

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 analysing 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 duration. The ability to reveal
17 finer-grained, “temporal states” of cognitive processes can have important consequences for
18 theory development by qualitatively changing the key inferences that are drawn from
19 psychological data. Luckily, well-established analytical approaches, such as event history
20 analysis (EHA), are able to evaluate the detailed shape of time-to-event distributions, and
21 thus characterise the time course of psychological states. One barrier to wider use of EHA,
22 however, is that the analytical workflow is typically more time-consuming and complex
23 than orthodox approaches. To help achieve broader uptake of EHA, in this paper we
24 outline a set of tutorials that detail one distributional method known as discrete-time
25 EHA. We touch upon several key aspects of the workflow, such as how to process raw data
26 and specify regression models, and we also consider the implications for experimental
27 design, as well as how to manage inter-individual differences. We finish the article by
28 considering the benefits of the approach for understanding psychological states, as well as
29 the limitations and future directions of this work. Finally, the project is written in R and
30 freely available, which means the approach can easily be adapted to other data sets.

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

33 Word count: 13574

34

1. Introduction

35 1.1 Motivation and background context: Comparing means versus 36 distributional shapes

37 In experimental psychology, it is standard practice to analyse response times (RTs),
38 saccade latencies, and fixation durations by calculating average performance across a series
39 of trials. Such mean-average comparisons have been the workhorse of experimental
40 psychology over the last century, and have had a substantial impact on theory development
41 as well as our understanding of the structure of cognition and brain function. However,
42 differences in mean RT conceal important pieces of information, such as when an
43 experimental effect starts, how long it lasts, how it evolves with increasing waiting time,
44 and whether its onset is time-locked to other events (Panis, 2020; Panis, Moran,
45 Wolkersdorfer, & Schmidt, 2020; Panis & Schmidt, 2016, 2022; Panis, Torfs, Gillebert,
46 Wagemans, & Humphreys, 2017; Panis & Wagemans, 2009; Wolkersdorfer, Panis, &
47 Schmidt, 2020). Such information is useful not only for the interpretation of experimental
48 effects under investigation, but also for cognitive psychophysiology and computational
49 model selection (Panis, Schmidt, Wolkersdorfer, & Schmidt, 2020).

50 As a simple illustration, Figure 1 shows the results of several simulated RT data sets,
51 which show how mean-average comparisons between two conditions can conceal the shape
52 of the underlying RT distributions. For instance, in examples 1-3, mean RT is always
53 comparable between two conditions, while the distributions differ (Figure 1, left). In
54 contrast, in examples 4-6, mean RT is lower in condition 2 compared to condition 1, but
55 the RT distributions differ in each case (Figure 1, right). Therefore, a comparison of means
56 would lead to a similar conclusion in examples 1-3, as well as examples 4-6, whereas a
57 comparison of the distributions would lead to a different conclusion in every case.

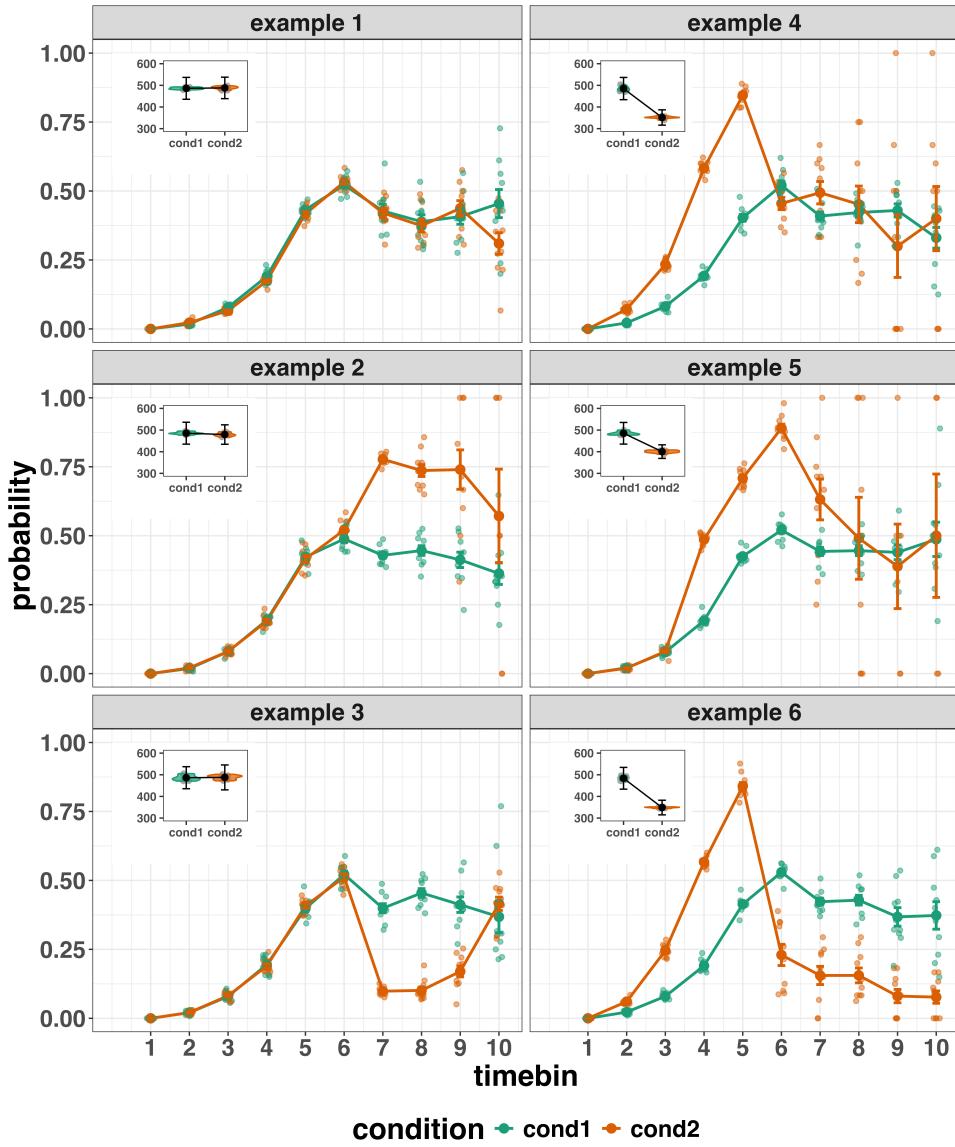


Figure 1. Means versus distributional shapes for six different simulated data set examples.

The first second after stimulus onset is divided in ten bins of 100 ms. Timebin indicates the bin rank. The first bin is (0,100], the last bin is (900,1000]. For our purposes here, it is enough to know that the distributions plotted represent the probability of an event occurring in that timebin, given that it has not yet occurred. Insets show mean response time per condition.

59 data across trials, a distributional approach offers the possibility to reveal the time course
60 of psychological states. As such, the approach permits different kinds of questions to be
61 asked, different inferences to be made, and it holds the potential to discriminate between
62 different theoretical accounts of psychological and/or brain-based processes. For example,
63 the distributions in example 4 show that the effect starts between 100 and 200 ms (in
64 timebin 2) and is gone when the waiting time reaches 500 ms or more. In contrast, in
65 example 5, the effect starts around 300 ms and is gone by 700 ms. And in example 6, the
66 effect reverses between 500 and 600 ms. What kind of theory or theories could account for
67 such effects? Are there new auxiliary assumptions that theories need to adopt? And are
68 there new experiments that need to be performed to test the novel predictions that follow
69 from these analyses? As we show later using published examples, for many psychological
70 questions, such “temporal states” information can be theoretically meaningful by leading to
71 more fine-grained understanding of psychological processes, as well as adding a relatively
72 under-used dimension – the passage of time – to the theory building toolkit.

73 From a historical perspective, it is worth noting that the development of analytical
74 tools that can estimate or predict whether and when events will occur is not a new
75 endeavour. Indeed, hundreds of years ago, analytical methods were developed to predict
76 the duration of time until people died (e.g., Halley, 1693; Makeham, 1860). The same logic
77 has been applied to psychological time-to-event data, as previously demonstrated (Panis,
78 Schmidt, et al., 2020).

79 1.2 Aims and structure of the paper

80 In this paper, we focus on a distributional method for time-to-event data known as
81 discrete-time Event History Analysis (EHA), a.k.a. survival analysis, hazard analysis,
82 duration analysis, failure-time analysis, and transition analysis (Singer & Willett, 2003).
83 We hope to show the added value of EHA for knowledge and theory building in cognitive
84 psychology and related areas of research, such as cognitive neuroscience. Most importantly,

85 we provide tutorials that provide step-by-step code and instructions in the hope that we
86 can enable others to use EHA in a more routine, efficient and effective manner.

87 We first provide a brief overview of EHA to orient the reader to the basic concepts
88 that we will use throughout the paper. However, this will remain relatively short, as this
89 has been covered in detail before (Allison, 1982, 2010; Singer & Willett, 2003). Indeed, our
90 primary aim here is to introduce the set of tutorials, which explain **how** to do such
91 analyses, rather than repeat in any detail **why** you may do them.

92 We provide seven different tutorials, which are written in the R programming
93 language and publicly available on our Github page ([https://github.com/sven-panis/
94 Tutorial_Event_History_Analysis](https://github.com/sven-panis/Tutorial_Event_History_Analysis)), along with all of the other code and material
95 associated with the project. The tutorials provide hands-on, concrete examples of key parts
96 of the analytical process, so that others can apply EHA to their own time-to-event data.
97 Each tutorial is provided as an RMarkdown file, so that others can download and adapt
98 the code to fit their own purposes. Additionally, each tutorial is made available as a .html
99 file, so that it can be viewed by any web browser, and thus available to those that do not
100 use R. Finally, the manuscript itself is written in R using the papaja package (Aust &
101 Barth, 2024), which makes it computationally reproducible, in terms of the underlying data
102 and figures.

103 In Tutorial 1a, we illustrate how to process or “wrangle” a previously published RT +
104 accuracy data set to calculate descriptive statistics when there is one independent variable.
105 The descriptive statistics are plotted, and we comment on their interpretation. In Tutorial
106 1b we provide a generalisation of this approach to illustrate how one can calculate the
107 descriptive statistics when using a more complex design, such as when there are two
108 independent variables.

109 In Tutorial 2a, we illustrate how one can fit Bayesian multilevel regression models to
110 RT data using the R package brms. We also perform prior predictive checks, compare

models, and interpret the plots of the predicted hazard functions for the selected model, and the posterior distributions of our contrasts of interest. In Tutorial 2b we fit Bayesian multilevel regression models to *timed* accuracy data to perform a micro-level speed-accuracy tradeoff (SAT) analysis, which complements the EHA of RT data for choice RT data.

In Tutorial 3a, we shortly illustrate how to fit similar multilevel regression models for RT data in a frequentist framework using the R package lme4. We then briefly compare and contrast these inferential frameworks when applied to EHA. In Tutorial 3b, we illustrate how to perform the SAT analysis in a frequentist framework.

In tutorial 4, we illustrate one approach to planning how much data to collect in an experiment using EHA. We use data simulation techniques to vary sample size and trial count per condition until a certain degree of statistical power or precision is reached.

In summary, even though EHA is a widely used statistical tool and there already exist many excellent reviews (e.g., Blossfeld & Rohwer, 2002; Box-Steffensmeier, 2004; Hosmer, Lemeshow, & May, 2011; Teachman, 1983) and tutorials (e.g., Allison, 2010; Landes, Engelhardt, & Pelletier, 2020) on its general use-cases, we are not aware of any tutorials that are aimed at psychological time-to-event data, and which provide worked examples of the key data processing and multilevel regression modelling steps. Therefore, our ultimate goal is twofold: first, we want to convince readers of the many benefits of using EHA when dealing with time-to-event data with a focus on psychological time-to-event data, and second, we want to provide a set of practical tutorials, which provide step-by-step instructions on how you actually perform a discrete-time EHA on time-to-event data such as RT data, as well as a complementary discrete-time SAT analysis on timed accuracy data.

134 **2. A brief introduction to event history analysis**

135 We recommend several excellent textbooks for a comprehensive background context
136 to EHA (Allison, 2010; Singer & Willett, 2003) and for a more general introduction to
137 understanding regression equations (Gelman, Hill, & Vehtari, 2020; Winter, 2019). Our
138 focus here is not on providing a detailed account of the underlying regression equations,
139 since this topic has been comprehensively covered many times before. Instead, we want to
140 provide an intuition regarding how EHA works in general, as well as in the context of
141 experimental psychology. As such, we only supply regression equations in section D of the
142 Supplemental Material.

143 **2.1 Basic features of event history analysis**

144 To apply EHA, one must be able to:

- 145 1. define an event of interest that represents a qualitative change that can be situated in
146 time (e.g., a button press, a saccade onset, a fixation offset, etc.);
- 147 2. define time point zero (e.g., target stimulus onset, fixation onset, etc.);
- 148 3. measure the passage of time between time point zero and event occurrence in discrete
149 or continuous time units.

150 In EHA, the definition of hazard and the type of models employed depend on
151 whether one is using continuous or discrete time units. Since our focus here is on hazard
152 models that use discrete time units, we describe that approach. After dividing time in
153 discrete, contiguous time bins indexed by t (e.g., $t = 1:10$ timebins), let RT be a discrete
154 random variable denoting the rank of the time bin in which a particular person's response
155 occurs in a particular experimental condition. For example, the first response might occur
156 at 546 ms and it would be in timebin 6 (any RTs from 501 ms to 600 ms).

157 Discrete-time EHA focuses on the discrete-time hazard function of event occurrence
158 and the discrete-time survivor function (Figure 2). The equations that define both of these
159 functions are reported in part A of the Supplemental Material. The discrete-time hazard
160 function gives you, for each time bin, the probability that the event occurs (sometime) in
161 bin t , given that the event does not occur in previous bins. In other words, it reflects the
162 instantaneous likelihood that the event occurs in the current bin, given that it has not yet
163 occurred in the past, i.e., in one of the prior bins. In contrast, the discrete-time survivor
164 function cumulates the bin-by-bin risks of event *nonoccurrence* to obtain the survival
165 probability, the probability that the event occurs after bin t . In other words, the survivor
166 function gives you for each time bin the likelihood that the event occurs in the future, i.e.,
167 in one of the subsequent timebins.

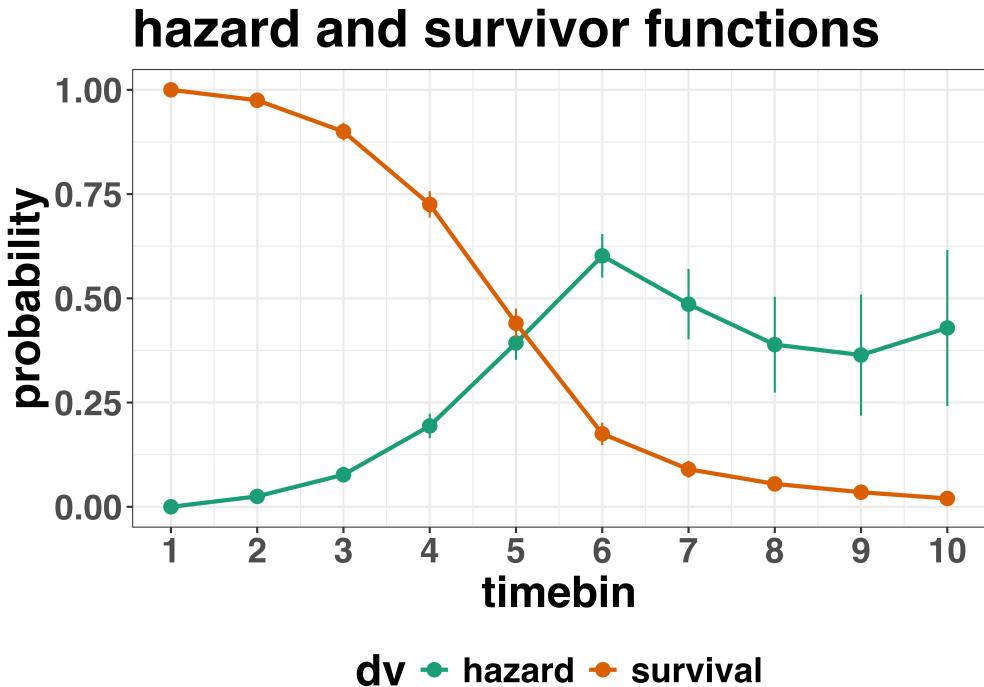


Figure 2. Discrete-time hazard and survivor functions. Discrete time-to-event data were simulated for 200 trials of 1 experimental condition. Error bars represent ± 1 standard error of the respective proportion. While the hazard function is the vehicle for inferring the time course of cognitive processes, the survival probability $S(t-1)$ can help to qualify or provide context to the interpretation of the hazard probability $h(t)$. For example, the high hazard of $.60 = h(t=6)$ is experienced only by 44 percent of the trials, as $S(t=5) = .44$. Because the survivor function is a decreasing function of time, the error bars in later parts of the hazard function will always be wider and less precise compared to earlier parts.

168 2.2 Benefits of event history analysis

169 Statisticians and mathematical psychologists recommend focusing on the hazard
 170 function when analyzing time-to-event data for various reasons. We do not cover these
 171 benefits in detail here, as these are more general topics that have been covered elsewhere in
 172 textbooks. Instead, we briefly list the benefits below, and refer the reader to section F of
 173 the Supplemental Material for more detailed coverage of the benefits. The benefits include:

- 174 1. Hazard functions are more diagnostic than density functions when one is interested in
175 studying the detailed shape of a RT distribution (Holden et al., 2009).
- 176 2. RT distributions may differ from each other in multiple ways, and hazard functions
177 allow one to capture these differences which mean-average comparisons may conceal
178 (Townsend, 1990).
- 179 3. EHA takes account of more of the data collected in a typical speeded response
180 experiment, by virtue of not discarding right-censored observations. Trials with very
181 long RTs are not discarded, but instead contribute to the risk set in each time bin
182 (see section 4.1.2 below).
- 183 4. Hazard modeling allows one to incorporate time-varying explanatory covariates, such
184 as heart rate, electroencephalogram (EEG) signal amplitude, gaze location, etc.
185 (Allison, 2010). This is useful for linking physiological effects to behavioral effects
186 when performing cognitive psychophysiology (Meyer, Osman, Irwin, & Yantis, 1988).
- 187 5. EHA can help to solve the problem of model mimicry, i.e., the fact that different
188 computational models can often predict the same mean RTs as observed in the
189 empirical data, but not necessarily the detailed shapes of the empirical RT hazard
190 distributions. As such, EHA can be a tool to help distinguish between competing
191 theories of cognition and brain function.

192 2.3 Event history analysis in the context of experimental psychology

193 To make EHA more relevant to researchers studying cognitive psychology and

194 cognitive neuroscience, in this section we provide a relevant worked example and consider
195 implications that are relevant to that domain of research.

196 **2.3.1 A worked example.** In the context of experimental psychology, it is

197 common for participants to be presented with either a 1-button detection task or a

198 discrimination task. For example, a task may involve choosing between two response
199 options with only one of them being correct. For such two-choice RT data, the
200 discrete-time EHA of the RT data (hazard and survivor functions) can be extended with a
201 discrete-time SAT analysis of the timed accuracy data. Specifically, the hazard function of
202 event occurrence can be extended with the discrete-time conditional accuracy function,
203 which gives you the probability that a response is correct given that it is emitted in time
204 bin t (Allison, 2010; Kantowitz & Pachella, 2021; Wickelgren, 1977). We refer to this
205 extended (hazard + conditional accuracy) analysis for choice RT data as EHA/SAT.

206 Integrating results between hazard and conditional accuracy functions for choice RT
207 data can be informative for understanding psychological processes. To illustrate, we
208 consider a hypothetical choice RT example that is inspired by real data (Panis & Schmidt,
209 2016), but simplified to make the main point clearer (Figure 3). In a standard priming
210 paradigm, there is a prime stimulus (e.g., an arrow pointing left or right) followed by a
211 target stimulus (another arrow pointing left or right). The prime can then be congruent or
212 incongruent with the target.

213 Figure 3 shows that the early upswing in hazard is equal for both priming conditions
214 (Figure 3, upper panel), and that early emitted responses are always correct in the
215 congruent condition and always incorrect in the incongruent condition (Figure 3, lower
216 panel). These results show that for short waiting times (< bin 6), responses always follow
217 the prime (and not the target, as instructed). During timebin 6 the target-triggered
218 response channel is activated and causes response competition – $ca(6) = .5$ – and a lower
219 hazard probability in the incongruent condition. For waiting times of 600 ms or more, the
220 hazard of response occurrence is lower in incongruent compared to congruent trials, and all
221 responses emitted in these late bins are correct.

222 This joint pattern of results is interesting because it can provide meaningfully
223 different conclusions about psychological processes compared to conventional analyses, such

as computing mean-average RT and accuracy across trials. Mean-average RT would only represent the overall ability of cognition to overcome interference, on average, across trials. For instance, if mean-average RT was higher in incongruent than congruent trials, one may conclude that cognitive mechanisms that support interference control are working as expected across trials, and are indexed by each recorded response. But such a conclusion is not supported when the effects are explored over a timeline. Instead, the psychological conclusion is much more nuanced and suggests that multiple states start, stop and possibly interact over a particular temporal window.

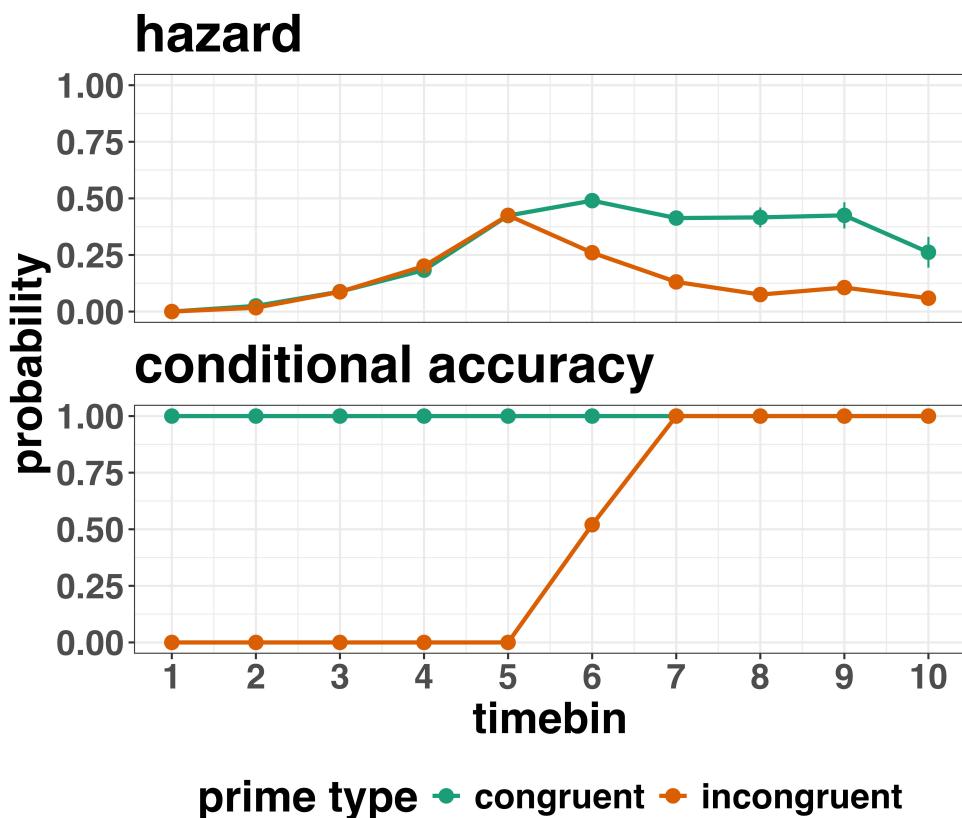


Figure 3. Discrete-time hazard and conditional accuracy functions. Discrete time-to-event and conditional accuracy data were simulated for 1000 trials for each of two priming conditions (congruent and incongruent prime stimuli). Error bars represent ± 1 standard error of the respective proportion. Bin width equals 100 ms.

Unlocking the temporal states of cognitive processes can be revealing for theory development and the understanding of basic psychological processes. Possibly more importantly, however, is that it simultaneously opens the door to address many new and previously unanswered questions. Do all participants show similar temporal states or are there individual differences? Do such individual differences extend to those individuals that have been diagnosed with some form of psychopathology? How do temporal states relate to brain-based mechanisms that might be studied using other methods from cognitive neuroscience? And how much of theory in cognitive psychology would be in need of revision if mean-average comparisons were supplemented with a temporal states approach?

2.3.2 Implications for designing experiments. Performing EHA in experimental psychology has implications for how experiments are designed. Indeed, if trials are categorised as a function of when responses occur, then each timebin will only include a subset of the total number of trials. For example, let's consider an experiment where each participant performs 2 conditions and there are 100 trial repetitions per condition. Those 100 trials must be distributed in some manner across the chosen number of bins.

In such experimental designs, since the number of trials per condition are spread across bins, it is important to have a relatively large number of trial repetitions per participant and per condition. Accordingly, experimental designs using this approach typically focus on factorial, within-subject designs, in which a large number of observations are made on a relatively small number of participants (so-called small- N designs). This approach emphasizes the precision and reproducibility of data patterns at the individual participant level to increase the inferential validity of the design (Baker et al., 2021; Smith & Little, 2018).

In contrast to the large- N design that typically average across many participants without being able to scrutinize individual data patterns, small- N designs retain crucial information about the data patterns of individual observers. This can be advantageous

259 whenever participants differ systematically in their strategies or in the time courses of their
260 effects, so that averaging them would lead to misleading data patterns. Note that because
261 statistical power derives both from the number of participants and from the number of
262 repeated measures per participant and condition, small- N designs can still achieve what
263 are generally considered acceptable levels of statistical power, if they have a sufficient
264 amount of data overall (Baker et al., 2021; Smith & Little, 2018).

265 **3. An overview of the general analytical workflow**

266 Although the focus is on EHA/SAT, we also want to briefly comment on broader
267 aspects of our general analytical workflow, which relate more to data science and data
268 analysis workflows.

269 **3.1 Data science workflow and descriptive statistics**

270 We perform data wrangling following tidyverse principles and a functional
271 programming approach (Wickham, Çetinkaya-Rundel, & Grolemund, 2023). In short,
272 functional programming means that you avoid writing your own loops and instead use
273 functions that have been built and tested by others. In addition, we also supply a set of
274 custom-built functions, which make the process of data wrangling in the context of data
275 preparation and descriptive statistics a lot quicker and more efficient.

276 **3.2 Inferential statistical approach**

277 Our lab adopts an estimation approach to multilevel regression (Kruschke & Liddell,
278 2018; Winter, 2019), which is heavily influenced by the Bayesian framework as suggested
279 by Richard McElreath (Kurz, 2023b; McElreath, 2020). We also use a “keep it maximal”
280 approach to specifying varying (or random) effects (Barr, Levy, Scheepers, & Tily, 2013).
281 This means that wherever possible we include varying intercepts and slopes per participant.

282 To make inferences, we use two main approaches. We compare models of different
 283 complexity, using information criteria (e.g., WAIC) and cross-validation (e.g., LOO), to
 284 evaluate out-of-sample predictive accuracy (McElreath, 2020). We also take the most
 285 complex model and evaluate key parameters of interest using point and interval estimates.

286 **3.3 Implementation**

287 We used R (Version 4.4.0; R Core Team, 2024)¹ for all reported analyses. The
 288 content of the tutorials, in terms of EHA and multilevel regression modelling, is mainly
 289 based on Allison (2010), Singer and Willett (2003), McElreath (2020), Heiss (2021), Kurz
 290 (2023a), and Kurz (2023b).

291 **4. Tutorials**

292 Tutorials 1a and 1b show how to calculate and plot the descriptive statistics of
 293 EHA/SAT when there are one or two independent variables, respectively. Tutorials 2a and
 294 2b illustrate how to use Bayesian multilevel modeling to fit hazard and conditional

¹ We, furthermore, used the R-packages *bayesplot* (Version 1.11.1; Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019), *brms* (Version 2.21.0; Bürkner, 2017, 2018, 2021), *citr* (Version 0.3.2; Aust, 2019), *cmdstanr* (Version 0.8.1.9000; Gabry, Češnovar, Johnson, & Brønner, 2024), *dplyr* (Version 1.1.4; Wickham, François, Henry, Müller, & Vaughan, 2023), *forcats* (Version 1.0.0; Wickham, 2023a), *ggplot2* (Version 3.5.1; Wickham, 2016), *lme4* (Version 1.1.35.5; Bates, Mächler, Bolker, & Walker, 2015), *lubridate* (Version 1.9.3; Grolemund & Wickham, 2011), *Matrix* (Version 1.7.0; Bates, Maechler, & Jagan, 2024), *nlme* (Version 3.1.166; Pinheiro & Bates, 2000), *papaja* (Version 0.1.2.9000; Aust & Barth, 2023), *patchwork* (Version 1.2.0; Pedersen, 2024), *purrr* (Version 1.0.2; Wickham & Henry, 2023), *RColorBrewer* (Version 1.1.3; Neuwirth, 2022), *Rcpp* (Eddelbuettel & Balamuta, 2018; Version 1.0.12; Eddelbuettel & François, 2011), *readr* (Version 2.1.5; Wickham, Hester, & Bryan, 2024), *RJ-2021-048* (Bengtsson, 2021), *standist* (Version 0.0.0.9000; Girard, 2024), *stringr* (Version 1.5.1; Wickham, 2023b), *tibble* (Version 3.2.1; Müller & Wickham, 2023), *tidybayes* (Version 3.0.6; Kay, 2023), *tidyverse* (Version 2.0.0; Wickham et al., 2019), and *tinylabels* (Version 0.2.4; Barth, 2023).

accuracy models, respectively. Tutorials 3a and 3b show how to implement, respectively, multilevel models for hazard and conditional accuracy in the frequentist framework. Additionally, to further simplify the process for other users, the first two tutorials rely on a set of our own custom functions that make sub-processes easier to automate, such as data wrangling and plotting functions (see section B in the Supplemental Material for a list of the custom functions).

Our list of tutorials is as follows:

- 1a. Wrangle raw data and calculate descriptive stats for one independent variable
- 1b. Wrangle raw data and calculate descriptive stats for two independent variables
- 2a. Bayesian multilevel modeling for $h(t)$
- 2b. Bayesian multilevel modeling for $ca(t)$
- 3a. Frequentist multilevel modeling for $h(t)$
- 3b. Frequentist multilevel modeling for $ca(t)$
- 4. Simulation and power analysis for planning experiments

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

4.1.1 Data wrangling aims.

Our data wrangling procedures serve two related purposes. First, we want to summarise and visualise descriptive statistics that relate to our main research questions about the time course of psychological processes, using a life table. A life table includes for each time bin, the risk set (i.e., the number of trials that are event-free at the start of the bin), the number of observed events, and the estimates of $h(t)$, $S(t)$, $P(t)$, possibly $ca(t)$, and their estimated standard errors (se).

Second, we want to produce two different data sets that can each be submitted to different types of inferential modelling approaches. The two types of data structure we label as ‘person-trial’ data and ‘person-trial-bin’ data. The ‘person-trial’ data (Table 1) will be familiar to most researchers who record behavioural responses from participants, as

³²⁰ it represents the measured RT and accuracy per trial within an experiment. This data set
³²¹ 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.

³²² In contrast, the ‘person-trial-bin’ data (Table 2) has a different, more extended
³²³ structure, which indicates in which bin a response occurred, if at all, in each trial.
³²⁴ Therefore, the ‘person-trial-bin’ data generates a 0 in each bin until an event occurs and
³²⁵ then it generates a 1 to signal an event has occurred in that bin. This data set is used
³²⁶ when fitting hazard models (Tutorials 2a and 3a). It is worth pointing out that there is no
³²⁷ requirement for an event to occur at all (in any bin), as maybe there was no response on
³²⁸ that trial or the event occurred after the time window of interest. Likewise, when the event

³²⁹ occurs in bin 1 there would only be one row of data for that trial in the person-trial-bin
³³⁰ 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.

³³¹ **4.1.2 A real data wrangling example.** To illustrate how to quickly set up life
³³² tables for calculating the descriptive statistics (functions of discrete time), we use a
³³³ published data set on masked response priming from Panis and Schmidt (2016). In their
³³⁴ first experiment, Panis and Schmidt (2016) presented a double arrow for 94 ms that
³³⁵ pointed left or right as the target stimulus with an onset at time point zero in each trial.
³³⁶ Participants had to indicate the direction in which the double arrow pointed using their
³³⁷ 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. Here we focus on the subset of trials in which no mask was presented. The 13-ms prime stimulus was a double arrow presented 187 ms before target onset in the congruent (same direction as target) and incongruent (opposite direction as target) prime conditions.

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

The required column names are as follows:

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

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

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

Next, we can set up the life tables and plots of the discrete-time functions $h(t)$, $S(t)$, $ca(t)$, and $P(t)$ – see section A of the Supplemental Material for their definitions. To do so

356 using a functional programming approach, one has to nest the data within participants
 357 using the group_nest() function, and supply a user-defined censoring time and bin width
 358 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))
```

359 Note that the censoring time should be a multiple of the bin width (both in ms). The
 360 censoring time should be a time point after which no informative responses are expected
 361 anymore. In experiments that implement a response deadline in each trial the censoring
 362 time can equal that deadline time point. Trials with a RT larger than the censoring time,
 363 or trials in which no response is emitted during the data collection period, are treated as
 364 right-censored observations in EHA. In other words, these trials are not discarded, because
 365 they contain the information that the event did not occur before the censoring time.
 366 Removing such trials before calculating the mean event time will result in underestimation
 367 of the true mean.

368 The person-trial-bin oriented data set is created by our custom function ptb(), and it
 369 has one row for each time bin (of each trial) that is at risk for event occurrence. The

370 variable “event” in the person-trial-bin oriented data set indicates whether a response
371 occurs (1) or not (0) for each bin.

372 The next step is to set up the life table using our custom function `setup_lt()`,
373 calculate the conditional accuracies using our custom function `calc_ca()`, add the `ca(t)`
374 estimates to the life table using our custom function `join_lt_ca()`, and then plot the
375 descriptive statistics using our custom function `plot_eha()`. When creating the plots, some
376 warning messages will likely be generated, like these:

- 377 • Removed 2 rows containing missing values or values outside the scale range
378 (`geom_line()`).
379 • Removed 2 rows containing missing values or values outside the scale range
380 (`geom_point()`).
381 • Removed 2 rows containing missing values or values outside the scale range
382 (`geom_segment()`).

383 The warning messages are generated because some bins have no hazard and `ca(t)`
384 estimates, and no error bars. They can thus safely be ignored. One can now inspect
385 different aspects, including the life table for a particular condition of a particular subject,
386 and a plot of the different functions for a particular participant. In general, it is important
387 to visually inspect the functions first for each participant, in order to identify individuals
388 that may be guessing (e.g., a flat conditional accuracy function at .5 indicates that
389 someone is just guessing), outlying individuals, and/or different groups with qualitatively
390 different behavior.

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

Table 3

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

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

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

Figure 4 displays the discrete-time hazard, survivor, conditional accuracy, and

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

Descriptive stats for subject 6

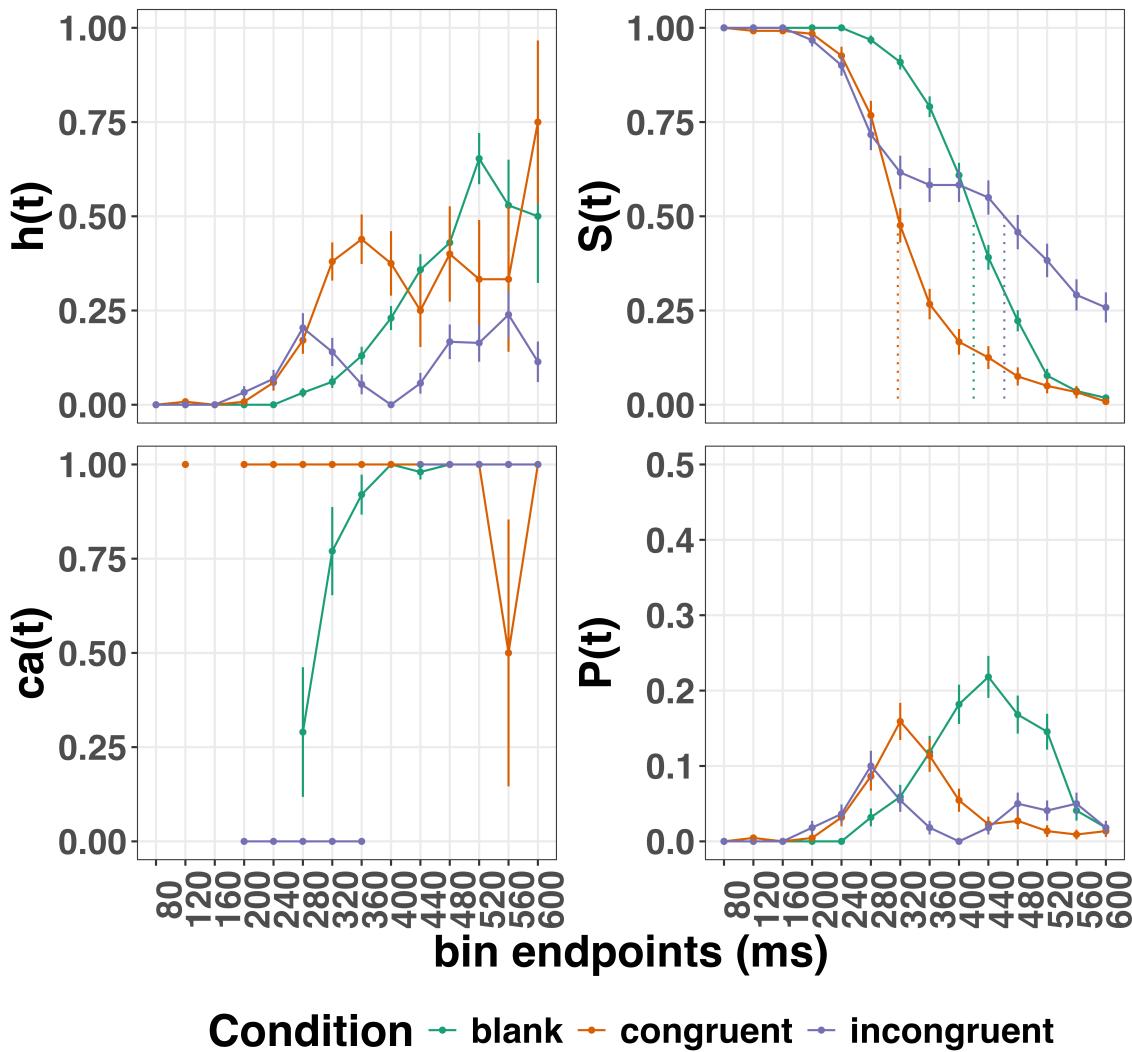


Figure 4. Estimated discrete-time hazard (h), survivor (S), conditional accuracy (ca) and probability mass (P) functions for participant 6. Vertical dotted lines indicate the estimated median RTs. Error bars represent ± 1 standard error of the respective proportion.

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

400 0.17, and 0.20 for the blank, congruent, and incongruent prime conditions, respectively. In

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

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

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

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

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

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

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

408 respectively.

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

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

411 0.36, 0.25, and 0.06 for the blank, congruent, and incongruent prime conditions,

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

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

414 conditions, respectively.

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

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

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

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

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

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

421 group-average distribution per condition (see Tutorial_1a.Rmd).

422 In general, these results go against the (often implicit) assumption in research on

423 priming that all observed responses are primed responses to the target stimulus. Instead,

424 the distributional data show that early responses are triggered exclusively by the prime

425 stimulus, while only later responses reflect primed responses to the target stimulus.

426 At this point, we have calculated, summarised and plotted descriptive statistics for
427 the key variables in EHA/SAT. As we will show in later Tutorials, statistical models for
428 $h(t)$ and $ca(t)$ can be implemented as generalized linear mixed regression models predicting
429 event occurrence (1/0) and conditional accuracy (1/0) in each bin of a selected time
430 window for analysis. But first we consider calculating the descriptive statistics for two
431 independent variables.

432 **4.2 Tutorial 1b: Generalising to a more complex design**

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

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

444 **4.3 Tutorial 2a: Fitting Bayesian hazard models to discrete time-to-event data**

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

449 Willett, 2003).

450 **4.3.1 Hazard model considerations.** There are several analytic decisions one
451 has to make when fitting a discrete-time hazard model. First, one has to select an analysis
452 time window, i.e., a contiguous set of bins for which there is enough data for each
453 participant. Second, given that the dependent variable (event occurrence) is binary, one
454 has to select a link function (see section C in the Supplemental Material). The cloglog link
455 is preferred over the logit link when events can occur in principle at any time point within
456 a bin, which is the case for RT data (Singer & Willett, 2003). Third, one has to choose
457 whether to treat TIME (i.e., the time bin index t) as a categorical or continuous predictor.
458 And when you treat a variable as a categorical predictor, you can choose between reference
459 coding and index coding. With reference coding, one defines the variable as a factor and
460 selects one of the k categories as the reference level. `Brm()` will then construct $k-1$
461 indicator variables (see model M1d in Tutorial_2a.Rmd for an example). With index
462 coding, one constructs an index variable that contains integers that correspond to different
463 categories (see models M0i and M1i below). As explained by McElreath (2020), the
464 advantage of index coding is that the same prior can be assigned to each level of the index
465 variable, so that each category has the same prior uncertainty.

466 In the case of a large- N design without repeated measurements, the parameters of a
467 discrete-time hazard model can be estimated using standard logistic regression software
468 after expanding the typical person-trial data set into a person-trial-bin data set (Allison,
469 2010). When there is clustering in the data, as in the case of a small- N design with
470 repeated measurements, the parameters of a discrete-time hazard model can be estimated
471 using population-averaged methods (e.g., Generalized Estimating Equations), and Bayesian
472 or frequentist generalized linear mixed models (Allison, 2010).

473 In general, there are three assumptions one can make or relax when adding
474 experimental predictor variables and other covariates: The linearity assumption for
475 continuous predictors (the effect of a 1 unit change is the same anywhere on the scale), the

476 additivity assumption (predictors do not interact), and the proportionality assumption
 477 (predictors do not interact with TIME).

478 In tutorial_2a.Rmd we fit several Bayesian multilevel models (i.e., generalized linear
 479 mixed models) that differ in complexity to the person-trial-bin oriented data set that we
 480 created in Tutorial 1a. We decided to select the analysis time window (200,600] and the
 481 cloglog link. Below, we shortly discuss two of these models. The person-trial-bin data set is
 482 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)))
```

483 **4.3.2 Prior distributions.** To get the posterior distribution of each model
 484 parameter given the data, we need to specify prior distributions for the model parameters
 485 which reflect our prior beliefs. In Tutorial_2a.Rmd we perform a few prior predictive
 486 checks to make sure our selected prior distributions reflect our prior beliefs (Gelman,
 487 Vehtari, et al., 2020).

488 The middle column of Supplementary Figure 2 (section E of the Supplemental
 489 Material) shows six examples of prior distributions for an intercept on the logit and/or
 490 cloglog scales. While a normal distribution with relatively large variance is often used as a

491 weakly informative prior for continuous dependent variables, rows A and B of
 492 Supplementary Figure 2 show that specifying such distributions on the logit and cloglog
 493 scales actually leads to rather informative distributions on the original probability scale, as
 494 most mass is pushed to probabilities of 0 and 1.

495 **4.3.3 Model M0i: A null model with index coding.** When you do not want to
 496 make assumptions about the shape of the hazard function, or its shape is not smooth but
 497 irregular, then you can use a general specification of TIME, i.e., fit one grand intercept per
 498 time bin. In this first model, we use a general specification of TIME using index coding,
 499 and do not include experimental predictors. We call this model “M0i”.

500 Before we fit model M0i, we select the necessary columns from the data, and specify
 501 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")
```

502 After selecting the bernoulli family and the cloglog link, the model formula is
 503 specified. The specification “0 + ...” removes the default intercept in brm(). The fixed
 504 effects include an intercept for each level of timebin. Each of these intercepts is allowed to

505 vary across individuals (variable pid). We request 2000 samples from the posterior
 506 distribution for each of four chains. Estimating model M0i took about 30 minutes on a
 507 MacBook Pro (Sonoma 14.6.1 OS, 18GB Memory, M3 Pro Chip).

508 **4.3.4 Model M1i: Adding the effects of prime-target congruency.** Previous
 509 research has shown that psychological effects typically change over time (Panis, 2020;
 510 Panis, Moran, et al., 2020; Panis & Schmidt, 2022; Panis et al., 2017; Panis & Wagemans,
 511 2009). In the next model, therefore, we use index coding for both TIME (variable
 512 “timebin”) and the categorical predictor prime-target-congruency (variable “prime”), so
 513 that we get 30 grand intercepts, one for each combination of timebin level and prime level.
 514 Here is the model formula of this model that we call “M1i”.

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

515 Estimating model M1i took about 124 minutes.

516 **4.3.5 Compare the models.** We can compare the two models using the Widely
 517 Applicable Information Criterion (WAIC) and Leave-One-Out (LOO) cross-validation, and
 518 look at model weights for both criteria (Kurz, 2023a; McElreath, 2020).

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

519 ## model_M0i model_M1i
 520 ## 0 1

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

521 ## model_M0i model_M1i
 522 ## 0 1

523 Clearly, both the loo and waic weighting schemes assign a weight of 1 to model M1i,
 524 and a weight of 0 to the other simpler model.

525 **4.3.6 Evaluating parameter estimates in model M1i.** To make inferences

526 from the parameter estimates in model M1i, we first plot the densities of the draws from
 527 the posterior distributions of its population-level parameters in Figure 5, together with
 528 point (median) and interval estimates (80% and 95% credible intervals).

Posterior distributions for population-level effects in Model M1i

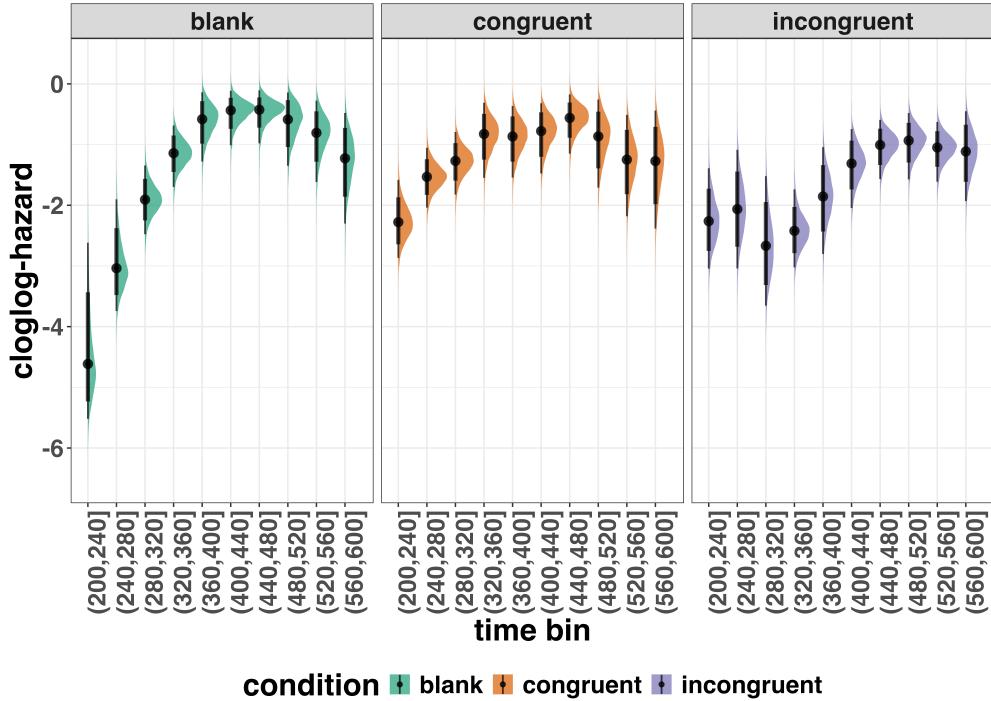


Figure 5. Medians and 80/95% credible intervals of the posterior distributions of the population-level parameters of model M1i.

529 Because the parameter estimates are on the cloglog-hazard scale, we can ease our

530 interpretation by plotting the expected value of the posterior predictive distribution – the
 531 predicted hazard values – for the grand average (Figure 6A), and for each participant in
 532 the data set (Figure 6B).

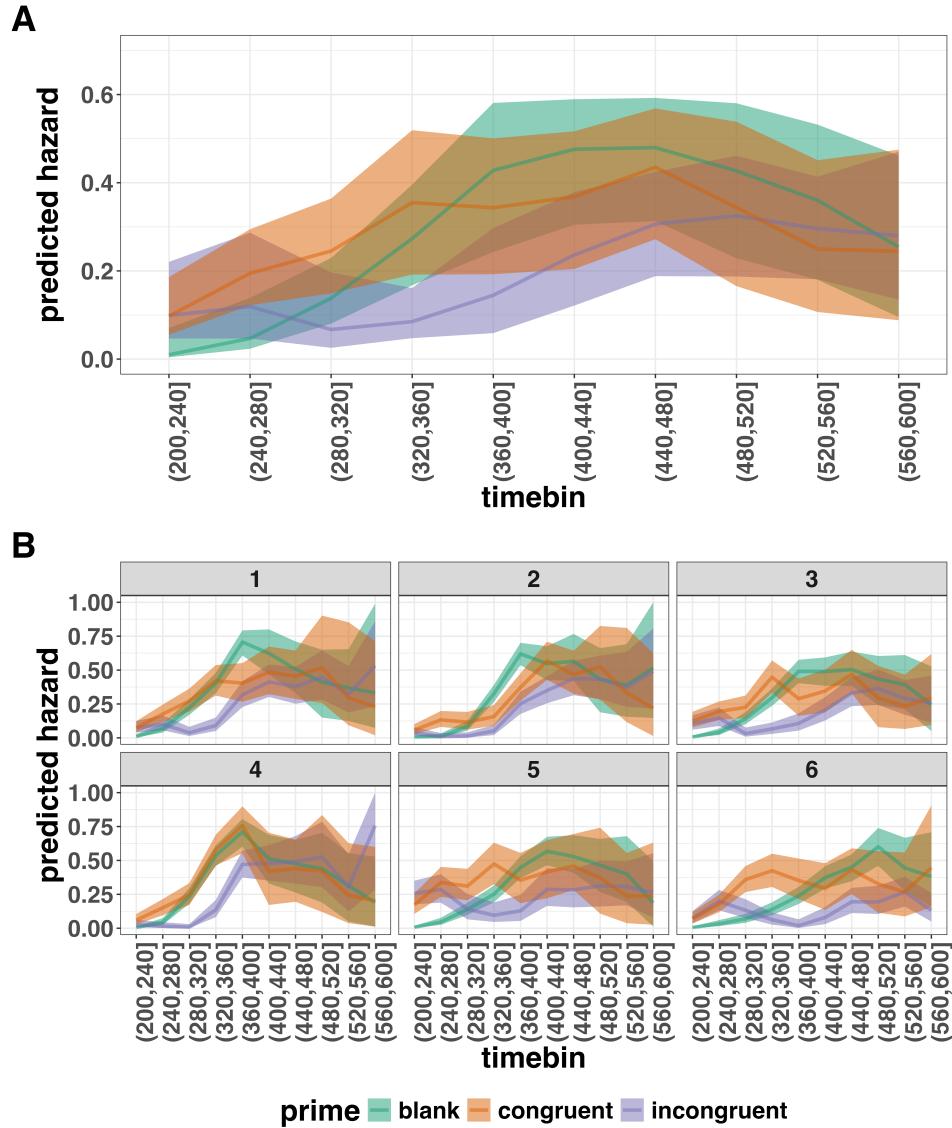


Figure 6. Point (median) and 80/95% credible interval summaries of the hazard estimates (expected values of the draws from the posterior predictive distributions) in each time bin for the grand average (A), and for each participant (B).

533 As we are actually interested in the effects of congruent and incongruent primes,
 534 relative to the blank prime condition, we can construct two contrasts (congruent-blank,
 535 incongruent-blank), and plot the posterior distributions of these contrast effects, both for
 536 the grand average (Figure 7A; grand average marginal effect) and for each participant in

537 the data set (Figure 7B; subject-specific average marginal effect).

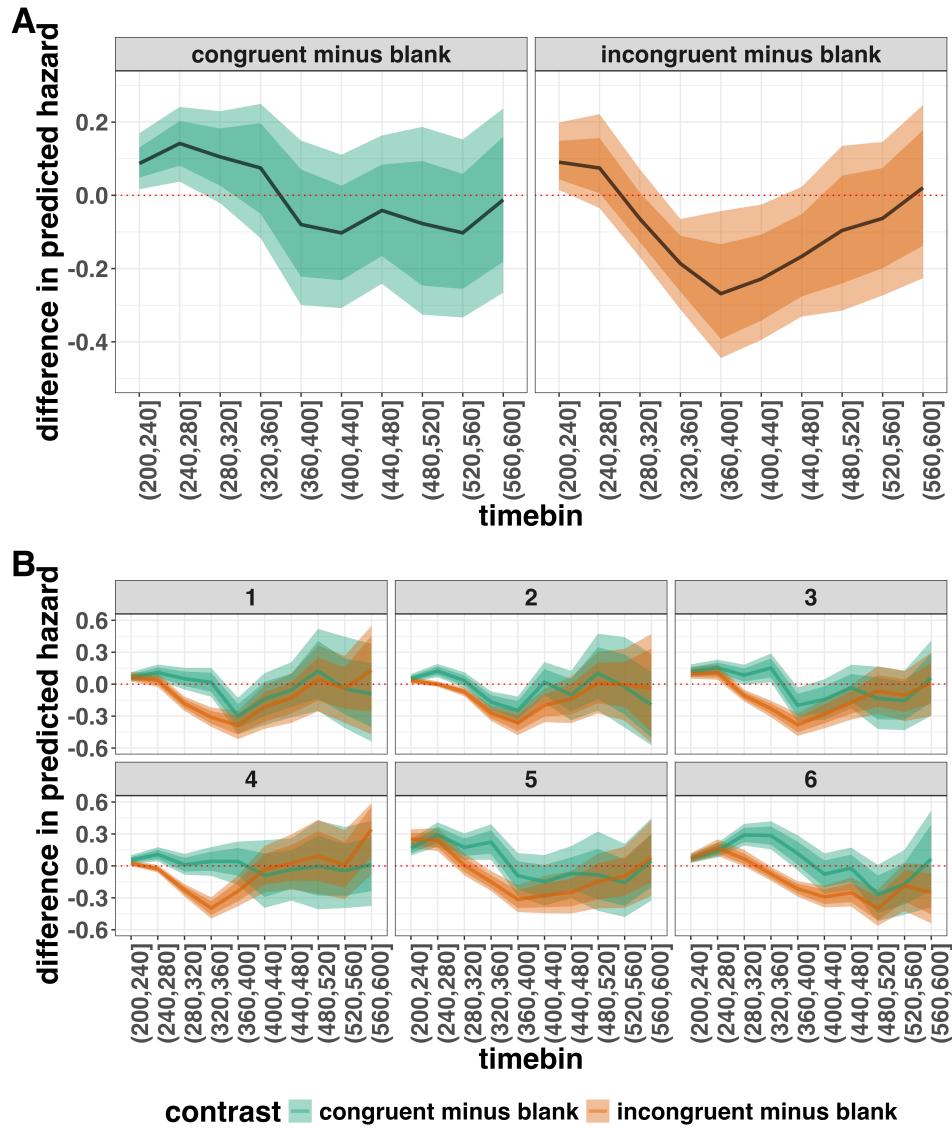


Figure 7. Point (mean) and 80/95% credible interval summaries of estimated differences in hazard in each time bin for the grand average (A), and for each participant (B).

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

540 **Example conclusions for M1i.** What can we conclude from model M1i about
541 our research question, i.e., the temporal dynamics of the effect of prime-target congruency

542 on RT? In other words, in which of the 40-ms time bins between 200 and 600 ms after
543 target onset does changing the prime from blank to congruent or incongruent affect the
544 hazard of response occurrence (for a prime-target SOA of 187 ms)?

545 If we want to estimate the population-level effect of prime type on hazard, we can
546 base our conclusion on Figure 7A. The contrast “congruent minus blank” was estimated to
547 be 0.09 hazard units in bin (200,240] (95% CrI = [0.02, 0.17]), and 0.14 hazard units in bin
548 (240,280]) (95% CrI = [0.04, 0.25]). For the other bins, the 95% credible interval contained
549 zero. The contrast “incongruent minus blank” was estimated to be 0.09 hazard units in bin
550 (200,240] (95% CrI = [0.01, 0.21]), -0.19 hazard units in bin (320,360] (95% CrI = [-0.31,
551 -0.06]), -0.27 hazard units in bin (360,400] (95% CrI = [-0.45, -0.04]), and -0.23 hazard
552 units in bin (400,440] (95% CrI = [-0.40, -0.03]). For the other bins, the 95% credible
553 interval contained zero.

554 There are thus two phases of performance for the average person between 200 and
555 600 ms after target onset. In the first phase, the addition of a congruent or incongruent
556 prime stimulus increases the hazard of response occurrence compared to blank prime trials
557 in the time period (200, 240]. In the second phase, only the incongruent prime decreases
558 the hazard of response occurrence compared to blank primes, in the time period (320,440].
559 The sign of the effect of incongruent primes on the hazard of response occurrence thus
560 depends on how much waiting time has passed since target onset.

561 If we want to focus more on inter-individual differences, we can study the
562 subject-specific hazard functions in Figure 7B. Note that three participants (1, 2, and 3)
563 show a negative difference for the contrast “congruent minus incongruent” in bin (360,400]
564 – subject 2 also in bin (320,360].

565 Future studies could (a) increase the number of participants to estimate the
566 proportion of “dippers” in the subject population, and/or (b) try to explain why this dip
567 occurs. For example, Panis and Schmidt (2016) concluded that active, top-down,

568 task-guided response inhibition effects emerge around 360 ms after the onset of the stimulus
569 following the prime (here: the target stimulus). Such a top-down inhibitory effect might
570 exist in our priming data set, because after some time participants will learn that the first
571 stimulus is not the one they have to respond to. To prevent a premature overt response to
572 the prime they thus might gradually increase a global response threshold during the
573 remainder of the experiment, which could result in a lower hazard in congruent trials
574 compared to blank trials, for bins after ~360 ms, and towards the end of the experiment.
575 This effect might be masked for incongruent primes by the response competition effect.

576 Interestingly, all subjects show a tendency in their mean difference (congruent minus
577 blank) to “dip” around that time (Figure 7B). Therefore, future modeling efforts could
578 incorporate the trial number into the model formula, in order to also study how the effects
579 of prime type on hazard change on the long experiment-wide time scale, next to the short
580 trial-wide time scale. In Tutorial_2a.Rmd we provide a number of model formulae that
581 should get you going.

582 4.4 Tutorial 2b: Fitting Bayesian conditional accuracy models

583 In this fourth tutorial, we illustrate how to fit a Bayesian multilevel regression model
584 to the timed accuracy data from the masked response priming data used in Tutorial 1a.
585 The general process is similar to Tutorial 2a, except that (a) we use the person-trial data,
586 (b) we use the logit link function, and (c) we change the priors. To keep the tutorial short,
587 we only fit one conditional accuracy model, which was based on model M1i from Tutorial
588 2a and labelled M1i_ca.

589 To make inferences from the parameter estimates in model M1i_ca, we first plot the
590 densities of the draws from the posterior distributions of its population-level parameters in
591 Figure 8, together with point (median) and interval estimates (80% and 95% credible
592 intervals).

Posterior distributions for population-level effects in Model M1i_ca

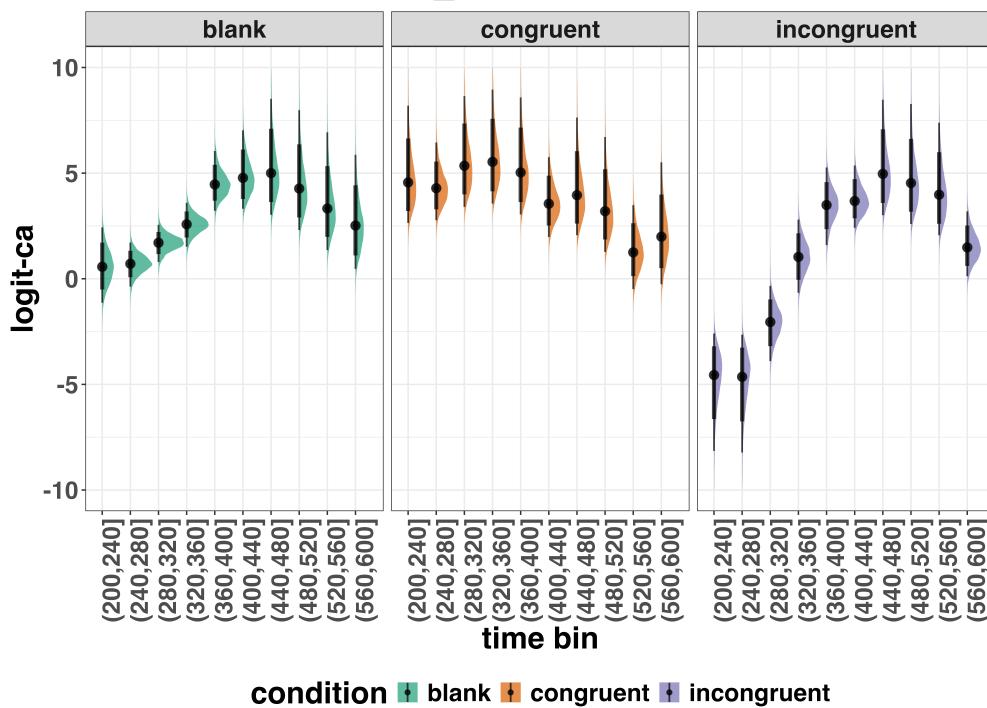


Figure 8. Medians and 80/95% credible intervals of the posterior distributions of the population-level parameters of model M1i_ca. ca = conditional accuracy.

Because the parameter estimates are on the logit-ca scale, we can ease our

interpretation by plotting the expected value of the posterior predictive distribution – the predicted conditional accuracies – for the grand average (Figure 9A), and for each participant in the data set (Figure 9B).

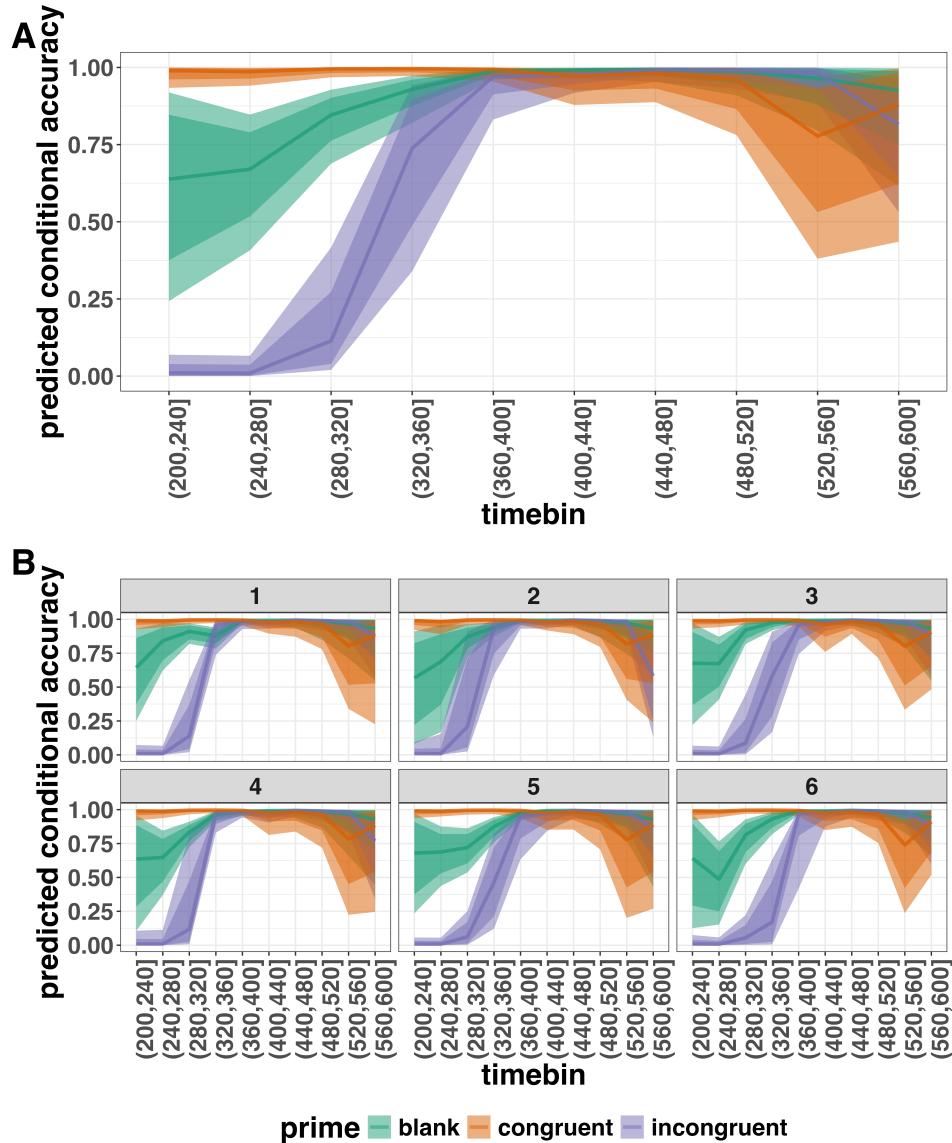


Figure 9. Point (median) and 80/95% credible interval summaries of the conditional accuracy estimates (expected values of the draws from the posterior predictive distributions) in each time bin for the grand average (A), and for each participant (B).

597 As we are actually interested in the effects of congruent and incongruent primes,

598 relative to the blank prime condition, we can construct two contrasts (congruent-blank,
 599 incongruent-blank), and plot the posterior distributions of these contrast effects for the
 600 grand average (Figure 10A; grand average marginal effect), and for each participant

601 (Figure 10B).

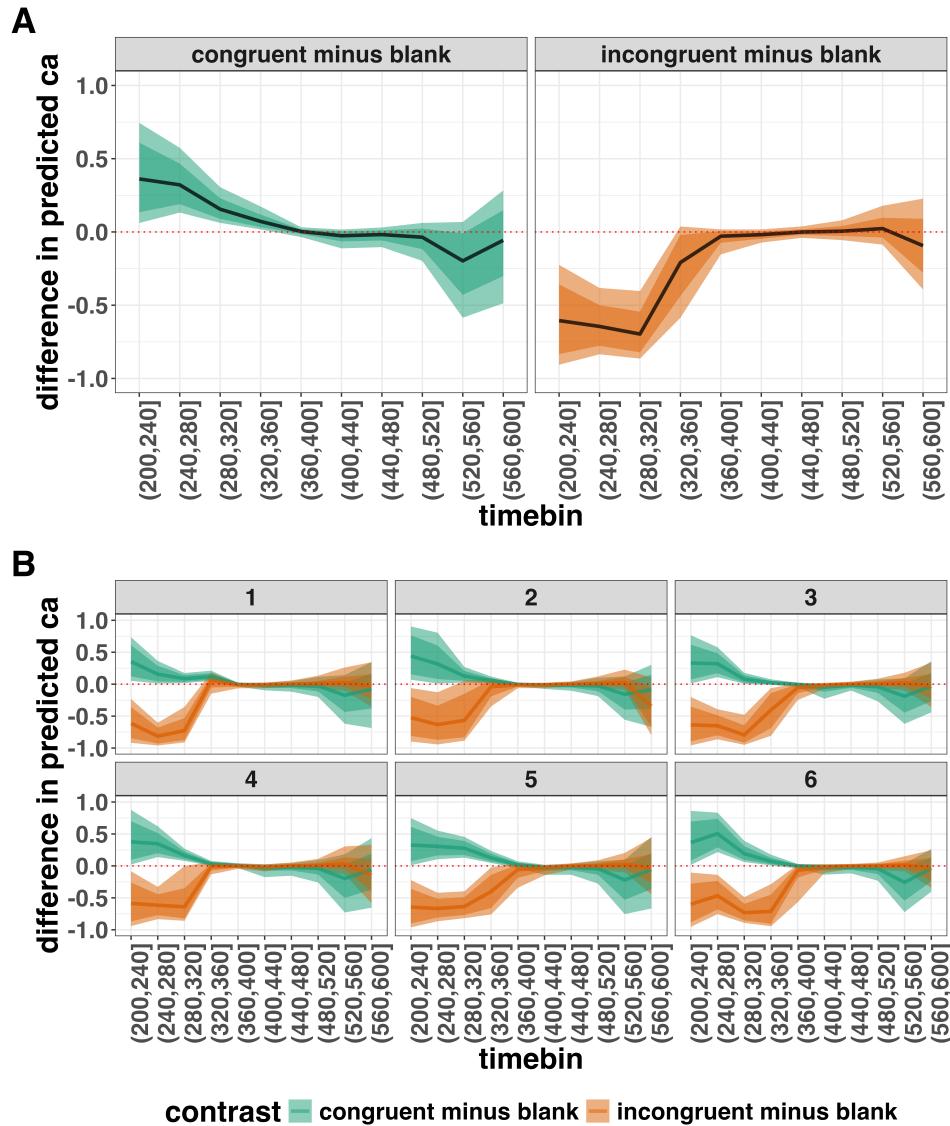


Figure 10. Point (mean) and 80/95% credible interval summaries of estimated differences in conditional accuracy in each time bin for the grand average (A), and for each participant (B).

602 Based on Figure 10A we see that congruent primes have a positive effect on the
 603 conditional accuracy of emitted responses in time bins (200,240], (240,280], (280,320], and
 604 (320,360], relative to the estimates in the baseline condition (blank prime; red dashed lines

605 in Figure 10A). Incongruent primes have a negative effect on the conditional accuracy of
606 emitted responses in the first time bins, relative to the estimates in the baseline condition.

607 **4.5 Tutorial 3a: Fitting Frequentist hazard models**

608 In this fifth tutorial we illustrate how to fit a multilevel regression model to RT data
609 in the frequentist framework, for the data used in Tutorial 1a. The general process is
610 similar to that in Tutorial 2a, except that there are no priors to set.

611 Again, to keep the tutorial concise, we only fit model M1i (see Tutorial 2a) using the
612 function glmer() from the R package lme4. Alternatively, one could also use the function
613 glmmPQL() from the R package MASS (Ripley et al., 2024). The resulting hazard model
614 is called M1i_f with the appended “_f” denoting a frequentist model.

615 In Figure 11 we compare the parameter estimates from the Bayesian regression model
616 M1i with those from the frequentist model M1i_f.

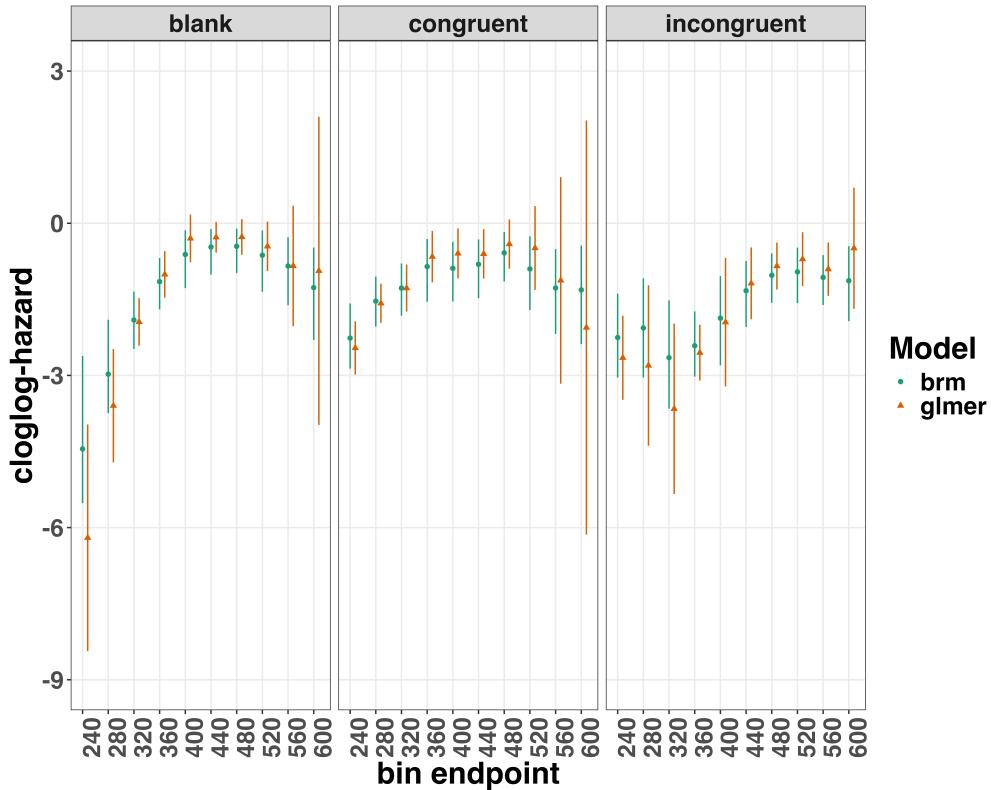


Figure 11. Parameter estimates for model M1i from brm() – means and 95% credible intervals – and model M1i_f from glmer() – maximum likelihood estimates and 95% confidence intervals.

617 Figure 11 confirms that the parameter estimates from both Bayesian and frequentist
 618 models are pretty similar, which makes sense given the close similarity in model structure.
 619 However, model M1i_f did not converge and resulted in a singular fit. This is of course one
 620 of the reasons why Bayesian modeling has become so popular in recent years. But the price
 621 you pay for being able to fit models with more complex varying effects structures via a
 622 Bayesian framework is increased computation time. In other words, as we have noted
 623 throughout, some of the Bayesian models in Tutorials 2a took several hours to build.

624 **4.6 Tutorial 3b: Fitting Frequentist conditional accuracy models**

625 In this sixth tutorial we illustrate how to fit a multilevel regression model to the
626 timed accuracy data in the frequentist framework, for the data used in Tutorial 1a. To be
627 concise, we only fit effects from model M1i_ca (see Tutorial 2b) using the function glmer()
628 from the R package lme4. Alternatively, one could also use the function glmmPQL() from
629 the R package MASS (Ripley et al., 2024). The resulting conditional accuracy model,
630 which we labelled M1i_ca_f, did not converge and resulted in a singular fit. Again, this
631 just highlights some of the difficulties in fitting reasonably complex varying/random effects
632 structures in frequentist workflows.

633 **4.7 Tutorial 4: Planning**

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

635 **4.7.1 Background.** The general approach to planning that we adopt here involves
636 simulating reasonably structured data to help guide what you might be able to expect from
637 your data once you collect it (Gelman, Vehtari, et al., 2020). The basic structure and code
638 follows the examples outlined by Solomon Kurz in his ‘power’ blog posts
639 (<https://solomonkurz.netlify.app/blog/bayesian-power-analysis-part-i/>) and Lisa
640 Debruine’s R package faux{} (<https://debruine.github.io/faux/>) as well as the related
641 paper (DeBruine & Barr, 2021).

642 **4.7.2 Basic workflow.** The basic workflow is as follows:

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

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

654 In the below, we only provide a high-level summary of the process and let readers

655 dive into the details within the tutorial should they feel so inclined.

656 **4.7.3 Fit a regression model and simulate one dataset.** We again use the
657 data from Panis and Schmidt (2016) to provide a worked example. We fit an index coding
658 model on a subset of timebins (six timebins in total) and for two prime conditions
659 (congruent and incongruent). We chose to focus on a subsample of the data to ease the
660 computational burden. We also used a full varying effects structure, with the model
661 formula as follows:

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

662 We then took parameters from this model and used them to create a single dataset

663 with 200 trials per condition for 10 individual participants. The raw data and the

664 simulated data are plotted in Figure 12 and show quite close correspondence, which is

665 re-assuring. But, this is only one dataset. What we really want to do is simulate many

666 datasets and vary parameters of interest, which is what we turn to in the next section.

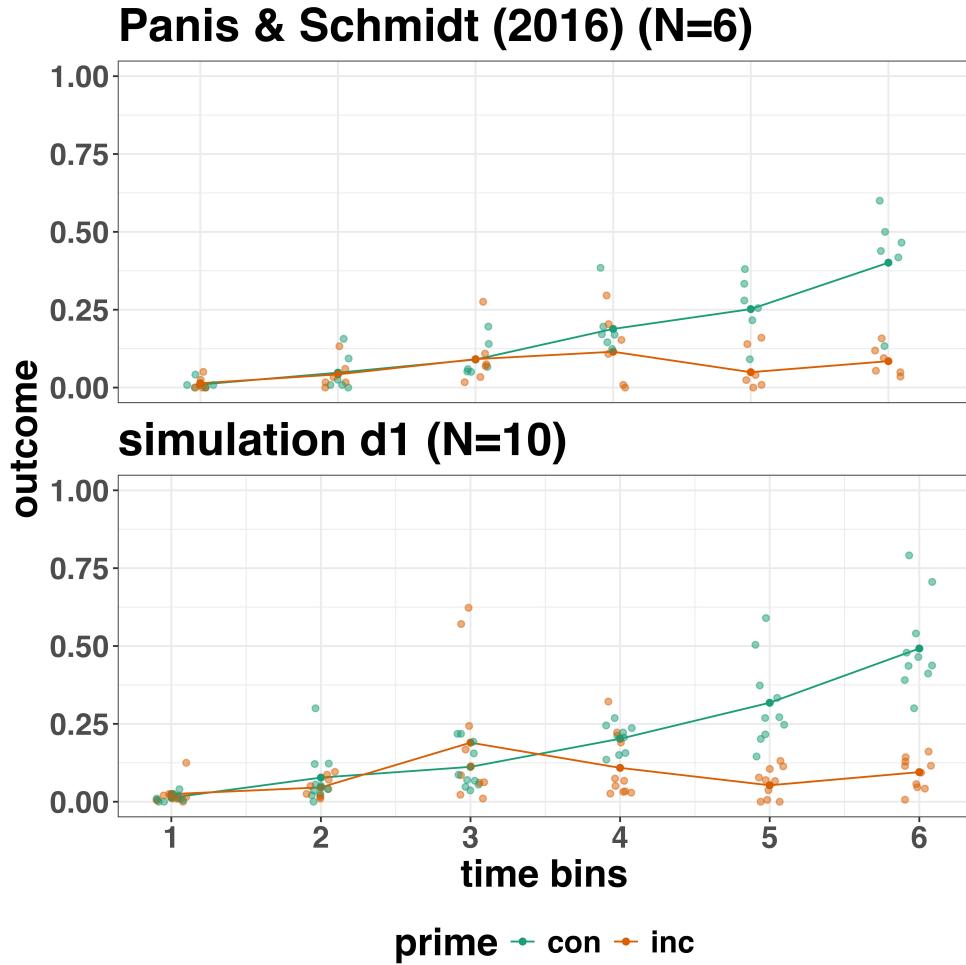


Figure 12. Raw data from Panis and Schmidt (2016) and simulated data from 10 participants.

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

Here we use the same data simulation process as used above, but instead of simulating one dataset, we simulate 1000 datasets per variation in parameter values. Specifically, in Simulation 1, we vary the number of trials per condition (100, 200, and 400), as well as the effect size in bin 6. We focus on bin 6 only, in terms of varying the effect size, just to make things simpler and easier to understand. The effect size observed in bin 6 in this subsample of data was a 79% reduction in hazard value from the congruent prime (0.401 hazard value) to the incongruent prime condition (0.085 hazard value). In other words, a hazard

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

Summary results from Simulation 1 are shown in Figure 13A. Figure 13A 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 13B. A summary of these power calculations might be as follows (trial count = 200 per condition in all cases):

- For a 70% reduction (0.3 hazard ratio), N=10 would give nearly 100% power.

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

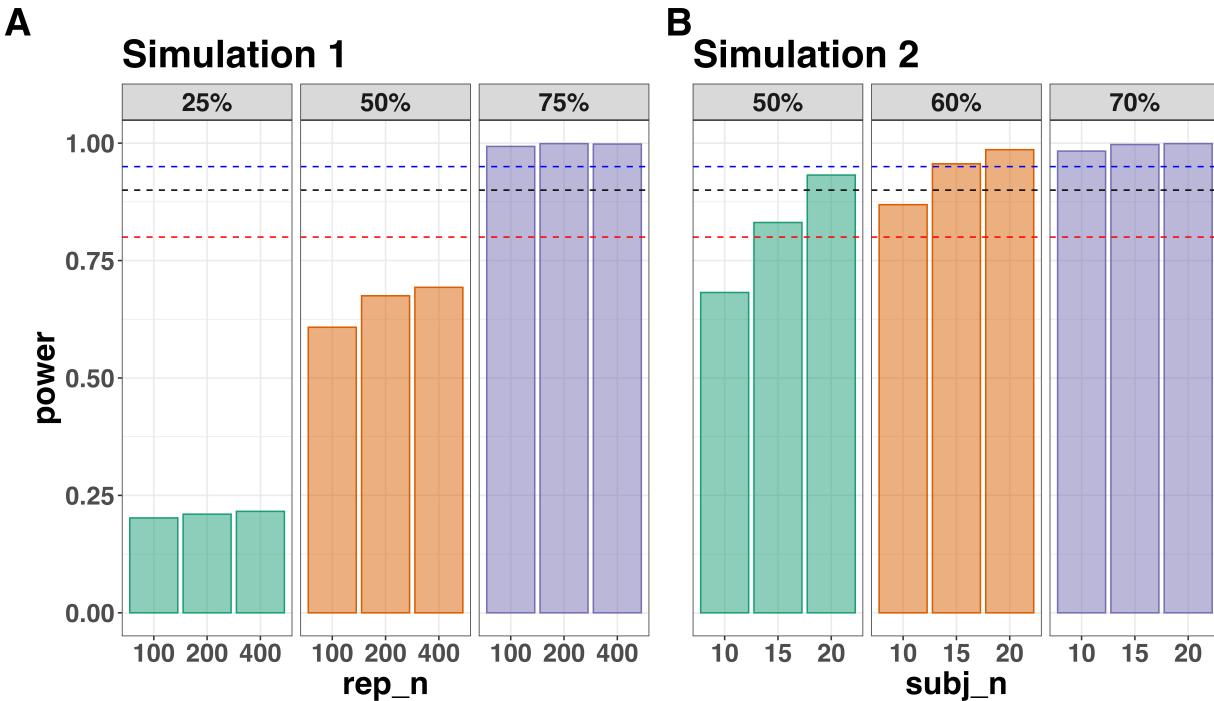


Figure 13. Statistical power across data Simulation 1 (A) and Simulation 2 (B). Power was calculated as the percentage of lower-bound 95% confidence intervals that exclude zero when the difference between prime condition is calculated (congruent - incongruent). In Simulation 1, the effect size was varied 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.

703 **4.7.5 Planning decisions.** Now that we have summarised our simulated data,

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

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

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

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

708 Some considerations might be as follows:

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

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

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

712 personnel?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

729

5. Discussion

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

745 **5.1 What are the main use-cases of EHA for understanding cognition and brain
746 function?**

747 For those researchers, like ourselves, who are primarily interested in understanding
748 human cognitive and brain systems, we consider two broadly-defined, main use-cases of
749 EHA. First, as we hope to have made clear by this point, EHA is one way to investigating
750 a “temporal states” approach to cognitive processes. EHA provides one way to uncover
751 when cognitive states may start and stop, as well as what they may be tied to or interact
752 with. Therefore, if your research questions concern **when** and **for how long** psychological
753 states occur, our EHA tutorials could be useful tools for you to use.

754 Second, even if you are not primarily interested in studying the temporal states of
755 cognition, EHA could still be a useful tool to consider using, in order to qualify inferences
756 that are being made based on mean-average comparisons. Given that distinctly different
757 inferences can be made from the same data based on whether one computes a
758 mean-average across trials or a RT distribution of events (Figure 1), it may be important
759 for researchers to supplement mean-average comparisons with EHA. One could envisage
760 scenarios where the implicit assumption of an effect manifesting across all of the time bins
761 measured would not be supported by EHA. Therefore, the conclusion of interest would not
762 apply to all responses, but instead it would be restricted to certain aspects of time.

763 5.2 Model complexity versus interpretability

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

780 4).

781 **5.3 Individual differences**

782 One important issue is that of possible individual differences in the overall location of
783 the distribution, and the time course of psychological effects. For example, when you wait
784 for a response of the participant on each trial, you allow the participant to have control
785 over the trial duration, and some participants might respond only when they are confident
786 that their emitted response will be correct. These issues can be avoided by introducing a
787 (relatively short) response deadline in each trial, e.g., 500 ms for simple detection tasks,
788 800 ms for more difficult discrimination tasks, or 2 s for tasks requiring extended high-level
789 processing. Because EHA can deal in a straightforward fashion with right-censored
790 observations (i.e., trials without an observed response in the analysis time window),
791 introducing a response deadline is recommended when designing RT experiments.
792 Furthermore, introducing a response deadline and asking participants to respond before the
793 deadline as much as possible, will also lead to individual distributions that overlap in time,
794 which is important when selecting a common analysis time window when fitting hazard
795 and conditional accuracy models.

796 But even when using a response deadline, participants can differ qualitatively in the
797 effects they display (see Panis, 2020). One way to deal with this is to describe and
798 interpret the different patterns. Another way is to run a clustering algorithm on the
799 individual hazard estimates across all bins and conditions. The obtained dendrogram can
800 then be used to identify a (hopefully big) cluster of participants that behave similarly, and
801 to identify a (hopefully small) cluster of participants with different behavioral patterns.
802 One might then exclude the smaller sub-group of participants before fitting a hazard model
803 or consider the possibility that different cognitive processes may be at play during task
804 performance across the different sub-groups.

805 Another approach to deal with individual differences is Bayesian prevalence (Ince,

806 Paton, Kay, & Schyns, 2021), which is a form of small- N approach (Smith & Little, 2018).

807 This method looks at effects within each individual in the study and asks how likely it

808 would be to see the same result if the experiment was repeated with a new person chosen

809 from the wider population at random. This approach allows one to quantify how typical or

810 uncommon an observed effect is in the population, and the uncertainty around this

811 estimate.

812 5.4 Limitations

813 Compared to the orthodox method – comparing mean-averages between conditions –

814 the most important limitation of multilevel hazard and conditional accuracy modeling is

815 that it might take a long time to estimate the parameters using Bayesian methods or the

816 model might have to be simplified significantly to use frequentist methods.

817 Another issue is that you need a relatively large number of trials per condition to

818 estimate the hazard function with high temporal resolution, which is required when testing

819 predictions of process models of cognition. Indeed, in general, there is a trade-off between

820 the number of trials per condition and the temporal resolution (i.e., bin width) of the

821 hazard function. Therefore, we recommend researchers to collect as many trials as possible

822 per experimental condition, given the available resources and considering the participant

823 experience (e.g., fatigue and boredom). For instance, if the maximum session length

824 deemed reasonable is between 1 and 2 hours, what is the maximum number of trials per

825 condition that you could reasonably collect? After consideration, it might be worth

826 conducting multiple testing sessions per participant and/or reducing the number of

827 experimental conditions. Finally, there is a user-friendly online tool for calculating

828 statistical power as a function of the number of trials as well as the number of participants,

829 and this might be worth consulting to guide the research design process (Baker et al., 2021).

We did not discuss continuous-time EHA, nor continuous-time SAT analysis. As indicated by Allison (2010), learning discrete-time EHA methods first will help in learning continuous-time methods. Given that RT is typically treated as a continuous variable, it is possible that continuous-time methods will ultimately prevail. However, they require much more data to estimate the continuous-time hazard (rate) function well. Thus, by trading a bit of temporal resolution for a lower number of trials, discrete-time methods seem ideal for dealing with typical psychological time-to-event data sets for which there are less than ~200 trials per condition per experiment.

5.5 Extensions

The hazard models in this tutorial assume that there is one event of interest. For RT data, this button-press event constitutes a single transition between an “idle” state and a “responded” state. However, in certain situations, more than one event of interest might exist. For example, in a medical or health-related context, an individual might transition back and forth between a “healthy” state and a “depressed” state, before being absorbed into a final “death” state. When you have data on the timing of these transitions, one can apply multi-state hazard models, which generalize EHA to transitions between three or more states (Steele, Goldstein, & Browne, 2004). Also, the predictor variables in this tutorial are time-invariant, i.e., their value did not change over the course of a trial. Thus, another extension is to include time-varying predictors, i.e., predictors whose value can change across the time bins within a trial (Allison, 2010). For example, when gaze position is tracked during a visual search trial, the gaze-target distance will vary during a trial when the eyes move around before a manual response is given; shorter gaze-target distances should be associated with a higher hazard of response occurrence. Note that the effect of a time-varying predictor (e.g., an occipital EEG signal) can itself vary over time.

854

6. Conclusions

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

863

References

- 864 Allison, P. D. (1982). Discrete-Time Methods for the Analysis of Event Histories.
865 *Sociological Methodology*, 13, 61. <https://doi.org/10.2307/270718>
- 866 Allison, P. D. (2010). *Survival analysis using SAS: A practical guide* (2. ed). Cary, NC:
867 SAS Press.
- 868 Aust, F. (2019). *Citr: 'RStudio' add-in to insert markdown citations*. Retrieved from
869 <https://github.com/crsh/citr>
- 870 Aust, F., & Barth, M. (2023). *papaja: Prepare reproducible APA journal articles with R*
871 *Markdown*. Retrieved from <https://github.com/crsh/papaja>
- 872 Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R*
873 *Markdown*. <https://doi.org/10.32614/CRAN.package.papaja>
- 874 Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., &
875 Andrews, T. J. (2021). Power contours: Optimising sample size and precision in
876 experimental psychology and human neuroscience. *Psychological Methods*, 26(3),
877 295–314. <https://doi.org/10.1037/met0000337>
- 878 Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for
879 confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*,
880 68(3), 10.1016/j.jml.2012.11.001. <https://doi.org/10.1016/j.jml.2012.11.001>
- 881 Barth, M. (2023). *tinylabes: Lightweight variable labels*. Retrieved from
882 <https://cran.r-project.org/package=tinylabes>
- 883 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects
884 models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
885 <https://doi.org/10.18637/jss.v067.i01>
- 886 Bates, D., Maechler, M., & Jagan, M. (2024). *Matrix: Sparse and dense matrix classes and*
887 *methods*. Retrieved from <https://CRAN.R-project.org/package=Matrix>
- 888 Bengtsson, H. (2021). A unifying framework for parallel and distributed processing in r
889 using futures. *The R Journal*, 13(2), 208–227. <https://doi.org/10.32614/RJ-2021-048>

- 890 Blossfeld, H.-P., & Rohwer, G. (2002). *Techniques of event history modeling: New*
891 *approaches to causal analysis, 2nd ed* (pp. x, 310). Mahwah, NJ, US: Lawrence
892 Erlbaum Associates Publishers.
- 893 Box-Steffensmeier, J. M. (2004). Event history modeling: A guide for social scientists.
894 Cambridge: University Press.
- 895 Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan.
896 *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- 897 Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms.
898 *The R Journal*, 10(1), 395–411. <https://doi.org/10.32614/RJ-2018-017>
- 899 Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal*
900 *of Statistical Software*, 100(5), 1–54. <https://doi.org/10.18637/jss.v100.i05>
- 901 DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through
902 Data Simulation. *Advances in Methods and Practices in Psychological Science*, 4(1),
903 2515245920965119. <https://doi.org/10.1177/2515245920965119>
- 904 Eddelbuettel, D., & Balamuta, J. J. (2018). Extending R with C++: A Brief Introduction
905 to Rcpp. *The American Statistician*, 72(1), 28–36.
906 <https://doi.org/10.1080/00031305.2017.1375990>
- 907 Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal*
908 *of Statistical Software*, 40(8), 1–18. <https://doi.org/10.18637/jss.v040.i08>
- 909 Gabry, J., Češnovar, R., Johnson, A., & Broder, S. (2024). *Cmdstanr: R interface to*
910 *'CmdStan'*. Retrieved from <https://github.com/stan-dev/cmdstanr>
- 911 Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization
912 in bayesian workflow. *J. R. Stat. Soc. A*, 182, 389–402.
913 <https://doi.org/10.1111/rssa.12378>
- 914 Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and Other Stories.
915 <https://www.cambridge.org/highereducation/books/regression-and-other-stories/DD20DD6C9057118581076E54E40C372C>; Cambridge University Press.

- 917 https://doi.org/10.1017/9781139161879
- 918 Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., ...
- 919 Modrák, M. (2020). *Bayesian Workflow*. arXiv.
- 920 https://doi.org/10.48550/arXiv.2011.01808
- 921 Girard, J. (2024). *Standist: What the package does (one line, title case)*. Retrieved from
922 https://github.com/jmgirard/standist
- 923 Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate.
- 924 *Journal of Statistical Software*, 40(3), 1–25. Retrieved from
925 https://www.jstatsoft.org/v40/i03/
- 926 Halley, E. (1693). VI. An estimate of the degrees of the mortality of mankind; drawn from
927 curious tables of the births and funerals at the city of breslaw; with an attempt to
928 ascertain the price of annuities upon lives. *Philosophical Transactions of the Royal
929 Society of London*, 17(196), 596–610. https://doi.org/10.1098/rstl.1693.0007
- 930 Heiss, A. (2021, November 10). A Guide to Correctly Calculating Posterior Predictions
931 and Average Marginal Effects with Multilevel Bayesian Models.
932 https://doi.org/10.59350/wbn93-edb02
- 933 Hosmer, D. W., Lemeshow, S., & May, S. (2011). *Applied Survival Analysis: Regression
934 Modeling of Time to Event Data* (2nd ed). Hoboken: John Wiley & Sons.
- 935 Ince, R. A., Paton, A. T., Kay, J. W., & Schyns, P. G. (2021). Bayesian inference of
936 population prevalence. *eLife*, 10, e62461. https://doi.org/10.7554/eLife.62461
- 937 Kantowitz, B. H., & Pachella, R. G. (2021). The Interpretation of Reaction Time in
938 Information-Processing Research 1. *Human Information Processing*, 41–82.
939 https://doi.org/10.4324/9781003176688-2
- 940 Kay, M. (2023). *tidybayes: Tidy data and geoms for Bayesian models*.
941 https://doi.org/10.5281/zenodo.1308151
- 942 Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing,
943 estimation, meta-analysis, and power analysis from a Bayesian perspective.

- 944 *Psychonomic Bulletin & Review*, 25(1), 178–206.
- 945 <https://doi.org/10.3758/s13423-016-1221-4>
- 946 Kurz, A. S. (2023a). *Applied longitudinal data analysis in brms and the tidyverse* (version
947 0.0.3). Retrieved from <https://bookdown.org/content/4253/>
- 948 Kurz, A. S. (2023b). *Statistical rethinking with brms, ggplot2, and the tidyverse: Second*
949 *edition* (version 0.4.0). Retrieved from <https://bookdown.org/content/4857/>
- 950 Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, 8(1), 33267.
951 <https://doi.org/10.1525/collabra.33267>
- 952 Landes, J., Engelhardt, S. C., & Pelletier, F. (2020). An introduction to event history
953 analyses for ecologists. *Ecosphere*, 11(10), e03238. <https://doi.org/10.1002/ecs2.3238>
- 954 Makeham, W. M. (1860). *On the Law of Mortality and the Construction of Annuity Tables*.
955 The Assurance Magazine, and Journal of the Institute of Actuaries.
- 956 McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and*
957 *STAN* (2nd ed.). New York: Chapman and Hall/CRC.
958 <https://doi.org/10.1201/9780429029608>
- 959 Meyer, D. E., Osman, A. M., Irwin, D. E., & Yantis, S. (1988). Modern mental
960 chronometry. *Biological Psychology*, 26(1-3), 3–67.
961 [https://doi.org/10.1016/0301-0511\(88\)90013-0](https://doi.org/10.1016/0301-0511(88)90013-0)
- 962 Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. Retrieved from
963 <https://CRAN.R-project.org/package=tibble>
- 964 Nelder, J. A. (1999). From Statistics to Statistical Science. *Journal of the Royal Statistical*
965 *Society. Series D (The Statistician)*, 48(2), 257–269. Retrieved from
966 <https://www.jstor.org/stable/2681191>
- 967 Neuwirth, E. (2022). *RColorBrewer: ColorBrewer palettes*. Retrieved from
968 <https://CRAN.R-project.org/package=RColorBrewer>
- 969 Panis, S. (2020). How can we learn what attention is? Response gating via multiple direct
970 routes kept in check by inhibitory control processes. *Open Psychology*, 2(1), 238–279.

- 971 https://doi.org/10.1515/psych-2020-0107
- 972 Panis, S., Moran, R., Wolkersdorfer, M. P., & Schmidt, T. (2020). Studying the dynamics
973 of visual search behavior using RT hazard and micro-level speed–accuracy tradeoff
974 functions: A role for recurrent object recognition and cognitive control processes.
975 *Attention, Perception, & Psychophysics*, 82(2), 689–714.
- 976 https://doi.org/10.3758/s13414-019-01897-z
- 977 Panis, S., Schmidt, F., Wolkersdorfer, M. P., & Schmidt, T. (2020). Analyzing Response
978 Times and Other Types of Time-to-Event Data Using Event History Analysis: A Tool
979 for Mental Chronometry and Cognitive Psychophysiology. *I-Perception*, 11(6),
980 2041669520978673. https://doi.org/10.1177/2041669520978673
- 981 Panis, S., & Schmidt, T. (2016). What Is Shaping RT and Accuracy Distributions? Active
982 and Selective Response Inhibition Causes the Negative Compatibility Effect. *Journal of*
983 *Cognitive Neuroscience*, 28(11), 1651–1671. https://doi.org/10.1162/jocn_a_00998
- 984 Panis, S., & Schmidt, T. (2022). When does “inhibition of return” occur in spatial cueing
985 tasks? Temporally disentangling multiple cue-triggered effects using response history
986 and conditional accuracy analyses. *Open Psychology*, 4(1), 84–114.
- 987 https://doi.org/10.1515/psych-2022-0005
- 988 Panis, S., Torfs, K., Gillebert, C. R., Wagemans, J., & Humphreys, G. W. (2017).
989 Neuropsychological evidence for the temporal dynamics of category-specific naming.
990 *Visual Cognition*, 25(1-3), 79–99. https://doi.org/10.1080/13506285.2017.1330790
- 991 Panis, S., & Wagemans, J. (2009). Time-course contingencies in perceptual organization
992 and identification of fragmented object outlines. *Journal of Experimental Psychology:*
993 *Human Perception and Performance*, 35(3), 661–687.
- 994 https://doi.org/10.1037/a0013547
- 995 Pedersen, T. L. (2024). *Patchwork: The composer of plots*. Retrieved from
996 https://CRAN.R-project.org/package=patchwork
- 997 Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-PLUS*. New York:

- 998 Springer. <https://doi.org/10.1007/b98882>
- 999 1000 1001 1002 1003 1004 R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- 1002 Ripley, B., Venables, B., Bates, D. M., ca 1998), K. H. (partial. port, ca 1998), A. G. (partial. port, & polr), D. F. (support. functions for. (2024). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*.
- 1003 1004 1005 1006 Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford, New York: Oxford University Press.
- 1007 1008 1009 Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, 25(6), 2083–2101.
<https://doi.org/10.3758/s13423-018-1451-8>
- 1010 1011 1012 1013 Steele, F., Goldstein, H., & Browne, W. (2004). A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modelling*, 4(2), 145–159.
<https://doi.org/10.1191/1471082X04st069oa>
- 1014 1015 Teachman, J. D. (1983). Analyzing social processes: Life tables and proportional hazards models. *Social Science Research*, 12(3), 263–301.
[https://doi.org/10.1016/0049-089X\(83\)90015-7](https://doi.org/10.1016/0049-089X(83)90015-7)
- 1017 1018 Tong, C. (2019). Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science. *The American Statistician*, 73(sup1), 246–261.
<https://doi.org/10.1080/00031305.2018.1518264>
- 1020 1021 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)
- 1022 1023 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <https://ggplot2.tidyverse.org>
- 1024 Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors)*.

- 1025 Retrieved from <https://forcats.tidyverse.org/>
- 1026 Wickham, H. (2023b). *Stringr: Simple, consistent wrappers for common string operations.*
- 1027 Retrieved from <https://stringr.tidyverse.org>
- 1028 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ...
- 1029 Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43),
- 1030 1686. <https://doi.org/10.21105/joss.01686>
- 1031 Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for data science: Import,*
- 1032 *tidy, transform, visualize, and model data* (2nd edition). Beijing Boston Farnham
- 1033 Sebastopol Tokyo: O'Reilly.
- 1034 Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A*
- 1035 *grammar of data manipulation.* Retrieved from <https://dplyr.tidyverse.org>
- 1036 Wickham, H., & Henry, L. (2023). *Purrr: Functional programming tools.* Retrieved from
- 1037 <https://purrr.tidyverse.org/>
- 1038 Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data.* Retrieved
- 1039 from <https://readr.tidyverse.org>
- 1040 Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data.* Retrieved from
- 1041 <https://tidyr.tidyverse.org>
- 1042 Winter, B. (2019). *Statistics for Linguists: An Introduction Using R.* New York:
- 1043 Routledge. <https://doi.org/10.4324/9781315165547>
- 1044 Wolkersdorfer, M. P., Panis, S., & Schmidt, T. (2020). Temporal dynamics of sequential
- 1045 motor activation in a dual-prime paradigm: Insights from conditional accuracy and
- 1046 hazard functions. *Attention, Perception, & Psychophysics*, 82(5), 2581–2602.
- 1047 <https://doi.org/10.3758/s13414-020-02010-5>