

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: 10094 (body) + 1764 (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). For example, Figure 1B shows a first state (up
58 to 400 ms after target onset) for which the early upswing in hazard is equal for both
59 conditions, and the emitted responses are always correct in condition 1 and always
60 incorrect in condition 2. In a second state (400 to 500 ms), hazard is higher in condition 1,

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

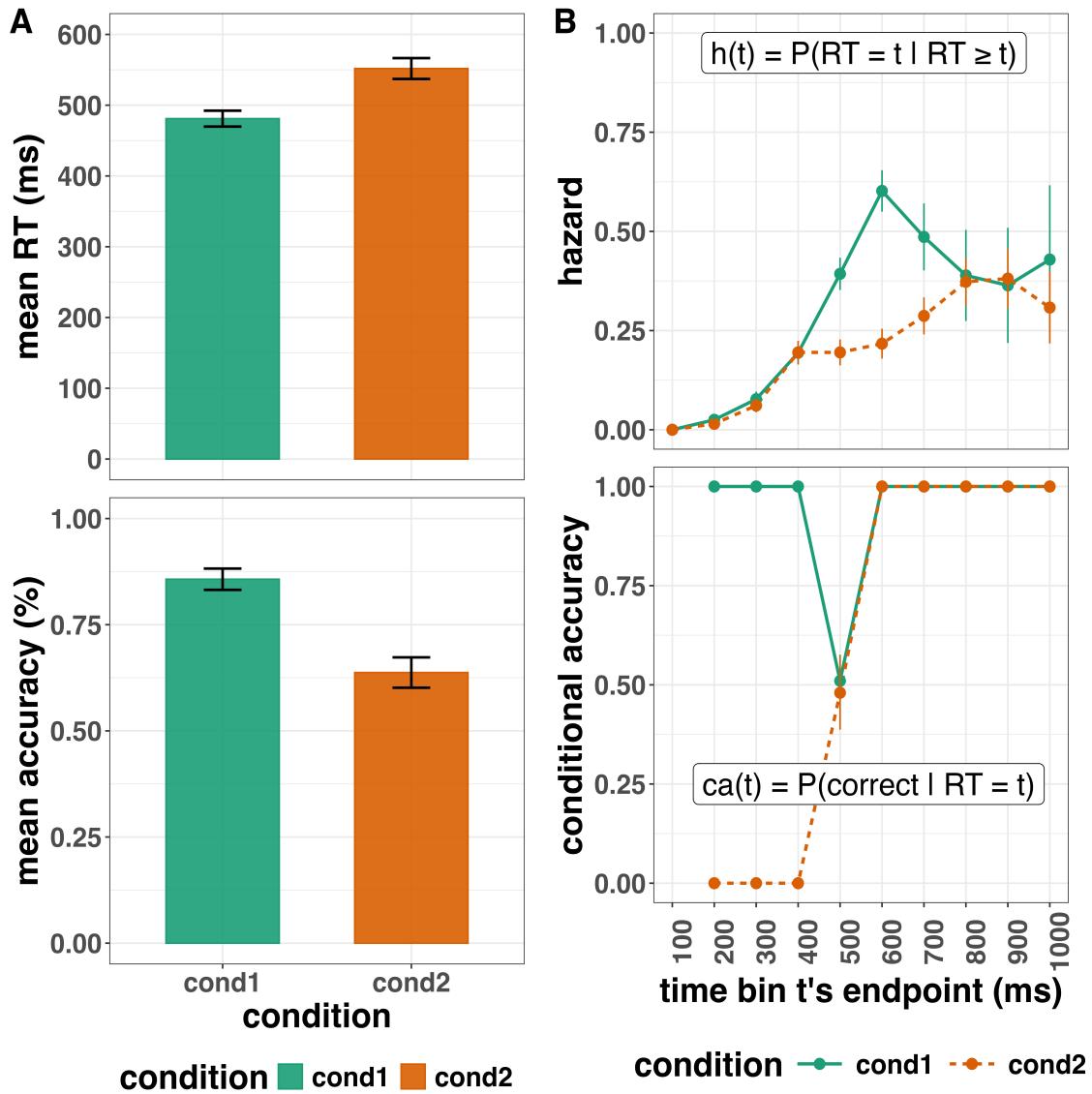


Figure 1. Simulated single-subject data showing mean performance versus distributional (EHA/SAT) analyses. (A) The mean RT (top) and overall accuracy (bottom) for two conditions are plotted. (B) The discrete-time hazard functions (top) and conditional accuracy functions (bottom) are plotted for the same data. The first second after target stimulus onset (time zero) is divided in ten bins of 100 ms ($t = 1$ to 10). The first bin is (0,100], the last bin is (900,1000]. Two hundred trials were simulated in each condition. Note that the hazard and conditional accuracy estimates are plotted at the endpoint of each time bin. The definitions of discrete-time hazard and conditional accuracy are further explained in section 2.1.2. Error bars represent ± 1 standard error of the mean (A) or proportion (B).

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

71 **1.2 Aims**

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

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

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

90 to benefit from using EHA in the future.

91 **1.3 Structure**

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

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

105 **2.1 A brief introduction to event history analysis**

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

114 **2.1.1 Terminology and minimum requirements for EHA.** To avoid possible

115 confusion in terminology used, it is worth noting that EHA is known by various labels,

116 such as survival analysis, hazard analysis, duration analysis, failure-time analysis, and

117 transition analysis (Singer & Willett, 2003). In this paper, we choose to use the term EHA

118 throughout.

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

120 1. define an event of interest that represents a qualitative change - a transition from one

121 discrete state to another - that can be situated in time (e.g., a button press, a

122 saccade onset, a fixation offset, etc.);

123 2. define time point zero in each trial (e.g., target stimulus onset, fixation onset, etc.);

124 3. measure the passage of time between time point zero and event occurrence in discrete

125 or continuous time units in each trial.

126 These minimal requirements are fulfilled by the RT data obtained in single-button

127 detection tasks, where the time-to-response is repeatedly measured in different trials in the

128 same individual. In section A of the Supplemental Material we visualize this and other

129 types of time-to-event data which are typically obtained in discrimination and bistable

130 perception tasks.

131 **2.1.2 Types of EHA.** There are different types of modeling approaches in EHA.

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

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

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

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

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

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

138 1984; Luce, 1991; Van Zandt, 2000). Thus, by trading a bit of temporal resolution for a

lower number of trials, discrete-time methods seem ideal for dealing with typical psychological RT data sets for which there are less than ~200 trials per condition per participant (Panis, Schmidt, et al., 2020). Moreover, as indicated by Allison (2010), learning discrete-time EHA methods first will help in learning continuous-time methods, so it seems like a good starting point.

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

In the context of experimental psychology, it is often (but not always), the case that responses can be classified as correct or incorrect. In those cases, one can also calculate the conditional accuracy function (Figure 1B 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). The conditional accuracy function is also known as the micro-level speed-accuracy tradeoff (SAT) function. We refer to this extended (hazard + conditional accuracy) analysis for choice RT data as EHA/SAT. The definitions of these and other discrete-time functions are given in section B of the Supplemental Material.

164 2.2 Benefits of event history analysis for research in experimental psychology

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

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

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

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

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

208 2.3 Implications for research design in experimental psychology

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

214 Second, since the number of trials per condition are spread across bins, it is

215 important to have a relatively large number of trial repetitions per participant and per
216 condition. Accordingly, experimental designs using this approach typically focus on
217 factorial, within-subject designs, in which a large number of observations are made on a
218 relatively small number of participants (so-called small-*N* designs). This approach
219 emphasizes the precision and reproducibility of data patterns at the individual participant
220 level to increase the inferential validity of the design (Baker et al., 2021; Smith & Little,
221 2018). Note that because statistical power derives both from the number of participants
222 and from the number of repeated measures per participant and condition, small-*N* designs
223 can still achieve what are generally considered acceptable levels of statistical power, if they
224 have a sufficient amount of data overall (Baker et al., 2021; Smith & Little, 2018).

225 Third, the width of each time bin will need to be determined. For instance, in Figure
226 1B we chose 100 ms in an arbitrary manner. In reality, however, bin width will need to be
227 set by considering a number of factors simultaneously. The optimal bin width will depend
228 on (a) the length of the observation period in each trial, (b) the rarity of event occurrence,
229 (c) the number of repeated measures (or trials) per condition per participant, and (d) the
230 shape of the hazard function. Finding an appropriate bin width in a given user case before
231 fitting models will require testing a number of options, when calculating and plotting the
232 descriptive statistics (see section 3.1). The goal is to find the smallest bin width that is
233 supported by the amount of data available. Based on our experience, a bin width of 50 ms
234 is a good starting value when the number of repeated measures is 100 or less. Overly small
235 bin widths will result in erratic hazard functions as many bins will have no events, and
236 thus hazard estimates of zero. Of note, however, is that time bins do not need to have the
237 same width. For example, Panis (2020) used larger bins towards the end of the observation
238 period, as fewer events occurred there.

239

3. Tutorials

240 Tutorials 1a and 1b show how to calculate and plot the descriptive statistics of
241 EHA/SAT when there are one or two independent variables, respectively. Tutorials 2a and
242 2b illustrate how to use Bayesian multilevel modeling to fit hazard and conditional
243 accuracy models, respectively. Tutorials 3a and 3b show how to implement, respectively,
244 multilevel models for hazard and conditional accuracy in the frequentist framework.
245 Tutorial 4 shows how to use simulation and power analysis for planning experiments.
246 Additionally, to further simplify the process for other users, the first two tutorials rely on a
247 set of our own custom functions that make sub-processes easier to automate, such as data
248 wrangling and plotting functions (see section C of the Supplemental Material for a list of
249 the custom functions).

250 The content of the tutorials, in terms of EHA and multilevel regression modelling, is
251 mainly based on Allison (2010), Singer and Willett (2003), McElreath (2020), Heiss (2021),
252 Kurz (2023a), and Kurz (2023b). We used R (Version 4.5.1; R Core Team, 2024) and the
253 R-packages *bayesplot* (Version 1.13.0; Gabry, Simpson, Vehtari, Betancourt, & Gelman,
254 2019), *brms* (Version 2.22.0; Bürkner, 2017, 2018, 2021), *citr* (Version 0.3.2; Aust, 2019),
255 *cmdstanr* (Version 0.9.0.9000; Gabry, Češnovar, Johnson, & Brønner, 2024), *dplyr* (Version
256 1.1.4; Wickham, François, Henry, Müller, & Vaughan, 2023), *forcats* (Version 1.0.0;
257 Wickham, 2023a), *futures* (Bengtsson, 2021b), *ggplot2* (Version 3.5.2; Wickham, 2016),
258 *lme4* (Version 1.1.37; Bates, Mächler, Bolker, & Walker, 2015), *lubridate* (Version 1.9.4;
259 Grolemund & Wickham, 2011), *Matrix* (Version 1.7.3; Bates, Maechler, & Jagan, 2024),
260 *nlme* (Version 3.1.168; Pinheiro & Bates, 2000), *papaja* (Version 0.1.3; Aust & Barth,
261 2024), *patchwork* (Version 1.3.0; Pedersen, 2024), *purrr* (Version 1.0.4; Wickham & Henry,
262 2023), *RColorBrewer* (Version 1.1.3; Neuwirth, 2022), *Rcpp* (Eddelbuettel & Balamuta,
263 2018; Version 1.0.14; Eddelbuettel & François, 2011), *readr* (Version 2.1.5; Wickham,
264 Hester, & Bryan, 2024), *RJ-2021-048* (Bengtsson, 2021a), *rstan* (Version 2.32.7; Stan

265 Development Team, 2024), *standist* (Version 0.0.0.9000; Girard, 2024), *StanHeaders*
266 (Version 2.32.10; Stan Development Team, 2020), *stringr* (Version 1.5.1; Wickham, 2023b),
267 *tibble* (Version 3.3.0; Müller & Wickham, 2023), *tidybayes* (Version 3.0.7; Kay, 2024), *tidyR*
268 (Version 1.3.1; Wickham, Vaughan, & Girlich, 2024), *tidyverse* (Version 2.0.0; Wickham et
269 al., 2019) and *tinylabels* (Version 0.2.5; Barth, 2023) for all reported analyses.

270 **3.1 Tutorial 1a: Calculating descriptive statistics using a life table**

271 **3.1.1 Data wrangling aims.** Our data wrangling procedures serve two related
272 purposes. First, we want to calculate descriptive statistics for each condition in each
273 individual using a life table. A life table includes for each time bin, the risk set (i.e., the
274 number of trials that are event-free at the start of the bin), the number of observed events,
275 and the estimates of the discrete-time hazard probability $h(t)$, survival probability $S(t)$,
276 probability mass $P(t)$, possibly the conditional accuracy $ca(t)$, and their estimated
277 standard errors (se). The definitions of these quantities are provided in section B of the
278 Supplemental Material.

279 Second, we want to produce two different data sets that can each be submitted to
280 different types of inferential modelling approaches. The two types of data structure we
281 label as ‘person-trial’ data and ‘person-trial-bin’ data. The ‘person-trial’ data (Table 1)
282 will be familiar to most researchers who record behavioural responses from participants, as
283 it represents the measured RT and accuracy per trial within an experiment. This data set
284 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.

285 In contrast, the ‘person-trial-bin’ data (Table 2) has a different, more extended
 286 structure, which indicates in which bin a response occurred, if at all, in each trial.
 287 Therefore, the ‘person-trial-bin’ data generates a 0 in each bin until an event occurs and
 288 then it generates a 1 to signal an event has occurred in that bin. This data set is used
 289 when fitting discrete-time hazard models (Tutorials 2a and 3a). It is worth pointing out
 290 that there is no requirement for an event to occur at all (in any bin), as maybe there was
 291 no response on that trial or the event occurred after the time window of interest. Likewise,
 292 when the event occurs in bin 1 there would only be one row of data for that trial in the
 293 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.

294 **3.1.2 A real data wrangling example.** To illustrate how to quickly set up life
 295 tables for calculating the descriptive statistics (functions of discrete time), we use a
 296 published data set on masked response priming from Panis and Schmidt (2016), who were
 297 interested in the temporal dynamics of the effect of prime-target congruency in RT and
 298 accuracy data. In their first experiment, Panis and Schmidt (2016) presented a double
 299 arrow for 94 ms that pointed left or right as the target stimulus with an onset at time
 300 point zero in each trial. Participants had to indicate the direction in which the double
 301 arrow pointed using their corresponding index finger, within 800 ms after target onset.
 302 Response time and accuracy were recorded on each trial. Prime type (blank, congruent,

303 incongruent) and mask type were manipulated across trials (i.e., repeated measures of
 304 time-to-response). Here we focus for each participant on the subset of 220 trials in which
 305 no mask was presented. The 13-ms prime stimulus was a double arrow presented 187 ms
 306 before target onset in the congruent (same direction as target) and incongruent (opposite
 307 direction as target) prime conditions.

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

311 The required column names are as follows:

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

318 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")))
```

319 Next, we can set up the life tables and plot for each condition the discrete-time hazard
 320 function $h(t)$, survivor function $S(t)$, probability mass function $P(t)$, and conditional

accuracy function `ca(t)`. To do so using a functional programming approach, one has to nest the person-trial data within participants using the `group_nest()` function, and supply a user-defined censoring time and bin width to our custom function “`censor()`”, as follows.

```
data_nested <- data_wr %>% group_nest(pid)

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

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

The person-trial-bin oriented data set is created by our custom function `ptb()`, and it

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

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

351 Table 3 shows the life table for condition “blank” (no prime stimulus presented) for
352 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.

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

Descriptive stats for subject 6

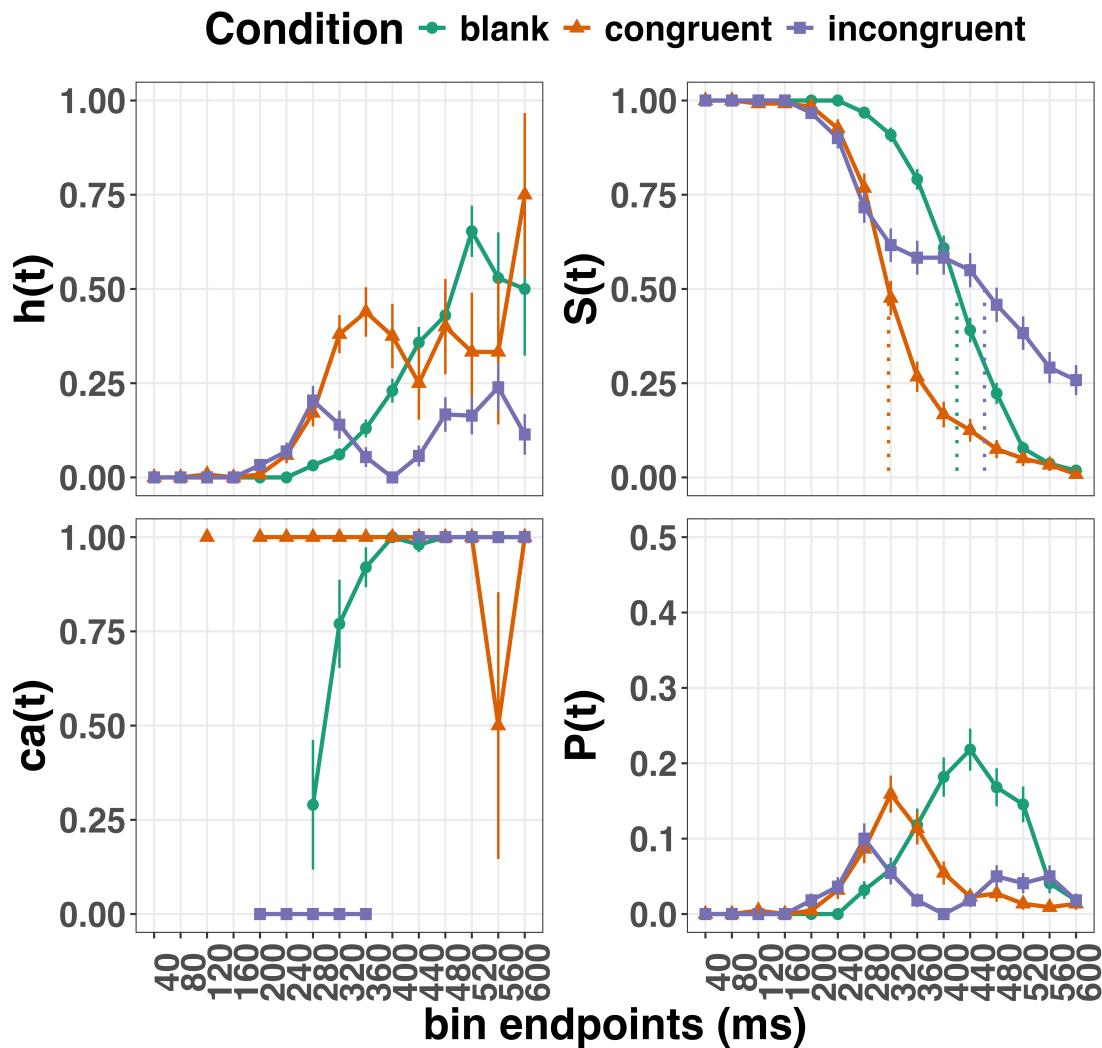


Figure 2. Estimated discrete-time hazard (h), survivor (S), conditional accuracy (ca) and probability mass (P) functions for participant 6. Vertical dotted lines indicate the estimated median RTs. Error bars represent ± 1 standard error of the respective proportion.

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

360 0.17, and 0.20 for the blank, congruent, and incongruent prime conditions, respectively. In

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

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

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

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

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

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

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

368 respectively.

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

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

371 0.36, 0.25, and 0.06 for the blank, congruent, and incongruent prime conditions,

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

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

374 conditions, respectively.

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

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

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

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

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

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

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

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

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

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

385 slower responses reflect primed responses to the target stimulus.

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

392 **3.2 Tutorial 1b: Generalising to a more complex design**

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

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

404 **3.3 Tutorial 2a: Fitting Bayesian hazard models to interval-censored RT data**

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

409 Willett, 2003).

410 In general, when fitting regression models, our lab adopts an estimation approach to
411 multilevel regression (Kruschke & Liddell, 2018; Winter, 2019), which is heavily influenced
412 by the Bayesian framework as suggested by Richard McElreath (Kurz, 2023b; McElreath,
413 2020). We also use a “keep it maximal” approach by specifying a full varying (or random)
414 effects structure (Barr, Levy, Scheepers, & Tily, 2013). This means that wherever possible
415 we include varying intercepts and slopes per participant. To make inferences, we use two
416 main approaches. We compare models of different complexity using information criteria
417 and cross-validation, to evaluate out-of-sample predictive accuracy (McElreath, 2020). We
418 also take the most complex model and evaluate key parameters of interest using point and
419 interval estimates.

420 **3.3.1 Hazard model considerations.** There are several analytic decisions one
421 has to make when fitting a discrete-time hazard model. First, because the first few bins
422 typically contain no responses, one has to select an analysis time window, i.e., a contiguous
423 set of bins for which there is data for each participant. Second, given that the dependent
424 variable (event occurrence) is binary, one has to select a link function (see section D of the
425 Supplemental Material). The cloglog link is preferred over the logit link when events can
426 occur in principle at any time point within a bin, which is the case for RT data (Singer &
427 Willett, 2003). Third, one has to choose whether to treat TIME (i.e., the time bin index t)
428 as a categorical or continuous predictor (see also section E of the Supplemental Material).
429 For example, when you want to know if cloglog-hazard is changing linearly or quadratically
430 over time, you should treat TIME as a continuous predictor. When you are only interested
431 in the effect of covariates on hazard, you can treat TIME as a categorical predictor (i.e., fit
432 an intercept for each bin), in which case you can choose between reference coding and
433 index coding. With reference coding, one defines the variable as a factor and selects one of
434 the k categories as the reference level. Brm() will then construct k-1 indicator variables
435 (see model M1d in Tutorial_2a.Rmd for an example). With index coding, one constructs

436 an index variable that contains integers that correspond to different categories (see models
 437 M0i and M1i below). As explained by McElreath (2020), the advantage of index coding is
 438 that the same prior can be assigned to each level of the index variable, so that each
 439 category has the same prior uncertainty.

440 In the case of a large- N design without repeated measurements, the parameters of a
 441 discrete-time hazard model can be estimated using standard logistic regression software
 442 after expanding the typical person-trial data set into a person-trial-bin data set (Allison,
 443 2010). When there is clustering in the data, as in the case of a small- N design with
 444 repeated measurements, the parameters of a discrete-time hazard model can be estimated
 445 using population-averaged methods (e.g., Generalized Estimating Equations), and Bayesian
 446 or frequentist generalized linear mixed models (Allison, 2010).

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

452 In tutorial_2a.Rmd we fit several Bayesian multilevel models (i.e., generalized linear
 453 mixed models) that differ in complexity to the person-trial-bin oriented data set that we
 454 created in Tutorial 1a. We decided to select the analysis time window (200,600] and the
 455 cloglog link. Below, we shortly discuss two of these models. The person-trial-bin data set is
 456 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)))

```

457 **3.3.2 Prior distributions.** To get the posterior distribution of each model

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

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

469 **3.3.3 Model M0i: A null model with index coding.** When you do not want to
 470 make assumptions about the shape of the hazard function, or its shape is not smooth but
 471 irregular, then you can use a general specification of TIME, i.e., fit one grand intercept per
 472 time bin. In this first baseline or reference model, we use a general specification of TIME
 473 using index coding, and do not include experimental predictors. We call this model “M0i”.
 474 The other model (see section 3.3.4) extends model M0i by including our experimental
 475 predictor prime type.

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

477 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")
```

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

484 **3.3.4 Model M1i: Adding the effects of prime-target congruency.** Previous
 485 research has shown that psychological effects typically change over time (Panis, 2020;
 486 Panis, Moran, et al., 2020; Panis & Schmidt, 2022; Panis et al., 2017; Panis & Wagemans,
 487 2009). In the next model, therefore, we use index coding for both TIME (variable
 488 “timebin”) and the categorical predictor prime-target-congruency (variable “prime”), so
 489 that we get 30 grand intercepts, one for each combination of timebin level and prime level.
 490 Here is the model formula of this model that we call “M1i”.

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

491 Estimating model M1i took about 124 minutes.

492 **3.3.5 Compare the models.** There are two popular strategies to evaluate how
 493 well models will perform in predicting new data on average: Leave-One-Out (LOO)
 494 cross-validation and the Widely Applicable Information Criterion or WAIC (McElreath,
 495 2020). LOO-weights represent the optimal linear combination of models for predictive
 496 performance, with higher weights for models with better out-of-sample predictive
 497 performance. WAIC-weights represent the relative evidence for each model, with higher
 498 weights for models with a better fit while accounting for model complexity (Kurz, 2023a;
 499 McElreath, 2020).

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

500 ## model_M0i model_M1i
501 ##     "0.00"     "1.00"

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

502 ## model_M0i model_M1i
503 ##     "0.00"     "1.00"
```

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

506 **3.3.6 Evaluating parameter estimates in model M1i.** To make causal
 507 inferences from the parameter estimates in model M1i (Frank et al., 2025), we first plot the
 508 densities of the draws from the posterior distributions of its population-level parameters in
 509 Figure 3A, together with point (median) and interval estimates (80% and 95% credible

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

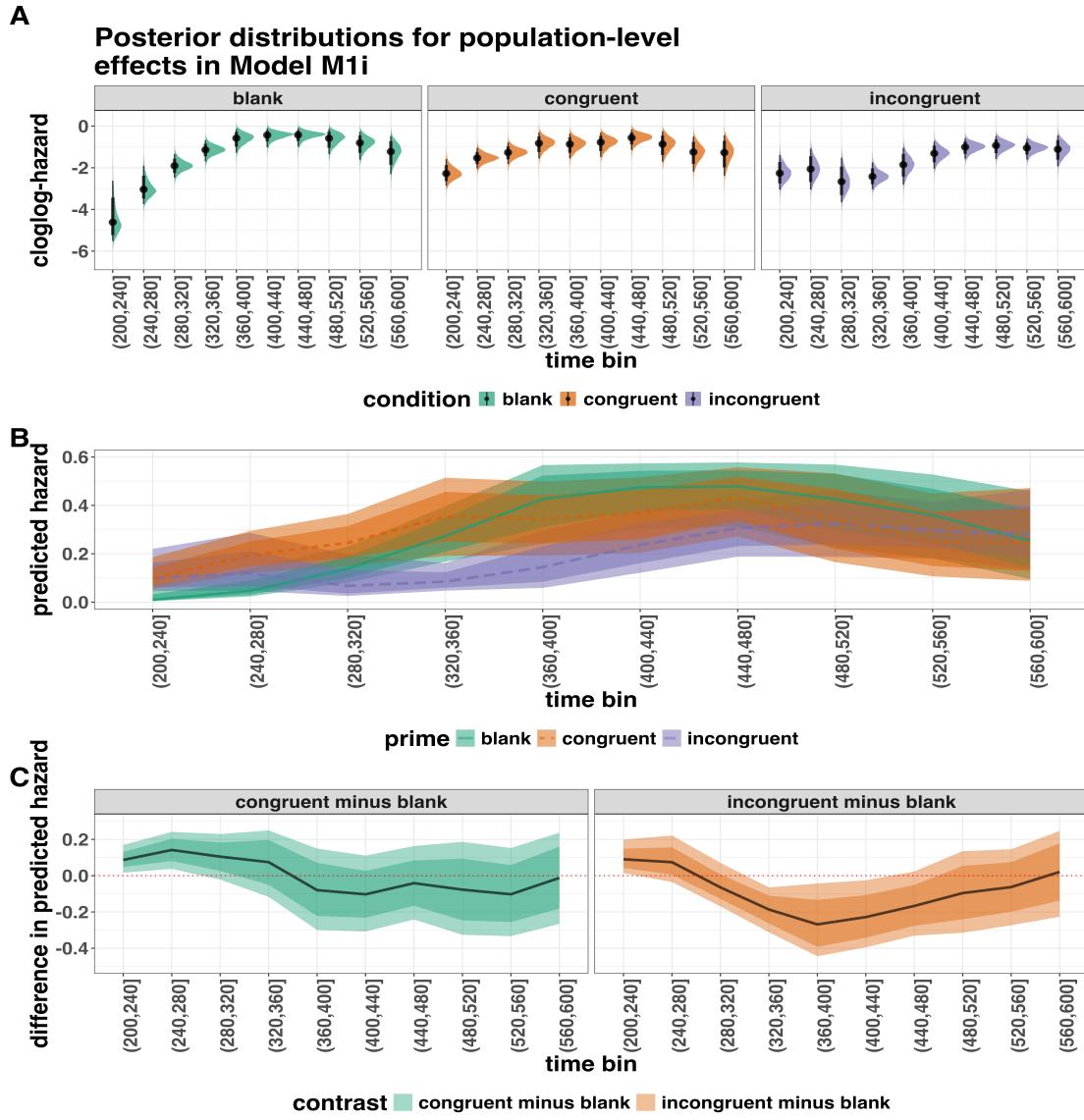


Figure 3. Discrete-time hazard modeling results. (A) Medians and 80/95% credible intervals of the posterior distributions of the population-level parameters of model M1i. (B) Point (median) and 80/95% credible interval summaries of the hazard estimates (expected values of the draws from the posterior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

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].

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

545 **3.4 Tutorial 2b: Fitting Bayesian conditional accuracy models**

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

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

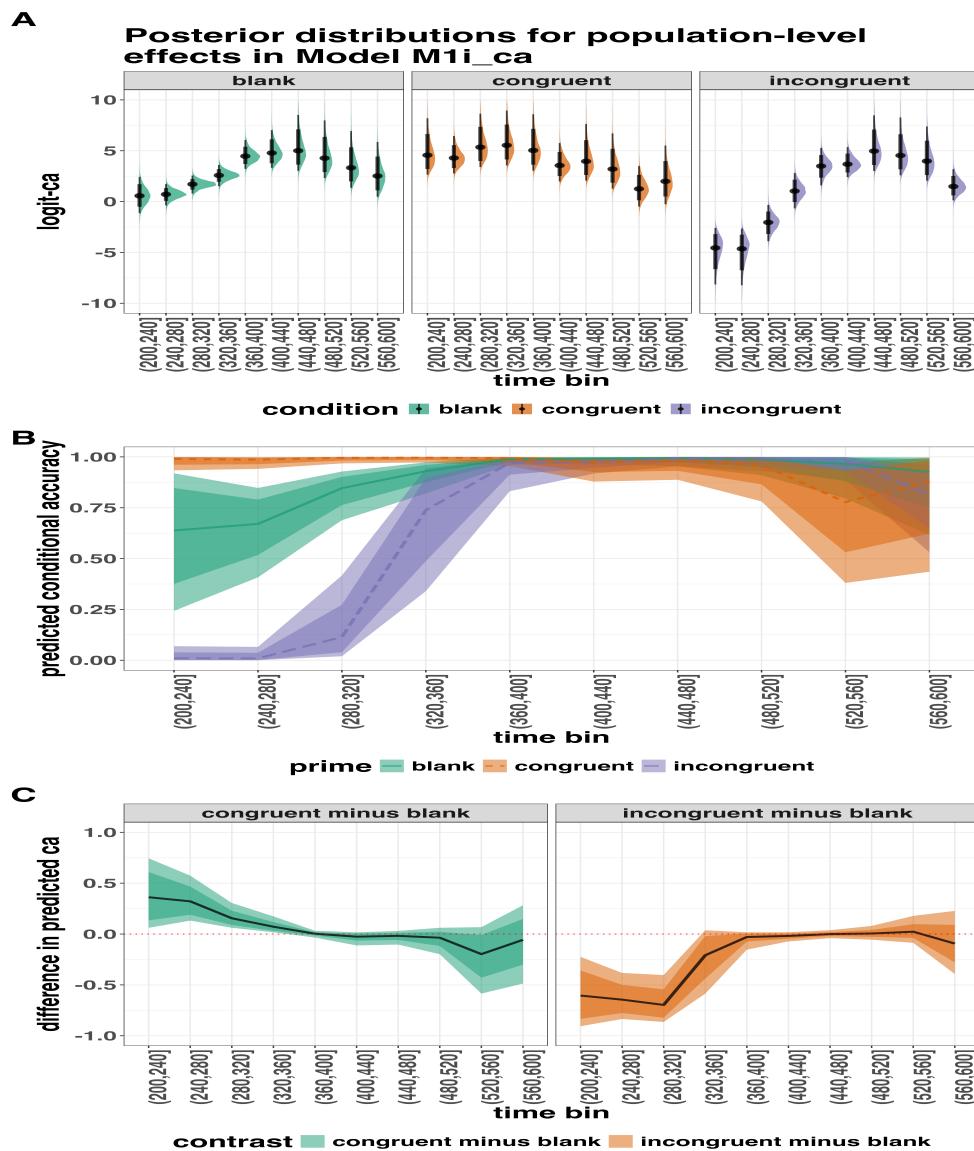


Figure 4. Conditional accuracy modeling results. (A) Medians and 80/95% credible intervals of the posterior distributions of the population-level parameters of model M1i_ca. (B) Point (median) and 80/95% credible interval summaries of the conditional accuracy estimates (expected values of the draws from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

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

563 Based on Figure 4C we see that on the population level congruent primes have a positive
564 effect on the conditional accuracy of emitted responses in time bins (200,240], (240,280],
565 (280,320], and (320,360], relative to the estimates in the baseline condition (blank prime;
566 red dashed lines in Figure 4C). Incongruent primes have a negative effect on the conditional
567 accuracy of emitted responses in the first three time bins, relative to blank primes.

568 Finally, because many researchers will be more familiar with frequentist statistics, we
569 also provide code to fit hazard and conditional accuracy models in the frequentist
570 framework using the R package lme4() (see Tutorial_3a.Rmd and Tutorial_3b.Rmd).

571 3.5 Tutorial 4: Planning

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

573 **3.5.1 Background.** The general approach to planning that we adopt here involves

574 simulating reasonably structured data to help guide what you might be able to expect from
575 your data once you collect it (Gelman, Vehtari, et al., 2020). The basic structure and code
576 follows the examples outlined by Solomon Kurz in his ‘power’ blog posts
577 (<https://solomonkurz.netlify.app/blog/bayesian-power-analysis-part-i/>) and Lisa
578 Debruine’s R package faux{} (<https://debruine.github.io/faux/>) as well as these related
579 papers (DeBruine & Barr, 2021; Pargent, Koch, Kleine, Lermer, & Gaube, 2024).

580 **3.5.2 Basic workflow.** The basic workflow is as follows:

- 581 1. Fit a regression model to existing data.

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

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

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

594 **3.5.3 Fit a regression model and simulate one dataset.** We again use the
595 data from Panis and Schmidt (2016) to provide a worked example. We fit an index coding
596 model on a subset of time bins (six time bins in total) and for two prime conditions
597 (congruent and incongruent). We chose to focus on a subsample of the data to ease the
598 computational burden. We also used a full varying effects structure, with the model
599 formula as follows:

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

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

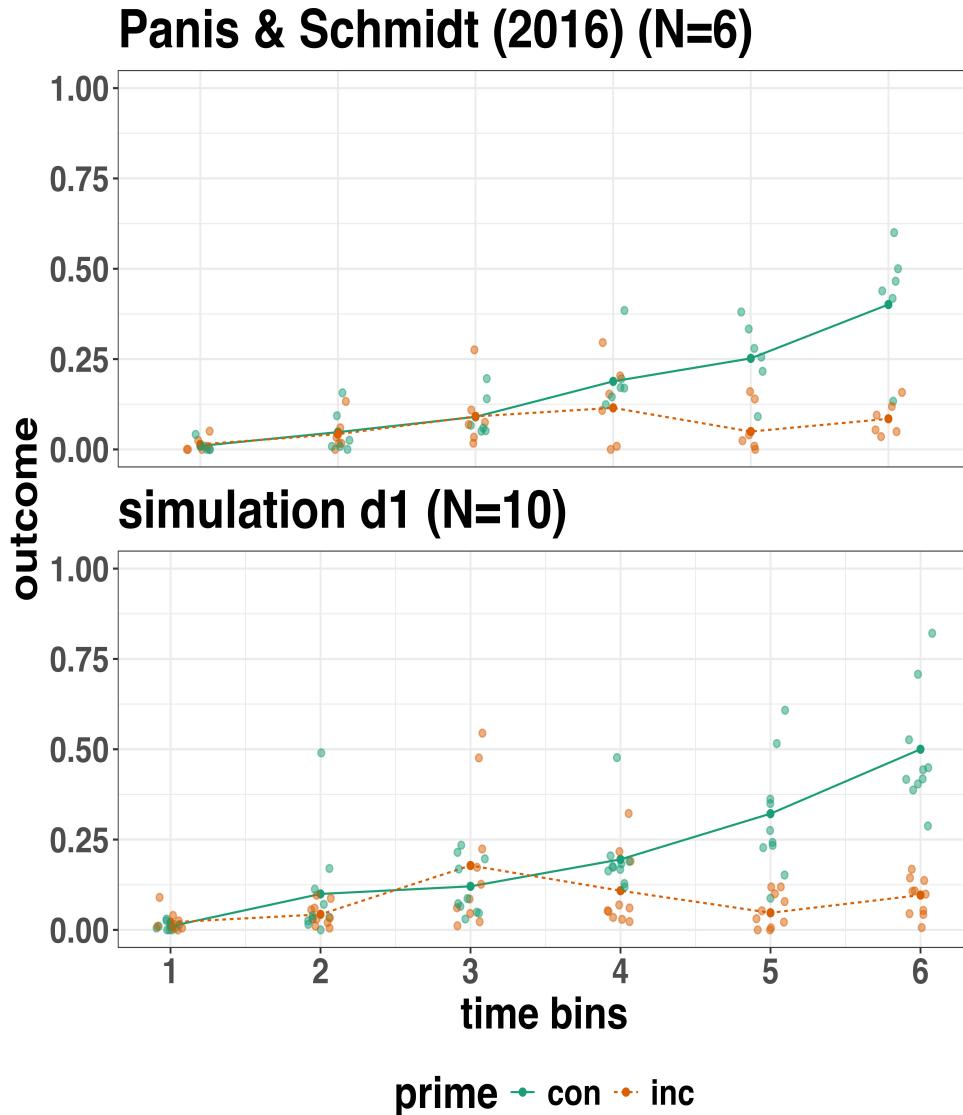


Figure 5. Raw data from Panis and Schmidt (2016) and simulated data from 10 participants.

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

605 Here we use the same data simulation process as used above, but instead of simulating one
 606 dataset, we simulate 1000 datasets per variation in parameter values. Specifically, in
 607 Simulation 1, we vary the number of trials per condition (100, 200, and 400), as well as the
 608 effect size in bin 6. We focus on bin 6 only, in terms of varying the effect size, just to make
 609 things simpler and easier to understand. The effect size observed in bin 6 in this subsample
 610

of data was a 79% reduction in hazard value from the congruent prime (0.401 hazard value) to the incongruent prime condition (0.085 hazard value). In other words, a hazard ratio of 0.21 (e.g., $0.085/0.401 = 0.21$). As a starting point, we chose three effect sizes, which covered a fairly broad range of hazard ratios (0.25, 0.5, 0.75), which correspond to a 75%, 50% and 25% reduction in hazard value as a function of prime condition.

Summary results from Simulation 1 are shown in Figure 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, what fraction of the simulated datasets generated an effect of prime that excludes the criterion mark of zero. We are aware that “power” is not part of a Bayesian analytical workflow, but we choose to include it here, as it is familiar to most researchers in experimental psychology.

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

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

637 calculations might be as follows (trial count = 200 per condition in all cases):

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

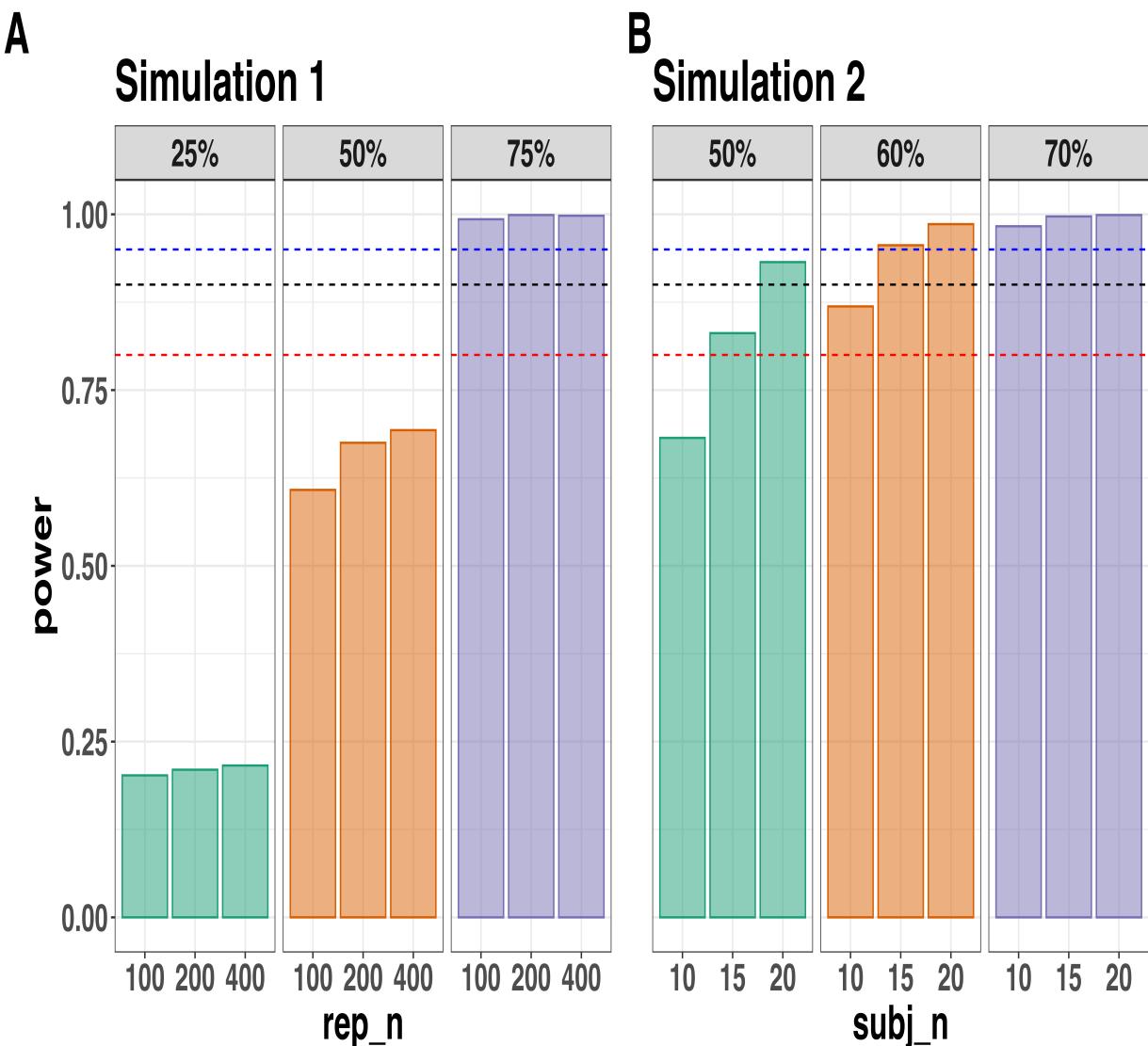


Figure 6. Statistical power across data Simulation 1 (A) and Simulation 2 (B). Power was calculated as the percentage of lower-bound 95% confidence intervals that exclude zero when the difference between prime condition is calculated (congruent - incongruent). In Simulation 1, the effect size was varied 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.

641 **3.5.5 Planning decisions.** Now that we have summarised our simulated data,

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

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

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

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

646 Some considerations might be as follows:

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

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

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

650 personnel?

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

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

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

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

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

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

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

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

659 **But**, and this is an important “but”, unless there are unavoidable reasons, no matter

660 what planning choices we made based on these data simulations, we would not solely rely

661 on data collected from one single study. Instead, we would run a follow-up experiment that

662 replicates and extends the initial result. By doing so, we would aim to avoid the Cult of

663 the Isolated Single Study (Nelder, 1999; Tong, 2019), and thus reduce the reliance on any

664 one type of planning tool, such as a power analysis. Then, we would look for common

665 patterns across two or more experiments, rather than trying to make the case that a single

666 study on its own has sufficient evidential value to hit some criterion mark.

667

4. Discussion

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

682 **4.1 What are the main use-cases of EHA for understanding cognition and brain
683 function?**

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

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

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

704 **4.2 Model complexity versus interpretability**

705 Hazard and conditional accuracy models can quickly become very complex when
706 adding more than one time scale, due to the many possible higher-order interactions. For
707 example, some of the models discussed in Tutorial 2a, which we did not focus on in the
708 main text, contain two time scales as covariates: the passage of time on the within-trial
709 time scale, and the passage of time on the across-trial (or within-experiment) time scale.
710 However, when trials are presented in blocks, and blocks of trials within sessions, and when
711 the experiment comprises a number of sessions, then four time scales can be defined
712 (within-trial, within-block, within-session, and within-experiment). From a theoretical
713 perspective, adding more than one time scale – and their interactions – can be important
714 to capture plasticity and other learning effects that may play out on such longer time
715 scales, and that are probably present in each experiment in general (Schöner & Spencer,
716 2016). From a practical perspective, therefore, some choices need to be made to balance
717 the amount of data that is being collected per participant, condition and across the varying

718 timescales. As one example, if there are several timescales of relevance, then it might be
719 prudent for interpretational purposes to limit the number of experimental predictor
720 variables (conditions). This is of course where planning and data simulation efforts would
721 be important to provide a guide to experimental design choices (see Tutorial 4 and section
722 2.3).

723 4.3 Limitations

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

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

⁷⁴⁴ participant, you can of course also try continuous-time methods (Allison, 2010; Elmer et
⁷⁴⁵ al., 2023).

⁷⁴⁶

5. Conclusions

⁷⁴⁷ Estimating the temporal distributions of RT and accuracy provide a rich source of
⁷⁴⁸ information on the time course of cognitive processing, which have been largely
⁷⁴⁹ undervalued in the history of experimental psychology and cognitive neuroscience. We
⁷⁵⁰ hope that by providing a set of hands-on, step-by-step tutorials, which come with
⁷⁵¹ custom-built and freely available code, researchers will feel more comfortable embracing
⁷⁵² EHA and investigating the shape of empirical hazard functions and the temporal profile of
⁷⁵³ cognitive states. On a broader level, we think that wider adoption of such approaches will
⁷⁵⁴ have a meaningful impact on the inferences drawn from data, as well as the development of
⁷⁵⁵ theories regarding the structure of cognition.

756

Author contributions

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

760

Conflicts of Interest

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

763

Prior versions

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

766

Supplemental Material

767

Disclosures**768 Data, materials, and online resources**

769 Link to public archive:
770 https://github.com/sven-panis/Tutorial_Event_History_Analysis
771 Supplemental Material: Panis_Ramsey_suppl_material.pdf

772 Ethical approval

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

775

References

- 776 Abney, D. H., Fausey, C. M., Suarez-Rivera, C., & Tamis-LeMonda, C. S. (2025).
777 Advancing a temporal science of behavior. *Trends in Cognitive Sciences*.
778 <https://doi.org/10.1016/j.tics.2025.05.010>
- 779 Allison, P. D. (1982). Discrete-Time Methods for the Analysis of Event Histories.
780 *Sociological Methodology*, 13, 61. <https://doi.org/10.2307/270718>
- 781 Allison, P. D. (2010). *Survival analysis using SAS: A practical guide* (2. ed.). Cary, NC:
782 SAS Press.
- 783 Aust, F. (2019). *Citr: 'RStudio' add-in to insert markdown citations*. Retrieved from
784 <https://github.com/crsh/citr>
- 785 Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R*
786 *Markdown*. <https://doi.org/10.32614/CRAN.package.papaja>
- 787 Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., &
788 Andrews, T. J. (2021). Power contours: Optimising sample size and precision in
789 experimental psychology and human neuroscience. *Psychological Methods*, 26(3),
790 295–314. <https://doi.org/10.1037/met0000337>
- 791 Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature*
792 *Reviews Neuroscience*, 22(6), 359–371. <https://doi.org/10.1038/s41583-021-00448-6>
- 793 Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for
794 confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*,
795 68(3), 10.1016/j.jml.2012.11.001. <https://doi.org/10.1016/j.jml.2012.11.001>
- 796 Barth, M. (2023). *tinylabes: Lightweight variable labels*. Retrieved from
797 <https://cran.r-project.org/package=tinylabes>
- 798 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects
799 models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
800 <https://doi.org/10.18637/jss.v067.i01>
- 801 Bates, D., Maechler, M., & Jagan, M. (2024). *Matrix: Sparse and dense matrix classes and*

- 802 methods. Retrieved from <https://Matrix.R-forge.R-project.org>
- 803 Bengtsson, H. (2021a). A unifying framework for parallel and distributed processing in r
804 using futures. *The R Journal*, 13(2), 208–227. <https://doi.org/10.32614/RJ-2021-048>
- 805 Bengtsson, H. (2021b). futures: A unifying framework for parallel and distributed
806 processing in r using futures. *The R Journal*, 13(2), 208–227.
807 <https://doi.org/10.32614/RJ-2021-048>
- 808 Berger, A., & Kiefer, M. (2021). Comparison of Different Response Time Outlier Exclusion
809 Methods: A Simulation Study. *Frontiers in Psychology*, 12, 675558.
810 <https://doi.org/10.3389/fpsyg.2021.675558>
- 811 Blossfeld, H.-P., & Rohwer, G. (2002). *Techniques of event history modeling: New*
812 *approaches to causal analysis*, 2nd ed (pp. x, 310). Mahwah, NJ, US: Lawrence
813 Erlbaum Associates Publishers.
- 814 Bloxom, B. (1984). Estimating response time hazard functions: An exposition and
815 extension. *Journal of Mathematical Psychology*, 28(4), 401–420.
816 [https://doi.org/10.1016/0022-2496\(84\)90008-7](https://doi.org/10.1016/0022-2496(84)90008-7)
- 817 Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in
818 psychology are heterogeneous. *Journal of Experimental Psychology: General*, 148(4),
819 601–618. <https://doi.org/10.1037/xge0000558>
- 820 Box-Steffensmeier, J. M. (2004). Event history modeling: A guide for social scientists.
821 Cambridge: University Press.
- 822 Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan.
823 *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- 824 Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms.
825 *The R Journal*, 10(1), 395–411. <https://doi.org/10.32614/RJ-2018-017>
- 826 Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal*
827 *of Statistical Software*, 100(5), 1–54. <https://doi.org/10.18637/jss.v100.i05>
- 828 DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through

- 829 Data Simulation. *Advances in Methods and Practices in Psychological Science*, 4(1),
830 2515245920965119. <https://doi.org/10.1177/2515245920965119>
- 831 Eddelbuettel, D., & Balamuta, J. J. (2018). Extending R with C++: A Brief Introduction
832 to Rcpp. *The American Statistician*, 72(1), 28–36.
833 <https://doi.org/10.1080/00031305.2017.1375990>
- 834 Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal
835 of Statistical Software*, 40(8), 1–18. <https://doi.org/10.18637/jss.v040.i08>
- 836 Elmer, T., Van Duijn, M. A. J., Ram, N., & Bringmann, L. F. (2023). Modeling
837 categorical time-to-event data: The example of social interaction dynamics captured
838 with event-contingent experience sampling methods. *Psychological Methods*.
839 <https://doi.org/10.1037/met0000598>
- 840 Frank, M. C., Braginsky, M., Cachia, J., Coles, N. A., Hardwicke, T. E., Hawkins, R. D.,
841 ... Williams, R. (2025). *Experimentology: An Open Science Approach to Experimental
842 Psychology Methods*. Stanford University. <https://doi.org/10.25936/3JP6-5M50>
- 843 Gabry, J., Češnovar, R., Johnson, A., & Broder, S. (2024). *Cmdstanr: R interface to
844 'CmdStan'*. Retrieved from <https://github.com/stan-dev/cmdstanr>
- 845 Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization
846 in bayesian workflow. *J. R. Stat. Soc. A*, 182, 389–402.
847 <https://doi.org/10.1111/rssa.12378>
- 848 Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and Other Stories.
849 [https://www.cambridge.org/highereducation/books/regression-and-other-
850 stories/DD20DD6C9057118581076E54E40C372C](https://www.cambridge.org/highereducation/books/regression-and-other-stories/DD20DD6C9057118581076E54E40C372C); Cambridge University Press.
851 <https://doi.org/10.1017/9781139161879>
- 852 Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., ...
853 Modrák, M. (2020). *Bayesian Workflow*. arXiv.
854 <https://doi.org/10.48550/arXiv.2011.01808>
- 855 Girard, J. (2024). *Standist: What the package does (one line, title case)*. Retrieved from

- 856 https://github.com/jmgirard/standist
- 857 Gromelund, G., & Wickham, H. (2011). Dates and times made easy with lubridate.
- 858 *Journal of Statistical Software*, 40(3), 1–25. Retrieved from
- 859 <https://www.jstatsoft.org/v40/i03/>
- 860 Heiss, A. (2021, November 10). A Guide to Correctly Calculating Posterior Predictions
- 861 and Average Marginal Effects with Multilevel Bayesian Models.
- 862 <https://doi.org/10.59350/wbn93-edb02>
- 863 Holden, J. G., Van Orden, G. C., & Turvey, M. T. (2009). Dispersion of response times
- 864 reveals cognitive dynamics. *Psychological Review*, 116(2), 318–342.
- 865 <https://doi.org/10.1037/a0014849>
- 866 Hosmer, D. W., Lemeshow, S., & May, S. (2011). *Applied Survival Analysis: Regression*
- 867 *Modeling of Time to Event Data* (2nd ed). Hoboken: John Wiley & Sons.
- 868 Kantowitz, B. H., & Pachella, R. G. (2021). The Interpretation of Reaction Time in
- 869 Information-Processing Research 1. *Human Information Processing*, 41–82.
- 870 <https://doi.org/10.4324/9781003176688-2>
- 871 Kay, M. (2024). *tidybayes: Tidy data and geoms for Bayesian models*.
- 872 <https://doi.org/10.5281/zenodo.1308151>
- 873 Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing,
- 874 estimation, meta-analysis, and power analysis from a Bayesian perspective.
- 875 *Psychonomic Bulletin & Review*, 25(1), 178–206.
- 876 <https://doi.org/10.3758/s13423-016-1221-4>
- 877 Kurz, A. S. (2023a). *Applied longitudinal data analysis in brms and the tidyverse* (version
- 878 0.0.3). Retrieved from <https://bookdown.org/content/4253/>
- 879 Kurz, A. S. (2023b). *Statistical rethinking with brms, ggplot2, and the tidyverse: Second*
- 880 *edition* (version 0.4.0). Retrieved from <https://bookdown.org/content/4857/>
- 881 Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, 8(1), 33267.
- 882 <https://doi.org/10.1525/collabra.33267>

- 883 Landes, J., Engelhardt, S. C., & Pelletier, F. (2020). An introduction to event history
884 analyses for ecologists. *Ecosphere*, 11(10), e03238. <https://doi.org/10.1002/ecs2.3238>
- 885 Lougheed, J. P., Benson, L., Cole, P. M., & Ram, N. (2019). Multilevel survival analysis:
886 Studying the timing of children's recurring behaviors. *Developmental Psychology*,
887 55(1), 53–65. <https://doi.org/10.1037/dev0000619>
- 888 Luce, R. D. (1991). *Response times: Their role in inferring elementary mental organization*
889 (1. issued as paperback). Oxford: Univ. Press.
- 890 McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and*
891 *STAN* (2nd ed.). New York: Chapman and Hall/CRC.
892 <https://doi.org/10.1201/9780429029608>
- 893 Mills, M. (2011). *Introducing Survival and Event History Analysis*. 1 Oliver's Yard, 55 City
894 Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd.
895 <https://doi.org/10.4135/9781446268360>
- 896 Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. Retrieved from
897 <https://CRAN.R-project.org/package=tibble>
- 898 Nelder, J. A. (1999). From Statistics to Statistical Science. *Journal of the Royal Statistical*
899 *Society. Series D (The Statistician)*, 48(2), 257–269. Retrieved from
900 <https://www.jstor.org/stable/2681191>
- 901 Neuwirth, E. (2022). *RColorBrewer: ColorBrewer palettes*. Retrieved from
902 <https://CRAN.R-project.org/package=RColorBrewer>
- 903 Panis, S. (2020). How can we learn what attention is? Response gating via multiple direct
904 routes kept in check by inhibitory control processes. *Open Psychology*, 2(1), 238–279.
905 <https://doi.org/10.1515/psych-2020-0107>
- 906 Panis, S., Moran, R., Wolkersdorfer, M. P., & Schmidt, T. (2020). Studying the dynamics
907 of visual search behavior using RT hazard and micro-level speed–accuracy tradeoff
908 functions: A role for recurrent object recognition and cognitive control processes.
909 *Attention, Perception, & Psychophysics*, 82(2), 689–714.

- 910 https://doi.org/10.3758/s13414-019-01897-z
- 911 Panis, S., Schmidt, F., Wolkersdorfer, M. P., & Schmidt, T. (2020). Analyzing Response
912 Times and Other Types of Time-to-Event Data Using Event History Analysis: A Tool
913 for Mental Chronometry and Cognitive Psychophysiology. *I-Perception*, 11(6),
914 2041669520978673. https://doi.org/10.1177/2041669520978673
- 915 Panis, S., & Schmidt, T. (2016). What Is Shaping RT and Accuracy Distributions? Active
916 and Selective Response Inhibition Causes the Negative Compatibility Effect. *Journal of*
917 *Cognitive Neuroscience*, 28(11), 1651–1671. https://doi.org/10.1162/jocn_a_00998
- 918 Panis, S., & Schmidt, T. (2022). When does “inhibition of return” occur in spatial cueing
919 tasks? Temporally disentangling multiple cue-triggered effects using response history
920 and conditional accuracy analyses. *Open Psychology*, 4(1), 84–114.
921 https://doi.org/10.1515/psych-2022-0005
- 922 Panis, S., Torfs, K., Gillebert, C. R., Wagemans, J., & Humphreys, G. W. (2017).
923 Neuropsychological evidence for the temporal dynamics of category-specific naming.
924 *Visual Cognition*, 25(1-3), 79–99. https://doi.org/10.1080/13506285.2017.1330790
- 925 Panis, S., & Wagemans, J. (2009). Time-course contingencies in perceptual organization
926 and identification of fragmented object outlines. *Journal of Experimental Psychology:*
927 *Human Perception and Performance*, 35(3), 661–687.
928 https://doi.org/10.1037/a0013547
- 929 Pargent, F., Koch, T. K., Kleine, A.-K., Lermer, E., & Gaube, S. (2024). A Tutorial on
930 Tailored Simulation-Based Sample-Size Planning for Experimental Designs With
931 Generalized Linear Mixed Models. *Advances in Methods and Practices in Psychological*
932 *Science*, 7(4), 25152459241287132. https://doi.org/10.1177/25152459241287132
- 933 Pedersen, T. L. (2024). *Patchwork: The composer of plots*. Retrieved from
934 https://patchwork.data-imaginist.com
- 935 Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-PLUS*. New York:
936 Springer. https://doi.org/10.1007/b98882

- 937 R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna,
938 Austria: R Foundation for Statistical Computing. Retrieved from
939 <https://www.R-project.org/>
- 940 Schöner, G., & Spencer, J. P. (2016). *Dynamic thinking: A primer on dynamic field theory*.
941 New York, NY: Oxford University Press.
942 <https://doi.org/10.1093/acprof:oso/9780199300563.001.0001>
- 943 Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling
944 Change and Event Occurrence*. Oxford, New York: Oxford University Press.
- 945 Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design.
946 *Psychonomic Bulletin & Review*, 25(6), 2083–2101.
947 <https://doi.org/10.3758/s13423-018-1451-8>
- 948 Stan Development Team. (2020). *StanHeaders: Headers for the R interface to Stan*.
949 Retrieved from <https://mc-stan.org/>
- 950 Stan Development Team. (2024). *RStan: The R interface to Stan*. Retrieved from
951 <https://mc-stan.org/>
- 952 Stoolmiller, M. (2015). *An Introduction to Using Multivariate Multilevel Survival Analysis
953 to Study Coercive Family Process* (Vol. 1; T. J. Dishion & J. Snyder, Eds.). Oxford
954 University Press. <https://doi.org/10.1093/oxfordhb/9780199324552.013.27>
- 955 Stoolmiller, M., & Snyder, J. (2006). Modeling heterogeneity in social interaction processes
956 using multilevel survival analysis. *Psychological Methods*, 11(2), 164–177.
957 <https://doi.org/10.1037/1082-989X.11.2.164>
- 958 Teachman, J. D. (1983). Analyzing social processes: Life tables and proportional hazards
959 models. *Social Science Research*, 12(3), 263–301.
960 [https://doi.org/10.1016/0049-089X\(83\)90015-7](https://doi.org/10.1016/0049-089X(83)90015-7)
- 961 Tong, C. (2019). Statistical Inference Enables Bad Science; Statistical Thinking Enables
962 Good Science. *The American Statistician*, 73(sup1), 246–261.
963 <https://doi.org/10.1080/00031305.2018.1518264>

- 964 Townsend, J. T. (1990). Truth and consequences of ordinal differences in statistical
965 distributions: Toward a theory of hierarchical inference. *Psychological Bulletin*, 108(3),
966 551–567. <https://doi.org/10.1037/0033-2909.108.3.551>
- 967 Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin &
968 Review*, 7(3), 424–465. <https://doi.org/10.3758/BF03214357>
- 969 Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics.
970 *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)
- 971 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New
972 York. Retrieved from <https://ggplot2.tidyverse.org>
- 973 Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors)*.
974 Retrieved from <https://forcats.tidyverse.org/>
- 975 Wickham, H. (2023b). *Stringr: Simple, consistent wrappers for common string operations*.
976 Retrieved from <https://stringr.tidyverse.org>
- 977 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ...
978 Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43),
979 1686. <https://doi.org/10.21105/joss.01686>
- 980 Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A
981 grammar of data manipulation*. Retrieved from <https://dplyr.tidyverse.org>
- 982 Wickham, H., & Henry, L. (2023). *Purrr: Functional programming tools*. Retrieved from
983 <https://purrr.tidyverse.org/>
- 984 Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data*. Retrieved
985 from <https://readr.tidyverse.org>
- 986 Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data*. Retrieved from
987 <https://tidyr.tidyverse.org>
- 988 Winter, B. (2019). *Statistics for Linguists: An Introduction Using R*. New York:
989 Routledge. <https://doi.org/10.4324/9781315165547>
- 990 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>