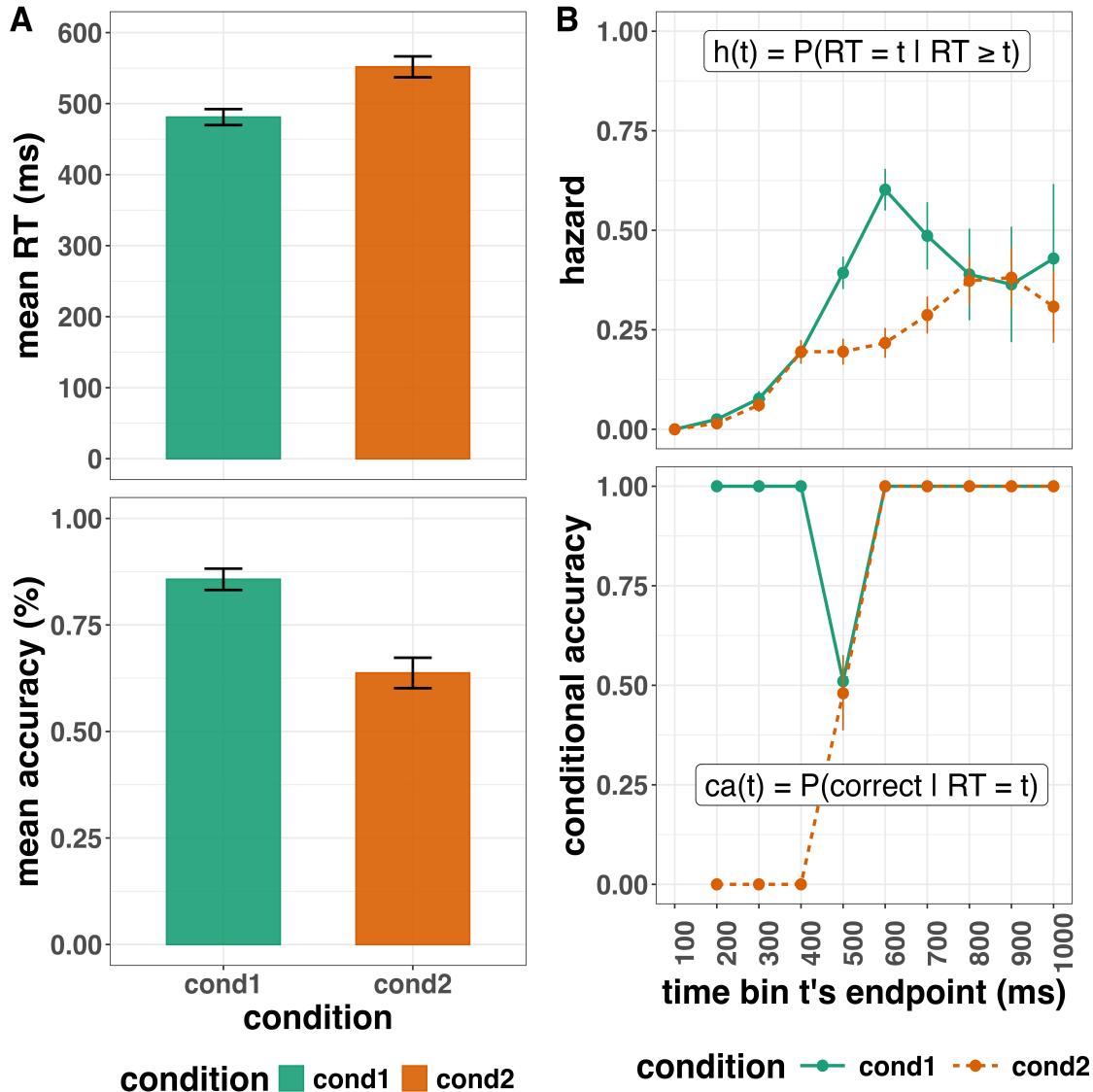


Figure 1. Simulated single-subject data showing mean performance versus a distributional analysis. (A) The mean RT (top) and overall accuracy (bottom) for two conditions are plotted. Two hundred trials are simulated in each condition. (B) The discrete-time hazard functions (top) and conditional accuracy functions (bottom) are plotted for the same data. The first second after target stimulus onset (time zero) is divided in ten time bins of 100 ms ($t = 1$ to 10). Each time bin is a half-open interval; for example, the first bin is the interval (0,100], with 0 excluded and 100 included. Note that the hazard and conditional accuracy estimates are plotted at the endpoint of each time bin. The definitions of discrete-time hazard and conditional accuracy are further explained in section 2.1.2. Error bars represent ± 1 standard error of the mean (A) or proportion (B).

67 Why does this matter for research in psychology? For many psychological questions,
 68 the estimation of such “temporal states” information can be theoretically meaningful by
 69 leading to more fine-grained understanding of psychological processes. Because EHA adds
 70 a relatively under-used but ever-present dimension – the passage of time – to the theory
 71 building toolkit, it provides one possible response to the recent call for a temporal science

72 of behavior (Abney, Fausey, Suarez-Rivera, & Tamis-LeMonda, 2025).

73 **1.2 Aims**

74 Our ultimate aim in this paper is twofold. First, we want to convince readers of the
75 many benefits of using EHA when dealing with psychological RT data. Second, we want to
76 provide a set of practical tutorials, which provide step-by-step instructions on how you
77 actually perform a (single event) discrete-time EHA on RT data, as well as a
78 complementary discrete-time SAT analysis on timed accuracy data in case of choice RT
79 data (Figure 1B).

80 Even though EHA is a widely used statistical tool and there already exist many
81 excellent reviews (Allison, 1982; Blossfeld & Rohwer, 2002; Box-Steffensmeier, 2004;
82 Hosmer, Lemeshow, & May, 2011; Mills, 2011; Singer & Willett, 2003; Teachman, 1983)
83 and tutorials (Allison, 2010; Elmer, Van Duijn, Ram, & Bringmann, 2023; Landes,
84 Engelhardt, & Pelletier, 2020; Lougheed, Benson, Cole, & Ram, 2019; Stoolmiller, 2015;
85 Stoolmiller & Snyder, 2006), we are not aware of any tutorials that are aimed specifically
86 at psychological RT (+ accuracy) data, and which provide worked examples of the key
87 data processing and Bayesian multilevel regression modelling steps.

88 Set within this context, our overall aim is to introduce a set of tutorials, which
89 explain **how** to do such analyses in the context of experimental psychology, rather than
90 repeat in any detail **why** you may do them. Therefore, we hope that our tutorials will
91 provide a pathway for research avenues in experimental psychology that have the potential
92 to benefit from using EHA in the future.

93 **1.3 Structure**

94 In what follows, the paper is organised in three main sections. In Section 2, we
95 provide a brief overview of EHA to orient the reader to the basic concepts that we will use

96 throughout the paper and why such an approach might be relevant for research in
97 experimental psychology. In Section 3, we outline a series of tutorials, which are written in
98 the R programming language and publicly available on our Github page
99 (https://github.com/sven-panis/Tutorial_Event_History_Analysis), along with all of the
100 other code and material associated with the project. The tutorials provide hands-on,
101 concrete examples of key parts of the analytical process, such as data wrangling, plotting
102 descriptive statistics, model fitting and planning future studies, so that others can apply
103 EHA to their own time-to-event data measured in RT tasks. In Section 4, we discuss the
104 strengths and weaknesses of the approach for researchers in experimental psychology.

105 **2. What is event history analysis and why is it relevant to research in
106 experimental psychology?**

107 **2.1 A brief introduction to event history analysis**

108 EHA is a class of statistical approaches to study the occurrence and timing of events,
109 such as disease onset, marriages, arrests, and job terminations (Allison, 2010). In this
110 section, we want to provide an intuition regarding how EHA works in general, as well as in
111 the context of experimental psychology. For those who want more detailed treatment of
112 EHA and/or regression equations, we refer the reader to several excellent textbooks on
113 these topics (Allison, 2010; Gelman, Hill, & Vehtari, 2020; Mills, 2011; Singer & Willett,
114 2003; Winter, 2019). We also supply relevant regression equations in section E of the
115 Supplemental Material.

116 **2.1.1 Terminology and minimum requirements for EHA.** To avoid possible
117 confusion in terminology used, it is worth noting that EHA is known by various labels,
118 such as survival analysis, hazard analysis, duration analysis, failure-time analysis, and
119 transition analysis (Singer & Willett, 2003). In this paper, we choose to use the term EHA
120 throughout.

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

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

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

133 **2.1.2 Types of EHA.** There are different types of modeling approaches in EHA.

134 For example, the definition of hazard and the type of models employed depend on whether

135 one is using continuous or discrete time units. As a lab, and mainly for practical reasons,

136 we have much more experience using discrete-time EHA, and that is the approach that we

137 describe and focus on in this paper. This choice may seem counter-intuitive, given that RT

138 is typically treated as a continuous variable. However, continuous forms of EHA require

139 much more data to reliably estimate the continuous-time hazard (rate) function (Bloxom,

140 1984; Luce, 1991; Van Zandt, 2000). Thus, by trading a bit of temporal resolution for a

141 lower number of trials, discrete-time methods seem ideal for dealing with typical

142 psychological RT data sets for which there are less than ~200 trials per condition per

143 participant (Panis, Schmidt, et al., 2020). Moreover, as indicated by Allison (2010),

144 learning discrete-time EHA methods first will help in learning continuous-time methods, so

145 it seems like a good starting point.

146 To apply discrete-time EHA, one divides the within-trial time in discrete, contiguous

147 time bins indexed by t (e.g., $t = 1$ to 10; Figure 1B). Then let RT be a discrete random

148 variable denoting the rank of the time bin in which a particular person's response occurs in

149 a particular trial across a repeated measures design. For example, a response in one trial

150 might occur at 546 ms and it would be in time bin 6 (any RTs from 501 ms to 600 ms).

151 One then calculates the sample-based estimate of the discrete-time hazard function of

152 event occurrence for each experimental condition (Figure 1B upper panel). The

153 discrete-time hazard function gives you, for each time bin, the conditional probability that

154 the event occurs (sometime) in bin t , given that the event does not occur in previous bins.

155 In other words, it reflects the instantaneous risk that the event occurs in the current bin t ,

156 given that it has not yet occurred in the past, i.e., in one of the prior bins ($t-1, t-2, \dots, 1$).

157 In the context of experimental psychology, it is often (but not always), the case that

158 responses can be classified as correct or incorrect. In those cases, one can also calculate the

159 conditional accuracy function (Figure 1B lower panel). The conditional accuracy function

160 gives you for each time bin the conditional probability that a response is correct given that

161 it is emitted in time bin t (Allison, 2010; Kantowitz & Pachella, 2021; Wickelgren, 1977).

162 The conditional accuracy function is also known as the micro-level speed-accuracy tradeoff

163 (SAT) function. We refer to this extended (hazard + conditional accuracy) analysis for

164 choice RT data as EHA/SAT. The definitions of these and other discrete-time functions are

165 given in section B of the Supplemental Material.

166 2.2 Benefits of event history analysis for research in experimental psychology

167 Statisticians and mathematical psychologists recommend focusing on the hazard

168 function when analyzing time-to-event data for various reasons (Holden, Van Orden, &

169 Turvey, 2009; Luce, 1991; Townsend, 1990). We do not cover these benefits in detail here,

170 as these are more general topics that have been covered elsewhere in textbooks (see also

171 section G of the Supplemental Material). Instead, here we focus on the benefits as we see

172 them for common research programmes in experimental psychology.

173 We highlight three benefits that we think are relevant to the domain of experimental
174 psychology. First, as illustrated in Figure 1, compared to averaging data across trials,
175 integrating results between hazard functions and their associated conditional accuracy
176 functions for choice RT data can be informative for understanding psychological processes,
177 in terms of inferences about the microgenesis and temporal organization of cognition and
178 theoretical development. As such, the approach permits different kinds of questions to be
179 asked, different inferences to be made, and it holds the potential to discriminate between
180 theoretical accounts of psychological and/or brain-based processes. For example, what kind
181 of theory or set of mechanisms could account for the shape of the functions and the
182 temporally localized effects reported in Figure 1B (Panis & Schmidt, 2016)? Are there new
183 auxiliary assumptions that computational models need to adopt (Panis, Moran, et al.,
184 2020)? Will the temporal effect patterns align nicely with EEG findings (Panis & Schmidt,
185 2022)? And are there new experiments that need to be performed to test the novel
186 predictions that follow from these analyses?

187 Second, compared to more conventional analytical approaches, EHA uses more of the
188 data because it deals with missing data differently. It is conventional with RT data to
189 either (a) use a response deadline and discard all trials without a response, or (b) wait in
190 each trial until a response occurs and then apply data trimming techniques, i.e., discarding
191 too short or too long RTs (and perhaps also erroneous responses) before calculating a mean
192 RT (Berger & Kiefer, 2021). Discarding data can introduce biases, however. Rather than
193 treat non-responses as missing data, EHA treats such trials as *right-censored* observations
194 on the variable RT, because all we know is that RT is greater than some value.
195 Right-censoring is a type of missing data problem and a nearly universal feature of survival
196 data including RT data. For example, if the censoring time was 1 second, then some trials
197 result in observed event times (those with a RT below 1 second), while the other trials
198 result in response times that are right-censored at 1 second. The fact that EHA can deal

199 with right-censoring, therefore, presents a analytical strength of the approach compared to
200 many common approaches in experimental psychology (e.g., ANOVA, linear regression,
201 delta plots).

202 Third, the approach is generalisable and applicable to many tasks that are commonly
203 used in experimental psychology, such as detection, discrimination and bistable perception
204 tasks, and to a range of common experimental manipulations, such as
205 stimulus-onset-asynchrony (see section A of the Supplemental Material). The upshot is
206 that one general analytical approach, which holds several potential advantages, is widely
207 applicable to many substantive use-cases in the RT domain of experimental psychology,
208 irrespective of the analyst's current view on the nature of cognition (Barack & Krakauer,
209 2021).

210 **2.3 Implications for research design in experimental psychology**

211 Performing EHA in experimental psychology has implications for how experiments
212 are designed. More specifically, we consider three implications that researchers will need to
213 consider when using discrete-time EHA. First, because EHA deals with right-censored
214 observations, one can use a fixed response deadline in each trial. This will increase design
215 efficiency as one does not need to wait for very long RTs that would be trimmed anyway.

216 Second, since the number of trials per condition are spread across bins, it is
217 important to have a relatively large number of trial repetitions per participant and per
218 condition. Accordingly, experimental designs using this approach typically focus on
219 factorial, within-subject designs, in which a large number of observations are made on a
220 relatively small number of participants (so-called small- N designs). This approach
221 emphasizes the precision and reproducibility of data patterns at the individual participant
222 level to increase the inferential validity of the design (Baker et al., 2021; Smith & Little,
223 2018). Note that because statistical power derives both from the number of participants

and from the number of repeated measures per participant and condition, small- N designs can still achieve what are generally considered acceptable levels of statistical power, if they have a sufficient amount of data overall (Baker et al., 2021; Smith & Little, 2018).

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

3. Tutorials

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

250 wrangling and plotting functions (see section C of the Supplemental Material for a list of
251 the custom functions).

252 The content of the tutorials, in terms of EHA and multilevel regression modelling, is
253 mainly based on Allison (2010), Singer and Willett (2003), McElreath (2020), Heiss (2021),
254 Kurz (2023a), and Kurz (2023b). We used R (Version 4.5.1; R Core Team, 2024)¹,
255 for all reported analyses.

256 **3.1 Tutorial 1a: Calculating descriptive statistics using a life table**

257 **3.1.1 Data wrangling aims.** Our data wrangling procedures serve two related
258 purposes. First, we want to calculate descriptive statistics for each condition in each
259 individual using a life table. A life table includes for each time bin, the risk set (i.e., the
260 number of trials that are event-free at the start of the bin), the number of observed events,
261 and the estimates of the discrete-time hazard probability $h(t)$, survival probability $S(t)$,
262 probability mass $P(t)$, possibly the conditional accuracy $ca(t)$, and their estimated
263 standard errors (se). The definitions of these quantities are provided in section B of the
264 Supplemental Material.

265 Second, we want to produce two different data sets that can each be submitted to
266 different types of inferential modelling approaches. The two types of data structure we
267 label as ‘person-trial’ data and ‘person-trial-bin’ data. The ‘person-trial’ data (Table 1)

¹ We, furthermore, used the R-packages *bayesplot* (Version 1.13.0; Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019), *brms* (Version 2.22.0; Bürkner, 2017, 2018, 2021), *citr* (Version 0.3.2; Aust, 2019), *cmdstanr* (Version 0.9.0.9000; Gabry, Češnovar, Johnson, & Brander, 2024), *dplyr* (Version 1.1.4; Wickham, François, Henry, Müller, & Vaughan, 2023), *forcats* (Version 1.0.0; Wickham, 2023a), *futures* (Bengtsson, 2021), *ggplot2* (Version 3.5.2; Wickham, 2016), *lme4* (Version 1.1.37; Bates, Mächler, Bolker, & Walker, 2015), *lubridate* (Version 1.9.4; Grolemund & Wickham, 2011), *Matrix* (Version 1.7.3; Bates, Maechler, & Jagan, 2024), *nlme* (Version 3.1.168; Pinheiro & Bates, 2000), *papaja* (Version 0.1.3; Aust & Barth, 2024), *patchwork* (Version 1.3.0; Pedersen, 2024), *purrr* (Version 1.0.4; Wickham & Henry, 2023), *RColorBrewer* (Version 1.1.3; Neuwirth, 2022), *Rcpp* (Eddelbuettel & Balamuta, 2018; Version 1.0.14; Eddelbuettel & François, 2011), *readr* (Version 2.1.5; Wickham, Hester, & Bryan, 2024), *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), *tidyrr* (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).

268 will be familiar to most researchers who record behavioural responses from participants, as
 269 it represents the measured RT and accuracy per trial within an experiment. This data set
 270 is used when fitting conditional accuracy models (Tutorials 2b and 3b).

Table 1
Data structure for ‘person-trial’ data

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

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

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

Table 2
Data structure for ‘person-trial-bin’ data

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

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

3.1.2 A real data wrangling example.

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

There are several data wrangling steps to be taken. First, we need to load the data

295 before we (a) supply required column names, and (b) specify the factor condition with the
 296 correct levels and labels.

297 The required column names are as follows:

- 298 • “pid”, indicating unique participant IDs;
 299 • “trial”, indicating each unique trial per participant;
 300 • “condition”, a factor indicating the levels of the independent variable (1, 2, ...) and
 301 the corresponding labels;
 302 • “rt”, indicating the response times in ms;
 303 • “acc”, indicating the accuracies (1/0).

304 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)))
```

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

```
data_nested <- data_wr %>% group_nest(pid)
data_final <- data_nested %>%
  # ! user input: censoring time, and bin width
```

```

  mutate(censored = map(data, censor, 600, 40)) %>%
  # create person-trial-bin data set
  mutate(ptb_data = map(censored, ptb)) %>%
  # create life tables without ca(t)
  mutate(lifetable = map(ptb_data, setup_lt)) %>%
  # calculate ca(t)
  mutate(condacc = map(censored, calc_ca)) %>%
  # create life tables with ca(t)
  mutate(lifetable_ca = map2(lifetable, condacc, join_lt_ca)) %>%
  # create plots
  mutate(plot = map2(.x = lifetable_ca, .y = pid, plot_eha,1))

```

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

320 The person-trial-bin oriented data set is created by our custom function `ptb()`, and it
 321 has one row for each time bin (of each trial) that is at risk for event occurrence. The
 322 variable “event” in the person-trial-bin oriented data set indicates whether a response
 323 occurs (1) or not (0) for each bin. The next steps are to set up the life table using our
 324 custom function `setup_lt()`, calculate the conditional accuracies using our custom function
 325 `calc_ca()`, add the $ca(t)$ estimates to the life table using our custom function `join_lt_ca()`,
 326 and then plot the descriptive statistics using our custom function `plot_eha()`. One can now

327 inspect different aspects, including the life table for a particular condition of a particular
 328 subject, and a plot of the different functions for a particular participant.

329 In general, it is important to visually inspect the functions first for each participant,
 330 in order to identify individuals that may not be following task instructions (e.g., a flat
 331 conditional accuracy function at .5 indicates that someone is just guessing), outlying
 332 individuals, and/or different groups with qualitatively different behavior. Also, to select a
 333 suited bin width for model fitting, one can test and compare various bin widths in the
 334 censor function, and select the smallest one that is supported by the data.

335 Table 3 shows the life table for condition “blank” (no prime stimulus presented) for
 336 participant 6.

Table 3
The life table for the blank prime condition of participant 6.

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

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

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

Descriptive stats for subject 6

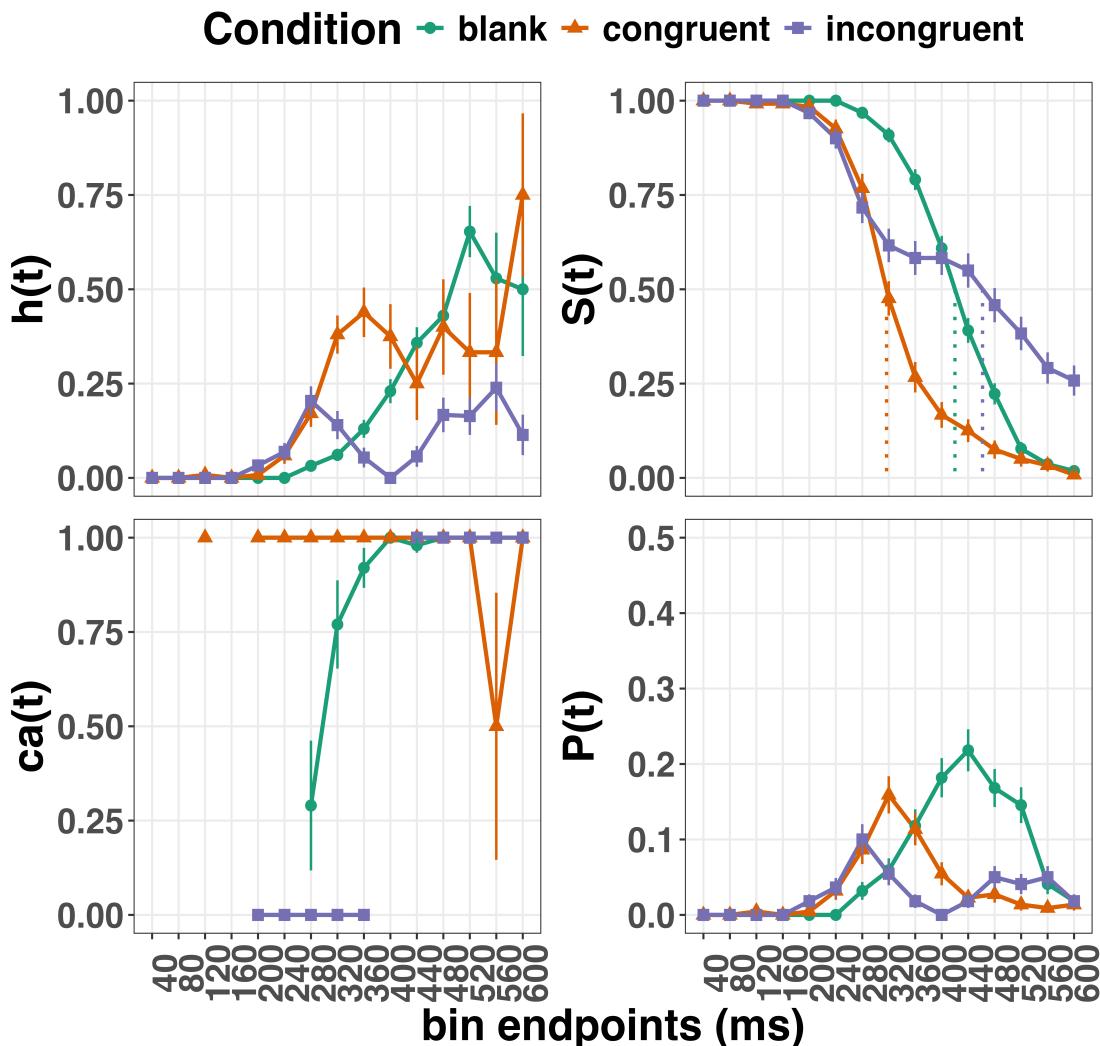


Figure 2. Estimated discrete-time hazard (h), survivor (S), conditional accuracy (ca) and probability mass (P) functions for participant 6. Vertical dotted lines indicate the estimated median RTs. Error bars represent ± 1 standard error of the respective proportion.

343 For example, for participant 6, the estimated hazard values in bin (240,280] are 0.03,

344 0.17, and 0.20 for the blank, congruent, and incongruent prime conditions, respectively. In

345 other words, when the waiting time has increased until *240 ms* after target onset, then the

346 conditional probability of response occurrence in the next 40 ms is more than five times

347 larger for both prime-present conditions, compared to the blank prime condition.

348 Furthermore, the estimated conditional accuracy values in bin (240,280] are 0.29, 1,

349 and 0 for the blank, congruent, and incongruent prime conditions, respectively. In other

350 words, if a response is emitted in bin (240,280], then the probability that it is correct is

351 estimated to be 0.29, 1, and 0 for the blank, congruent, and incongruent prime conditions,

352 respectively.

353 However, when the waiting time has increased until *400 ms* after target onset, then

354 the conditional probability of response occurrence in the next 40 ms is estimated to be

355 0.36, 0.25, and 0.06 for the blank, congruent, and incongruent prime conditions,

356 respectively. And when a response does occur in bin (400,440], then the probability that it

357 is correct is estimated to be 0.98, 1, and 1 for the blank, congruent, and incongruent prime

358 conditions, respectively.

359 These distributional results suggest that participant 6 is initially responding to the

360 prime even though (s)he was instructed to only respond to the target, that response

361 competition emerges in the incongruent prime condition around 300 ms, and that only

362 slower responses are fully controlled by the target stimulus. Qualitatively similar results

363 were obtained for the other five participants. When participants show qualitatively similar

364 distributional patterns, one might consider aggregating their data and plotting the

365 group-average distribution per condition (see Tutorial_1a.Rmd). More generally, these

366 results go against the (often implicit) assumption in research on priming that all observed

367 responses are primed responses to the target stimulus. Instead, the distributional data

368 show that fast responses are triggered exclusively by the prime stimulus, while only the

369 slower responses reflect primed responses to the target stimulus.

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

376 **3.2 Tutorial 1b: Generalising to a more complex design**

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

383 To this end, Tutorial 1b illustrates how to calculate and plot the descriptive statistics
384 for the full data set of Experiment 1 of Panis and Schmidt (2016), which includes two
385 independent variables: mask type and prime type. As we use the same functional
386 programming approach as in Tutorial 1a, we simply refer the reader to Tutorial_1b.Rmd.

387 **3.3 Tutorial 2a: Fitting Bayesian hazard models to interval-censored RT data**

388 In this third tutorial, we illustrate how to fit Bayesian multilevel regression models to
389 the RT data of the masked response priming data used in Tutorial 1a. Fitting (Bayesian or
390 non-Bayesian) regression models to time-to-event data is important when you want to
391 study how the shape of the hazard function depends on various predictors (Singer &
392 Willett, 2003).

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

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

420 M0i and M1i below). As explained by McElreath (2020), the advantage of index coding is
 421 that the same prior can be assigned to each level of the index variable, so that each
 422 category has the same prior uncertainty.

423 In the case of a large- N design without repeated measurements, the parameters of a
 424 discrete-time hazard model can be estimated using standard logistic regression software
 425 after expanding the typical person-trial data set into a person-trial-bin data set (Allison,
 426 2010). When there is clustering in the data, as in the case of a small- N design with
 427 repeated measurements, the parameters of a discrete-time hazard model can be estimated
 428 using population-averaged methods (e.g., Generalized Estimating Equations), and Bayesian
 429 or frequentist generalized linear mixed models (Allison, 2010).

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

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

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

ptb_data <- ptb_data %>%
  # select analysis time range: (200,600] with 10 bins (time bin ranks 6 to 15)
  filter(period > 5) %>%
  # define categorical predictor TIME as index variable named timebin
```

```

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

```

3.3.2 Prior distributions. To get the posterior distribution of each model

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

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

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

461 Before we fit model M0i, we select the necessary columns from the data, and specify

462 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")
```

463 After selecting the bernoulli family and the cloglog link, the model formula is

464 specified. The specification “0 + …” removes the default intercept in brm(). The fixed

465 effects include an intercept for each level of timebin. Each of these intercepts is allowed to

466 vary across individuals (variable pid). We request 2000 samples from the posterior

467 distribution for each of four chains. Estimating model M0i took about 30 minutes on a

468 MacBook Pro (Sonoma 14.6.1 OS, 18GB Memory, M3 Pro Chip).

469 3.3.4 Model M1i: Adding the effects of prime-target congruency.

Previous

470 research has shown that psychological effects typically change over time (Panis, 2020;

471 Panis, Moran, et al., 2020; Panis & Schmidt, 2022; Panis et al., 2017; Panis & Wagemans,

472 2009). In the next model, therefore, we use index coding for both TIME (variable

473 “timebin”) and the categorical predictor prime-target-congruency (variable “prime”), so

474 that we get 30 grand intercepts, one for each combination of timebin level and prime level.

475 Here is the model formula of this model that we call “M1i”.

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

476 Estimating model M1i took about 124 minutes using the same MacBook Pro.

477 **3.3.5 Compare the models.** There are two popular strategies to evaluate how
 478 well models will perform in predicting new data on average: Leave-One-Out (LOO)
 479 cross-validation and the Widely Applicable Information Criterion or WAIC (McElreath,
 480 2020). LOO-weights represent the optimal linear combination of models for predictive
 481 performance, with higher weights for models with better out-of-sample predictive
 482 performance. WAIC-weights represent the relative evidence for each model, with higher
 483 weights for models with a better fit while accounting for model complexity (Kurz, 2023a;
 484 McElreath, 2020).

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

485 ## model_M0i model_M1i
 486 ## "0.00" "1.00"

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

487 ## model_M0i model_M1i
 488 ## "0.00" "1.00"

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

491 **3.3.6 Evaluating parameter estimates in model M1i.** To make causal
 492 inferences from the parameter estimates in model M1i (Frank et al., 2025), we first plot the
 493 densities of the draws from the posterior distributions of its population-level parameters in
 494 Figure 3A, together with point (median) and interval estimates (80% and 95% credible

⁴⁹⁵ intervals). A credible interval is a range of values that contains a parameter's true value
⁴⁹⁶ with a specified probability, given the observed data and model.

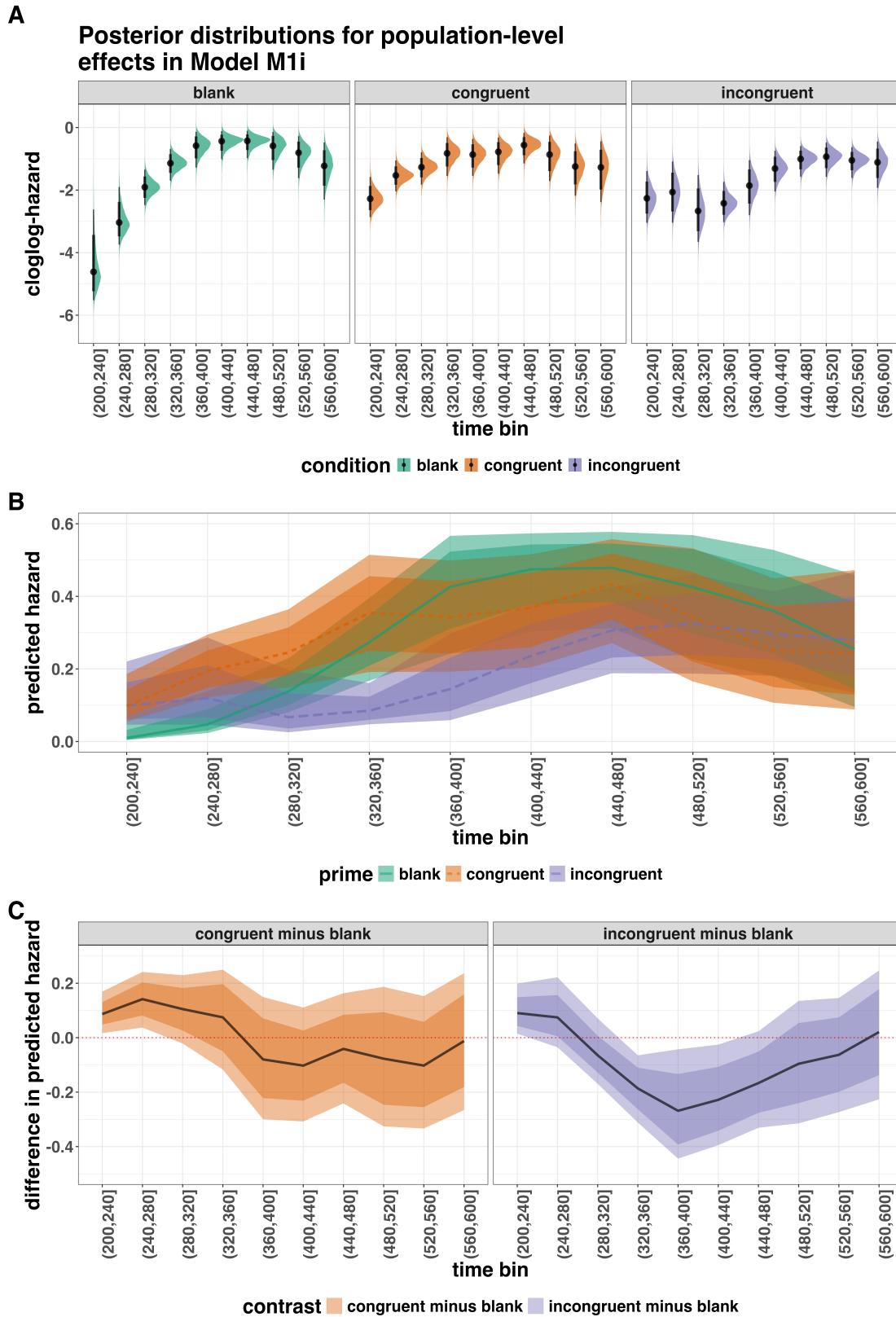


Figure 3. Discrete-time hazard modeling results at the population level. (A) Medians and 80/95% credible intervals of the posterior distributions of the population-level parameters of model M1i. (B) Point (median) and 80/95% credible interval summaries of the hazard estimates (expected values of the 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.

Because the parameter estimates are on the cloglog-hazard scale, we can ease our interpretation by plotting the expected value of the posterior predictive distribution – the predicted hazard values – at the population level (Figure 3B). As we are actually interested in the effects of congruent and incongruent primes, relative to the blank prime condition, we can construct two contrasts (congruent-blank, incongruent-blank), and plot the posterior distributions of these contrast effects at the population level (Figure 3C). The point estimates and quantile intervals can also be reported in a table (see Tutorial_2a.Rmd for details).

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

If we want to estimate the population-level effect of prime type on hazard, we can base our conclusion on the credible Intervals (CrIs) in Figure 3C. The contrast “congruent minus blank” was estimated to be 0.09 hazard units in bin (200,240] (95% CrI = [0.02, 0.17]), and 0.14 hazard units in bin (240,280]) (95% CrI = [0.04, 0.25]). For the other bins, the 95% credible interval contained zero. The contrast “incongruent minus blank” was estimated to be 0.09 hazard units in bin (200,240] (95% CrI = [0.01, 0.21]), -0.19 hazard units in bin (320,360] (95% CrI = [-0.31, -0.06]), -0.27 hazard units in bin (360,400] (95% CrI = [-0.45, -0.04]), and -0.23 hazard units in bin (400,440] (95% CrI = [-0.40, -0.03]). For the other bins, the 95% credible interval contained zero.

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

524 The sign of the effect of incongruent primes on the hazard of response occurrence thus
525 depends on how much waiting time has passed since target onset. Future modeling efforts
526 could incorporate the trial number into the model formula, in order to also study how the
527 effects of prime type on hazard change on the long experiment-wide time scale, next to the
528 short trial-wide time scale. In Tutorial_2a.Rmd we provide a number of model formulae
529 that should get you going.

530 **3.4 Tutorial 2b: Fitting Bayesian conditional accuracy models**

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

538 To make inferences from the parameter estimates in model M1i_ca, we first plot the
539 densities of the draws from the posterior distributions of its population-level parameters in
540 Figure 4A, together with point (median) and interval estimates (80% and 95% credible
541 intervals).

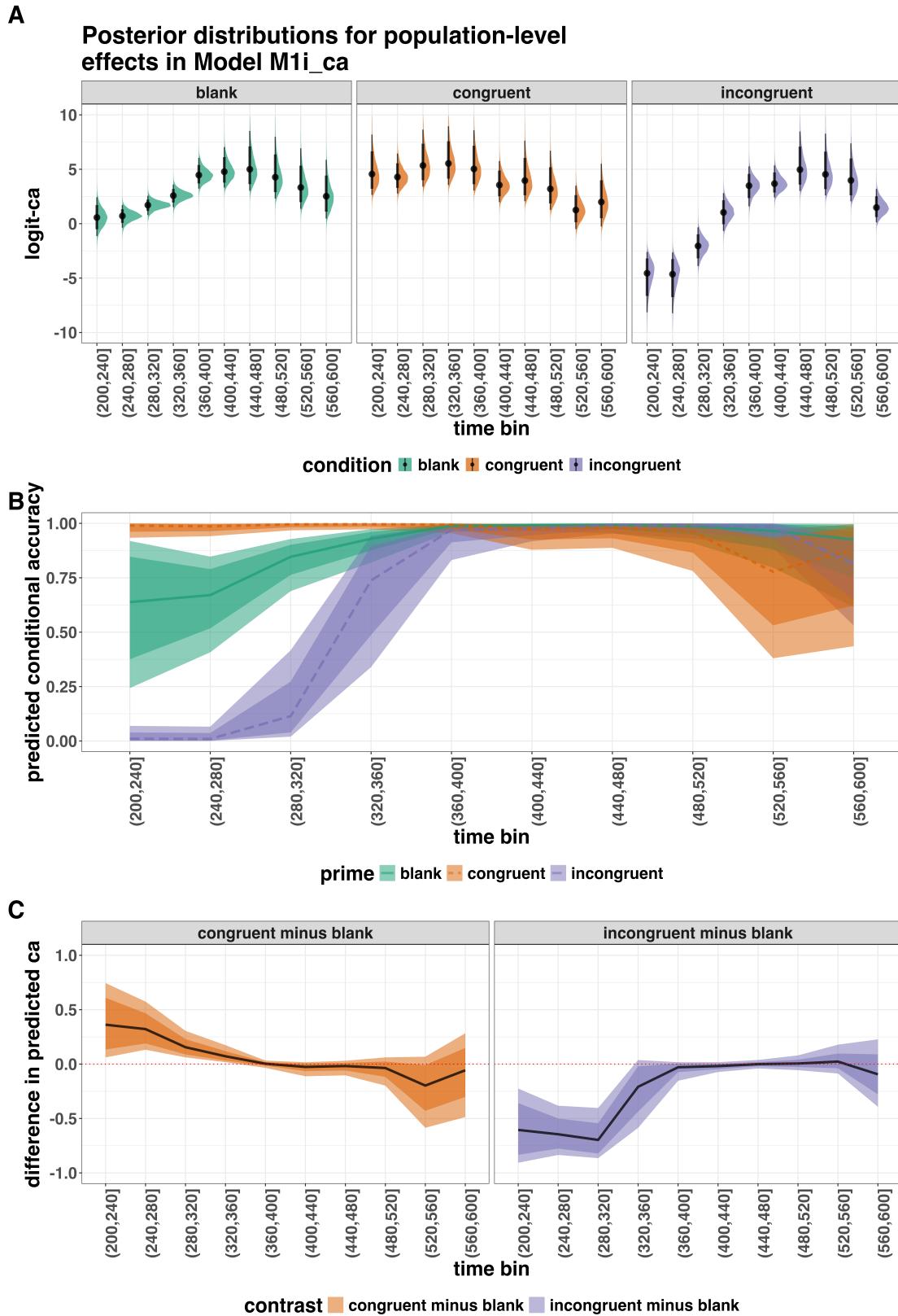


Figure 4. Conditional accuracy modeling results at the population level. (A) Medians and 80/95% credible intervals of the posterior distributions of the population-level parameters of model M1i_ca. (B) Point (median) and 80/95% credible interval summaries of the conditional accuracy (ca) estimates (expected 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.

542 Because the parameter estimates are on the logit-ca scale, we can ease our
543 interpretation by plotting the expected value of the posterior predictive distribution – the
544 predicted conditional accurcies – at the population level (Figure 4B). As we are actually
545 interested in the effects of congruent and incongruent primes, relative to the blank prime
546 condition, we can construct two contrasts (congruent-blank, incongruent-blank), and plot
547 the posterior distributions of these contrast effects at the population level (Figure 4C).

548 Based on Figure 4C we see that on the population level congruent primes have a positive
549 effect on the conditional accuracy of emitted responses in time bins (200,240], (240,280],
550 (280,320], and (320,360], relative to the estimates in the baseline condition (blank prime;
551 red dashed lines in Figure 4C). Incongruent primes have a negative effect on the conditional
552 accuracy of emitted responses in the first three time bins, relative to blank primes.

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

557 3.5 Tutorial 4: Planning

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

559 **3.5.1 Background.** The general approach to planning that we adopt here involves
560 simulating reasonably structured data to help guide what you might be able to expect from
561 your data once you collect it (Gelman, Vehtari, et al., 2020). The basic structure and code
562 follows the examples outlined by Solomon Kurz in his ‘power’ blog posts
563 (<https://solomonkurz.netlify.app/blog/bayesian-power-analysis-part-i/>) and Lisa
564 Debruine’s R package faux{} (<https://debruine.github.io/faux/>), as well as these related
565 papers (DeBruine & Barr, 2021; Pargent, Koch, Kleine, Lermer, & Gaube, 2024).

566 **3.5.2 Basic workflow.** The basic workflow is as follows:

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

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

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

580 **3.5.3 Fit a regression model and simulate one dataset.** We again use the
 581 data from Panis and Schmidt (2016) to provide a worked example. We fit an index coding
 582 model on a subset of time bins (six time bins in total) and for two prime conditions
 583 (congruent and incongruent). We chose to focus on a subsample of the data to ease the
 584 computational burden. We also used a full varying effects structure, with the model
 585 formula as follows:

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

586 We then took parameters from this model and used them to create a single dataset
 587 with 200 trials per condition for 10 individual participants. The raw data and the
 588 simulated data are plotted in Figure 5 and show quite close correspondence, which is
 589 re-assuring. But, this is only one dataset. What we really want to do is simulate many
 590 datasets and vary parameters of interest, which is what we turn to in the next section.

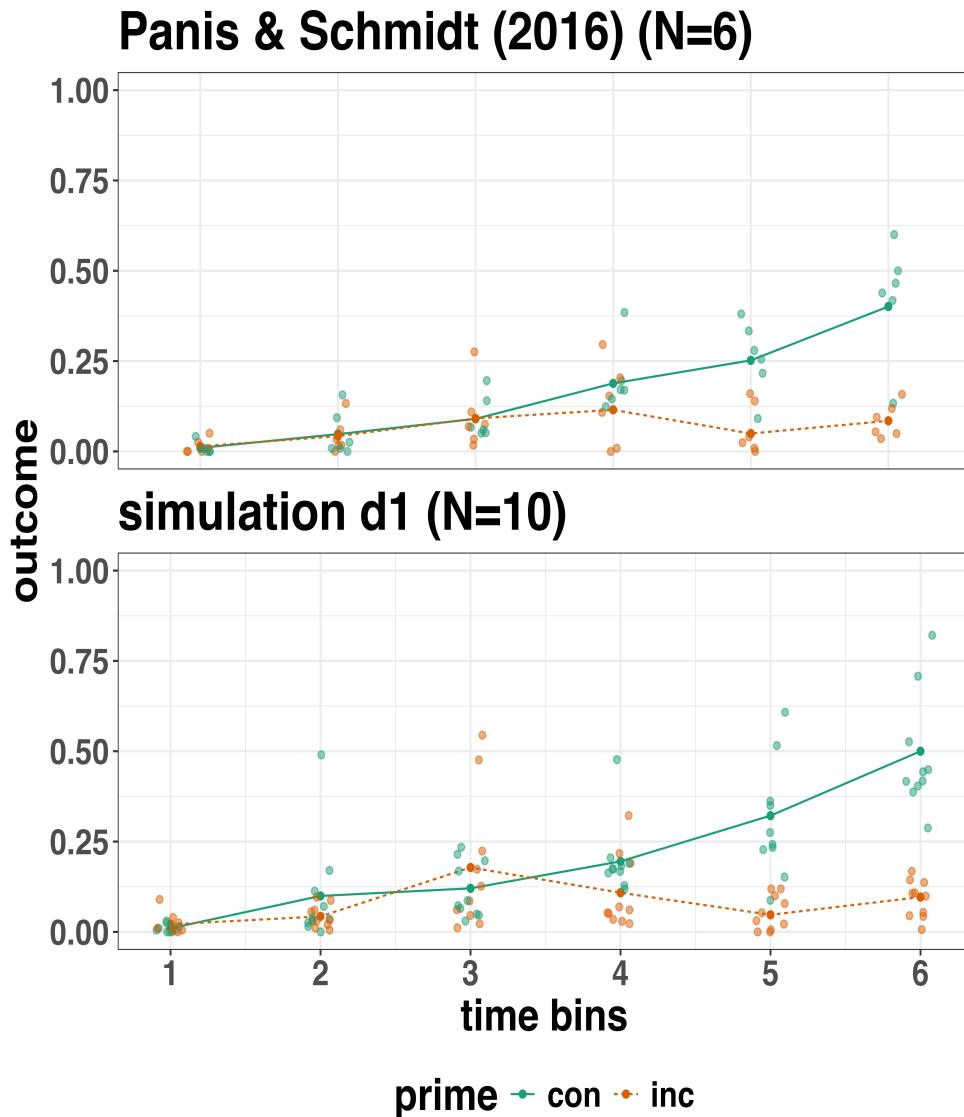


Figure 5. Raw data from Panis and Schmidt (2016) and simulated data from 10 participants.

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

591 Here we use the same data simulation process as used above, but instead of simulating one
 592 dataset, we simulate 1000 datasets per variation in parameter values. Specifically, in
 593 Simulation 1, we vary the number of trials per condition (100, 200, and 400), as well as the
 594 effect size in bin 6. We focus on bin 6 only, in terms of varying the effect size, just to make
 595 things simpler and easier to understand. The effect size observed in bin 6 in this subsample
 596 of data was a 79% reduction in hazard value from the congruent prime (0.401 hazard
 597

598 value) to the incongruent prime condition (0.085 hazard value). In other words, a hazard
599 ratio of 0.21 (e.g., $0.085/0.401 = 0.21$). As a starting point, we chose three effect sizes,
600 which covered a fairly broad range of hazard ratios (0.25, 0.5, 0.75), which correspond to a
601 75%, 50% and 25% reduction in hazard value as a function of prime condition.

602 Summary results from Simulation 1 are shown in Figure 6A. Figure 6A depicts
603 statistical “power” as calculated by the percentage of lower-bound 95% confidence intervals
604 that exclude zero when the difference between prime condition is calculated (congruent -
605 incongruent). In other words, we calculate the fraction of simulated datasets that
606 generated an effect of prime that excludes the criterion mark of zero. We are aware that
607 “power” is not part of a Bayesian analytical workflow, but we choose to include it here, as
608 it is familiar to most researchers in experimental psychology.

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

617 In Simulation 2, we varied the effect size between a different range of values (0.5, 0.4,
618 0.3), which correspond to a 50%, 60% and 70% reduction in hazard value as a function of
619 prime condition. In addition, we varied the number of participants per experiment between
620 10, 15, and 20 participants. Given that trial count per condition made little difference to
621 power in Simulation 1, we fixed trial count at 200 trials per condition in Simulation 2.
622 Summary results from Simulation 2 are shown in Figure 6B. A summary of these power
623 calculations might be as follows (trial count = 200 per condition in all cases):

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

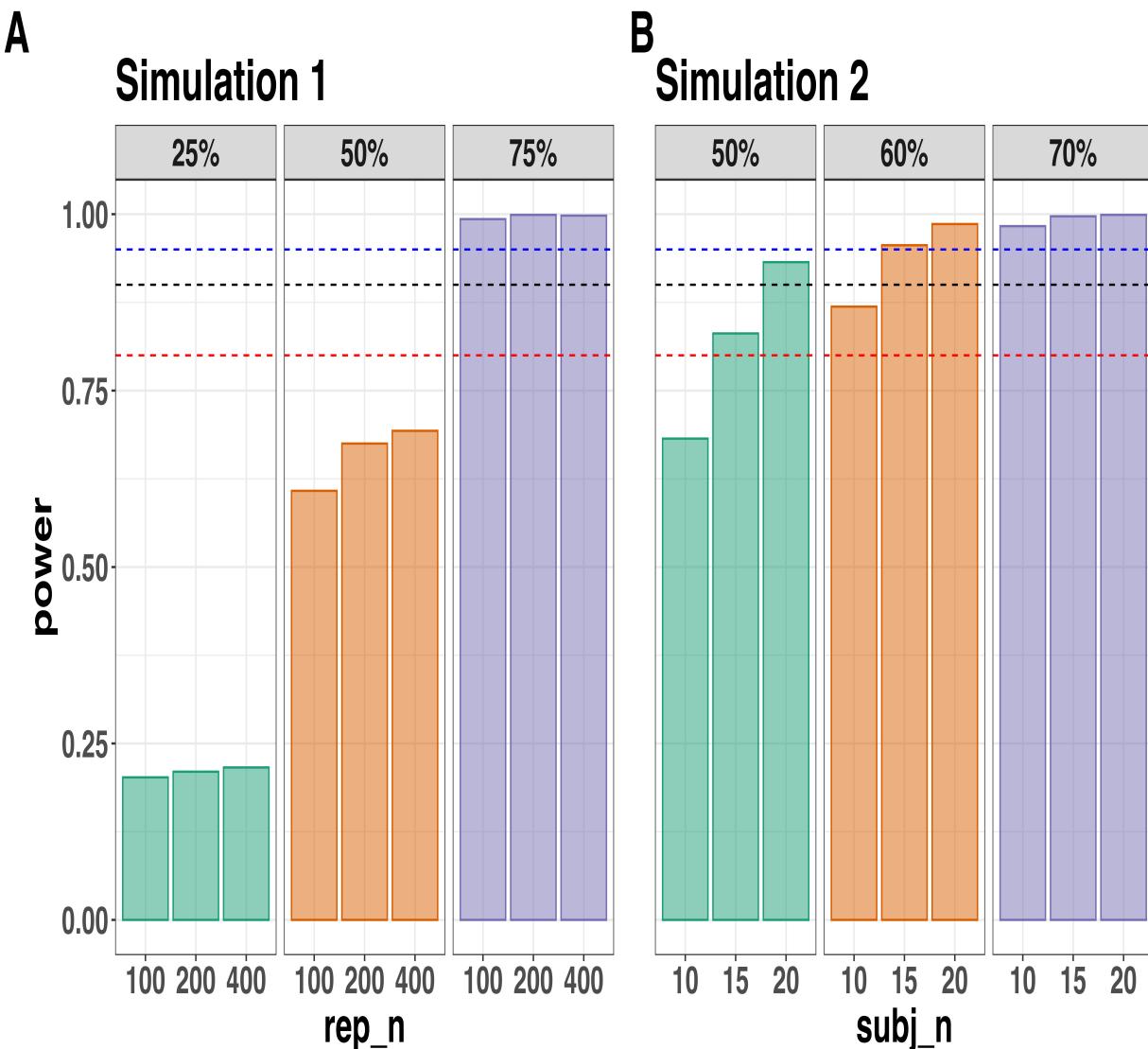


Figure 6. Statistical power across data Simulation 1 (A) and Simulation 2 (B). Power was calculated as the percentage of lower-bound 95% confidence intervals that exclude zero when the difference between prime condition is calculated (congruent - incongruent). In Simulation 1, the effect size was varied between a 25%, 50% and 75% reduction in hazard value, whereas the trial count was varied between 100, 200 and 400 trials per condition (the number of participants was fixed at N=10). In Simulation 2, the effect size was varied between a 50%, 60% and 70% reduction in hazard value, whereas the number of participants was varied between N=10, 15 and 20 (the number of trials per condition was fixed at 200). The dashed lines represent 80% (red), 90% (black) and 95% (blue) power. Abbreviations: rep_n = the number of trials per experimental condition; subj_n = the number of participants per simulated experiment.

627 **3.5.5 Planning decisions.** Now that we have summarised our simulated data,

628 what planning decisions could we make about a future study? More concretely, how many

629 trials per condition should we collect and how many participants should we test? Like

630 almost always when planning future studies, the answer depends on your objectives, as well

631 as the available resources (Lakens, 2022). There is no straightforward and clear-cut answer.

632 Some considerations might be as follows:

- 633 • How much power or precision are you looking to obtain in this particular study?

- 634 • Are you running multiple studies that have some form of replication built in?

- 635 • What level of resources do you have at your disposal, such as time, money and

636 personnel?

- 637 • How easy or difficult is it to obtain the specific type of sample?

638 If we were running this kind of study in our lab, what would we do? We might pick a

639 hazard ratio of 0.4 or 0.5 as a target effect size since this is much smaller than that

640 observed previously (Panis & Schmidt, 2016). Then we might pick the corresponding

641 combination of trial count per condition (e.g., 200) and participant sample size (e.g., N=10

642 or N=15) that takes you over the 80% power mark. If we wanted to maximise power based

643 on these simulations, and we had the time and resources available, then we would test

644 N=20 participants, which would provide >90% power for an effect size of 0.5.

645 But, and this is an important caveat, unless there are unavoidable reasons, no matter

646 what kind of planning choices we made based on these data simulations, we would not

647 solely rely on data collected from one single study. Instead, we would run a follow-up

648 experiment that replicates and extends the initial result. By doing so, we would aim to

649 avoid the Cult of the Isolated Single Study (Nelder, 1999; Tong, 2019), and thus reduce the

650 reliance on any one type of planning tool, such as a power analysis. Then, we would look

651 for common patterns across two or more experiments, rather than trying to make the case

652 that a single study on its own has sufficient evidential value to hit some criterion mark.

653

4. Discussion

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

668 **4.1 What are the main use-cases of EHA for understanding cognition and brain
669 function?**

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

678 **how they are temporally organized**, our EHA tutorials could be useful tools to use for
679 basic knowledge development, as well as theory building.

680 Second, even if you are not primarily interested in studying the temporal organization
681 of cognitive states, EHA could still be a useful tool to consider using, in order to qualify
682 inferences that are being made based on comparisons between means. Given that distinctly
683 different inferences can be made from the same data based on whether one computes a
684 mean across trials or a RT distribution of events (Figure 1), it may be important for
685 researchers to supplement comparisons between means with EHA. For instance, EHA
686 might reveal that the conclusion of interest based on averaging across trials does not apply
687 to all responses, but is instead restricted to certain periods of within-trial time.

688 4.2 Model complexity versus interpretability

689 Hazard and conditional accuracy models can quickly become very complex when
690 adding more than one time scale, due to the many possible higher-order interactions. For
691 example, some of the models discussed in Tutorial 2a, which we did not focus on in the
692 main text, contain two time scales as covariates: the passage of time on the within-trial
693 time scale, and the passage of time on the across-trial (or within-experiment) time scale.
694 However, when trials are presented in blocks, and blocks of trials within sessions, and when
695 the experiment comprises a number of sessions, then four time scales can be defined
696 (within-trial, within-block, within-session, and within-experiment). From a theoretical
697 perspective, adding more than one time scale – and their interactions – can be important
698 to capture plasticity and other learning effects that may play out on such longer time
699 scales, and that are probably present in each experiment in general (Schöner & Spencer,
700 2016). From a practical perspective, therefore, some choices need to be made to balance
701 the amount of data that is being collected per participant, condition and across the varying
702 timescales. As one example, if there are several timescales of relevance, then it might be
703 prudent for interpretational purposes to limit the number of experimental predictor

704 variables (conditions). This is of course where planning and data simulation efforts would
705 be important to provide a guide to experimental design choices (see Tutorial 4 and section
706 2.3).

707 **4.3 Limitations**

708 Compared to the orthodox method – comparing means between conditions – the
709 most important limitation of multilevel hazard and conditional accuracy modeling is that it
710 might take a long time to estimate the parameters using Bayesian methods or the model
711 might have to be simplified significantly to use frequentist methods. Relatedly, as these
712 models can be quite complex in terms of the number of possible parameters, more thought
713 is required at the model specification and model building stages.

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

730

5. Conclusions

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

740

Author contributions

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

744

Conflicts of Interest

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

747

Prior versions

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

750

Supplemental Material

751

Disclosures

752 **Data, materials, and online resources**

753 Link to public archive:

754 https://github.com/sven-panis/Tutorial_Event_History_Analysis

755 Supplemental Material: Panis_Ramsey_suppl_material.pdf

756 **Ethical approval**

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

759

References

- 760 Abney, D. H., Fausey, C. M., Suarez-Rivera, C., & Tamis-LeMonda, C. S. (2025).
761 Advancing a temporal science of behavior. *Trends in Cognitive Sciences*.
762 <https://doi.org/10.1016/j.tics.2025.05.010>
- 763 Allison, P. D. (1982). Discrete-Time Methods for the Analysis of Event Histories.
764 *Sociological Methodology*, 13, 61. <https://doi.org/10.2307/270718>
- 765 Allison, P. D. (2010). *Survival analysis using SAS: A practical guide* (2. ed.). Cary, NC:
766 SAS Press.
- 767 Aust, F. (2019). *Citr: 'RStudio' add-in to insert markdown citations*. Retrieved from
768 <https://github.com/crsh/citr>
- 769 Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R*
770 *Markdown*. <https://doi.org/10.32614/CRAN.package.papaja>
- 771 Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., &
772 Andrews, T. J. (2021). Power contours: Optimising sample size and precision in
773 experimental psychology and human neuroscience. *Psychological Methods*, 26(3),
774 295–314. <https://doi.org/10.1037/met0000337>
- 775 Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature*
776 *Reviews Neuroscience*, 22(6), 359–371. <https://doi.org/10.1038/s41583-021-00448-6>
- 777 Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for
778 confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*,
779 68(3), 10.1016/j.jml.2012.11.001. <https://doi.org/10.1016/j.jml.2012.11.001>
- 780 Barth, M. (2023). *tinylabes: Lightweight variable labels*. Retrieved from
781 <https://cran.r-project.org/package=tinylabes>
- 782 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects
783 models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
784 <https://doi.org/10.18637/jss.v067.i01>
- 785 Bates, D., Maechler, M., & Jagan, M. (2024). *Matrix: Sparse and dense matrix classes and*

- 786 methods. Retrieved from <https://Matrix.R-forge.R-project.org>
- 787 Bengtsson, H. (2021). futures: A unifying framework for parallel and distributed
788 processing in r using futures. *The R Journal*, 13(2), 208–227.
789 <https://doi.org/10.32614/RJ-2021-048>
- 790 Berger, A., & Kiefer, M. (2021). Comparison of Different Response Time Outlier Exclusion
791 Methods: A Simulation Study. *Frontiers in Psychology*, 12, 675558.
792 <https://doi.org/10.3389/fpsyg.2021.675558>
- 793 Blossfeld, H.-P., & Rohwer, G. (2002). *Techniques of event history modeling: New*
794 *approaches to causal analysis*, 2nd ed (pp. x, 310). Mahwah, NJ, US: Lawrence
795 Erlbaum Associates Publishers.
- 796 Bloxom, B. (1984). Estimating response time hazard functions: An exposition and
797 extension. *Journal of Mathematical Psychology*, 28(4), 401–420.
798 [https://doi.org/10.1016/0022-2496\(84\)90008-7](https://doi.org/10.1016/0022-2496(84)90008-7)
- 799 Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in
800 psychology are heterogeneous. *Journal of Experimental Psychology: General*, 148(4),
801 601–618. <https://doi.org/10.1037/xge0000558>
- 802 Box-Steffensmeier, J. M. (2004). Event history modeling: A guide for social scientists.
803 Cambridge: University Press.
- 804 Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan.
805 *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- 806 Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms.
807 *The R Journal*, 10(1), 395–411. <https://doi.org/10.32614/RJ-2018-017>
- 808 Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal*
809 *of Statistical Software*, 100(5), 1–54. <https://doi.org/10.18637/jss.v100.i05>
- 810 DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through
811 Data Simulation. *Advances in Methods and Practices in Psychological Science*, 4(1),
812 2515245920965119. <https://doi.org/10.1177/2515245920965119>

- 813 Eddelbuettel, D., & Balamuta, J. J. (2018). Extending R with C++: A Brief Introduction
814 to Rcpp. *The American Statistician*, 72(1), 28–36.
815 <https://doi.org/10.1080/00031305.2017.1375990>
- 816 Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal
817 of Statistical Software*, 40(8), 1–18. <https://doi.org/10.18637/jss.v040.i08>
- 818 Elmer, T., Van Duijn, M. A. J., Ram, N., & Bringmann, L. F. (2023). Modeling
819 categorical time-to-event data: The example of social interaction dynamics captured
820 with event-contingent experience sampling methods. *Psychological Methods*.
821 <https://doi.org/10.1037/met0000598>
- 822 Frank, M. C., Braginsky, M., Cachia, J., Coles, N. A., Hardwicke, T. E., Hawkins, R. D.,
823 ... Williams, R. (2025). *Experimentology: An Open Science Approach to Experimental
824 Psychology Methods*. Stanford University. <https://doi.org/10.25936/3JP6-5M50>
- 825 Gabry, J., Češnovar, R., Johnson, A., & Broder, S. (2024). *Cmdstanr: R interface to
826 'CmdStan'*. Retrieved from <https://github.com/stan-dev/cmdstanr>
- 827 Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization
828 in bayesian workflow. *J. R. Stat. Soc. A*, 182, 389–402.
829 <https://doi.org/10.1111/rssa.12378>
- 830 Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and Other Stories*.
831 [https://www.cambridge.org/highereducation/books/regression-and-other-
833 stories/DD20DD6C9057118581076E54E40C372C](https://www.cambridge.org/highereducation/books/regression-and-other-
832 stories/DD20DD6C9057118581076E54E40C372C); Cambridge University Press.
834 <https://doi.org/10.1017/9781139161879>
- 835 Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., ...
836 Modrák, M. (2020). *Bayesian Workflow*. arXiv.
837 <https://doi.org/10.48550/arXiv.2011.01808>
- 838 Girard, J. (2024). *Standist: What the package does (one line, title case)*. Retrieved from
839 <https://github.com/jmgirard/standist>
- 839 Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate.

- 840 *Journal of Statistical Software*, 40(3), 1–25. Retrieved from
841 <https://www.jstatsoft.org/v40/i03/>
- 842 Heiss, A. (2021, November 10). A Guide to Correctly Calculating Posterior Predictions
843 and Average Marginal Effects with Multilevel Bayesian Models.
844 <https://doi.org/10.59350/wbn93-edb02>
- 845 Holden, J. G., Van Orden, G. C., & Turvey, M. T. (2009). Dispersion of response times
846 reveals cognitive dynamics. *Psychological Review*, 116(2), 318–342.
847 <https://doi.org/10.1037/a0014849>
- 848 Hosmer, D. W., Lemeshow, S., & May, S. (2011). *Applied Survival Analysis: Regression*
849 *Modeling of Time to Event Data* (2nd ed). Hoboken: John Wiley & Sons.
- 850 Kantowitz, B. H., & Pachella, R. G. (2021). The Interpretation of Reaction Time in
851 Information-Processing Research 1. *Human Information Processing*, 41–82.
852 <https://doi.org/10.4324/9781003176688-2>
- 853 Kay, M. (2024). *tidybayes: Tidy data and geoms for Bayesian models*.
854 <https://doi.org/10.5281/zenodo.1308151>
- 855 Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing,
856 estimation, meta-analysis, and power analysis from a Bayesian perspective.
857 *Psychonomic Bulletin & Review*, 25(1), 178–206.
858 <https://doi.org/10.3758/s13423-016-1221-4>
- 859 Kurz, A. S. (2023a). *Applied longitudinal data analysis in brms and the tidyverse* (version
860 0.0.3). Retrieved from <https://bookdown.org/content/4253>
- 861 Kurz, A. S. (2023b). *Statistical rethinking with brms, ggplot2, and the tidyverse: Second*
862 *edition* (version 0.4.0). Retrieved from <https://bookdown.org/content/4857/>
- 863 Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, 8(1), 33267.
864 <https://doi.org/10.1525/collabra.33267>
- 865 Landes, J., Engelhardt, S. C., & Pelletier, F. (2020). An introduction to event history
866 analyses for ecologists. *Ecosphere*, 11(10), e03238. <https://doi.org/10.1002/ecs2.3238>

- 867 Lougheed, J. P., Benson, L., Cole, P. M., & Ram, N. (2019). Multilevel survival analysis:
868 Studying the timing of children's recurring behaviors. *Developmental Psychology*,
869 55(1), 53–65. <https://doi.org/10.1037/dev0000619>
- 870 Luce, R. D. (1991). *Response times: Their role in inferring elementary mental organization*
871 (1. issued as paperback). Oxford: Univ. Press.
- 872 McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and*
873 *STAN* (2nd ed.). New York: Chapman and Hall/CRC.
874 <https://doi.org/10.1201/9780429029608>
- 875 Mills, M. (2011). *Introducing Survival and Event History Analysis*. 1 Oliver's Yard, 55 City
876 Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd.
877 <https://doi.org/10.4135/9781446268360>
- 878 Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. Retrieved from
879 <https://CRAN.R-project.org/package=tibble>
- 880 Nelder, J. A. (1999). From Statistics to Statistical Science. *Journal of the Royal Statistical*
881 *Society. Series D (The Statistician)*, 48(2), 257–269. Retrieved from
882 <https://www.jstor.org/stable/2681191>
- 883 Neuwirth, E. (2022). *RColorBrewer: ColorBrewer palettes*. Retrieved from
884 <https://CRAN.R-project.org/package=RColorBrewer>
- 885 Panis, S. (2020). How can we learn what attention is? Response gating via multiple direct
886 routes kept in check by inhibitory control processes. *Open Psychology*, 2(1), 238–279.
887 <https://doi.org/10.1515/psych-2020-0107>
- 888 Panis, S., Moran, R., Wolkersdorfer, M. P., & Schmidt, T. (2020). Studying the dynamics
889 of visual search behavior using RT hazard and micro-level speed–accuracy tradeoff
890 functions: A role for recurrent object recognition and cognitive control processes.
891 *Attention, Perception, & Psychophysics*, 82(2), 689–714.
892 <https://doi.org/10.3758/s13414-019-01897-z>
- 893 Panis, S., Schmidt, F., Wolkersdorfer, M. P., & Schmidt, T. (2020). Analyzing Response

- 894 Times and Other Types of Time-to-Event Data Using Event History Analysis: A Tool
895 for Mental Chronometry and Cognitive Psychophysiology. *I-Perception*, 11(6),
896 2041669520978673. <https://doi.org/10.1177/2041669520978673>
- 897 Panis, S., & Schmidt, T. (2016). What Is Shaping RT and Accuracy Distributions? Active
898 and Selective Response Inhibition Causes the Negative Compatibility Effect. *Journal of*
899 *Cognitive Neuroscience*, 28(11), 1651–1671. https://doi.org/10.1162/jocn_a_00998
- 900 Panis, S., & Schmidt, T. (2022). When does “inhibition of return” occur in spatial cueing
901 tasks? Temporally disentangling multiple cue-triggered effects using response history
902 and conditional accuracy analyses. *Open Psychology*, 4(1), 84–114.
903 <https://doi.org/10.1515/psych-2022-0005>
- 904 Panis, S., Torfs, K., Gillebert, C. R., Wagemans, J., & Humphreys, G. W. (2017).
905 Neuropsychological evidence for the temporal dynamics of category-specific naming.
906 *Visual Cognition*, 25(1-3), 79–99. <https://doi.org/10.1080/13506285.2017.1330790>
- 907 Panis, S., & Wagemans, J. (2009). Time-course contingencies in perceptual organization
908 and identification of fragmented object outlines. *Journal of Experimental Psychology:
909 Human Perception and Performance*, 35(3), 661–687.
910 <https://doi.org/10.1037/a0013547>
- 911 Pargent, F., Koch, T. K., Kleine, A.-K., Lermer, E., & Gaube, S. (2024). A Tutorial on
912 Tailored Simulation-Based Sample-Size Planning for Experimental Designs With
913 Generalized Linear Mixed Models. *Advances in Methods and Practices in Psychological
914 Science*, 7(4), 25152459241287132. <https://doi.org/10.1177/25152459241287132>
- 915 Pedersen, T. L. (2024). *Patchwork: The composer of plots*. Retrieved from
916 <https://patchwork.data-imaginist.com>
- 917 Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-PLUS*. New York:
918 Springer. <https://doi.org/10.1007/b98882>
- 919 R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna,
920 Austria: R Foundation for Statistical Computing. Retrieved from

- 921 https://www.R-project.org/
- 922 Schöner, G., & Spencer, J. P. (2016). *Dynamic thinking: A primer on dynamic field theory*.
923 New York, NY: Oxford University Press.
- 924 https://doi.org/10.1093/acprof:oso/9780199300563.001.0001
- 925 Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling*
926 *Change and Event Occurrence*. Oxford, New York: Oxford University Press.
- 927 Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design.
928 *Psychonomic Bulletin & Review*, 25(6), 2083–2101.
929 https://doi.org/10.3758/s13423-018-1451-8
- 930 Stan Development Team. (2020). *StanHeaders: Headers for the R interface to Stan*.
931 Retrieved from https://mc-stan.org/
- 932 Stan Development Team. (2024). *RStan: The R interface to Stan*. Retrieved from
933 https://mc-stan.org/
- 934 Stoolmiller, M. (2015). *An Introduction to Using Multivariate Multilevel Survival Analysis*
935 *to Study Coercive Family Process* (Vol. 1; T. J. Dishion & J. Snyder, Eds.). Oxford
936 University Press. https://doi.org/10.1093/oxfordhb/9780199324552.013.27
- 937 Stoolmiller, M., & Snyder, J. (2006). Modeling heterogeneity in social interaction processes
938 using multilevel survival analysis. *Psychological Methods*, 11(2), 164–177.
939 https://doi.org/10.1037/1082-989X.11.2.164
- 940 Teachman, J. D. (1983). Analyzing social processes: Life tables and proportional hazards
941 models. *Social Science Research*, 12(3), 263–301.
942 https://doi.org/10.1016/0049-089X(83)90015-7
- 943 Tong, C. (2019). Statistical Inference Enables Bad Science; Statistical Thinking Enables
944 Good Science. *The American Statistician*, 73(sup1), 246–261.
945 https://doi.org/10.1080/00031305.2018.1518264
- 946 Townsend, J. T. (1990). Truth and consequences of ordinal differences in statistical
947 distributions: Toward a theory of hierarchical inference. *Psychological Bulletin*, 108(3),

- 948 551–567. <https://doi.org/10.1037/0033-2909.108.3.551>
- 949 950 Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, 7(3), 424–465. <https://doi.org/10.3758/BF03214357>
- 951 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](https://doi.org/10.1016/0001-6918(77)90012-9)
- 952 953 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <https://ggplot2.tidyverse.org>
- 954 955 Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from <https://forcats.tidyverse.org/>
- 956 957 Wickham, H. (2023b). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from <https://stringr.tidyverse.org>
- 958 959 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- 960 961 Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. Retrieved from <https://dplyr.tidyverse.org>
- 962 963 Wickham, H., & Henry, L. (2023). *Purrr: Functional programming tools*. Retrieved from [https://purrr.tidyverse.org/](https://purrr.tidyverse.org)
- 964 965 Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data*. Retrieved from <https://readr.tidyverse.org>
- 966 967 Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data*. Retrieved from <https://tidyr.tidyverse.org>
- 968 969 Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data*. Retrieved from <https://tidyr.tidyverse.org>
- 970 971 Winter, B. (2019). *Statistics for Linguists: An Introduction Using R*. New York: Routledge. <https://doi.org/10.4324/9781315165547>
- 972 973 974 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.

975 <https://doi.org/10.3758/s13414-020-02010-5>